

**POLITECNICO DI TORINO**

Collegio di Ingegneria Gestionale e della Produzione

**Corso di Laurea Magistrale**

**INGEGNERIA DELLA PRODUZIONE INDUSTRIALE E  
DELL'INNOVAZIONE TECNOLOGICA**

Tesi di Laurea Magistrale



**Politecnico  
di Torino**

**Qualità dei dati nei database bibliometrici:  
Analisi empirica e classificazione degli errori più rilevanti**

**Relatore**

Prof. Domenico A. MAISANO

**Candidato**

Lucrezia FERRARA

Anno Accademico 2023-2024

# INDICE

<b>Indice delle Tabelle.....</b>	<b>4</b>
<b>Indice delle Figure.....</b>	<b>5</b>
<b>INTRODUZIONE.....</b>	<b>7</b>
<b>1. CONCETTI INTRODUTTIVI SULLA BIBLIOMETRIA.....</b>	<b>10</b>
<b>1.1 LE TRE METRICHE.....</b>	<b>10</b>
<b>1.2 FINALITÀ DELLA BIBLIOMETRIA.....</b>	<b>14</b>
<b>1.3 BANCHE DATI CITAZIONALI.....</b>	<b>19</b>
<b>1.3.1 Scopus.....</b>	<b>19</b>
<b>1.3.2 Web of Science.....</b>	<b>23</b>
<b>2. ANALISI DELL'ACCURATEZZA NELLA CLASSIFICAZIONE DELLE TIPOLOGIE DI DOCUMENTO.....</b>	<b>26</b>
<b>2.1 IDENTIFICAZIONE DEL PROBLEMA E OBIETTIVO DELLA RICERCA....</b>	<b>26</b>
<b>2.2 METODOLOGIA.....</b>	<b>29</b>
<b>2.3 APPLICAZIONE DEGLI STEP METODOLOGICI.....</b>	<b>32</b>
<b>3. RISULTATI.....</b>	<b>39</b>
<b>3.1 ANALISI DEGLI ERRORI.....</b>	<b>39</b>
<b>3.2 TIPOLOGIE DI ERRORI.....</b>	<b>43</b>
<b>3.2.1 Esempi di errori e anomalie nel <i>corpus</i> analizzato.....</b>	<b>44</b>

<b>CONCLUSIONI</b> .....	54
<b>BIBLIOGRAFIA E SITOGRAFIA</b> .....	57
<b>RINGRAZIAMENTI</b> .....	60

## Indice delle Tabelle

Tabella 1 – Principali termini utilizzati come metriche di valutazione della ricerca (Das, 2015).....	13
Tabella 2 - Valutazione della ricerca (Cassella & Bozzarelli, 2011).....	17
Tabella 3 - Sintesi dei DT classificati da Scopus e WoS con riferimento al sottoinsieme “intersezione” nell'esempio applicativo .....	34
Tabella 4 - Esempio della matrice di concordanza .....	35
Tabella 5 - Esempio di tabella degli errori per Scopus .....	37
Tabella 6 - Esempio di tabella degli errori per Web of Science .....	38

## Indice delle Figure

Figura 1 – Le tre metriche nella metrologia .....	12
Figura 2 - Diagramma dell'infometria (Björneborn & Ingwersen, 2004).....	13
Figura 3 - Estratto dell'articolo di Eugene Garfield “Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas” (Garfield, 1955).....	15
Figura 4 - Misura dell'impatto dei contributi scientifici (Mattarella, 2023) .....	18
Figura 5 - Scopus in numeri (Elsevier, 2024) .....	20
Figura 6 - Esempi di DT non trattati da Scopus (Elsevier, 2024) .....	21
Figura 7 - DT trattati da Scopus (Parte 1/2) (Elsevier, 2024) .....	22
Figura 8 - DT trattati da Scopus (Parte 2/2) (Elsevier, 2024) .....	23
Figura 9 - DT attualmente non trattati da WoS (Clarivate, 2024).....	24
Figura 10 - Fasi della metodologia.....	29
Figura 11 - Step metodologici (Step 1-5 analisi automatica; Step 6-7 analisi manuale) .....	30
Figura 12 - Descrizione degli step metodologici .....	31
Figura 13 - Mappa delle statistiche degli errori nel corpus analizzato.....	42
Figura 14 - Tipologie di errore riscontrate nel corpus analizzato .....	43
Figura 15 - Esempio di documento non accessibile per DOI inattivo ( <a href="https://www.doi.org/">https://www.doi.org/</a> )	44
Figura 16 - Esempio di documento classificato erroneamente da Scopus come documento da conferenza anziché articolo ( <a href="https://doi.org/10.1016/j.ejor.2018.12.016">https://doi.org/10.1016/j.ejor.2018.12.016</a> ). .....	45
Figura 17 - Esempio di documento classificato erroneamente da Scopus come articolo anziché review (doi: 10.1016/j.jweia.2018.04.021) .....	46
Figura 18 - Esempio di documento classificato erroneamente da Scopus come articolo anziché conference paper (doi: 10.1016/j.promfg.2018.01.017).....	47
Figura 19 - Esempio di documento classificato erroneamente da Scopus come articolo anziché conference paper e da WoS come review (doi: 10.1111/eea.12967) .....	48

Figura 20 - Esempio di documento classificato da Scopus come review e da WoS come articolo (doi: 10.5194/hess-25-3937-2021).....	50
Figura 21 - Esempio di documento classificato da Scopus come articolo e da WoS come review (doi: 10.4254/wjh.v14.i5.923 ).....	51
Figura 22 – Esempio di documento classificato da Scopus come articolo e da WoS come review (doi: 10.1007/s12024-019-00105-6) .....	53

# INTRODUZIONE

## **Premessa: la centralità dei dati**

I dati sono il nuovo petrolio, con la differenza che il petrolio è una risorsa limitata mentre i dati sono potenzialmente infiniti finché ci saranno utenti che li generano. La raccolta, l'archiviazione e l'elaborazione dei dati diventano quindi attività cruciali e devono avvenire attraverso sistemi informativi adeguati e qualitativamente efficienti. Al fine di garantire accuratezza, corretta gestione dei dati e per trasformare dati grezzi in informazioni utili e contestualizzate, uno degli strumenti di gestione principale è la base dati. Anche conosciuta come database, una base dati è uno strumento essenziale per gestire grandi quantità di dati in modo ordinato e funzionale, adattandosi agli obiettivi delle varie organizzazioni, indipendentemente dal settore di attività o dalla loro dimensione.

## **Obiettivo della trattazione**

Nello specifico l'obiettivo della tesi è quello di valutare la qualità dei dati presenti nei database bibliometrici, come Scopus e Web of Science (WoS). I database bibliometrici sono raccolte di pubblicazioni scientifiche per analisi bibliometriche. In particolare, viene analizzata l'accuratezza nella classificazione del Document Type (DT), elencando e classificando gli errori più rilevanti secondo una nuova metodologia semi-automatica. Tra le molteplici variabili, le pubblicazioni scientifiche sono solitamente classificate anche secondo i loro DT, categorie predefinite che ne delineano la natura, come ad esempio, *articoli di ricerca*, *atti di conferenze*, *recensioni*, *note*, *lettere*, *capitoli di libri*, ecc.

## **Implicazioni pratiche e sociali**

Migliorare l'accuratezza e la precisione dei dati presenti nei database bibliometrici contribuisce e influenza gli indicatori bibliometrici, uno dei principali fattori su cui viene valutata la produzione scientifica. La comunità scientifica potrebbe beneficiare di tali miglioramenti, almeno in parte, sia in termini di processi decisionali in merito alla reputazione accademica sia in merito ai finanziamenti per la ricerca.

## **Approccio metodologico**

La prima fase della metodologia proposta nella tesi è automatizzata e si basa sull'individuazione delle discrepanze nelle classificazioni del DT tra due database concorrenti, Scopus e WoS. Questo primo passo mira a identificare un sottoinsieme di documenti potenzialmente soggetti a errori di classificazione del DT. La seconda fase prevede, invece, un'analisi manuale del sottoinsieme di documenti identificato nella precedente fase, con l'obiettivo di individuare e correggere gli errori di classificazione del DT. Nel presente elaborato la metodologia viene illustrata attraverso un esempio di applicazione pratica.

## **Struttura della trattazione**

L'elaborato è strutturato in tre capitoli:

- *Capitolo I – Concetti introduttivi sulla bibliometria.* Vengono presentati i principali concetti sulla bibliometria e sui database bibliometrici come Scopus e Web of Science. L'obiettivo è fornire una panoramica generale sul contesto e sulla disciplina in cui è inserita la ricerca.

- *Capitolo II – Analisi dell’accuratezza nella classificazione delle tipologie di documenti su database bibliometrici.* Viene descritta la ricerca condotta, introducendo la metodologia adottata, la descrizione degli strumenti utilizzati e l’applicazione degli step metodologici.
- *Capitolo III – Risultati.* Viene presentata l’analisi dei risultati ottenuti e gli errori riscontrati nell’esempio applicato. La trattazione si conclude presentando le implicazioni pratiche, i limiti e le possibili direzioni per la ricerca futura.

# Capitolo I

## 1. CONCETTI INTRODUTTIVI SULLA BIBLIOMETRIA

### 1.1 LE TRE METRICHE

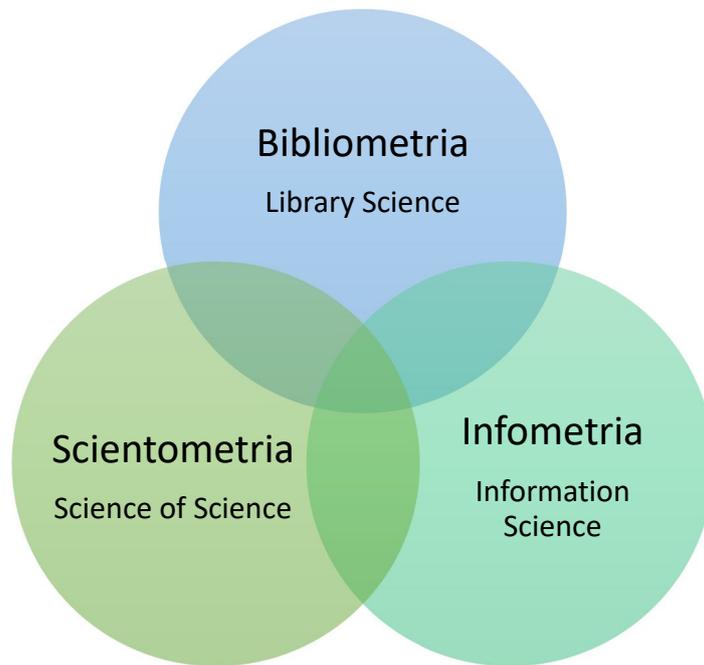
La valutazione della produzione scientifica rappresenta un passaggio imprescindibile in sede di attribuzione dei fondi di ricerca, nonché di selezione del personale docente e ricercatore delle Università. Uno dei principali strumenti utilizzati per la valutazione della ricerca sono gli indicatori bibliometrici, utilizzati nell'ambito dell'analisi scientometrica, per esaminare i modelli di diffusione delle pubblicazioni scientifiche e per valutare il loro impatto sulle comunità scientifiche (Castiglia & Tranchina, 2016).

Il termine scientometria è stato usato per la prima volta come traduzione del termine russo "naukomtriya" (misurazione della scienza) coniato nel 1969 dal filosofo e umanista Nalimov e da Mulchenko. Secondo Nalimov e Mulchenko la scientometria comprende «quei metodi quantitativi che si occupano dell'analisi della scienza vista come processo informativo» (Schubert, 2001). Il termine ha acquisito una maggiore notorietà grazie alla seconda edizione di "Science Since Babylon" di De Solla Price nel 1975, che definisce la scientometria come la disciplina che deve «puntare gli strumenti della scienza sulla scienza stessa» (Yagi, Badash, & de B.Beaver, 1996). Tuttavia, l'uso del termine si diffuse soprattutto con la creazione della rivista *Scientometrics* da parte del chimico ungherese Tibor Braun nel 1978.

La scientometria è considerata lo studio quantitativo, principalmente statistico, di qualsiasi aspetto misurabile dell'attività scientifica, con l'obiettivo di comprendere e, se possibile, migliorare il suo meccanismo di funzionamento (Schubert, 2001). Tale disciplina si distingue da approcci più umanistici e qualitativi per l'uso di strumenti quantitativi e di tecniche più vicine a quelle delle scienze fisiche rispetto a quelle delle scienze umane e sociali.

La definizione dell'obiettivo e dello scopo di una ricerca ha un particolare significato nel distinguere la scientometria dai campi strettamente correlati della bibliometria e dell'infometria. Queste ultime condividono con la scientometria sia l'oggetto che la metodologia, ma l'obiettivo principale della bibliometria è il miglioramento dei servizi bibliotecari e informativi (Pritchard, 1969), mentre l'infometria è interessata soprattutto a questioni teoriche e metodologiche (Egghe & Rosseau, 1990) (Schubert, 2001).

La bibliometria, la scientometria e l'infometria sono anche conosciute come le tre metriche e sono tre termini correlati nella metrologia (*Figura 1*). Va sottolineato che in letteratura non è presente un consenso generale sulle linee di demarcazione tra i diversi campi metrici. Di fatto risulta esserci una significativa sovrapposizione tra di essi e, a volte, vengono utilizzati in modo quasi intercambiabile (Schubert, 2001).



