https://doi.org/10.51408/issi2025_053

Impact of Web of Science and Scopus Policies on Multiple Document-Type Classification

Domenico A. Maisano[1], Lucrezia Ferrara[2], Fiorenzo Franceschini[3]

[1]*domenico.maisano@polito.it*, [2]*lucrezia.ferrara@polito.it*, [3]*fiorenzo.franceschini@polito.it*
Politecnico di Torino, DIGEP (Department of Management and Production Engineering),
Corso Duca degli Abruzzi 24, 10129, Torino (Italy)

Abstract

Document-type (DT) classification – i.e., the assignment of conventional labels such as *article*, *review*, *proceedings paper*, etc., to scientific documents – is crucial for information retrieval in bibliometric databases, but its incomplete objectivity can lead to errors with implications on indicators and research evaluations. This study focuses on a portion of the documents (with a relatively small incidence ~4%) with dual-DT assignment in Web of Science (WoS) – a feature that is absent in Scopus, which applies only single-DT assignments – to assess their characteristics and classification accuracy.

A manual analysis of more than a thousand documents revealed three main scenarios of dual-DT assignment in WoS: (i) the combination of one DT describing the *content* and another describing the *container* (e.g., *book chapters*, *proceedings papers*), (ii) the handling of specialized DTs (e.g., *data paper*, *retracted* paper), and (iii) the combination of a DT related to journal publication with a temporary DT for the *early-access* designation.

Documents with dual-DT assignment in WoS exhibit higher error rates, confirming the greater difficulty of classification for both databases, even for Scopus, regardless of its single-DT policy. WoS's dual-DT classification policy offers more detail and potentially greater accuracy but also shows some inconsistencies. Conversely, Scopus's single-DT policy reduces the level of detail and increases the risk of misclassification, particularly for papers from *conference proceedings* or journal *special issues*.

This study highlights the need for clearer DT definitions and recommends that bibliometric databases consider adopting more flexible multiple-DT classification policies to enhance both detail and accuracy in document classification. A limitation of this research is the relatively small *corpus* of documents analysed, which will be expanded in future studies.

Introduction

Document types (DTs) – such as research *articles*, *reviews*, *proceedings papers*, and *book chapters* – are conventional labels applied to scientific documents to describe their nature and main characteristics, facilitating information retrieval (Donner, 2017; Yeung, 2021). Depending on the publication context, DTs can be assigned by various stakeholders, including authors, editorial boards, publishers and bibliometric databases. However, because there are no universally accepted definitions or standardized rules for DT classification, a degree of subjectivity is unavoidable. This subjectivity often leads to questionable or even erroneous classifications. For instance, a *review* or a *note* might be misclassified as a research *article*, leading to several potential consequences. Beyond misleading researchers during document searches, these classification errors can distort bibliometric indicators for journals, individual researchers, and entire research institutions. Such distortions arise because bibliometric indicators often depend on the DT classification of the documents under

analysis. For instance, if a journal mislabels a substantial number of documents as *articles* rather than, say, *editorials* or *letters*, its *Journal Impact Factor* could be distorted: citations to those misclassified items would still be counted in the numerator, while the denominator (which includes only *articles* and *reviews*) might be inappropriately inflated or deflated (Haupka et al., 2024).

In some cases, these errors may even impact research evaluation exercises, which frequently include or exclude documents based on their DTs. For example, certain DTs – such as *proceedings papers*, *notes*, and *book chapters* – are often deemed less significant and are excluded from evaluations (García-Pérez, 2010; Franceschini et al., 2015; Yeung, 2019; Mokhnacheva, 2023). A *conference paper* erroneously classified as a journal *article* might grant a researcher undue credit in evaluations that prioritize journal publications, potentially influencing hiring, promotion, or funding decisions. Conversely, an important research contribution misclassified as a less prestigious DT (e.g., an *article* mislabelled as a *note*) could be undervalued in performance assessments.

Additionally, different research disciplines may be affected by misclassification of DTs in distinct ways. Fields that make heavy use of conference proceedings (e.g., *computer science* and *engineering*) may be particularly susceptible to misclassification between *conference papers* and journal *articles*, whereas disciplines that focus primarily on journal articles (e.g., *biology* and *medicine*) may be more concerned with distinguishing research *articles* from *reviews* or *editorial materials*.

Scientific literature on DT-classification errors is relatively sparse, primarily because such investigations typically involve samples of only a few hundred or thousand documents, requiring labour-intensive manual analysis. Recent studies suggest that DT-classification errors in general-purpose bibliometric databases, such as Web of Science (WoS) and Scopus, are non-negligible and account for a few percentage points (Franceschini et al., 2016a; Yeung, 2021; Donner, 2023; Zhu et al., 2024). These findings are corroborated by a recent study by Maisano et al. (2025), which introduces a semi-automated approach to detect potentially misclassified documents. This approach utilizes discrepancies between DT classifications assigned by competing databases, WoS and Scopus, to automatically identify subsets of potentially misclassified documents. Manual analysis, which is inherently time-consuming, can then be concentrated on this subset while excluding most documents, which are presumed to be correctly classified. This approach allows for an approximately two-order-of-magnitude increase in the size of analysed samples – e.g., from a few thousand to hundreds of thousands – without requiring additional manual-analysis effort. Maisano et al. (2025) analysed a sample of nearly 28,000 documents recently published by over 2,000 researchers affiliated with the two largest universities in Turin, Politecnico di Torino (PoliTO) and Università di Torino (UniTO). The study estimated overall error rates of approximately 2.3% for WoS and 2.7% for Scopus.

