XXIV SEMEAD
Seminários em Administração
novembro de 2021
ISSN 2177-3866

CHALLENGES IN IDENTIFYING STUDIES FOR A LITERATURE REVIEW

RICARDO MARQUES DE ALMEIDA DANTAS
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO (UFRJ)

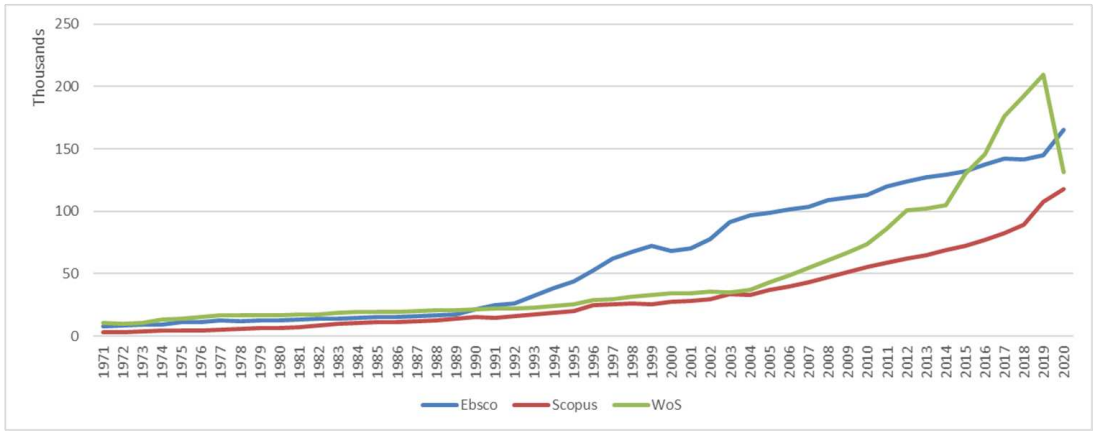
DENISE LIMA FLECK
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO (UFRJ)

CHALLENGES IN IDENTIFYING STUDIES FOR A LITERATURE REVIEW

INTRODUCTION

As one considers embarking on a research topic, one's typical conjectures may include the likely relevance and feasibility of conducting research on the topic (Creswell, 2014), as well as whether one's studies might be of interest to others (Davis, 1971). Well-crafted literature reviews may not only help to address these conjectures, but also become a foundation for advancing knowledge on a certain domain (Snyder, 2019). However, synthesizing past research findings has become progressively complex (Zupic and Cater, 2015), due to the increasing amount of published papers (Aria and Cuccurullo, 2017), as Figure 1 illustrates.

Figure 1 – Number of publications by year and by bibliographic database.



- **Ebsco**: filtered by Academic journals and english language
- **Scopus**: filtered by SUBJAREA (busi OR econ) AND (LIMIT-TO (SRCTYPE,"j")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English"))
- **WoS**: filtered by SU= (Business & Economics) RESEARCH DOMAIN: (SOCIAL SCIENCES) AND LANGUAGE: (ENGLISH) AND TYPES OF DOCUMENT: (ARTICLE OR REVIEW) AND RESEARCH AREA: (BUSINESS ECONOMICS) Databases: WOS, DIIDW, KJD, RSCI, SCIELO

To complicate things further, relying on a single database may leave out relevant works on the investigated topic. For instance, regarding the organizational decline topic, seminal articles such as Weitzel & Jonsson (1989) and Whetten (1980) are not available in all three most representative databases (refer to Table 1). While the former is included in the Ebsco database only, the latter is not available in the Scopus database, indicating that by relying on one specific database alone precludes researchers from performing comprehensive reviews. As Wanyama et al. (2021) verified, 25.7% of his search were found in two of these three bibliographic databases – Scopus, Web of Science and Ebsco.