*Figura 1 – Le tre metriche nella metrologia*

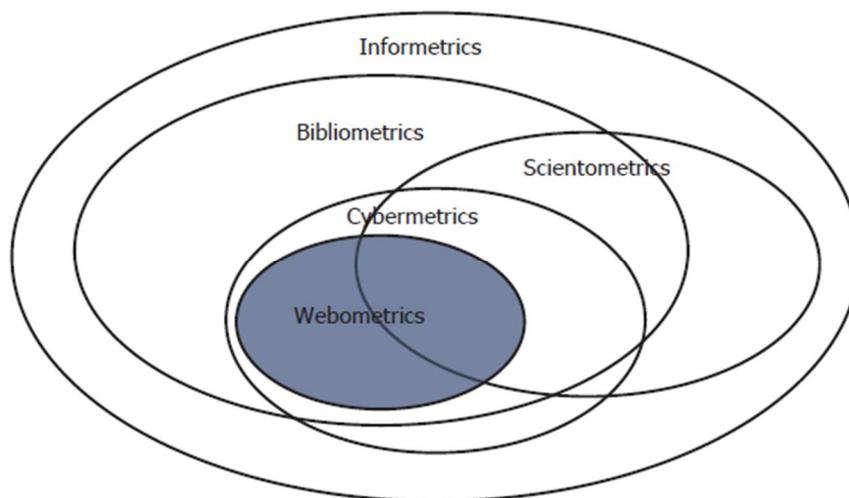
In particolare, il termine bibliometria è stato usato per la prima volta nel 1969 dal ricercatore Alan Pritchard, che l'ha definita come «l'applicazione di metodi matematici e statistici ai libri e alle altre forme di comunicazione scritta». La bibliometria è dunque l'applicazione di tecniche matematiche e statistiche allo studio della diffusione e dell'impatto della letteratura scientifica e ai prodotti della comunicazione scientifica e tecnologica volti al progresso della conoscenza (Bagnasco, 2023).

L'infometria, invece, è lo studio degli aspetti quantitativi dell'informazione in qualunque forma (non solo libri, articoli...) e presso qualunque gruppo sociale (non solo scienziati). Successivamente si è vista un'estensione dei metodi biblio-sciento-infometrici ai flussi informativi della rete internet (web, posta elettronica...) con le discipline della Webmetrica, Cybermetrica e Netmetrica (Almind & Ingwersen, 1997).

Nella *Tabella 1* e nella *Figura 2* viene proposta una panoramica sulle definizioni dei principali termini usati come metriche di valutazione della ricerca.

*Tabella 1 – Principali termini utilizzati come metriche di valutazione della ricerca (Das, 2015)*

<b>TERMINE</b>	<b>DEFINIZIONE</b>
<i>Bibliometrics (Bibliometria)</i>	Applicazione di metodi matematici e statistici per analizzare quantitativamente la letteratura accademica (libri e altri strumenti di comunicazione).
<i>Informetrics (Infometria)</i>	Studio degli aspetti quantitativi dell'informazione, che comprende la produzione, la dissertazione e l'utilizzo di tutte le forme di informazione, indipendentemente dalla loro forma o origine.
<i>Scientometrics (Scientometria)</i>	Studio quantitativo dello sviluppo della scienza
<i>Cybermetrics (Cybermetria)</i>	Analisi quantitativa di contenuti e applicazioni dell'intero web, dei suoi collegamenti ipertestuali e delle risorse internet. Termine alternativo a Webometria.
<i>Webometrics (Webometria)</i>	Analisi quantitativa di contenuti web, indirizzata a risolvere problemi bibliometrici
<i>Altmetrics</i>	Analisi e impiego di indicatori d'impatto in ambito scientifico basati su attività e strumenti online



*Figura 2 - Diagramma dell'infometria (Björneborn & Ingwersen, 2004)*

## 1.2 FINALITÀ DELLA BIBLIOMETRIA

La bibliometria, essendo una disciplina basata sull'analisi quantitativa delle pubblicazioni, è diventata un efficace strumento nella valutazione della ricerca, soprattutto a seguito dello sviluppo e dell'implementazione delle banche dati citazionali.

Per secoli la valutazione della qualità della letteratura scientifica è stata ritenuta di competenza esclusiva della comunità dei pari, un gruppo di esperti nella disciplina in cui rientra la pubblicazione, anche conosciuto come sistema di peer reviewing (Castiglia & Tranchina, 2016). Come per ogni procedura che prevede un giudizio finale, tale metodo poggia sul presupposto della totale imparzialità dei revisori e sulla necessità di una procedura obiettiva e non gravata da giudizi personali. La difficoltà di trovare l'equilibrio tra l'esigenza di imparzialità del metodo e i problemi di soggettività, ha spinto il mondo della scienza a esplorare alternative al sistema di revisione tra pari.

È proprio in questo contesto che l'impiego di strumenti bibliometrici basati sull'analisi delle citazioni ha cominciato a ricevere grande attenzione da parte del mondo scientifico. Rispetto alla valutazione basata sul giudizio dei pari, quella condotta con indicatori bibliometrici, risulta essere più rapida, economica e i dati su cui si basa sono facilmente estraibili tramite gli indici citazionali (Cassella & Bozzarelli, 2011).

La prima intuizione in merito al metodo bibliometrico, per definire il peso e l'importanza di un prodotto di comunicazione, risale al 1955 ed è stata introdotta dal bibliografo Garfield (Garfield, 1955). Il chimico statunitense propose, sulle pagine di Science, la creazione di un indice generale delle citazioni tra articoli scientifici che permettesse di navigare la rete costituita

dalle pubblicazioni scientifiche su un determinato argomento. Pochi anni dopo, la società fondata da Garfield, l'Institute for Scientific Information (ISI), pubblicò per la prima volta il Science Citation Index (SCI), un indice che raccoglie le citazioni tra articoli scientifici.



Figura 3 - Estratto dell'articolo di Eugene Garfield "Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas" (Garfield, 1955)

In realtà, gli indici come SCI e la bibliometria in generale, non vennero originariamente concepiti come strumenti per la valutazione della ricerca. Piuttosto vennero creati, come scritto dallo stesso Garfield (*Figura 3*), per connettere tra loro studi sulla base di affinità concettuali, creando una rete di «associazioni di idee» tra riviste e discipline diverse (Garfield, 1955).

Tuttavia, l'idea sottostante alla bibliometria, ovvero che analizzando questa rete di citazioni sia possibile ottenere informazioni quantitative sul processo di costruzione della conoscenza scientifica, si rivelò troppo allettante per non estenderne l'applicazione alla valutazione della ricerca: identificare gli studi più influenti, individuare gli istituti di ricerca più significativi, valutare i ricercatori più brillanti ecc. (Bagnasco, 2023).

Dopo la pubblicazione di un contributo scientifico è dunque di fondamentale importanza, sia per l'autore che per l'ente di appartenenza, rilevare quanto e in che modo i prodotti finali della ricerca siano considerati dalla comunità scientifica. La valutazione dell'impatto della ricerca segue quindi, oltre al sistema di revisione alla pari, anche un sistema basato sugli strumenti bibliometrici. Avremo quindi un duplice approccio: qualitativo e quantitativo (*Tabella 2*).

Tabella 2 - Valutazione della ricerca (Cassella & Bozzarelli, 2011)

	ANALISI QUALITATIVA	ANALISI QUANTITATIVA
<i>Strumento</i>	PEER-REVIEW	INDICATORI
<i>In cosa consiste</i>	<p>Metodologia per la valutazione di merito su prodotti di ricerca mediante gli esiti di peer-review (revisione dei pari), esperti nelle medesime discipline accademiche e di ricerca in cui ricadono i prodotti esaminati. Si tratta di un'analisi non economica e soggettiva per definizione, ma frutto di competenze scientifiche.</p>	<p>Metodologia strutturata su indicatori matematici volti a misurare l'impatto di un lavoro di ricerca sulle comunità scientifiche. Tali indicatori quantitativi si distinguono in:</p> <ul style="list-style-type: none"> <li>• Indicatori di tipo bibliometrico (ad esempio Impact Factor, H-Index);</li> <li>• Indicatori non bibliometrici (ad esempio i dati brevettuali, i dati sui fondi di ricerca ottenuti, il numero di partecipazioni a conferenze, il numero di partecipazioni a comitati scientifici di riviste ecc.).</li> </ul>

In *Figura 4* viene presentato un diagramma raffigurante il processo per la misurazione dell'impatto dei contributi scientifici.

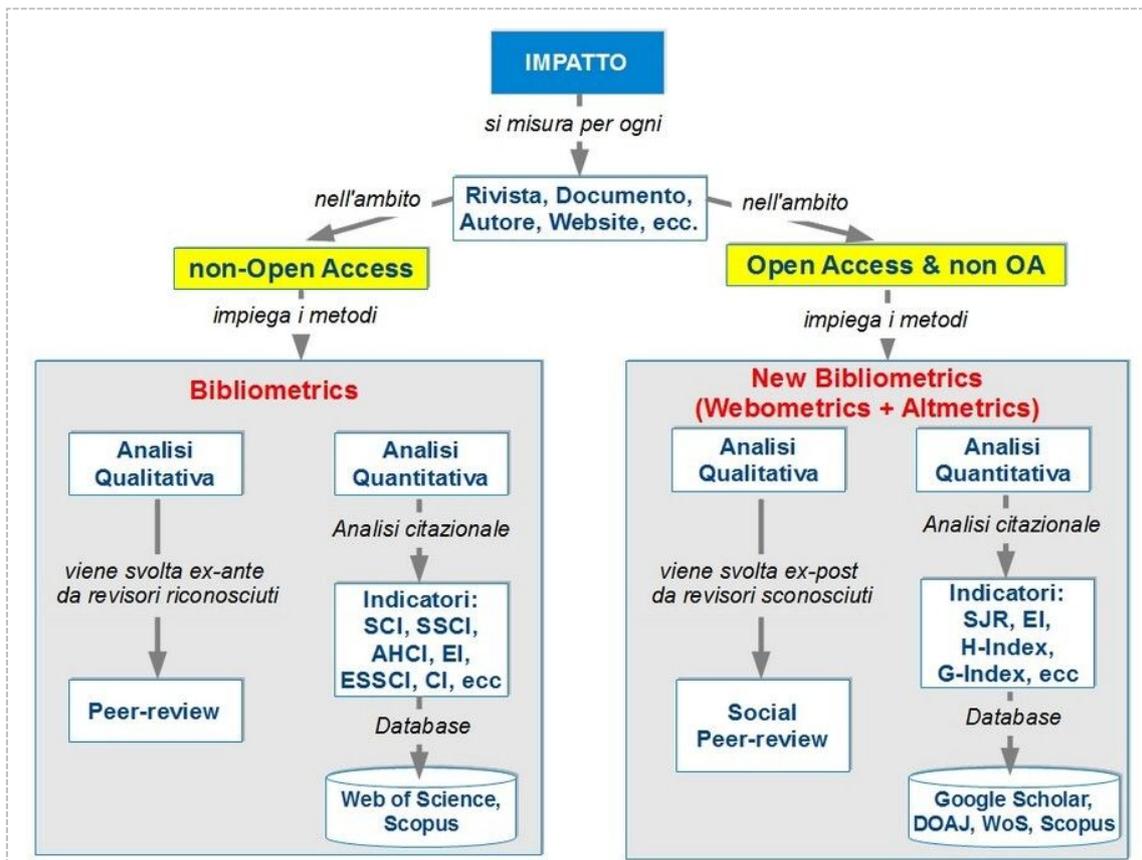


Figura 4 - Misura dell'impatto dei contributi scientifici (Mattarella, 2023)

I dati provenienti dall'analisi bibliometrica sono dunque diventati uno strumento efficace e in continua espansione e servono principalmente per:

- Selezionare, acquisire, accedere, utilizzare e conservare le collezioni bibliografiche;
- Determinare l'impatto e il ranking di documenti e riviste;
- Individuare i revisori alla pari delle pubblicazioni, da parte dei redattori delle riviste.

## 1.3 BANCHE DATI CITAZIONALI

Un database bibliometrico è un sistema organizzato che raccoglie e fornisce accesso a informazioni bibliometriche, ovvero dati relativi alla produzione scientifica e accademica. Questi database sono spesso utilizzati per analizzare e valutare l'impatto della ricerca, identificare tendenze nel campo scientifico e monitorare la produzione accademica di singoli autori, istituzioni o paesi. Tra i principali database bibliometrici troviamo Scopus e Web of Science (WoS).

### 1.3.1 Scopus



Scopus è una banca dati bibliografica citazionale multidisciplinare creata dalla casa editrice Elsevier nel 2004. Scopus copre una vasta gamma di discipline scientifiche, tecniche, mediche e sociali, rendendolo un database estremamente versatile per ricercatori in diversi campi. Inoltre, include pubblicazioni provenienti da tutto il mondo, il che lo rende una risorsa fondamentale per la ricerca internazionale e la collaborazione tra istituzioni di diversi paesi. In *Figura 5* viene fornita una panoramica sui numeri del database aggiornata al mese di gennaio 2024.