During the data collection for the research in (Maisano et al., 2025), an intriguing fact emerged: while most documents indexed by WoS and Scopus featured a single-DT classification, approximately 4% exhibited dual-DT classifications in WoS –

e.g., documents classified simultaneously as *editorial material; book chapter* or *article; proceedings paper* – whereas Scopus consistently applied single-DT classifications. These dual-DT classifications in WoS were excluded from the earlier study to avoid complicating the analysis.

It is important to clarify that in a dual-DT classification by WoS, up to two DTs in combination can be assigned to a single document. The most frequent combination involves one DT indicating the *content* type of the document (e.g., *article, review, letter*) and the other indicating the corresponding publication *container* (e.g., *proceedings paper, book chapter, journal*). However, other combinations of DTs are also possible. In contrast, Scopus' policy limits each document to a single DT label, forcing a choice even in cases where multiple DTs would be appropriate. This conceptual distinction is important, as some DTs are not mutually exclusive; in other words, sometimes a single document may legitimately fall under two DTs.

Building upon these observations, this study specifically focuses on this portion of documents with dual-DT classifications in WoS. The objectives are twofold: (i) to explore the reasons behind WoS's dual-DT assignments, likely indicative of greater classification challenges for these documents, and (ii) to compare WoS's dual-DT-assignment policy with Scopus's single-DT-assignment approach. Formally, the study addresses the following research questions, respectively:

RQ#1: *Is the error rate (for both WoS and Scopus) higher for documents with dual-DT assignments compared to those with single-DT assignments, confirming that the former are inherently more challenging to classify?*

RQ#2: *Based on the analysis, which approach – WoS's dual-DT assignments or Scopus's single-DT assignments – appears more reasonable?*

Methodologically, the study will conduct an exhaustive manual analysis of a *corpus* of documents of interest, assessing the accuracy of DT classifications in WoS and Scopus and attributing errors where detected. The remainder of this study is organized in three sections. The "Methodology" section details the methodological approach, including the sample selection, manual analysis procedure, and statistical measures to be constructed. The "Results" section presents the findings and relevant statistics, accompanied by descriptions, interpretations of the results, and practical examples. Finally, the "Conclusions" section summarizes the key findings, highlights practical implications for the scientific community, discusses limitations, and suggests directions for future research.

Methodology

As outlined in the "Introduction", this study builds on the dataset used in Maisano et al. (2025), which combines publications authored by researchers affiliated with UniTO (a generalist university in Turin) and PoliTO (a technical university in Turin) during the 2019–2023 period. The choice of these two medium-to-large universities – with a combined total of over 100,000 students and approximately 2,000 tenured researchers, covering a wide range of scientific disciplines – ensures that the dataset of publications is diverse in terms of subjects, DTs, journals and publishers, making

it relatively representative of the entire recent scientific literature. These publications are indexed by both WoS and Scopus.

From an initial set of nearly 30,000 documents, 1,085 were identified as having dual-DT assignments by WoS. These documents constitute the *corpus* under investigation in this study. Table 1 provides a detailed classification of these documents by both databases. Notably, the DT labels assigned by WoS and Scopus do not always align, due to minor differences in DT naming conventions (e.g., *conference paper* in Scopus versus *proceedings paper* in WoS) and the inclusion or exclusion of certain specialized DTs (e.g., *expression of concern* and *meeting abstract* in WoS but not in Scopus). For further details, refer to the official DT lists provided by Scopus and WoS (Clarivate, 2025; Elsevier, 2025).

Table 1. Summary of DTs classified by WoS and Scopus for the 1,085 publications analysed in this study. DTs in each database are sorted in descending order based on the number of documents they include.

(a) DTs classified by WoS	No. of docs	(b) DTs classified by Scopus	No. of docs
Article; Proceedings paper	423	*Article*	767
Article; Early access	394	*Book chapter*	152
Article; Book chapter	146	*Conference paper*	54
Article; Data paper	45	*Review*	53
Review; Early access	44	*Data paper*	40
Editorial material; Book chapter	10	*Letter*	7
Letter; Early access	7	*Editorial*	6
Editorial material; Early access	5	*Erratum*	4
Review; Book chapter	5	*Note*	1
Correction; Early access	4	*Retracted*	1
Article; Expression of concern	1		
Article; Retracted publication	1		
Total no. of documents	**1,085**	**Total no. of documents**	**1,085**

Table 2 presents a matrix that highlights the similarities and discrepancies between the DT classifications in WoS (with DTs listed in the rows) and Scopus (with DTs listed in the columns) for the analysed documents. While some classifications appear consistent (e.g., the four documents classified as *correction; early access* in WoS and *erratum* in Scopus), others exhibit clear incompatibilities (e.g., the eight documents classified as *review; early access* in WoS but as *article* in Scopus).

Table 2. Matrix of DT classifications for the analysed documents, according to WoS (rows) and Scopus (columns). All 1,085 documents were manually analysed to identify potential classification errors.

DT classifications →		by Scopus										
	↓	*Article*	*Book chapter*	*Conf. paper*	*Review*	*Data paper*	*Letter*	*Editorial*	*Erratum*	*Note*	*Retracted*	**Row total**
by WoS	*Article; Proceedings paper*	359	-	52	12	-	-	-	-	-	-	**423**
	Article; Early access	389	-	2	3	-	-	-	-	-	-	**394**
	Article; Book chapter	3	143	-	-	-	-	-	-	-	-	**146**
	Article; Data paper	5	-	-	-	40	-	-	-	-	-	**45**
	Review; Early access	8	-	-	36	-	-	-	-	-	-	**44**
	Editorial material; Book chapter	-	7	-	-	-	-	3	-	-	-	**10**
	Letter; Early access	-	-	-	-	-	7	-	-	-	-	**7**
	Editorial material; Early access	1	-	-	-	-		3	-	1	-	**5**
	Review; Book chapter	1	2	-	2	-	-	-	-	-	-	**5**
	Correction; Early access	-	-	-	-	-	-	-	4	-	-	**4**
	Article; Expression of concern	1	-	-	-	-	-	-	-	-	-	**1**
	Article; Retracted publication	-	-	-	-	-	-	-	-	-	1	**1**
	Column total	**767**	**152**	**54**	**53**	**40**	**7**	**6**	**4**	**1**	**1**	**1,085**