Table 1 – Sample of articles about decline (X = availability on database)

Article	Scopus	Web of Science	Ebsco
Whetten (1980) - Organizational Decline: A Neglected Topic in Organizational Science			X
Weitzel and Jonsson (1989) - Decline in Organizations: A Literature Integration and Extension		X	X
Mellahi and Wilkinson (2004) - Organizational failure: a critique of recent research and a proposed integrative framework	X	X	X
Pretorius (2008) - Critical variables of business failure: A review and classification framework	X	X	

Serra et al. (2017) - Organizational Decline Research Review: Challenges and Issues for a Future Research Agenda	X		X

Identifying studies constitutes an important step in the literature review process, and it may significantly affect the review outcome, despite rigorous protocols of the review method itself (Wanyama et al., 2021). Considering that only a small fraction of the vast amount of available literature will likely be relevant in a study (Wanyama et al., 2021), some procedures are usually employed. Typical examples include the strategies of inclusion and exclusion such as focusing on peer-reviewed journals (Hiebl, 2021), and basing the search for academic work on convenience and/or availability (Wanyama et al., 2021).

While acknowledging the importance of every step along the literature review process, this paper seeks to pinpoint some challenges authors may face during the "Identifying Studies" step (Jesson et al., 2011). These include the staggering amount of published academic work; the absence of shared standards among different database providers, as well as the nonexistence of keyword catalogs to help distinguishing and classifying studies, to name a few. Aiming at illustrating the nature of challenges one faces during this phase, this paper provides practical examples of difficulties on identifying academic work related to the organizational growth and decline topic.

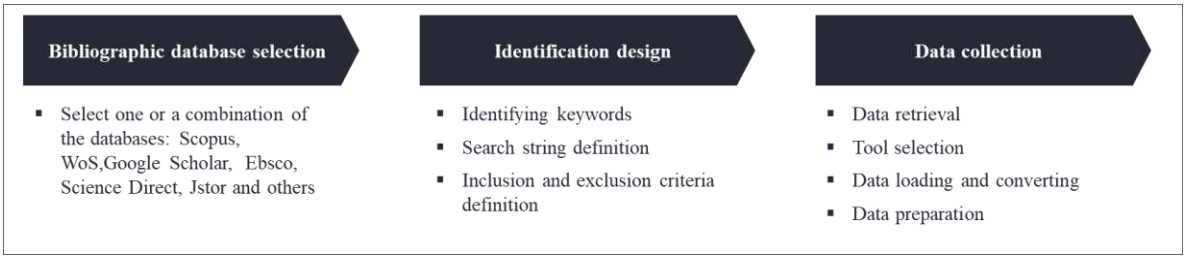
In what follows, the next section provides an overview of the identifying studies process and the profile of bibliometric databases. Thereafter, there follow three sections. The first describes the method we employed to select data on organizational growth and decline from the three databases; the other discusses the findings; and finally, the concluding section, puts forward the main contributions of this study and suggests next steps.

BACKGROUND ON IDENTIFYING STUDIES

Different literature review methods are available, which enable researchers to position the review within an already established domain of study or to redirect to new ideas, bringing together less obvious connections (Breslin and Gatrell, 2020). However, as critical as it is defining the review method, so it is identifying the literature to be reviewed, which has been referred as sample selection (Hiebl, 2021), identifying studies (Jesson et al., 2011), locating studies (Creswell, 2014), article selection (Aguinis et al., 2020; Snyder, 2019), search for literature (Jesson el al., 2011), among others. Yet, recommended practices regarding sample selection are still scarce (Hiebl, 2021).

For instance, a simple query using the word "decline" has identified 1,307,311 items in the Scopus database, 560,186 items in the Web of Science database, 304,329 articles in the Ebsco (Business Source Ultimate) database. Hence, a studies identification method should be selected among many examples the literature provides (Jesson el al., 2011; Aguinis et al., 2020; Hiebl, 202; Creswell, 2014). Drawing on Zupic and Carter (2015), Figure 2 depicts a three-step workflow of procedures to conduct the sample selection, namely: bibliographic database selection, identification design and data collection. There follows a brief description of these procedures.

Figure 2 – Identifying studies workflow.



Bibliographic database selection

Thompson Reuters' Web of Science and Elsevier's Scopus are the most popular databases used by scholars to run bibliometric studies, since both databases contain information on cited references. Alternatively, Google Scholar has gained visibility among researchers due to its broad coverage, as well as to the citation data it includes. Nonetheless, Google Scholar poses limitations to the automation of data extraction.