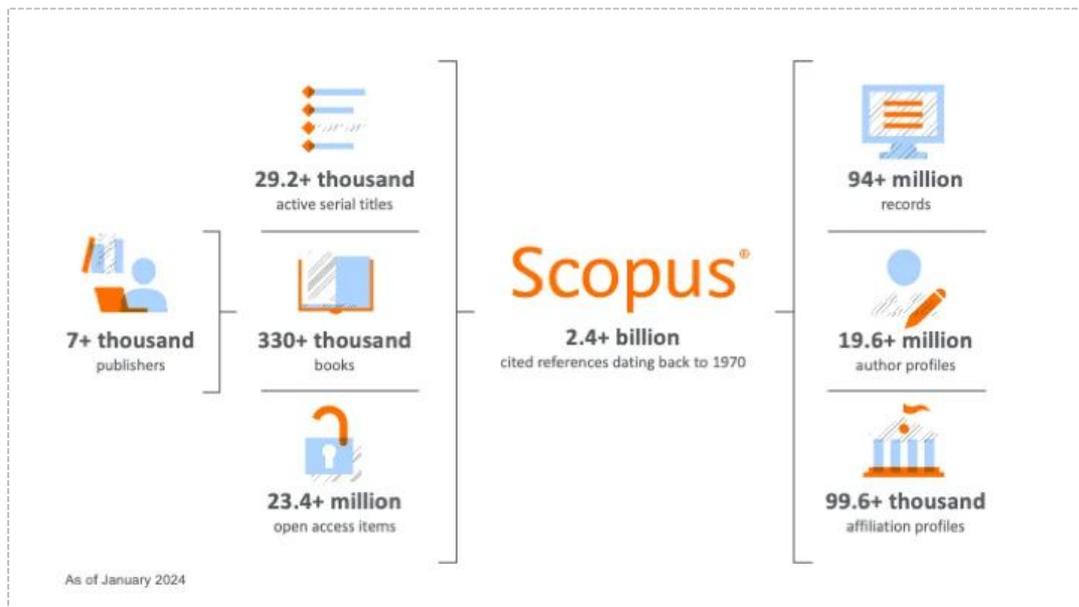


Figura 5 - Scopus in numeri (Elsevier, 2024)

Ai fini di questa trattazione, la cui ricerca si basa sull'analisi dei documenti secondo la loro classificazione per tipologia di documento, è necessario fare un focus sui diversi DT trattati dal database.

La copertura di Scopus si concentra sui tipi di documenti primari, ovvero documenti in cui l'autore della pubblicazione è identico al ricercatore responsabile dei risultati presentati. Scopus non include dunque tipologie di documenti secondari, in cui l'autore non risulta identico alla persona che sta dietro alla ricerca presentata, come i necrologi (*obituaries*), le recensioni di libri (*book reviews*) e gli *abstracts di conferenze* (Figura 6) (Elsevier, 2024).

### Examples of document types not covered in Scopus

Document type	Definition
<b>Book reviews</b>	Scopus does not cover book reviews. The reason for this is that they do not represent primary literature, and the publishers in whose journals they appear often regard them as full-text. As a full-text article, Scopus would only be able to display the title of the book review, which is often identical to the actual book, causing confusion to Scopus users. Lastly, book reviews are not often cited in research literature. As an example of this, the average citation per item for the “Journal of Academic Librarianship” drops by 50% (2.13 to 1.12) when book reviews are included.
<b>Conference meeting abstracts</b>	Scopus does not cover conference meeting abstracts. Meeting abstracts are not a complete and unique record of research and can drive ambiguity or duplication. Furthermore, the scientific content and substance of meeting abstracts is generally not peer-reviewed and is limited as they are typically written before the actual research is complete.

*Figura 6 - Esempi di DT non trattati da Scopus (Elsevier, 2024)*

In *Figura 7* e *Figura 8* vengono invece elencate le tipologie di documento trattate dal database Scopus con le relative definizioni.

## Document types covered in Scopus

Document type	Definition
<b>Article</b>	<p>Original research or opinion.</p> <p><b>Characteristics:</b> Articles in peer-reviewed journals are usually several pages in length, most often subdivided into sections: abstract, introduction, materials &amp; methods, results, conclusions, discussion and references. However, case reports, technical and research notes and short communications are also considered to be articles and may be as short as one page in length. Articles in trade journals are typically shorter than in peer-reviewed journals, and may also be as brief as one page in length.</p>
<b>Article-in-Press (AiP)</b>	<p>Accepted article made available online before official publication (see section 5).</p>
<b>Book</b>	<p>A whole monograph or entire book.</p> <p><b>Characteristics:</b> Book type is assigned to the whole. Additionally, for books with individual chapters, each chapter, along with a general item summarizing the book, is also indexed with the source type Book.</p>
<b>Chapter</b>	<p>A book chapter.</p> <p><b>Characteristics:</b> A complete chapter in a book or book series volume where the item is identified as a chapter by a heading or section indicator.</p>
<b>Conference paper</b>	<p>Original article reporting data presented at a conference or symposium.</p> <p><b>Characteristics:</b> Conference papers are of any length reporting data from a conference, with the exception of conference abstracts. Conference papers may range in length and content from full papers and published conference summaries to short items as short as one page in length (also see section 2.1).</p>
<b>Data paper</b>	<p>Searchable metadata documents describing an online accessible dataset, or group of datasets.</p> <p><b>Characteristics:</b> The intent of a data paper is to offer descriptive information on the related dataset(s) focusing on data collection, distinguishing features, access, and potential reuse rather than information on data processing and analysis.</p>
<b>Editorial</b>	<p>Summary of several articles, or provides editorial opinions or news.</p> <p><b>Characteristics:</b> Editorials are typically identified as editorial, introduction, leading article, preface or foreword, and are usually listed at the beginning of the table of contents.</p>
<b>Erratum</b>	<p>Report of an error, correction or retraction of a previously published paper. The erratum notice is linked to the original published paper this concerns and vice versa.</p> <p><b>Characteristics:</b> Errata are short items citing errors in, corrections to, or retractions of a previously published article in the same journal to which a citation is provided.</p>

Figura 7 - DT trattati da Scopus (Parte 1/2) (Elsevier, 2024)

Document type	Definition
<b>Letter</b>	Letter to or correspondence with the editor. <b>Characteristics:</b> Letters are individual letters or replies. Each individual letter or reply is processed as a single item.
<b>Note</b>	Note, discussion or commentary. <b>Characteristics:</b> Notes are short items that are not readily suited to other item types. They may or may not share characteristics of other item types, such as author, affiliation and references. Discussions and commentaries that follow an article are defined as notes and considered to be items in their own right. Notes also include questions and answers, as well as comments on other (often translated) articles. In trade journals, notes are generally shorter than half a page in length.
<b>Retracted article</b>	Published articles that the author(s) or publisher has requested to retract. The erratum or retraction notice announcing the retraction is linked to the retracted article. <b>Characteristics:</b> Articles with a published retraction note will be updated to the document type "Retracted." Usually, these articles are indicated with the words "retracted" or "retraction".
<b>Review</b>	Significant review of original research, also includes conference papers. <b>Characteristics:</b> Reviews typically have an extensive bibliography. Educational items that review specific issues within the literature are also considered to be reviews. As non-original articles, reviews lack the most typical sections of original articles such as materials & methods and results.
<b>Short survey</b>	Short or mini-review of original research. <b>Characteristics:</b> Short surveys are similar to reviews, but usually are shorter (not more than a few pages) and with a less extensive bibliography.

*The Scopus editorial team is responsible for the document type classification of records.*

Figura 8 - DT trattati da Scopus (Parte 2/2) (Elsevier, 2024)

### 1.3.2 Web of Science



Web of Science (WoS) è una banca dati citazionale e multidisciplinare online dal 2002. WoS è l'evoluzione del primo database citazionale realizzato in ambito accademico, Science Citation Index (SCI) dell'Institute of Scientific Information (ISI), oggi proprietà del gruppo Clarivate Analytics. È una piattaforma composta da diversi database di ricerca bibliografica progettati

per supportare la ricerca scientifica e accademica e la copertura disciplinare comprende i campi della scienza, tecnologia, medicina, scienze sociali, arti e discipline umanistiche.

Nonostante il gran numero di tipologie di documento che vengono trattate da WoS, il cui elenco completo è disponibile online sul sito di Clarivate al seguente link <https://webofscience.help.clarivate.com/en-us/Content/document-types.html>, in *Figura 9* vengono elencate le categorie che sono state ritirate dal database. Infatti, tali DT non vengono attualmente trattati da WoS e questo crea ulteriori discrepanze in termini di nomenclatura, definizioni e categorie tra i due database analizzati (Scopus e WoS).

<b>Retired Document Types</b>	
While retired document types are no longer added to items indexed in the Web of Science Core Collection, you can use these document types for searching or refining/analyzing search results.	
<b>Document Type</b>	<b>Description</b>
<b>Abstract of a Published Item</b>	Bibliographic-only data on a published paper. Generally, finds records dating back to 1974 or before.
<b>Chronology</b>	A review of events on a specific topic or subject in their order of occurrence in time. . Generally, finds records dating back to 2012 or before. A chronology is currently processed as an Article.
<b>Discussion</b>	An article or paper that discusses questions in an open and usually informal debate. Generally, finds records dating back to 1996 or before. Discussions are currently processed as Editorial Material or an Article depending on the content.
<b>Item About an Individual</b>	A review of the work(s) of a celebrated person in a particular field of study. As of 1999, all Items about an Individual are coded as a Biographical Item.
<b>Note</b>	A paper that mentions or remarks on a published paper on a specific subject. Generally, finds records dating back to 1996 or before. Notes are currently processed as an article depending on the content
<b>TV Review, Radio Review, Video Review</b>	Reviews of television, radio broadcasts, and videos. This document type was used until 1997. It is currently processed as TV Review, Radio Review.

*Figura 9 - DT attualmente non trattati da WoS (Clarivate, 2024)*

Si noti poi che entrambi i DB sono caratterizzati dall'essere multidisciplinari, dal momento che indicizzano pubblicazioni relative a molteplici discipline. Tale aspetto aumenta la complessità degli strumenti visto che dovranno avere la capacità di gestire l'eterogeneità delle fonti di

acquisizione dati. L'assegnazione dei DT alle pubblicazioni è un compito svolto dagli editori e/o dai gestori dei database bibliometrici che le indicizzano (Donner, Document type assignment accuracy in the journal citation index data of Web of Science, 2017).

Inoltre, in relazione alla tipologia di documento, è importante sottolineare che per WoS è possibile assegnare contemporaneamente, alla stessa pubblicazione, molteplici e spesso contraddittori Document Type (Mokhnacheva, 2023). Per esempio, lo stesso documento può essere classificato simultaneamente come *articolo di ricerca* e *editorial material*. Al contrario, Scopus non prevede l'assegnazione multipla su uno stesso documento.

## Capitolo II

### **2. ANALISI DELL'ACCURATEZZA NELLA CLASSIFICAZIONE DELLE TIPOLOGIE DI DOCUMENTO**

#### **2.1 IDENTIFICAZIONE DEL PROBLEMA E OBIETTIVO DELLA RICERCA**

Per assistere i ricercatori a selezionare e recuperare in modo efficiente le informazioni rilevanti e per aiutare il calcolo di varie statistiche bibliometriche, i database bibliometrici categorizzano le pubblicazioni scientifiche in base al loro tipo di documento, categorie predefinite che ne delineano la natura e le caratteristiche principali (Yeung, 2021).

Una classificazione inaccurata delle pubblicazioni nei database può generare notevoli complicazioni pratiche, inclusa la distorsione delle statistiche bibliometriche a vari livelli di aggregazione, come intere riviste scientifiche o gruppi di ricerca all'interno di istituzioni e dipartimenti. Inoltre, questo fenomeno può danneggiare i ricercatori stessi durante la valutazione dei loro risultati di ricerca o, ad esempio, nelle procedure di promozione accademica. Questo è dovuto al fatto che non tutti i DT contribuiscono egualmente alla costruzione di indicatori e valutazioni bibliometriche. Ad esempio, i DT considerati meno

"prestigiosi", come *atti di conferenze, note brevi e lettere*, possono essere addirittura esclusi in alcune valutazioni.

L'assegnazione errata del DT può derivare dall'assenza di regole standardizzate, che determina la soggettività nel processo di assegnazione del DT. Inoltre, le regole di assegnazione utilizzate sia dalle banche dati bibliometriche sia dagli editori, non sono completamente trasparenti, né lo sono i metadati inviati dagli editori alle banche dati bibliometriche (Donner, Document type assignment accuracy in the journal citation index data of Web of Science, 2017). Dunque, dal momento che l'assegnazione del DT non è sempre accurata, sono presenti casi comuni di errori di etichettatura dei contributi scientifici, come l'errata classificazione di *articoli di ricerca* anziché *recensioni* oppure *lettere e note brevi* anziché *articoli di ricerca*, ecc.

Come brevemente presentato nel *Capitolo I, paragrafo 1.3.1 e 1.3.2* di questa trattazione, ci sono spesso discrepanze nella nomenclatura e nelle definizioni dei DT tra i diversi editori e database. Queste discrepanze diventano evidenti quando si confrontano gli elenchi ufficiali dei DT dei principali database generalisti, come Scopus e Web of Science (Elsevier, 2024) (Clarivate, 2024).

In letteratura sono documentati relativamente pochi studi sugli errori di attribuzione del DT, che fanno parte di una categoria di ricerca più ampia sugli errori nei database bibliometrici (Donner, 2023); (Franceschini et al., 2013; 2015); (García-Pérez, 2010); (Moed, 2005); (Olensky, Schmidt, & van Eck, 2016); (Valderrama-Zurián, Aguilar, Melero-Fuentes, & Aleixandre-Benavent, 2015). Tali studi presentano principalmente due limitazioni: (1) si basano su un campione di poche centinaia o migliaia di documenti e (2) richiedono dispendiose analisi manuali.