All 1,085 documents in the matrix were manually analysed to determine their "true" (or correct) DTs and identify any potential classification errors by the databases. The manual analysis was conducted shortly after data retrieval, in February 2024. Depending on the need, the following information was considered to determine the "true" DT(s) for each document, with a progressively deeper manual analysis where required:

- Title and abstract;
- Information and metadata provided on the journal and/or publisher's webpage;
- Formal structure of the document;
- Number of references cited within the document;
- Full text.

For each document, the accuracy of the DT classification provided by each database was assessed, also considering their respective DT definitions and presumed assignment rules. A logic of internal consistency was applied to establish whether a database's DT classification was correct or erroneous.

Results

This section is divided into two subsections: (i) the presentation of results from the perspective of WoS and Scopus, using so-called "error tables" (Maisano et al., 2025) and associated error statistics, and (ii) the practical interpretation of these results, supported by several pedagogical examples.

Error tables and error statistics

Error! Reference source not found. presents the error table for WoS, a contingency table that displays the DT classifications assigned by WoS in its columns and the "true" (or correct) DTs determined through manual analysis in its rows (Maisano et al., 2025). The main diagonal of the error table contains the correct DT classifications

– with the corresponding document counts shown in round parentheses "(·)" – while the off-diagonal elements represent incorrect classifications. Notably, while the columns include only the DTs listed in Table 1(a) for WoS, additional DTs appear in the last rows of **Error! Reference source not found.**, denoted by the symbol "(*)" – including one with dual-DT classification "*review; proceedings paper*" at the top. These additional DTs were introduced because the manual analysis revealed misclassifications that were corrected by assigning more appropriate DT classifications, following WoS's declared rules (Clarivate, 2025).

Table 3. Error table for WoS. Quantities in "(·)" represent correctly classified documents, "[·]" denote *partial* errors (with weight ½), and "{·}" indicate *full* errors. Statistics α_i and β_j were calculated only for groups with at least 30 documents for statistical reliability. The symbol "(*)" denotes additional DTs added following manual analysis.

"True" DT classifications	DT classification by WoS: *Article; Proceed. paper*	*Article; Early access*	*Article; Book chapter*	*Article; Data paper*	*Review; Early access*	*Editorial material; Book chapter*	*Letter; Early access*	*Editorial material; Early access*	*Review; Book chapter*	*Correction, Early access*	*Article; Expression of concern*	*Article; Retracted publicat.*	Row total	Total "{·}" row errors	Total "[·]" row errors	α_i
Article; Proceedings paper	(409)	[2]	-	-	-	-	-	-	-	-	-	-	411	0	2	**0.2%**
Article; Early access	-	(329)	-	-	-	-	-	-	-	-	-	-	329	0	0	**0.0%**
Article; Book chapter	-	-	(145)	-	-	-	-	-	-	-	-	-	145	0	0	**0.0%**
Article; Data paper	-	-	-	(45)	-	-	-	-	-	-	-	-	45	0	0	**0.0%**
Review; Early access	-	[1]	-	-	(33)	-	-	-	-	-	-	-	34	0	1	**1.5%**
Editorial material; Book chapter	-	-	-	-	-	(8)	-	-	-	-	-	-	8	0	0	-
Letter; Early access	-	-	-	-	-	-	(6)	-	-	-	-	-	6	0	0	-
Editorial material; Early access	-	-	-	-	-	-	-	(3)	-	-	-	-	3	0	0	-
Review; Book chapt.	-	-	-	-	-	-	-	-	(5)	-	-	-	5	0	0	-
Correction, Early access	-	-	-	-	-	-	-	-	-	(4)	-	-	4	0	0	-
Article; Expression of concern	-	-	-	-	-	-	-	-	-	-	(1)	-	1	0	0	-
Article; Retracted publication	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	-
Review; Proceedings paper(*)	[12]	-	-	-	[1]	-	-	-	-	-	-	-	13	0	13	-
Article(*)	-	[58]	-	-	{2}	-	-	-	-	-	-	[1]	61	2	59	-
Review(*)	-	{3}	-	-	[8]	-	-	-	-	-	-	-	11	3	8	-
Letter(*)	-	-	-	-	-	-	[1]	-	-	-	-	-	1	0	1	-
Editorial material(*)	{1}	-	{1}	-	-	-	-	-	-	-	-	-	2	2	0	-
Book chapter(*)	-	-	-	-	-	[2]	-	-	-	-	-	-	2	0	2	-
Proceedings paper(*)	[1]	-	-	-	-	-	-	-	-	-	-	-	1	0	1	-
Other(*)	-	{1}	-	-	-	-	-	{2}	-	-	-	-	3	3	0	-
Column total	423	394	146	45	44	10	7	5	5	4	1	1	1,085			
Total "{·}" column errors	1	4	1	0	2	0	0	2	0	0	0	0		10		
Total "[·]" column errors	13	61	0	0	9	2	1	0	0	0	0	1			87	
β_j	**1.8%**	**8.8%**	**0.7%**	**0.0%**	**14.8%**	-	-	-	-	-	-	-				$\varepsilon \cong$ **4.9%**

Among the off-diagonal elements, two types of errors can be distinguished:

- *Full* errors (quantities denoted in curly brackets "{·}"), representing cases where both DTs assigned by WoS are incorrect.
- *Partial* errors (quantities denoted in square brackets "[·]"), involving cases where one of the two assigned DTs is correct, while the other is incorrect.