Additional bibliographical databases include Ebsco, Science Direct and Jstor. However, because they do not contain citation information (Waltman, 2016), running analysis such as citation, co-citation and coupling is quite difficult. Instead of scrutinizing some bibliographical database, some authors identify top journals in their research field and prioritize them based on JCR impact factor. The latter strategy may not only secure the relevance of the collected data, but it may also limit the number of articles under analysis. But this comes at the expense of leaving important articles out of the scope of analysis.

Identification design

Identifying and collecting data from the selected bibliographic database requires defining a search string using keywords related to the theme the literature review investigates (Cobo et al., 2011a). Calibrating the search string and applying it to scrutinize the title, abstract or database keywords play an important role in reducing redundancy. Nonetheless, this may pose some problems in management research, where search is hardly ever precisely defined (Wanyama et al., 2021).

Data collection

Data collection comprises data retrieval, data loading and converting, and data preparation. There are several tools at the disposal of the researcher to be used for this stage of the process (Cobo et al., 2001b; Aria and Cuccurullo, 2017; Zupic and Cater, 2015). Aria and Cuccurullo (2017) present an example of this process using the set of functions of bibliometrix from language R.

Data preparation, specifically, is a critical task during data collection due to duplicate articles, misspelling items (Cobo et al., 2011a) and lack of uniformity across and within databases (Wanyama et al., 2021). Another critical issue refers to cited references, because multiple versions of one single work may present author's name, journal and title in different formats (Zupic and Carter, 2015; Aria and Cuccurullo, 2017).

METHOD USED TO APPLY THE SAMPLE SELECTION PROCESS

Seeking to identify the challenges a researcher likely faces when defining the sample selection (Hiebl, 2021), this paper uses the topic organizational growth and decline to illustrate this process.

Bibliographic database selection

Bibliographic databases hold different sets of literature, as Figure 1 illustrates. In line with Wanyama et al.'s (2021) work, which has examined out how complementary Scopus, Web of Science (WoS) and Ebsco databases are, the present study has selected these three databases.

Identification design

Finding the right terms to use on a search is almost a handicraft work within the management field, because while the search is not precisely defined (Wanyama et al., 2021), it should cover not only the literature already known by the authors, but also unknown fields (Hiebl, 2021). Therefore, the search string definition took a two-step or a back-and-forth process. For this study, the word decline has many correlates and a first understanding of it was done by searching the words "decline AND organizational AND review" on all fields of Ebsco database. Based on this preliminary analysis, the terms failure, failing, bankruptcy, death, mortality, growth, growing, success and survival were identified either as correlated or associated with decline.

In this context, a search using the string "decline OR failure OR bankruptcy OR survival OR growth OR success OR failing OR mortality OR growing OR death" on the field "article title" was run on these databases. Additionally, the search string was refined to include the words "organization, corporate, firm or organisation" on article title, abstract or keywords fields to establish the organization as the level of analysis. Moreover, to narrow down the search, only articles within the subareas of business and economics were considered. Articles should also be of type "article" or "review", be written in English and had their source type as "Journal". The author's and database's keywords were used to exclude articles related to the words "Economic Growth", "economic development", "Gross Domestic Product", "Foreign Direct Investment", "Income Distribution", "Economic And Social Effects", "Urban Growth", "Environmental Economics", "Economic Policy", "Population Growth", "Economic History", "Labor Market", "Monetary Policy", "Public Policy" and "Political Economy". And finally, articles presenting the following terms were excluded: "career success", "project success", "product success", "implementation success", " "system success", "success factor", "service failure", "project failure", "failure factor", "learning failure", "power failure", "maintenance failure", "system failure", "systemic failure", "productivity growth", "productivity decline", "environmental policy", "system errors", "information system", "control systems". "employment growth", "infant mortality".

Data collection

Once the ".bib" files were exported from the selected databases, they were uploaded and converted into data frames format, using bibliometrix functions on R Studio, in order to prepare the data for analysis. R Studio is a free and open source tool for language R and, although it is necessary to learn to program on it, it is a very flexible tool and it provides the functions to automatize the process of uploading the files resulted from the export process. Once the files were uploaded into a data frame, it was possible to handle duplicates, eliminate articles without author information and standardize the cited references.