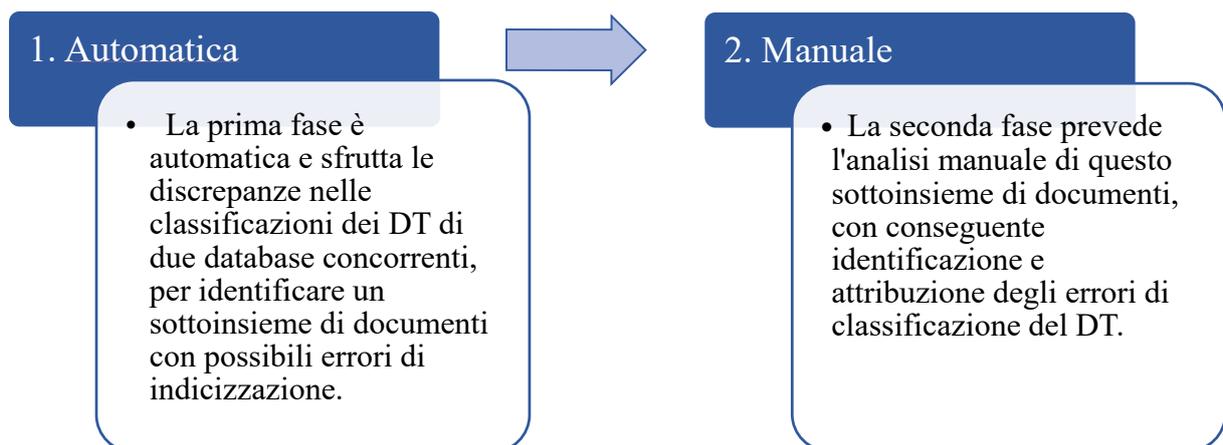
Questa ricerca, la cui metodologia è stata presentata alla conferenza “6th International Conference on Quality Engineering and Management” (Maisano, Mastrogiacomo, Ferrara, & Franceschini, 2024), amplia la letteratura esistente introducendo una nuova metodologia semi-automatica per l'analisi di un numero relativamente elevato di assegnazioni di DT da parte dei database bibliometrici Scopus e WoS. In particolare, tale ricerca fornisce un duplice contributo:

- (i) estende la dimensione del campione in analisi a decine di migliaia di documenti analizzati;
- (ii) utilizza una procedura adattiva che non richiede la verifica delle assegnazioni di DT per tutti i documenti, ma si concentra su quelli in cui le banche dati mostrano discrepanze.

## 2.2 METODOLOGIA

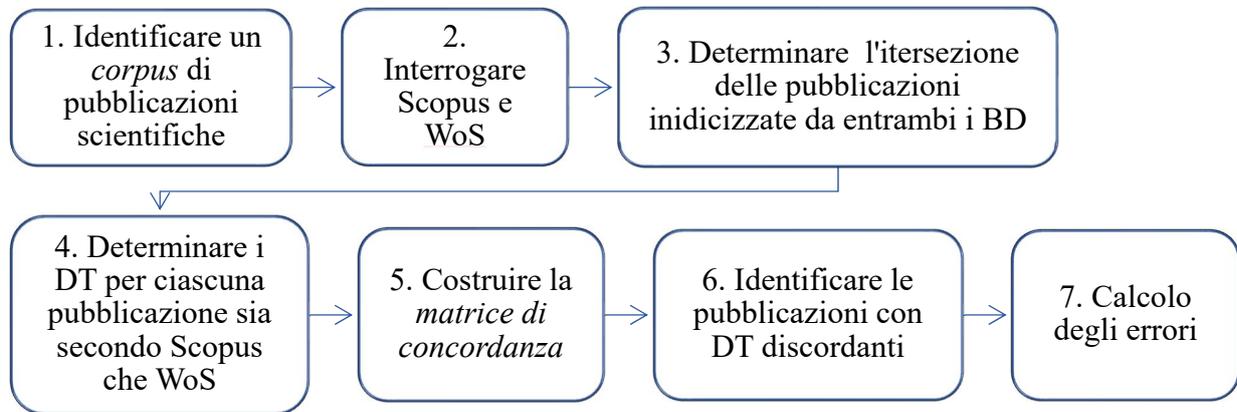
La metodologia proposta può essere definita semi-automatica (*Figura 10*) perché prevede una prima fase di analisi automatica di una grande quantità di documenti, seguita da un'analisi manuale di un sottoinsieme relativamente piccolo di documenti, con assegnazioni di DT potenzialmente errate, al fine di identificare l'errore e attribuirlo al database responsabile o in alcuni casi a entrambi i database.

La procedura utilizza le discrepanze tra le banche dati Scopus e WoS come indicatori di anomalie ed è ispirata a un precedente lavoro sullo studio semi-automatico degli errori delle banche dati bibliometriche (Franceschini , Maisano, & Mastrogiacomo, A novel approach for estimating the omitted citation rate of bibliometric databases with an application to the field of bibliometrics, 2013) (Franceschini, Maisano, & Mastrogiacomo, Empirical analysis and classification of database errors in Scopus and Web of Science, 2016).



*Figura 10 - Fasi della metodologia*

In *Figura 11* e *Figura 12* viene presentato il processo operativo composto dai sette passaggi da implementare per svolgere l'analisi.



*Figura 11 - Step metodologici (Step 1-5 analisi automatica; Step 6-7 analisi manuale)*

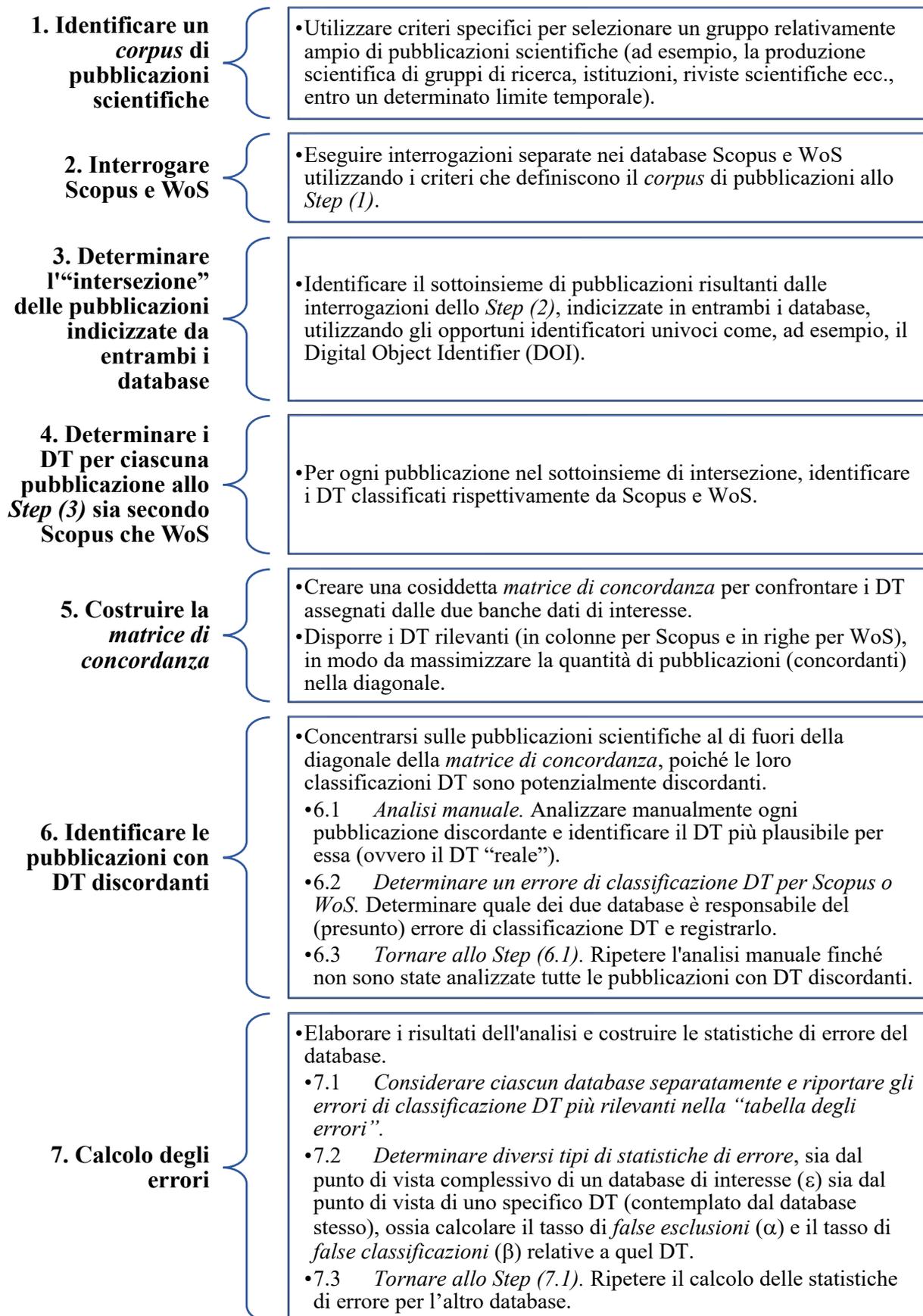


Figura 12 - Descrizione degli step metodologici

## 2.3 APPLICAZIONE DEGLI STEP METODOLOGICI

In questa sezione, verranno applicati gli step metodologici precedentemente descritti nel *paragrafo 2.2* in *Figura 12*. La metodologia sarà illustrata attraverso un esempio pratico, il cui *corpus* di documenti analizzato comprende diverse decine di migliaia di pubblicazioni scientifiche e garantisce una copertura di un ampio spettro di discipline.

### ***Step 1: Identificare un corpus di pubblicazioni scientifiche.***

In primo luogo, è necessario definire i criteri di selezione del *corpus*. Il primo passo consiste nell'individuare un campione relativamente ampio di pubblicazioni scientifiche e, in questo contesto, sono state identificate le due principali università di Torino: il Politecnico di Torino e l'Università di Torino. È stato inoltre determinato un limite temporale compreso tra il 2019 e il 2023. Di conseguenza, sono state prese in considerazione le pubblicazioni prodotte durante questi cinque anni da più di duemila ricercatori affiliati a queste due università.

### ***Step 2: Interrogare Scopus e WoS.***

Successivamente, vengono effettuate interrogazioni separate a Scopus e WoS per estrarre i dati sulle pubblicazioni di interesse, in base ai criteri definiti nello *Step 1*. Ciò produce due insiemi di pubblicazioni, un insieme restituito da Scopus e uno da WoS. In particolare, da Scopus sono stati ottenuti circa 34.000 documenti, mentre da WoS ne sono stati ottenuti 38.000. È fondamentale sottolineare che i due insiemi non hanno identiche dimensioni poiché le basi di dati considerate differiscono in termini di copertura delle fonti (ad esempio, un database potrebbe indicizzare alcune riviste o atti di convegni che non sono necessariamente indicizzati dall'altro database e viceversa).

**Step 3: Determinare l'“intersezione” delle pubblicazioni indicizzate da entrambi i database.**

Utilizzando gli identificatori univoci appropriati, come ad esempio il codice DOI<sup>1</sup>, si identifica il sottoinsieme di pubblicazioni presenti in entrambi i database, ovvero i documenti comuni sia a Scopus che a WoS. Le pubblicazioni prive di codice DOI sono state escluse dall'analisi in quanto, per queste, non è stato possibile verificare l'“intersezione”. Incrociando i due set di documenti restituiti da Scopus e WoS si ottiene un sottoinsieme di intersezione, composto da 26,405 pubblicazioni totali indicizzate da entrambi i database (*Tabella 3*).

**Step 4: Determinare i DT per ciascuna pubblicazione allo Step 3 sia secondo Scopus che WoS.**

Per ogni pubblicazione nel sottoinsieme di intersezione, sono stati identificati i DT classificati rispettivamente da Scopus e WoS. Nell'esempio applicativo, l'analisi è stata circoscritta ai DT relativi agli articoli di ricerca (*research article*), atti di conferenza (*conference/proceedings paper*), lettere (*letter*) e recensioni (*review*), tralasciando deliberatamente altre categorie come capitoli di libro (*book chapters*) o monografie (*monographs*), al fine di semplificare il processo.

Attraverso l'incrocio dei due insiemi di documenti ottenuti da Scopus e WoS, è stato generato un sottoinsieme di intersezione composto da un totale di 26.405 documenti (*Step 3*), ciascuno classificato nei rispettivi DT dai rispettivi database, come mostrato nella *Tabella 3*.

---

<sup>1</sup> DOI, acronimo di Digital Object Identifier, in italiano “identificatore di un oggetto digitale”, consente di identificare in modo univoco oggetti digitali e di accedervi in modo affidabile <https://www.doi.org/index.html>

Tabella 3 - Sintesi dei DT classificati da Scopus e WoS con riferimento al sottoinsieme "intersezione" nell'esempio applicativo

Scopus DT		WoS DT	
<i>Articolo</i>	21.793	<i>Articolo</i>	21.800
<i>Review</i>	2.445	<i>Review</i>	2.472
<i>Conference paper</i>	1.810	<i>Proceedings paper</i>	1.756
<i>Lettera</i>	357	<i>Lettera</i>	377
<b>Totale</b>	<b>26.405</b>	<b>Totale</b>	<b>26.405</b>

Le quantità di pubblicazioni classificate nei DT delle due banche dati sono molto vicine. Ad esempio, se consideriamo la riga relativa ad *articolo* in *Tabella 3*, vediamo che risultano esserci 21.793 articoli classificati in Scopus, mentre per WoS sono 21.800. Tale differenza indica possibili discrepanze nella classificazione dei DT, che verranno meglio evidenziate nella cosiddetta *matrice di concordanza*, introdotta nello *Step 5*.

Inoltre, bisogna evidenziare la differente nomenclatura delle tipologie di documento. In questo caso, le etichette dei DT dei due database in *Tabella 3* coincidono, ad eccezione del termine *conference paper*, utilizzato da Scopus, e *proceedings paper*, utilizzato da WoS.