Partial errors are weighted with a score of ½, as they represent an intermediate level of error between fully incorrect DT assignments (score of 1) and correct DT assignments (score of 0). The content of the error table can be summarized using an overall (weighted) error rate:

$$\varepsilon = \frac{d^{\{\cdot\}} + ½ \cdot d^{[\cdot]}}{d^{\{\cdot\}} + d^{[\cdot]} + d^{(\cdot)}}, \quad (1)$$

where:

$d^{\{\cdot\}} = \sum_{i,j} d_{i,j}^{\{\cdot\}}$ is the total number of documents with *full* errors in the error table;

$d^{[\cdot]} = \sum_{i,j} d_{i,j}^{[\cdot]}$ is the total number of documents with *partial* errors in the error table;

$d^{(\cdot)} = \sum_{i,j} d_{i,j}^{(\cdot)}$ is the total number of correctly classified documents in the error table.

The denominator of the fraction in Eq. 1 represents the sum of all three document categories, which amounts to 1,085. For WoS, the number of *partial* errors ($d^{[\cdot]} = 87$) is significantly higher than the number of *full* errors ($d^{\{\cdot\}}$=10). Moreover, the error rate for WoS (~4.9%) is markedly higher than the rate observed in the previous study (~2.3%) for documents with single-DT assignments (Maisano et al., 2025). A statistical test on the difference between the two proportions confirmed this rigorously (Ross, 2017). Addressing RQ#1, it can be concluded that documents with dual-DT classifications in WoS are inherently more challenging to classify, as they exhibit a significantly higher propensity for misclassification.

Beyond the overall error rate (ε), additional error statistics can be constructed. Specifically, for each row *i*, the probability that a document belonging to a given DT is wrongly classified into another DT (i.e., *missing assignment to the DT of interest*) is:

$$\alpha_i = \frac{\sum_j d_{i,j}^{\{\cdot\}} + ½ \cdot \sum_j d_{i,j}^{[\cdot]}}{\sum_j d_{i,j}^{\{\cdot\}} + \sum_j d_{i,j}^{[\cdot]} + \sum_j d_{i,j}^{(\cdot)}}. \quad (2)$$

For each column *j*, the probability of misclassifying a document into the specific DT of that column (i.e., *false classification into the DT of interest*) is:

$$\beta_j = \frac{\sum_i d_{i,j}^{\{\cdot\}} + ½ \cdot \sum_i d_{i,j}^{[\cdot]}}{\sum_i d_{i,j}^{\{\cdot\}} + \sum_i d_{i,j}^{[\cdot]} + \sum_i d_{i,j}^{(\cdot)}}. \quad (3)$$

These statistics (α_i and β_j) were calculated only for groups with at least 30 total documents (in rows or columns) to ensure statistically reliable estimates. The statistics indicate that several errors involve journal documents with dual-DT assignments including the *early-access* designation, despite being already published in their final form (i.e., with specific volume/issue numbers and definitive page numbers). However, these errors are not particularly severe for two reasons:

1. They do not fundamentally alter the "true" nature of the document; they simply attach an erroneous (temporary) designation of *early access*;

2. From a follow-up investigation conducted approximately 10 months after data collection (December 2024), it was found that over 80% of these inaccuracies had been corrected by WoS.

A smaller proportion of errors involves *conference proceedings papers* misclassified as *articles*, although they are actually *reviews* or *surveys*.
For Scopus (see the relevant error table in
Table **4**), no dual-DT assignments appear, as this database only permits single-DT classifications (Elsevier, 2025). Consequently, only *full* errors are observed, with an error rate of $\varepsilon = \frac{86}{1{,}085} \cong 7.9\%$, significantly higher than the 2.7% reported in the previous study (Maisano et al., 2025). Statistical testing confirmed the significance of this difference. Scopus appears to encounter even greater challenges than WoS when classifying these particularly delicate documents. The distribution of errors in Scopus reveals that many articles are misclassified as *conference papers*, while others classified as *articles* belong to more specific DTs (e.g., *reviews*, *book chapters*, *data papers*). As explored in the next subsection, the root cause of these errors often lies in the limitations of Scopus's DT definitions and its strict single-DT-assignment policy.

Interpretation of results

This subsection provides an interpretation of the most significant results of the analysis, supported by numerous practical examples. Three typical scenarios were observed in which WoS assigns dual-DT classifications, each of which is analysed individually below:

1. Combination of a DT related to a document's *content* and a DT related to the *container* (or dissemination context);
2. *Early-access* documents, typically linked to scientific journals;
3. Classification of uncommon, specialized documents from scientific journal.

Table 4. Error table for Scopus. Quantities in "(·)" represent correctly classified documents, while "{·}" denote *full* errors. Statistics α_i and β_j were calculated only for groups with at least 30 documents for statistical reliability. The symbol "(*)" denotes additional DTs added following manual analysis.