Concerning duplicates, a case in point is the article "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy (Altman, 1968)" listed on Scopus database, for instance. On WOS, this article is written without "the" on the title. Those duplications occur frequently intra and inter databases. We have eliminated the duplicates by first comparing the titles and then by comparing a key composed by the first 45 characters of the title plus the first two characters of the author, with stop words, white spaces and punctuation removed. Duplicates inter bibliographic databases give rise to an additional issue because they present different information about the same article. For instance, the article "Organizational failure: a critique of recent research and a proposed integrative framework (Mellahi and Wilkinson, 2004)" has 229 citations on Scopus and 213 citations on Web of Science. Finally, we decided to prioritize the articles in the following order: Scopus, Web of Science and Ebsco.

The next step was to standardize the knowledge base. The same cited reference may appear in different formats, even within the same bibliographic database. For instance, "(Altman, 1968)" was cited using more than two hundred formats, such as "ALTMAN EI, 1968, J FINANC, V23, P589, DOI 10.2307/2978933" or "ALTMAN, E.I., FINANCIAL RATIOS, DISCRIMINANT ANALYSIS AND THE PREDICTION OF CORPORATE BANKRUPTCY (1968) J. FINANC., 23, PP. 589-609". This issue was handled by introducing a standard in the knowledge base, which comprised only the main author surname from the cited reference, plus the first letter of the author's first name and plus the year of publication. For example, all citations formats referred to "Altman (1968)" becames "altman e 1968".

FINDINGS AND DISCUSSION

This section highlights and discusses three main learning lessons one should pay attention to when identifying studies in a literature review. Every step across the sample selection holds one major challenge that may affect the quality of a literature review. At the "Bibliographic database selection" step, the researcher faces a staggering amount of published academic work, missing information and lack of standardized content across bibliographic databases. At the "Identification design" step, the nonexistence of keyword catalogs makes the search activity difficult in distinguishing and classifying studies within complex, fragmented and not consolidated domains of literature. Finally, at the "Data collection" step, the absence of shared standards among different database providers makes it difficult to combine their collections of items.

Staggering amount of published academic work

Reviewing academic work is becoming increasingly challenging due to the amount of published work (Antons et al., 2021), as illustrated in Figure 1. In 1990, none of the three bibliographic databases considered on this study had reached more than 30 thousand academic works. 30 years later, each of those databases reached more than 100 thousand works published, presenting an average of 7% of annual growth in published work combined.

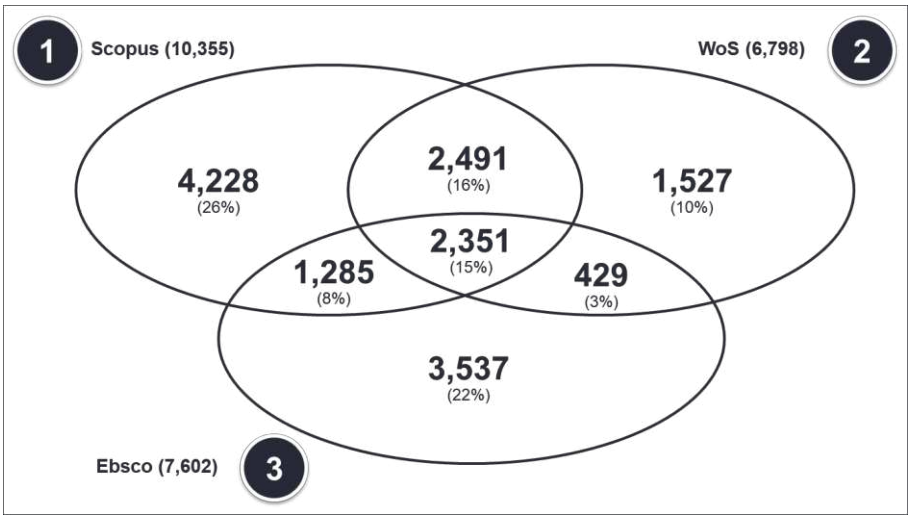
To complicate things further, as already demonstrated at Wanyama et al. (2021), combining bibliographic databases provides the researcher with a broader list of articles. Scopus, WOS and Ebsco databases combined collected more than 400 thousand academic works in 2020. So, based on this context, this study built a comprehensive database of 15,848 unique items extracted from those databases. Nonetheless, only 15% of all unique articles appeared on the

intersection of those three databases. Moreover, 42% of all unique articles appeared in at least two of those databases. In other words, Scopus, WoS and Ebsco exclusively contribute to 58% of all unique articles on the comprehensive database of this study, as Figure 3 shows. Hence, the step of identification design is critical to overcome the challenge the amount of published work poses to the researcher.