Se infatti si espande lo studio anche ad altre tipologie di documento più specializzate, lo scenario risulta più complesso e le differenze tra le basi di dati più pronunciate. Per esempio, si possono osservare DT coperti da un database ma non dall'altro, come *biographical-item*, *book review*, *expression of concern*, *fiction*, *meeting abstract*, *retraction*, *data paper* e molti altri. (Clarivate, 2024) (Elsevier, 2024).

**Step 5: Costruire la matrice di concordanza.**

Questo step prevede di realizzare una cosiddetta *matrice di concordanza* per confrontare i DT assegnati dalle due banche dati. Si tratta di una tabella in cui vengono disposti i DT rilevanti, delle pubblicazioni di interesse, in colonne per Scopus e in righe per WoS.

Le righe e le colonne sono ordinate in modo tale che i totali delle righe e i totali delle colonne siano entrambi ordinati in modo decrescente, permettendo così alla maggior parte degli elementi di essere posizionati sulla diagonale principale. Allo stesso tempo, questo tipo di diagonalizzazione stabilisce una corrispondenza empirica tra i DT trattati da uno e dall'altro database. Questa corrispondenza può essere definita “adattiva” in quanto non richiede alcun legame a priori tra i DT di Scopus e WoS (Owen, 2001). In altre parole, tra le possibili permutazioni tra i DT di Scopus e WoS, la diagonalizzazione proposta assicura il massimo grado di concordanza.

Tabella 4 - Esempio della matrice di concordanza

Classificazione DT		da Scopus				Totale riga
		Article	Review	Confer. paper	Letter	
da WoS	Article	[21.418]	297	71	14	<b>21.800</b>
	Review	323	[2.146]	2	1	<b>2.472</b>
	Proceed. paper	19		[1.737]	-	<b>1.756</b>
	Letter	33	2	-	[342]	<b>377</b>
	Totale colonna	<b>21.793</b>	<b>2.445</b>	<b>1.810</b>	<b>357</b>	<b>26.405</b>

In *Tabella 4* è rappresentata la *matrice di concordanza* relativa ai dati raccolti per l'esempio applicativo. Mentre gli elementi nella diagonale (tra parentesi quadre) rappresentano classificazioni del DT che sono concordanti tra Scopus e WoS, quelli fuori dalla diagonale indicano classificazioni discordanti dei DT, indicando potenziali errori di classificazione da parte di almeno una banca dati.

### **Step 6: Identificare le pubblicazioni con DT discordanti.**

Questo step prevede un processo di analisi manuale iterativo volto a identificare il “reale” DT della pubblicazione che, in *Tabella 4*, si trovano al di fuori della diagonale. Si vuole determinare quale database è responsabile del presunto errore di classificazione e ripetere l’analisi finché non sono stati analizzati tutti i documenti con DT discordanti. Questo ciclo iterativo, nonostante sia un processo manuale e richieda tempo, assicura un’analisi accurata di tutte le pubblicazioni, con classificazioni del DT potenzialmente discordanti, e permette di individuare e correggere gli errori di classificazione nei database.

Per quanto riguarda l’esempio presentato, è stata eseguita l’analisi manuale di ciascuna delle pubblicazioni discordanti (*sotto-step 6.1*) esaminando ciascun documento (abstract, struttura, keywords, ecc.) e sfruttando le informazioni disponibili sul sito dell’editore.

In particolare, sono stati analizzati manualmente 762 documenti, numero dato dalla somma degli elementi fuori dalla diagonale in *Tabella 4*. Per gli elementi sulla diagonale (tra parentesi quadre), le classificazioni del DT in entrambi i database sono automaticamente considerate corrette, dato l’accordo generale tra i database corrispondenti. Per testare la plausibilità di questa ipotesi, è stata condotta un’analisi di un campione dello 0,5% di questi documenti concordanti (ovvero  $[0,5\% \cdot 25.643] = 129$ ), senza rilevare alcun errore di classificazione DT.

Nel *Capitolo 3* al paragrafo 3.2 di questa trattazione verranno presentati esempi di documenti analizzati sia con errori di classificazione da parte di un solo database sia con errori di classificazione su entrambe le banche dati simultaneamente.

### Step 7: Calcolo degli errori.

Dopo aver completato l'analisi manuale dei documenti discordanti (*Step 6* e relativi *sotto-step*), vengono elaborati i risultati e costruite le statistiche di errore. Per ciascun database i risultati vengono riassunti in una *tabella degli errori*, ovvero una tabella che illustra nelle colonne le assegnazioni DT fatte dal database di interesse e nelle righe i DT "reali" risultanti dall'analisi manuale. In pratica, la diagonale principale della tabella degli errori contiene le classificazioni DT corrette, mentre gli elementi fuori dalla diagonale corrispondono alle classificazioni DT errate.

*Tabella 5 e Tabella 6* mostrano le *tabelle degli errori* relative a Scopus e WoS, con riferimento all'esempio applicativo. Si noti che per una piccola parte degli errori di classificazione, il "vero" DT appartiene alla categoria denominata *other*, che comprende tutti gli altri DT non trattati in questo studio semplificato (ad esempio, recensioni di libri, note, brevi sondaggi, ecc.). In altre parole, la categoria *other* raccoglie i DT che sono diversi dai quattro DT (*article*, *proceedings paper*, *letter* e *review*) considerati in questa ricerca. Le statistiche di errore verranno spiegate nel *Capitolo 3, paragrafo 3.1*.

Tabella 5 - Esempio di tabella degli errori per Scopus

		Classificazione DT da Scopus				Totale riga	$\alpha_{\text{Scopus, DT}}$
		<i>Article</i>	<i>Review</i>	<i>Confer. paper</i>	<i>Letter</i>		
Classificazione DT "Vero"	<i>Article</i>	[21.503]	102	53		21.658	<b>0,7%</b>
	<i>Review</i>	233	[2.331]			2.564	<b>9,1%</b>
	<i>Confer. paper</i>	20	5	[1.757]		1.782	<b>1,4%</b>
	<i>Letter</i>	30	2		[357]	389	<b>8,2%</b>
	<i>Other DT(s)</i>	7	5			12	-
Totale colonna		21.793	2.445	1.810	357	26.405	
$\beta_{\text{Scopus, DT}}$		<b>1,3%</b>	<b>4,7%</b>	<b>2,9%</b>	<b>0,0%</b>		$\epsilon_{\text{Scopus}} \cong$ <b>1,7%</b>

Tabella 6 - Esempio di tabella degli errori per Web of Science

		Classificazione DT da WoS				Totale riga	$\alpha_{\text{WoS}, DT}$
		<i>Article</i>	<i>Review</i>	<i>Confer. paper</i>	<i>Letter</i>		
Classificazione DT "Vero"	<i>Article</i>	[21.584]	75		3	21.662	<b>0,4%</b>
	<i>Review</i>	176	[2.386]			2.562	<b>6,9%</b>
	<i>Proc. paper</i>	23	3	[1.756]		1.782	<b>1,5%</b>
	<i>Letter</i>	16	1		[374]	391	<b>4,3%</b>
	<i>Other DT(s)</i>	5	7			12	-
Totale colonna		21.804	2.472	1.756	377	26.405	
<b><math>\beta_{\text{WoS}, DT}</math></b>		<b>1,0%</b>	<b>3,5%</b>	<b>0,0%</b>	<b>0,8%</b>		<b><math>\varepsilon_{\text{WoS}} \cong 1,2\%</math></b>

## Capitolo III

### 3. RISULTATI

#### 3.1 ANALISI DEGLI ERRORI

Le analisi sono costruite a partire dalle *tabelle degli errori* (Tabella 5 e Tabella 6) presentate nello Step 7 del *Capitolo 2, paragrafo 2.3*. Queste tabelle forniscono una panoramica dettagliata delle discrepanze tra le classificazioni dei documenti nei due database (i.e. Scopus e WoS), permettendo di identificare e quantificare le diverse tipologie di errori, limitatamente ai dati usati nell'esempio applicato. L'obiettivo è approfondire la natura e la frequenza di tali errori, al fine di migliorare l'accuratezza delle classificazioni bibliometriche e, di conseguenza, la validità degli indicatori utilizzati per la valutazione della ricerca.

In particolare, sono stati costruiti alcuni indicatori che identificano gli errori nei database presi in considerazione. Un indicatore complessivo degli errori di classificazione del DT da parte di un certo database ( $\varepsilon_{db}$ ) è espresso dal rapporto tra il numero totale di elementi fuori diagonale (ovvero che rappresentano classificazioni errate del DT risultanti dall'analisi manuale) e tutti gli elementi della *tabella degli errori* (cioè il numero totale di classificazioni del DT):

$$\varepsilon_{db} = \frac{\sum_{\forall(i,j) \mid i \neq j} d_{db,(i,j)}}{\sum_{\forall(i,j)} d_{db,(i,j)}}, \quad (1)$$

dove  $d_{db,(i,j)}$  è numero di documenti riportati nella  $i$ -esima riga e nella  $j$ -esima colonna della tabella degli errori, per il database di interesse ( $db$ ).

I valori di  $\varepsilon_{db}$  risultanti dall'analisi, riportati anche nel vertice in basso a destra delle *tabelle degli errori* in *Tabella 5* e *Tabella 6*, sono  $\varepsilon_{\text{Scopus}} \cong 1,7\%$  e  $\varepsilon_{\text{WoS}} \cong 1,2\%$ . Nonostante questi risultati siano difficilmente generalizzabili, poiché si riferiscono a un campione piuttosto ridotto di categorie di documenti (*article, proceedings paper, letter* e *review*), è interessante notare alcune tendenze generali.

Altre statistiche più dettagliate sugli errori del database possono essere costruite dal punto di vista dei singoli DT. A questo proposito, viene definita l'*ipotesi nulla* ( $H_0$ ) secondo cui: "*Il DT assegnato dal database in uso è corretto*" e si definiscono due tipi di errore legati al DT: *Type-I error* e *Type-II error*.

*Type-I error* corrisponde alla probabilità di rifiutare erroneamente  $H_0$  quando è vera, ovvero alla probabilità che un documento appartenente a uno specifico DT venga erroneamente classificato in un altro DT. In altre parole, si tratta della falsa esclusione dal DT di interesse:

$$\alpha_{db,DT} = \frac{\sum_{\forall j | j \neq DT} d_{db,(DT,j)}}{\sum_{\forall j} d_{db,(DT,j)}}, \quad (2)$$

dove  $DT$  è il DT per il database di interesse ( $db$ ), mentre  $\alpha_{db,DT}$  è calcolato riga per riga attraverso il rapporto tra gli errori di classificazione (ovvero le false esclusioni dal DT di interesse) e il totale della riga.

*Type-II error* corrisponde alla probabilità di non scartare  $H_0$  quando è falsa, ovvero alla probabilità di classificare erroneamente un documento in uno specifico DT di interesse. Questo porta alla falsa classificazione nel DT di interesse:

$$\beta_{db, DT} = \frac{\sum_{\forall i | i \neq DT} d_{db, (i, DT)}}{\sum_{\forall i} d_{db, (i, DT)}}, \quad (3)$$

$\beta_{db, DT}$  è calcolato colonna per colonna come rapporto tra gli errori di classificazione (ovvero le false classificazioni nel DT di interesse) e il totale della colonna. È interessante notare che lo stesso elemento ( $d_{db, (i, j)}$ ) nella *tabella degli errori* contribuisce a entrambi le tipologie di errore.

Con riferimento all'esempio applicativo, i valori di  $\alpha_{db, DT}$  e  $\beta_{db, DT}$  sono riportati rispettivamente nella parte destra e nella parte bassa delle corrispondenti *tabelle di errore* (*Tabella 5 e Tabella 6*).

La mappa in *Figura 13* fornisce una rappresentazione grafica delle statistiche sugli errori di cui sopra in *Tabella 5 e Tabella 6*. Le linee tratteggiate rappresentano i valori  $\varepsilon_{Scopus}$  e  $\varepsilon_{WoS}$ , mentre per ogni DT di un determinato database, viene tracciato un punto di coordinate corrispondente ( $\alpha_{db, DT}, \beta_{db, DT}$ ).