		DT classification by Scopus										Row total	Total row errors	α_i
		Article	*Book chapter*	*Conf. paper*	*Review*	*Data paper*	*Letter*	*Editorial*	*Erratum*	*Note*	*Retracted*			
"True" DT classifications	*Article*	(745)	-	{52}	{4}	-	-	-	-	-	{1}	802	57	**7.1%**
	Book chapter	{3}	(147)	-	-	-	-	-	-	-	-	150	3	**2.0%**
	Conf. paper	-	-	(1)	-	-	-	-	-	-	-	1	0	-
	Review	{11}	{1}	{1}	(48)	-	-	-	-	-	-	61	13	**21.3%**
	Data paper	{5}	-	-	-	(40)	-	-	-	-	-	45	5	**11.1%**
	Letter	-	-	-	-	-	(7)	-	-	-	-	7	0	-
	Editorial	{1}	{4}	-	-	-	-	(6)	-	-	-	11	5	-
	Erratum	-	-	-	-	-	-	-	(4)	-	-	4	0	-
	Note	{2}	-	-	-	-	-	-	-	(1)	-	3	2	-
	Retracted	-	-	-	-	-	-	-	-	-	-	0	0	-
	Short survey(*)	-	-	-	{1}	-	-	-	-	-	-	1	1	-
Column total		767	152	54	53	40	7	6	4	1	1	1,085		
Total column errors		22	5	53	5	0	0	0	0	0	1		86	
β_j		**2.9%**	**3.3%**	**98.1%**	**9.4%**	**0.0%**	-	-	-	-	-			$\varepsilon \cong$ **7.9%**

(1) Combination of a DT related to *content* and a DT related to the *container*. A common pairing observed involves (i) a DT describing the document's *content* in terms of objectives and structure (e.g., research *article, review, letter*), and (ii) a DT related to the *container*, representing the dissemination context (e.g., *journal, conference proceedings, book chapter*). Specifically, WoS seems to include the container DT only for scientific publications that differ from *journal* contributions[1]. This practice is consistent with the WoS definition of *proceedings paper*, which states, "*proceedings papers will have a dual document type: article; proceedings paper*", though no similar rule exists for *book chapters*, which are defined only as "*a monograph or publication written on a specific topic within a main division in a book*" (Clarivate, 2025). On the other hand, Scopus, constrained by its single-DT-assignment policy, provides systematically less detailed classifications and occasionally misleading ones. The examples below document some of the most common and/or curious errors observed for both databases.

- For example, Scopus defines a *book chapter* as "*a complete chapter in a book or book-series volume, identified as a chapter by a heading or section indicator*". However, some special *book chapters*, such as book series introductions, are often classified by Scopus as *editorial*. Additionally, inconsistencies arise because some book-series *editorials* are still classified as *book chapters*. These internal inconsistencies in Scopus are generally avoided by WoS due to its dual-DT-assignment policy.

  For example, documents 1.1 to 1.5 in Table 5 pertain to *book chapters*. The first three are from the same book but differ in content: the first is an *introduction* to the whole book, the second is a classic research *article* (complete with methodology, results, discussion, etc.), and the third contains concluding notes related to the whole book. WoS assigns the container-DT *book chapter* to all three, pairing it with a content-DT: *article* for the second and *editorial material* for the introduction and conclusions, consistent with its definition of *editorial material* (Clarivate, 2025). Conversely, Scopus classifies the first document as *editorial* (but not as *book chapter*), while the other two are classified as *book chapters* (but not as *article* or *editorial*). Although this DT classification is not exactly wrong, it is undoubtedly less detailed and potentially more misleading than that of WoS.

  Focusing on documents 1.4 and 1.5 in Table 5, both from another book, WoS not only classifies them correctly as *book chapters* but also distinguishes their content by assigning the additional DTs *review* and *article*, respectively. In contrast, Scopus's classification, while accurate, assigns only the single DT *book chapter* to both contributions.

[1] In fact, the DT *journal* is not envisaged by WoS as it is implied when the database makes a single-DT assignment related to the document's content, such as *article*, *review* or *letter*.

Table 5. Examples of documents with peculiar DT classifications, discussed in the analysis.

Ref.	DOI	DT classification		Brief description
		WoS	Scopus	
1.1	https://doi.org/10.1007/978-3-319-71837-8_1	*Editorial material; book chapter*	*Editorial*	Introductory chapter of a book divided into chapters.
1.2	https://doi.org/10.1007/978-3-319-71837-8_10	*Article; book chapter*	*Book chapter*	Chapter corresponding to a research *article*.
1.3	https://doi.org/10.1007/978-3-319-71837-8_12	*Editorial material; book chapter*	*Book chapter*	Concluding chapter of a book divided into chapters.
1.4	https://doi.org/10.1007/978-3-319-79084-8_2	*Review; book chapter*	*Book chapter*	Same as 1.2, but corresponding to a *review*.
1.5	https://doi.org/10.1007/978-3-319-79084-8_3	*Article; book chapter*	*Book chapter*	Same as 1.2.
2.1	https://doi.org/10.1111/odi.13076	*Article; proceedings paper*	*Conf. paper*	*Review* published in a journal special issue dedicated to a medical workshop (7th WWOM).
3.1	https://doi.org/10.5004/dwt.2018.22308	*Article; proceedings paper*	*Article*	*Article* in a journal special issue dedicated to an international conference (CEST 2017).
3.2	https://doi.org/10.5004/dwt.2018.22995	*Article; proceedings paper*	*Article*	Same as 3.1.
3.3	https://doi.org/10.5004/dwt.2019.24424	*Article; proceedings paper*	*Article*	*Article* in a journal special issue dedicated to an international conference (NAXOS 2018).
3.4	https://doi.org/10.5506/APhysPolB.51.1627	*Article; proceedings paper*	*Article*	*Article* in a journal (Acta Physica Polonica B) exclusively dedicated to conference proceedings.
3.5	https://doi.org/10.5506/APhysPolB.51.655	*Article; proceedings paper*	*Article*	Same as 3.4.
3.6	https://doi.org/10.5506/APhysPolB.51.661	*Article; proceedings paper*	*Article*	Same as 3.4.
3.7	https://doi.org/10.1200/JCO.18.00053	*Article; proceedings paper*	*Article*	In the "Prior presentation" section of this journal *article*, it is stated that the contribution was presented in three different conferences held in 2017 and 2018.
4.1	https://doi.org/10.1002/bmb.21490	*Editorial material; early access*	*Article*	Christmas song called "*Oxidosqualene (OS) cyclase—Lanosterol synthase*", appearing in the scientific journal (Biochemistry and Molecular Biology Education), which classifies the paper with the specialized DT *the lighter side*.
4.2	https://doi.org/10.1002/alz.13526	*Article; early access*	*Article*	In the "Presentations" section of this journal *article*, it is stated that the contribution was presented at three different conferences held in 2021 and 2023. Following its final publication, WoS's *early-access* designation was removed, leaving the single DT *article*.
4.3	https://doi.org/10.1002/hon.3184	*Article; early access*	*Article*	Before the abstract, it is stated that "*Preliminary results were presented as an abstract and oral presentation at the 63rd ASH Annual Meeting & Exposition in 2021*". Following its final publication, WoS's *early-access* designation was removed, leaving the single DT *article*.
5.1	https://doi.org/10.1002/ecy.2448	*Article; data paper*	*Article*	Document published in a journal as a *data paper*, containing detailed information on the dataset used for a research *article* with its DOI provided.