Therefore, not only the amount of published work is increasing but the variation of content across bibliographic databases is also increasing, as Figure 1 also suggests and Figure 3 corroborates. This fact itself would not represent an issue if it were possible to evaluate the content each database holds. For instance, relying on authors' and database's keywords should be a relevant way to identify items; instead, incomplete or inexistent information on keywords are quite often the case, which contributes to complicate the identifying studies efforts.

From the 15,848 articles within the comprehensive database of this study, 4,164 have no authors' keywords and only 3,427 carry database's keywords. This situation has improved in recent years concerning authors' keywords, if you consider that in 1993, only 39% of the articles exported included this information, whereas in 2020, this percentage has raised to 92%. On the other hand, and differently from the authors' keywords, this situation has not improved over time regarding the database's keyword. While in 1993, 26% of the articles exported carry this information, in 2020, this percentage is still no more than 25%.

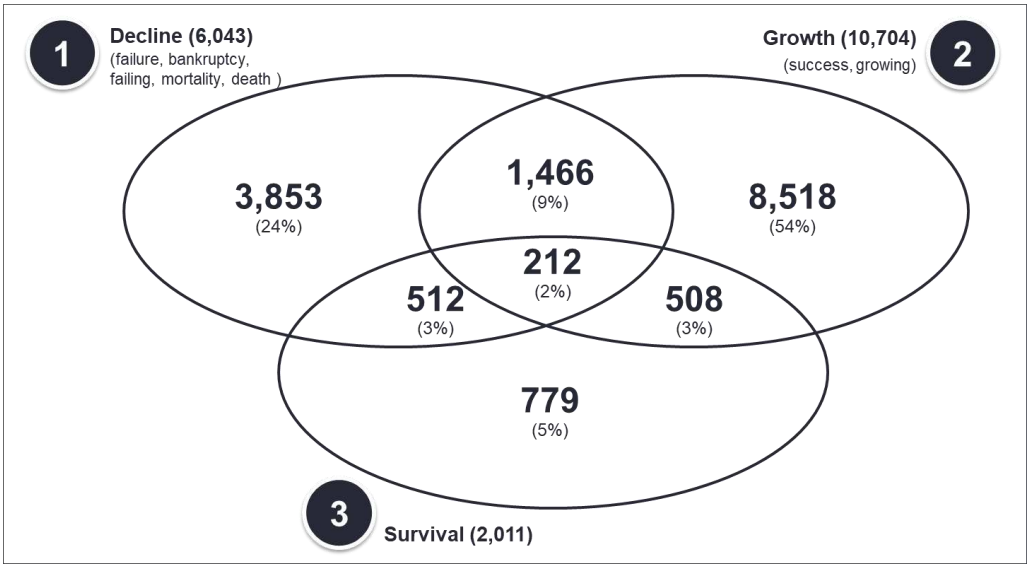
Figure 3 – Venn diagram and intersections of Scopus, Web of Science and Ebsco.



Nonexistence of keyword catalogs

The search definition step during the process of sample selection has also revealed the importance of carefully selecting the search words. Depending on which words one chooses, important literature items may not be included in one's study. Even though the decline, growth and survival topics are more often than not addressed separately, Figure 4 illustrates how decline, growth and survival are intertwined. For instance, 17% of the 15,848 unique articles within the comprehensive database are shared with at least two of the decline, growth and survival domains. It suggests, thus, that one should not study domains such as decline by themselves; quite on the contrary, growth and survival are an intrinsic part of the understanding of such a complex phenomenon.