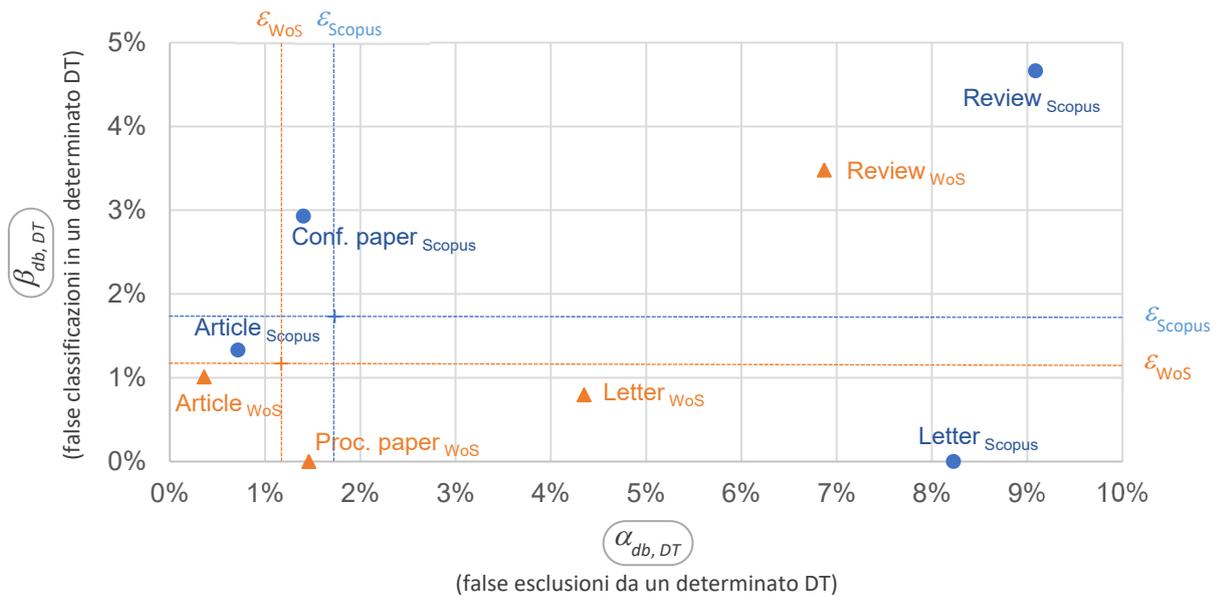


Figura 13 - Mappa delle statistiche degli errori nel corpus analizzato

## 3.2 TIPOLOGIE DI ERRORI

Le tipologie di errori nella classificazione dei documenti nei database bibliometrici possono variare significativamente. Ai fini della tesi sono state identificate principalmente due tipologie di classificazione errata: (1) errore di classificazione da parte di un solo database, (2) errore di classificazione simultaneo in entrambi i database.

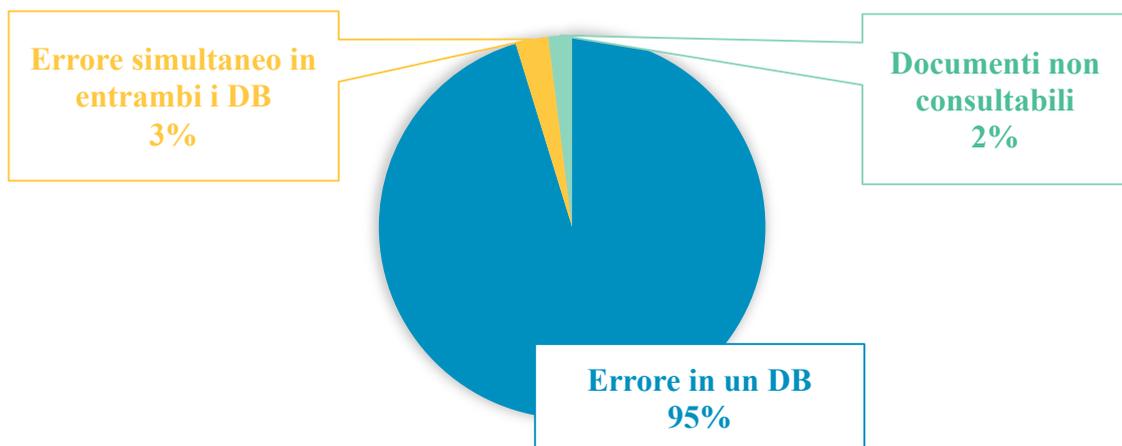


Figura 14 - Tipologie di errore riscontrate nel corpus analizzato

Inoltre, come rappresentato nel grafico in *Figura 14*, sulla base dell'esempio applicato, sono presenti un numero di documenti a cui non si è potuto accedere. Si tratta di documenti che non è stato possibile consultare per verificare manualmente la tipologia di pubblicazione. Tale problematica è principalmente dovuta al fatto che i documenti sono pubblicati su riviste per le quali è richiesta una licenza aggiuntiva o per errori di collegamento al DOI di riferimento (*Figura 15*).

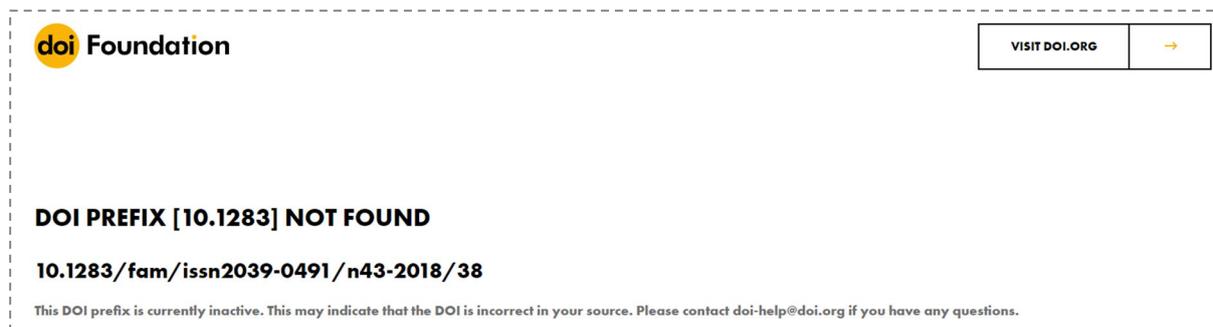


Figura 15 - Esempio di documento non accessibile per DOI inattivo (<https://www.doi.org/>)

Le percentuali presentate nel grafico a torta in *Figura 14*, sono state calcolate prendendo in considerazione il numero di documenti in cui era presente il singolo errore (o l'errore simultaneo o la mancata verifica) rapportato al numero totale di elementi discordanti (ovvero quelli fuori dalla diagonale nella *matrice di concordanza* presentata allo *Step 5* della metodologia). Ad esempio, il valore di circa 95%, caso in cui è stato rilevato l'errore in un solo DB, indipendentemente se associato a Scopus o WoS, è il risultato del rapporto tra 726 (numero di pubblicazioni con errore di classificazione su un singolo database) e 762 (numero totale di elementi discordanti identificati attraverso la *matrice di concordanza*).

### 3.2.1 Esempi di errori e anomalie nel *corpus* analizzato

Tra gli esempi di documenti con errore di classificazione su un database, la *Figura 16* esemplifica l'intestazione di una pubblicazione classificata da Scopus come *documento di conferenza* e da WoS come *articolo*. L'analisi manuale ha rivelato che il "vero" DT è *articolo*, poiché il documento è un contributo di ricerca completo, non legato ad alcuna conferenza. In questo caso, quindi, l'errore di classificazione del DT è attribuito a Scopus.



Figura 16 - Esempio di documento classificato erroneamente da Scopus come documento da conferenza anziché articolo (<https://doi.org/10.1016/j.ejor.2018.12.016>).

Altro esempio di documento appartenente alla categoria di errori più riscontrata è presentato in Figura 17. La pubblicazione è stata indicizzata da Scopus come *articolo* e da WoS come *review*. Come esplicitato però nel titolo e dichiarato dagli autori stessi nell'*abstract*, la pubblicazione ha lo scopo principale di esaminare lo stato dell'arte e offrire un quadro di riferimento per impostare correttamente le future attività di ricerca. L'errore di classificazione del DT è quindi attribuito a Scopus che considera erroneamente la pubblicazione come *articolo* e non come *review*.



Contents lists available at ScienceDirect

Journal of Wind Engineering &amp; Industrial Aerodynamics

journal homepage: [www.elsevier.com/locate/jweia](http://www.elsevier.com/locate/jweia)

## Windblown sand along railway infrastructures: A **review** of challenges and mitigation measures

Luca Bruno<sup>a,b,\*</sup>, Marko Horvat<sup>a</sup>, Lorenzo Raffaele<sup>a,b</sup><sup>a</sup> Politecnico di Torino, Department of Architecture and Design, Viale Mattioli 39, I-10125, Torino, Italy<sup>b</sup> Windblown Sand Modelling and Mitigation Joint Research Group, Italy

## ARTICLE INFO

## Keywords:

Windblown sand  
Sand limit states  
Sand mitigation measures  
Railway infrastructures

## ABSTRACT

The engineering interest about windblown sand is dictated by the harmful interactions between sand and a number of human infrastructures in arid environments. Particularly, the ongoing grand railway projects in the deserts of Far East, Middle East and North Africa regions require robust technical solutions to guarantee the efficient railway performance. The huge competences of the railway industry, traditionally developed in non-arid regions, should be developed and complemented to face mentioned challenges. The rationale problem setting, design, quantitative analysis and verification of sand mitigation measures are at present not sufficiently developed. The paper introduces original categorizations of both the windblown sand-induced performance deficiencies of the railway systems (windblown Sand Ultimate and Serviceability Limit States) and the prevention techniques to mitigate the windblown sand effects (Source-Path-Receiver categorization of the Sand Mitigation Measures). The state of the art is reviewed in the attempt to present the classification as accurately as possible. The main goal of the classification is to provide an orienting framework for scholars, railway owners, designers, general contractors and operators. We suggest the presented framework as a structured and organic base to properly set up future research activities, project terms of reference, most suited design solution, plan maintenance practices.

Figura 17 - Esempio di documento classificato erroneamente da Scopus come articolo anziché review (doi: 10.1016/j.jweia.2018.04.021)

Un'altra categoria che possiamo considerare dubbia e soggetta a errori è quella dei documenti/atti di conferenza. Infatti, come mostrato in Figura 18, il documento è classificato da Scopus come *articolo* e da WoS come *proceedings paper*. Emergono così le discrepanze e i sottili confini presentati dalle definizioni dei DT nei database (riferimenti al *Capitolo I – paragrafo 1.3*). Il documento è pubblicato su un *journal special issue* e tratta una ricerca presentata ad una conferenza internazionale. Dunque, stando alla definizione pubblicata da Scopus, il documento dovrebbe rientrare nella categoria *conference paper* che recita: “*original article reporting data presented at a conference or symposium*”. Risulta essere quindi un errore di classificazione da parte di Scopus. WoS, invece, lo indicizza correttamente come *proceedings paper* come proposto dalla sua definizione.



Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

ScienceDirect

Procedia Manufacturing 19 (2018) 119–126

Procedia  
MANUFACTURING

[www.elsevier.com/locate/procedia](http://www.elsevier.com/locate/procedia)

6th International Conference on Through-life Engineering Services, TESConf 2017, 7-8  
November 2017, Bremen, Germany

## Cost benefit evaluation of maintenance options for aging equipment using monetised risk values: a practical application

M.C. Leva<sup>a</sup>, G. Baldissone<sup>b</sup>, R. Caso<sup>b</sup>, M. Demichela<sup>b</sup>, L. Lawlor<sup>c</sup>, B. McAleer<sup>c\*</sup>

<sup>a</sup>Dublin Institute of Technology Health And Safety Department, Dublin, Ireland

<sup>b</sup>Politecnico di Torino, DISAT Dipartimento Scienza Applicata E Tecnologia, Torino, Italy, <sup>c</sup>ESB Electricity Supply Board, Dublin, Ireland

### Abstract

With constant pressure to reduce maintenance costs as well as short-term budget constraints in a changing market environment, asset managers are compelled to continue operating aging assets while deferring maintenance and investment. The scope of the paper is to get an overview of the methods used to evaluate risks and opportunities for deferred maintenance interventions on aging equipment, and underline the importance to include monetised risk considerations and timeline considerations, to evaluate different scenarios connected with the possible options. Monetised risk values offer the opportunity to support risk-based decision-making using the data collected from the field. The paper presents examples of two different methods and their practical applicability in two case studies in the energy sector for a company managing power stations. The use of the existing and the new proposed solutions are discussed on the basis of their applicability to the concrete examples

© 2018 The Authors. Published by Elsevier B.V.

Peer-review under responsibility of the scientific committee of the 6th International Conference on Through-life Engineering Services.

**Keywords:** Cost benefit evaluation, monetised risk values,

Figura 18 - Esempio di documento classificato erroneamente da Scopus come articolo anziché conference paper (doi: 10.1016/j.promfg.2018.01.017)

Altro esempio relativo alla categoria *documenti/atti di conferenza* è rappresentato in *Figura 19*. Tale esempio propone, inoltre, una ulteriore anomalia/discrepanza, che solitamente WoS risolve con l'assegnazione multipla (riferimenti al *Capitolo I – paragrafo 1.3*). Il documento presentato è classificato da Scopus come *articolo* e da WoS come *review*. L'errore di Scopus, indicizzarlo come articolo, potrebbe derivare dal fatto che si tratti di un *journal special issue*. Tuttavia, però il documento è stato presentato ad una conferenza internazionale e dovrebbe essere indicizzato da Scopus come *conference paper* o al più, come da definizione di *review* di Scopus "*significant review of original research, also includes conference papers*". Altro scenario è invece legato a Web of Science che lo indicizza come *review*. In effetti, sia nel titolo

che nella conclusione è presente la parola *review*. Facendo riferimento ad una nota presente nella definizione di *review* di WoS: “*if an article is not assigned a review by the journal but Review, Systematic Review or Mini review appears in the title, it must also appear someplace else in the article in order to be assigned the document type review*”. Classificarlo quindi come *review* risulta corretto secondo la definizione. Per completezza, però, dato che WoS consente una molteplice assegnazione ad un singolo documento, il database avrebbe dovuto anche indicizzarlo come *proceedings paper* poiché il documento è stato presentato alla 6<sup>th</sup> International Entomophagous Insects Conference.