Ref.	DOI	DT classification		Brief description
		WoS	Scopus	
5.2	https://doi.org/10.1016/j.dib.2018.10.142	*Article; data paper*	*Article*	Same as 5.1, but classified as a *data article*.
5.3	https://doi.org/10.1016/j.dib.2018.11.129	*Article; data paper*	*Article*	Same as 5.2.
6.1	https://doi.org/10.1007/978-1-0716-0603-2_5	*Article; book chapter*	*Book chapter*	Special document consisting of a *protocol* published in a book dedicated to operational *methods and protocols* in medicine/biology.
6.2	https://doi.org/10.1007/978-1-0716-0611-7_11	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.3	https://doi.org/10.1007/978-1-0716-0978-1_14	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.4	https://doi.org/10.1007/978-1-0716-0978-1_25	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.5	https://doi.org/10.1007/978-1-0716-0978-1_27	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.6	https://doi.org/10.1007/978-1-0716-0978-1_38	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.7	https://doi.org/10.1007/978-1-0716-0978-1_40	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.8	https://doi.org/10.1007/978-1-0716-1174-6_14	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.9	https://doi.org/10.1007/978-1-4939-7584-6_8	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.10	https://doi.org/10.1007/978-1-4939-8837-2_3	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.11	https://doi.org/10 1007/978-1-4939-8982-9_15	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.12	https://doi.org/10.1007/978-1-4939-9873-9_5	*Article; book chapter*	*Book chapter*	Same as 6.1.

- Another common pairing of content-related and container-related DTs concerns contributions derived from conferences. These contributions, primarily articles, are commonly classified by WoS as *article; proceedings paper*, accounting for ~40% of the documents with dual-DT assignments (i.e., 423 out of 1,085; see Table 1). Although WoS's official definitions for *article* and *proceedings paper* are somewhat convoluted and appear to reference dual DTs only in specific cases, empirical observation shows that WoS systematically applies dual-DT assignments for *articles* published in special issues derived from conferences.
- An exception to the previous point arises for *reviews* originating from conferences. According to WoS's official definition: "*Review articles that were presented at symposium or conference will be processed as proceedings papers*" (Clarivate, 2025), which means that such papers are classified as pure *proceedings papers* without dual-DT assignments. This choice appears inconsistent with the dual-DT-assignment policy for *article; proceedings paper*. It would likely be more consistent to also allow dual-DT assignments such as *review; proceedings paper*. For example, document 2.1 in Table 5 is a *review* clearly derived from a conference, as explicitly stated in the journal special issue where it appears.

However, WoS erroneously classifies it as *article; proceedings paper* instead of *review; proceedings paper*.

- From Scopus's perspective, articles from journal special issues linked to conferences are simply classified as *articles*, effectively equating them with "pure" journal articles, which are typically subjected to a more rigorous selection process. For instance, documents 3.1 to 3.7 in Table 5 come from three different conferences: the first three, linked to CEST 2017 and NAXOS 2018, were published in special issues of the journal *Desalination and Water Treatment*. The next three, from the *Random Matrix Theory* conference in Kraków, appeared in *Acta Physica Polonica B*, a journal exclusively dedicated to conference proceedings. The final document (3.7) was published in a regular issue of the *Journal of Clinical Oncology* but had been presented at three different conferences[2]. Both WoS and Scopus correctly classify these seven documents based on their internal criteria. However, as with *book chapters*, Scopus's convention for journal special issues results in a loss of information regarding their conference origin, effectively conflating them with pure journal articles. This sort of "promotion" can impact bibliometric indicators – at the journal, researcher, or institutional level – which may not always differentiate between regular and special issues of journals (Franceschini et al., 2019).

(2) *Early-access* documents. A substantial portion of the analysed documents (i.e., 454 out of 1,085, corresponding to ~42%; see Table 1) received dual-DT assignments in WoS, where the primary DT refers to the content of the contribution, typically in a scientific journal (*article*, *correction*, *review*, or *editorial material*), and the secondary DT corresponds to the temporary designation of *early access*. This designation indicates that the contribution has been accepted and made publicly available online but has not yet appeared in its final editorial format (e.g., with volume, issue number, and definitive page numbers). The points below summarise some curious aspects observed regarding this category of documents.

- Although Scopus officially includes *article in press* among its defined DTs (Elsevier, 2025), it does not appear to use this designation in practice. As a result, documents labelled as *early access* in WoS generally do not pose classification issues for Scopus, except for occasional misclassifications between *article* and *review* – a phenomenon already observed in previous studies (Donner, 2023; Haupka et al., 2024; Zhu et al., 2024; Maisano et al., 2025).
- Returning to WoS, the manual analysis revealed that *early-access* documents are generally classified correctly, except for some misclassifications of the primary DT, particularly between *article* and *review*. Regarding the secondary *early-access* DT, ~85% of the analysed documents were found to be correctly classified.