## Entomologia Experimentalis et Applicata



DOI: 10.1111/eea.12967

**SPECIAL ISSUE: 6TH INTERNATIONAL ENTOMOPHAGOUS INSECTS  
CONFERENCE**

### Biological control of invasive stink bugs: review of global state and future prospects

Eric Conti<sup>1\*</sup> , Gonzalo Avila<sup>2,3</sup> , Barbara Barratt<sup>3,4</sup> , Fernanda Cingolani<sup>5</sup> , Stefano Colazza<sup>6</sup> , Salvatore Guarino<sup>7</sup> , Kim Hoelmer<sup>8</sup> , Raul Alberto Laumann<sup>9</sup> , Lara Maistrello<sup>10</sup> , Guillaume Martel<sup>11,12</sup> , Ezio Peri<sup>6</sup> , Cesar Rodriguez-Saona<sup>13</sup> , Gabriele Rondoni<sup>1</sup> , Michael Rostás<sup>14</sup> , Pio Federico Roversi<sup>15</sup> , René F.H. Sforza<sup>11</sup> , Luciana Tavella<sup>16</sup>  & Eric Wajnberg<sup>17,18</sup> 

<sup>1</sup>Department of Agricultural, Food and Environmental Sciences, University of Perugia, Borgo XX Giugno, Perugia 06121, Italy, <sup>2</sup>The New Zealand Institute for Plant & Food Research Limited, Private Bag 92169, Auckland 1142, New Zealand, <sup>3</sup>Better Border Biosecurity, Christchurch New Zealand, <sup>4</sup>AgResearch, Invermay Research Centre, Private Bag 50034, Mosgiel 9053, New Zealand, <sup>5</sup>CEPAVE (CONICET-UNLP), Boulevard 120 e/ 60 y 64 s/n, La Plata 1900, Argentina, <sup>6</sup>Dipartimento di Scienze Agrarie, Alimentari e Forestali, Università degli Studi di Palermo, Viale delle Scienze, edificio 5, Palermo 90128, Italy, <sup>7</sup>Institute of Biosciences and Bioresources (IBBR), National Research Council of Italy (CNR), Corso Calatafimi 414, Palermo 90129, Italy, <sup>8</sup>USDA-ARS, Beneficial Insects Introduction Research Unit, 501 S. Chapel St, Newark DE, USA, <sup>9</sup>Embrapa Recursos Genéticos e Biotecnologia. PqEB Avda W5 Norte (Final), CEP 70770-917, Brasília DF, Brazil, <sup>10</sup>Dipartimento di Scienze della Vita, Centro BIOGEST-SITEIA, Università di Modena e Reggio Emilia, Via G. Amendola 2, Reggio-Emilia 42122, Italy, <sup>11</sup>USDA-ARS-European Biological Control Laboratory, Campus International de Baillarguet, 810, avenue du Campus Agropolis, Montferrier-sur-Lez 34980, France, <sup>12</sup>Montpellier SupAgro, Place Pierre Viala, 2, Montpellier 34000, France, <sup>13</sup>Department of Entomology, Rutgers University, New Brunswick NJ, 08901, USA, <sup>14</sup>Agricultural Entomology, Department of Crop Sciences, University of Göttingen, Grisebachstr. 6, Göttingen 37077, Germany, <sup>15</sup>CREA Consiglio per la Ricerca in Agricoltura e l'analisi dell'economia agraria, Via di Lanciola, 12/a, Cascine del Riccio, Firenze 50125, Italy, <sup>16</sup>Dipartimento di Scienze Agrarie, Forestali e Alimentari (DISAFA), Università degli Studi di Torino, Largo Paolo Braccini 2, Grugliasco TO, 10095, Italy, <sup>17</sup>INRA, 400 Route des Chappes, BP 167, Sophia Antipolis Cedex 06903, France, and <sup>18</sup>INRIA, Sophia Antipolis, Project Hephaistos, 2004 Route des Lucioles, BP 93, Sophia Antipolis Cedex 06902, France

Accepted: 11 June 2020

**Key words:** biocontrol, chemical ecology, Hemiptera, invasive species, landscape management, pre-emptive classical biological control, semiochemicals, parasitoid, Pentatomidae, risk assessment, Tachinidae, Scelionidae

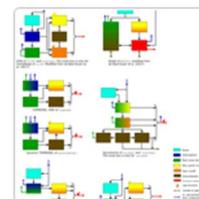
Figura 19 - Esempio di documento classificato erroneamente da Scopus come articolo anziché conference paper e da WoS come review (doi: 10.1111/eea.12967)

Nella trattazione sono state considerate quattro tipologie di documento, tuttavia dall'analisi manuale sono risultate, con corretta assegnazione, anche altre categorie non appartenenti a quelle in esame. Al fine di questo elaborato, tutti i documenti la cui corretta classificazione appartiene a categorie che non siano *articolo*, *review*, *atto di conferenza* o *lettera*, ricadono nella categoria *other DT(s)*, visibile in *Tabella 5* e *Tabella 6*.

Un esempio è in *Figura 20*. Il documento è definito come *technical note* dagli stessi autori nel titolo, ma l'esempio è classificato da Scopus come *review* e da WoS come *articolo*. Con riferimento alle definizioni presentate nel *Capitolo I - paragrafo 1.3*, una *technical note* è considerata da Scopus nella categoria degli *articoli*, nonostante abbia una lunghezza inferiore ad una pagina. Anche stando alle definizioni di WoS dovrebbe essere un *articolo* poiché si tratta di una *technical note* pubblicata in un *journal*. Tuttavia, il documento è stato erroneamente indicizzato da Scopus come *review*.

Nonostante, il contributo non abbia la consistenza e la struttura per essere considerato alla pari di un articolo (di ricerca), né fornisca una rassegna sufficientemente approfondita dello stato dell'arte per essere considerato una *review*, attenendoci alle definizioni presentate dagli editori, uno dei due database (Scopus), risulta comunque in errore. Secondo l'analisi manuale condotta i DT di Scopus e WoS più plausibili sarebbero rispettivamente *note* e *altro*, che convergerebbero nella classe *other DT(s)* nella presente analisi semplificata.

## Technical note: Hydrology modelling R packages – a unified analysis of models and practicalities from a user perspective



Paul C. Astagneau , Guillaume Thirel, Olivier Delaigue, Joseph H. A. Guillaume, Juraj Parajka, Claudia C. Brauer, Alberto Viglione, Wouter Buytaert, and Keith J. Beven

### Abstract

Following the rise of R as a scientific programming language, the increasing requirement for more transferable research and the growth of data availability in hydrology, R packages containing hydrological models are becoming more and more available as an open-source resource to hydrologists. Corresponding to the core of the hydrological studies workflow, their value is increasingly meaningful regarding the reliability of methods and results. Despite package and model distinctiveness, no study has ever provided a comparison of R packages for conceptual rainfall–runoff modelling from a user perspective by contrasting their philosophy, model characteristics and ease of use. We have selected eight packages based on our ability to consistently run their models on simple hydrology modelling examples. We have uniformly analysed the exact structure of seven of the hydrological models integrated into these R packages in terms of conceptual storages and fluxes, spatial discretisation, data requirements and output provided. The analysis showed that very different modelling choices are associated with these packages, which emphasises various hydrological concepts. These specificities are not always sufficiently well explained by the package documentation. Therefore a synthesis of the package functionalities was performed from a user perspective. This synthesis helps to inform the selection of which packages could/should be used depending on the problem at hand. In this regard, the technical features, documentation, R implementations and computational times were investigated. Moreover, by providing a framework for package comparison, this study is a step forward towards supporting more transferable and reusable methods and results for hydrological modelling in R.

*Figura 20 - Esempio di documento classificato da Scopus come review e da WoS come articolo (doi: 10.5194/hess-25-3937-2021)*

Altro esempio è quello presentato in *Figura 21*. La pubblicazione è stata indicizzata da Scopus come *articolo* e da WoS come *review*. Tuttavia, nel paper viene espressamente dichiarato che si tratta di una *minireview*. Tale categoria, stando alle definizioni fornite dai database, dovrebbe essere considerata come *short survey* per Scopus (“*short or mini-review of original research*”) e *review* per WoS (“*review includes review of literature, mini-review and systematic review*”). Ai fini di questa analisi questo documento ricadrà dunque nella riga other DTs riferita al “vero” DT nelle *tabelle degli errori*.



## Benign focal liver lesions: The role of magnetic resonance imaging

Marco Gatti, Cesare Maino, Davide Tore, Andrea Carisio, Fatemeh Darvizeh, Eleonora Tricarico, Riccardo Inchingolo, Davide Ippolito, Riccardo Faletti

**Specialty type:** Gastroenterology and hepatology

**Provenance and peer review:** Invited article; Externally peer reviewed

**Peer-review model:** Single blind

**Peer-review report's scientific quality classification**

Grade A (Excellent): 0  
Grade B (Very good): B  
Grade C (Good): 0  
Grade D (Fair): 0  
Grade E (Poor): 0

**P-Reviewer:** Gamarra LF, Brazil

**Received:** April 18, 2021

**Peer-review started:** April 18, 2021

**First decision:** July 27, 2021

**Revised:** August 7, 2021

**Accepted:** April 9, 2022

**Article in press:** April 9, 2022

**Published online:** May 27, 2022



**Marco Gatti, Davide Tore, Andrea Carisio, Riccardo Faletti,** Department of Surgical Sciences, University of Turin, Turin 10126, Italy

**Cesare Maino, Davide Ippolito,** Department of Diagnostic Radiology, Ospedale San Gerardo, Monza 20900, Italy

**Fatemeh Darvizeh,** School of Medicine, Vita-Salute San Raffaele University, Milan 20121, Japan

**Eleonora Tricarico,** Department of Radiology, "F. Perinci" Hospital, Altamura 70022, Italy

**Riccardo Inchingolo,** Interventional Radiology Unit, "F. Miulli" Regional General Hospital, Acquaviva delle Fonti 70021, Italy

**Davide Ippolito,** Department of Diagnostic Radiology, School of Medicine, University of Milano-Bicocca, Monza 20900, Italy

**Corresponding author:** Marco Gatti, MD, Research Fellow, Department of Surgical Sciences, University of Turin, Via Genova 3, Turin 10126, Italy. [marcogatti17@gmail.com](mailto:marcogatti17@gmail.com)

### Abstract

Liver lesions are common findings in radiologists' daily routine. They are a complex category of pathology that range from solitary benign lesions to primary liver cancer and liver metastases. Benign focal liver lesions can arise from different liver cell types: Epithelial (hepatocytes and biliary cells) and non-epithelial (mesenchymal cells). Liver magnetic resonance imaging (MRI) is a fundamental radiological method in these patients as it allows with its multiparametric approach optimal non-invasive tissue characterization. Furthermore, advanced liver MRI techniques such as diffusion-weighted imaging and hepatobiliary contrast agents have improved the detection of focal liver lesions and can be highly effective in differentiating pseudotumor from tumors, as well as benign from malignant lesions, and can also be used for differential diagnosis. Although histological examination can be useful in making a definitive diagnosis, MRI is an important modality in the diagnosis of liver lesions with a significant impact on patient care. This aim of this review is to provide a comprehensive overview of benign liver lesions on MRI.

**Key Words:** Magnetic resonance imaging; Liver neoplasms; Biliary tract; Hepatocytes

Figura 21 - Esempio di documento classificato da Scopus come articolo e da WoS come review (doi: 10.4254/wjh.v14.i5.923)

Un ulteriore esempio riguarda il documento in *Figura 22*, che esemplifica una pubblicazione nel campo della psichiatria forense, contenente un breve resoconto di un *caso clinico* con alcuni riferimenti alla letteratura esistente sull'argomento. Il contributo non ha la consistenza e l'originalità per essere considerato un *articolo di ricerca*, né fornisce una rassegna sufficientemente approfondita dello stato dell'arte per essere considerato una *review*. La rivista lo classifica infatti come un *case report*. Tuttavia, se si considerano le definizioni presentate dai DB, i *case report* sono considerati *articoli* sia per Scopus che per WoS. Pertanto, almeno un DB (in questo caso WoS) presenta un parziale errore di classificazione. Inoltre, è stato notato che Scopus tende a classificare le pubblicazioni di *case report* come *articoli*, mentre WoS tende a classificarle come *review*, con conseguente errore sistematico di classificazione.

L'errore legato a Web of Science nell'assegnazione di documenti come *review* può essere dovuto alla presenza di parole chiave come "*review*" o "*literature review*". Spesso questi termini, comunemente utilizzati nei titoli, negli abstract o nelle conclusioni per indicare una revisione della letteratura o una sintesi di casi, possono indurre l'algoritmo di classificazione automatica di WoS a categorizzare i *case report* come *review*. Tuttavia, l'analisi manuale condotta ha mostrato che i DT di Scopus e WoS più plausibili sarebbero rispettivamente *nota* e *altro*, che convergerebbero nella classe *other DT(s)* nella presente analisi semplificata.



## The risk of assault against mental health professionals: a fatal case report and literature review

Lucia Tattoli<sup>1</sup> · Caterina Bosco<sup>2</sup> · Ignazio Grattagliano<sup>3</sup> · Giancarlo Di Vella<sup>2</sup>

Accepted: 24 February 2019 / Published online: 26 April 2019  
© Springer Science+Business Media, LLC, part of Springer Nature 2019

### Abstract

Assaults by patients against healthcare providers are an increasing phenomenon worldwide. Mental health professionals in acute facilities and rehabilitation wards have the highest risk of being attacked at work. Verbal abuse or intimidating behaviors represent the most common types of violence. Fatal assault by psychiatric patients has been rarely reported in the literature. We present a case of a female psychiatrist who was fatally stabbed in her office in a Mental Health Center. At autopsy seventy stab wounds were found: four wounds of the neck, fifty penetrating wounds of the thorax, three wounds of the abdomen, six wounds of the lumbar region, and seven wounds of the upper arms including defense injuries. The cause of death was massive blood loss due to multiple stab wounds. The perpetrator was a 44-year-old male patient who had been referred to the victim after a previous admission to hospital following experiences of suicidal ideation and confusion. The extreme and unmotivated violence in a non-acute setting were notable. A borderline-antisocial personality disorder was later diagnosed by forensic experts. This case emphasizes the significant occupational risk for mental healthcare staff to sustain life threatening injuries or death, with implications for training of clinicians, and strategies for preventing aggressive behaviors.

**Keywords** Workplace violence · Psychiatry · Violence · Stab wounds

*Figura 22 – Esempio di documento classificato da Scopus come articolo e da WoS come review (doi: 10.1007/s12024-019-00105-6)*

Attraverso gli esempi presentati si vuole evidenziare come le ambiguità nelle definizioni e le discrepanze, sia interne allo stesso database che tra database avversari, rendano l'analisi ancora più complessa. Una modalità adottata da WoS per far fronte a queste ambiguità, come accennato nel *Capitolo I - paragrafo 1.3.2*, è quella di fornire una molteplice assegnazione di DT ad una singola pubblicazione. Tale metodo però, seppure apparentemente risolutivo, oltre a generare discrepanze di classificazione nei confronti del database competitor, crea anche “errori intrinseci” sullo stesso WoS. Infatti, alcuni errori di classificazione in WoS sono dovuti alla mancanza di aderenza alle regole e definizioni presentate dal database stesso, come principalmente visto negli esempi in *Figura 19* e *Figura 20*.

## CONCLUSIONI

Questa trattazione ha come obiettivo quello di presentare una nuova metodologia di analisi delle tipologie di documento presenti nei database bibliometrici, mettendo in luce alcuni di quelli che possono essere i diversi errori nell'indicizzazione. Questa analisi vuole essere un punto di partenza per proporre miglioramenti e suggerire maggior accuratezza nella classificazione dei documenti scientifici raccolti in queste basi dati.

Il punto centrale del documento è quello di proporre una metodologia per una analisi semi-automatica di un numero di pubblicazioni scientifiche maggiori rispetto a quelli analizzati nell'attuale letteratura. La metodologia di analisi semi-automatica è utile a identificare gli errori di classificazione del DT nei database come Scopus e WoS e, sfruttando la discordanza nella classificazione del DT tra i due database, fornisce automaticamente un sottoinsieme di documenti potenzialmente classificati in modo errato. Tale sottoinsieme verrà poi sottoposto all'analisi manuale.

La metodologia si è dimostrata efficace nell'identificare gli errori di classificazione del DT, nel fornire un'analisi dell'accuratezza e nel suggerire modi per migliorare la qualità e l'affidabilità dei database bibliometrici. Potrebbe essere utilizzata da singoli scienziati, istituzioni e da gestori di database per individuare e correggere gli errori di classificazione dei DT in tempi relativamente brevi e, avendo la licenza di utilizzo di entrambi i database, la nuova metodologia può essere facilmente implementata utilizzando fogli di calcolo e tabelle pivot.

Nell'esempio considerato, a fronte di un *corpus* totale di 26.405 documenti in entrambe le banche dati, sono stati analizzati manualmente solo i 762 documenti “fuori diagonale” nella

*matrice di concordanza*, corrispondenti a circa il 2,9% del totale. Inoltre, lo 0,5% dei 25.643 documenti della diagonale (ovvero 129) è stato analizzato in modo casuale, non riscontrando alcun errore di assegnazione del DT e confermando la plausibilità dell'ipotesi secondo cui è improbabile che le classificazioni concordanti del DT siano errate.

Per quanto riguarda le tempistiche di analisi nella fase manuale della metodologia, considerando che l'analisi di un singolo errore richiede solitamente non più di cinque minuti, l'analisi totale ha richiesto circa  $\frac{(762+129) \cdot 5 \text{ min}}{60 \text{ min/h}} = \frac{891 \cdot 5 \text{ min}}{60 \text{ min/h}} = 74,25 \text{ h}$  di lavoro.

### **Limiti della ricerca**

Arrivare a conclusioni generali basate sul campione di documenti attualmente analizzato è complicato. Sebbene questo campione sia maggiore rispetto a quelli analizzati nella letteratura esistente, esso rimane comunque limitato, specialmente per quanto riguarda il ristretto campione delle categorie di documenti (solo quattro categorie coinvolte nell'esempio applicato). Un ulteriore ostacolo è dato dalla sottile differenza di definizione tra una categoria e l'altra, che spesso genera incoerenza nella classificazione all'interno di uno stesso database.

La complessità e le limitazioni attuali rendono quindi difficile trarre conclusioni definitive e generali. Risulta però evidente la necessità di standardizzazione delle definizioni e un miglioramento nella coerenza della classificazione dei documenti per la gestione delle pubblicazioni scientifiche. Questo consentirebbe di adottare uno standard più chiaro e condiviso tra i diversi stakeholder (tra cui i vari database, gli editori, i curatori e i ricercatori stessi, ecc.).

## **Sviluppi futuri della ricerca**

Al fine di presentare analisi più approfondite è necessario continuare a espandere il campione di documenti analizzati e ampliare le categorie di documenti per identificare pattern e discrepanze.

Inoltre, un ulteriore sviluppo potrebbe essere quello di includere la categorizzazione e la diversificazione degli errori. In altre parole, creare diversi livelli di errori e assegnare loro pesi diversi. Infatti, alcuni errori possono essere definiti come puri errori di classificazione, mentre altri derivano da ambiguità nelle definizioni e nella nomenclatura dei DT introdotte dagli editori e dai redattori. Potrebbe essere interessante anche dare una priorità all'assegnazione dei DT nei casi di DB con assegnazione multipla, come accade per WoS.

Nonostante la metodologia consenta di identificare in automatico un sottoinsieme di documenti per l'analisi manuale, riducendo così nettamente i tempi di analisi rispetto agli studi precedenti, un ulteriore miglioramento potrebbe essere quello di esplorare la possibilità di assistere la lunga analisi manuale con tecniche basate sull'intelligenza artificiale. Questo è solo un punto di partenza, ma sarà molto interessante vedere come gli ulteriori sviluppi continueranno a portare valore alla comunità scientifica.

## BIBLIOGRAFIA E SITOGRAFIA

- Valderrama-Zurián, J., Aguilar, R., Melero-Fuentes, D., & Aleixandre-Benavent, R. (2015). A systematic analysis of duplicate records in Scopus. *Journal of Informetrics*, 9(3), 570-576. doi:10.1016/j.joi.2015.05.002
- Almind, T., & Ingwersen, P. (1997). Informetric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of Documentation*, 53(4), 404-426. doi:10.1108/EUM0000000007205
- Bagnasco, S. (2023, Aprile). *Comitato Italiano per il Controllo delle Affermazioni sulle Pseudoscienze (CICAP)*. Tratto da Gli indici bibliometrici: alla ricerca della misura ideale: <https://www.cicap.org/n/articolo.php?id=1801069>
- Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), 1216-1227. doi:10.1002/asi.20077
- Cassella, M., & Bozzarelli, O. (2011). Nuovi scenari per la valutazione della ricerca tra indicatori bibliometrici citazionali e metriche alternative nel contesto digitale. *Biblioteche Oggi*, XXIX(2), p. 66-78. Tratto da <https://core.ac.uk/reader/11888700>
- Castiglia, M., & Tranchina, V. (2016, Ottobre). *Università degli Studi di Palermo - Sistema bibliotecario e Archivio storico di Ateneo - Settore Biblioteca Digitale*. Tratto da <https://www.unipa.it/biblioteche/.content/documenti/servizi-ricerca/libro-Bibliometria.pdf>
- Clarivate. (2024). *Clarivate*. Tratto il giorno Maggio 2024 da Web of Science Help - Document Types: <https://webofscience.help.clarivate.com/en-us/Content/document-types.html>
- Das, A. K. (2015). *Research evaluation metrics*. (B. K. Sen, A cura di) Paris, France: United Nations Educational, Scientific and Cultural Organization (UNESCO Digital Library). Tratto da <https://unesdoc.unesco.org/ark:/48223/pf0000232210>

- Donner, P. (2017, October ). Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics*, *113*(1), 219–236. doi:10.1007/s11192-017-2483-y
- Donner, P. (2023). Data inaccuracy quantification and uncertainty propagation for bibliometric indicators. *arXiv*. doi:10.48550/arXiv:2303.16613
- Elsevier. (2024). *Elsevier*. Tratto il giorno Maggio 2024 da Scopus Content: <https://www.elsevier.com/products/scopus/content>
- Franceschini , F., Maisano, D. A., & Mastrogiacomo, L. (2013, October ). A novel approach for estimating the omitted citation rate of bibliometric databases with an application to the field of bibliometrics. *Journal of the American Society for Information Science and Technology*, *64*(10), 2149-2156. doi:10.1002/asi.22898
- Franceschini, F., Maisano, D. A., & Mastrogiacomo, L. (2016, November). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Infometrics*, *10*(4), 933-953. doi:10.1016/j.joi.2016.07.003
- García-Pérez, M. (2010, June 2). Accuracy and Completeness of Publication and Citation Records in the Web of Science, PsycINFO, and Google Scholar: A Case Study for the Computation of h Indices in Psychology. *Journal of the American Society for Information Science and Technology*, *61*(10), 2070-85. doi:10.1002/asi.21372
- Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science*, *122*(3159), 108-111. doi:10.1126/science.122.3159.108
- Maisano, D., Mastrogiacomo, L., Ferrara, L., & Franceschini, F. (2024). Quality of Bibliometric Databases: Accuracy in Classification of Document Types. *6th International Conference on Quality Engineering and Management*. Girona, Spain: Article-in-Press. Tratto da [https://icqem.dps.uminho.pt/wp-content/uploads/2024/06/ICQEM24\\_ProceedingsBook.pdf](https://icqem.dps.uminho.pt/wp-content/uploads/2024/06/ICQEM24_ProceedingsBook.pdf)

- Mattarella, F. (2023). *pensierocritico.eu*. Tratto da Dalla Bibliometria allo sviluppo della Webometria: <https://www.pensierocritico.eu/webometria.html>
- Moed, H. (2005). *Citation Analysis in Research Evaluation*. Springer Science & Business Media . doi:10.1007/1-4020-3714-7
- Mokhnacheva, Y. (2023, May). Document Types Indexed in WoS and Scopus: Similarities, Differences, and Their Significance in the Analysis of Publication Activity. *Scientific and Technical Information Processing*, 50(1), 40-46. doi:10.3103/S0147688223010033
- Olensky, M., Schmidt, M., & van Eck, N. (2016). Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to Web of Science. *Journal of the Association for Information Science and Technology (JASIST)*, 67(10), 2550-64. doi:arXiv:1507.03314
- Owen, A. (2001). *Empirical Likelihood* (1st Edition ed.). (C. a. Hall/CRC, A cura di) New York: Taylor & Francis. doi:10.1201/9781420036152
- Schubert, A. (2001). Scientometrics: The Research Field and its Journal. In *Organizations and Strategies in Astronomy: Volume II. Astrophysics and Space Science Library* (p. 179–195). Springer Netherlands, Dordrecht. doi:10.1007/978-94-010-0666-8\_12
- Yagi, E., Badash, L., & de B.Beaver, D. (1996). Derek J. de S. Price (1922-83): Historian of science and herald of scientometric. *Interdisciplinary Science Reviews*, 21(1), 64-84. Tratto da Derek J. de S. Price (1922-83): Historian of science and herald of scientometrics: <https://garfield.library.upenn.edu/price/priceintdis1996.pdf>
- Yeung, A. (2021, December). Document type assignment by Web of Science, Scopus, PubMed, and publishers to “Top 100” papers. *Malaysian Journal of Library & Information Science*, 26(3), 97-103. doi:10.22452/mjlis.vol26no3.5

## RINGRAZIAMENTI

Alla determinazione,  
al coraggio e alla sensibilità  
e a chi ha contribuito  
a trasmettermi questi valori.

Un ringraziamento a tutti i docenti, mentori e tutor che mi hanno accompagnato in questi anni. In particolare, desidero esprimere profonda gratitudine al Prof. Domenico Augusto Maisano e al Prof. Luca Mastrogiacomo per aver creduto in me e avermi concesso questa opportunità di arricchimento e avvicinamento al mondo della ricerca.

Un grazie infinito alla mia famiglia e a tutte le persone che mi hanno, e stanno ancora, accompagnando nel cammino di crescita, fonte di supporto continuo e per i preziosi valori trasmessi. Agli Amici straordinari che mi sono stati vicini con affetto, aiuto reciproco e senza mai far mancare il sorriso. Sono estremamente fortunata e grata che siate parte della mia vita!

A ogni viaggio compiuto, ogni libro letto, ogni brano ascoltato, ogni persona incontrata e ogni esperienza affrontata lungo il percorso che mi ha reso (si spera) una persona migliore. *“Che tu possa continuare a incontrare persone ed esperienze che ti arricchiscano ulteriormente e ti aiutino a diventare sempre più chi desideri essere e che tu possa essere altrettanto per loro.”*

P.S. La premessa sulla centralità dei dati presente nell'introduzione della Tesi riprende il concetto dei “dati come nuovo petrolio” espresso nelle conclusioni della Prova Finale di Laurea

Triennale. Seppur i due elaborati trattino argomenti apparentemente lontani, ho percepito questo collegamento e ho voluto creare una connessione tra le due prove finali del mio percorso di crescita e formazione in questi cinque anni tra Politecnico e TUS, tra Italia e Irlanda.

Questo è sì un punto di arrivo, un traguardo, ma è contemporaneamente la partenza verso altre prove e sfide che non vedo l'ora di intraprendere, con tutti i soliti timori del caso, ma con estrema determinazione, curiosità e sempre con il sorriso.