[2] In fact, a dedicated section of the paper, named "Prior Presentation", reads: "*Presented at the 59th Annual Meeting of the American Society of Hematology, Atlanta, GA, December 9-12, 2017; the meeting of the American Society of Blood and Marrow Transplantation, Salt Lake City, UT, February 21-25, 2018; and the 16th International Umbilical Cord Blood Symposium, San Diego, CA, June 14-16, 2018*".

However, ~15% were already in their final editorial format, meaning the temporary *early-access* designation should have been removed. As noted in the previous subsection, this inaccuracy does not appear to be particularly severe. Furthermore, a follow-up check conducted in December 2024 (approximately eight months after the initial data retrieval and analysis) revealed that nearly all previous anomalies had been corrected by WoS (Franceschini et al., 2016b).

- As an anecdote, one document (document 4.1 in Table 5) highlights a rare double misclassification. This unique document – a *Christmas song* published in a scientific journal – was not only no longer *early access* but was also misclassified by WoS as *editorial material*. A more appropriate classification might have been *other*. Scopus also misclassified it too as an *article*, whereas *note* would probably have been a more suitable designation.
- Another curious observation involves *early-access* designations for articles in journal special issues or extended versions of conference contributions. Occasionally, this temporary designation seems to "overwrite" a potential secondary DT of *proceedings paper*. Once the article, initially classified as *article; early access*, is published in its final form, WoS does not appear to replace *early access* with *proceedings paper*, which would seem appropriate as it is typically associated with journal articles originating from conferences, as previously documented. For instance, consider documents 4.2 and 4.3 in Table 5. Both are extended versions of articles originally presented in conference proceedings. However, they do not carry the dual-DT assignment *article; proceedings paper* once published in their final form. It appears that the temporary *early-access* designation displaces the secondary *proceedings paper* DT, and the database fails to reinstate it after final publication. This observation warrants further investigation in future studies. While Scopus's policy of assigning only a single DT introduces potential inaccuracies, WoS's limit of a maximum of two DTs may sometimes lead to inaccuracies, as exemplified by the issue discussed above.

(3) Uncommon, specialized journal documents. A less frequent scenario in which WoS assigns dual DTs involves documents published in journals that differ from traditional contributions (*articles*, *reviews*, *letters*, etc.). Below are some of the most interesting cases observed.

- Among these less common documents, we identified forty-five so-called *data papers*, which are essentially documents containing detailed datasets that support other scientific contributions (typically journal articles) to which they are linked. According to its internal rule, WoS classifies these contributions with a dual-DT assignment: *article; data paper* (cf. the definition: "*A data paper will have a dual document type: article; data paper*" (Clarivate, 2025)). Conversely, Scopus, despite having a dedicated *data-paper* DT (Elsevier, 2025), sometimes classifies these documents simply as *articles*. Table 5 lists three examples of such documents (5.1 to 5.3). The most critical consequence of these inaccuracies is the undue "promotion" of *data papers* to the level as journal *articles*.

- Among the less common documents with dual DTs in WoS, we also found one classified as *article; retracted* and another as *article; expression of concern*. Scopus, in comparison, categorized these documents into its dedicated *retracted* and *article* categories, respectively.
- Finally, we draw attention to another uncommon category of documents: *protocols* or *methods and protocols*. These documents, which primarily detail best practices in medicine/biology, do not have a direct counterpart in the DT categories of either WoS or Scopus. Since these contributions appear almost exclusively in book series, Scopus indexed them – correctly, in our view – as *book chapters*, while WoS assigned them the dual DT *article; book chapter*. While WoS's classifications were not deemed erroneous in our analysis, it might have been more appropriate to classify these documents as *other; book chapter*. In fact, the content of these *protocols* often lacks the structure of canonical articles, making the *article* designation less fitting. Table 5 exemplifies the twelve documents (6.1 to 6.12) identified during the analysis.

Conclusions

This research focused on scientific documents with dual-DT assignments in WoS, which – based on a preliminary estimate – constitute ~4% of all indexed documents. The aim was to identify potential issues in DT classification, not only from the perspective of WoS but also Scopus. Manual analysis of a *corpus* of 1,085 documents revealed that documents with dual DTs are more prone to classification errors than those with single DTs: error rate of 4.9% versus 2.3% for WoS and 7.9% versus 2.7% for Scopus. Thus, addressing **RQ#1**, it can be concluded that these documents significantly differ from those with single DTs, as confirmed by appropriate statistical tests.

In general, three main scenarios were identified where WoS uses dual-DT classification:

1. Cases where the primary DT specifies the *content* type (e.g., *article, review, letter*), while the secondary DT specifies the *container* (e.g., *book chapter, proceedings paper*), if different from the journal container implicitly referenced by WoS for single-DT classifications.
2. Less common and specialized documents usually published in journals. These documents are relatively few and do not significantly impact overall error statistics. In cases where the specific DT is not covered by the database's predefined categories, we suggest avoiding overuse of the DT *article* and instead replacing it with a "catch-all" DT *other*. This would avoid various undue "promotions".
3. Documents temporarily assigned the *early-access* secondary DT, while awaiting their final published format. A notable number of errors stemmed from the failure to update *early-access* journal documents (*articles, reviews, letters*, etc.) in WoS after their final publication.

The analysis also revealed that some DT-classification errors may stem from inconsistencies or ambiguities in DT definitions. For instance, WoS's definition of

book chapter and Scopus's definitions of *article* exhibit certain ambiguities. Nevertheless, allowing dual-DT assignments in WoS serves as a "safety net" to avoid relatively severe errors, which Scopus sometimes encounters. For WoS, cases where both assigned DTs are incorrect (*full* errors) represent only 10 out of 1,085 documents, whereas in 87 cases, one of the assigned DTs was correct (*partial* errors). Scopus's strict single-DT assignment policy appears to be a relevant cause of its misclassifications. This limitation arises from the simple fact that DTs are not always mutually exclusive; in some cases, multiple DTs may be valid simultaneously (e.g., a *review* and a *conference paper*). Although forcing a single-DT assignment might seem like a simplification, this approach can lead to potential errors, such as undue "promotion" (e.g., from *proceedings paper* to journal *article*) or, at least, a loss of information about the documents in question. For this reason, WoS's policy of allowing dual-DT assignments seems more prudent (**RQ#2**). From a practical standpoint, it might even make sense to use up to three DTs in certain cases: one for *content*, one for the *container*, and one for an *accessory* designation (e.g., *early access*, *retracted*, etc.).

The findings of this study have practical implications for several stakeholders. For individual researchers, they may provide additional guidance for collecting and selecting documents from scientific literature through databases. For bibliometric indicator developers, this study raises awareness of potential distortions caused by DT classification errors, which have been at least preliminarily quantified here. For database managers, the comparative analysis of current DT-assignment policies could inform future improvements in DT definitions and their assignment logic.

In general, we recommend that database providers refine and clarify their DT classification guidelines to minimize ambiguities (e.g., clearly distinguishing an *article* from an *editorial material* or *proceedings paper*) and consider integrating AI-based tools to assist in the DT classification process. Automated checks – using machine learning trained on document metadata and *full texts* – could help flag inconsistent or unlikely DT assignments for human review, thus improving the overall accuracy of the databases. Additionally, relaxing the current WoS-imposed limit of two DTs to allow a third DT could be a reasonable step forward.

Furthermore, it would be useful to consider how other bibliometric databases handle DT classification to put these findings in context. For example, the *Dimensions* bibliometric platform (developed by Digital Science) combines publisher metadata and machine learning to assign DTs and links them to research grants, patents, and policy outputs (Digital Science, 2025). Open scholarly platforms like *OpenAlex* rely largely on publisher-provided metadata (via *Crossref*) for DTs, resulting in a broader but less standardized set of DTs. Studies have shown that DTs can differ considerably between providers, and what counts as a "research" document versus "non-research" can vary by database. These discrepancies underscore the absence of a universal standard for DT classification across database, which can complicate cross-database comparisons. Our findings and recommendations align with recent calls for richer and more consistent DT metadata in bibliometric data sources (Haupka et al., 2024).

The primary limitation of this research is the relatively small sample size (1,085 documents), which may hinder the generalizability of the findings. In future work, we aim to extend the sample size to provide a more comprehensive analysis.

Acknowledgments

This research was carried out under the MICS (Made in Italy, Circular and Sustainable) Extended Partnership and partially funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza – Missione 4, Componente 2, Investimento 1.3, D.D. 1551.11-10-2022, PE00000004). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

Clarivate (2025). Web of Science Help – Document Types, https://webofscience.help.clarivate.com/en-us/Content/document-types.html [accessed April 2025].

Digital Science (2025) Which publication and document types are available in Dimensions? https://dimensions.freshdesk.com/support/solutions/articles/23000018866-which-publication-and-document-types-are-available-in-dimensions- [accessed April 2025].

Donner, P. (2017). Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics*, 113(1), 219-236.

Donner, P. (2023). Data inaccuracy quantification and uncertainty propagation for bibliometric indicators. *arXiv preprint*, arXiv:2303.16613

Elsevier (2025). Scopus Content Coverage Guide, https://www.elsevier.com/products/scopus/content [accessed April 2024].

Franceschini, F., Maisano, D., Mastrogiacomo, L. (2015). Errors in DOI indexing by bibliometric databases. *Scientometrics*, 102, 2181-2186.

Franceschini, F., Maisano, D., Mastrogiacomo, L. (2016a). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics*, 10(4), 933-953.

Franceschini, F., Maisano, D., Mastrogiacomo, L. (2016b). Do Scopus and WoS correct "old" omitted citations? *Scientometrics*, 107, 321-335.

Franceschini, F., Galetto, M., Maisano, D. (2019) *Designing Performance Measurement Systems: Theory and Practice of Key Performance Indicators*, Springer International Publishing. Cham, Switzerland, ISBN: 978-3-030-01191-8.

García-Pérez, M.A. (2010). 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.

Haupka, N., Culbert, J. H., Schniedermann, A., Jahn, N., Mayr, P. (2024). Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, PubMed and Semantic Scholar. *arXiv preprint*, arXiv:2406.15154.

Maisano, D., Mastrogiacomo, L., Ferrara, L., Franceschini, F. (2025). A large-scale semi-automated approach for assessing document-type classification errors in bibliometric databases. *Scientometrics*, 130, 1901-1938.

Mokhnacheva, Y.V. (2023). 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.

Ross, S.M. (2017). *Introductory statistics*. Academic Press, London, ISBN 978-0-12-804317-2.

Yeung, A.W.K. (2019). Comparison between Scopus, Web of Science, PubMed and publishers for mislabelled review papers. *Current Science*, 116(11), 1909-1914.

Yeung, A.W.K. (2021). 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.

Zhu, M., Lu, X., Chen, F., Yang, L., Shen, Z. (2024). An explorative study on document type assignment of review articles in Web of Science, Scopus and journals' websites. *Journal of Data and Information Science*, 9(1), 11-36.