Figure 4 – Venn diagram and intersections of literature domains.



Despite the complementarity of domains, their integration is anything but simple, since authors seem not to follow a standard procedure when choosing the words that will represent their work on title and abstract. Actually, regarding some specific topic, there is quite a diversity of words authors employ in their works. As an example, and illustrated on Figure 5, let's take the subset of 1,631 articles where the main search word "bankruptcy" appears in the title field (A). Within this subset, the word "bankruptcy" appears 86% of times in the abstract field. In other words, 14% of times authors don't mention "bankruptcy" when writing their abstracts (B). This might be because "bankruptcy" is not the relevant concept in their academic work, or it might also be a bit of disregard concerning the elaboration of their abstracts. Important to note that this rate is even lower in author's keywords (C) field where the word "bankruptcy" appears 67% of times. Not to mention in database's keywords (D) field, where the word "bankruptcy" is not even the most frequent term within the subset.

Still exploring the "bankruptcy" subset, any word stands out when associated with "bankruptcy" in the title field. For instance, the terms "corporate" and "prediction" go along with "bankruptcy" in title field 15% and 13% of times, respectively. In one hand this might suggest that the "bankruptcy" literature comprehend different subgroups of literature such as "prediction models" or the "process of bankruptcy". And on the other hand, this might also suggest that authors use alternative words to represent "corporate", such as "business" and "organization". In the end, this wording exercise makes the search activity a relevant step toward the quality of the literature review.

Figure 5 – Statistics of most frequent words within the subset "Bankruptcy".

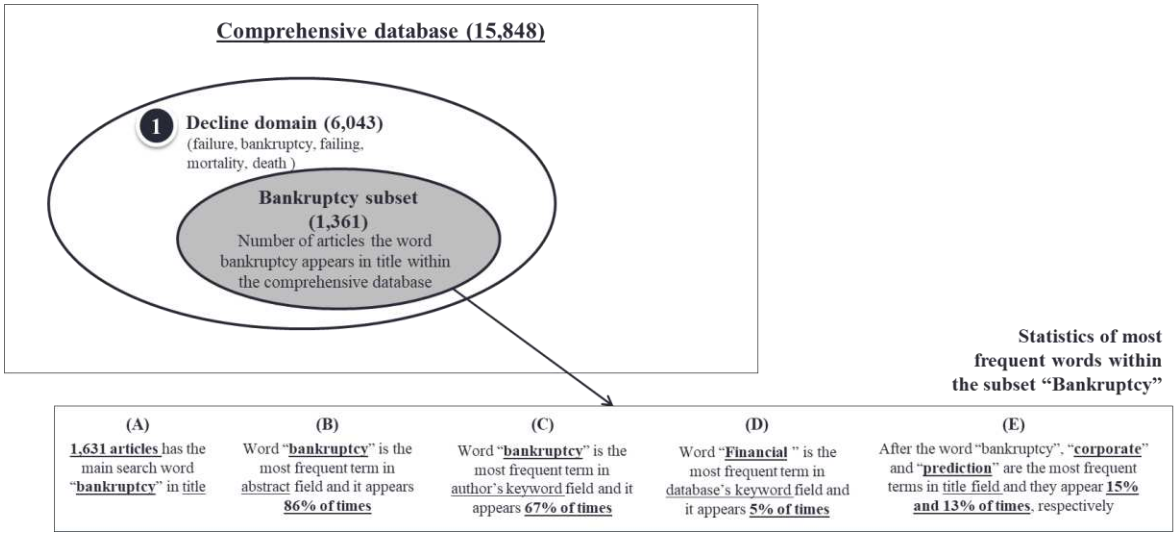


Table 2 gathers the statistics of most frequent words of other subsets of articles within the comprehensive database of this study. And opposed to "bankruptcy", the subsets of "failing" and "growing" have their search word showed up less than 50% of times in the abstract (B). Authors' (C) and database's (D) keywords have even lower rates, not to mention that in some cases the most frequent term in those fields is different from the search term. And finally, words that represent the level of analysis, such as "firm", "organization", "business" or "corporate", are in most cases the most frequent terms that appear on the title after the main search word (E).

Table 2 – Statistics of most frequent words within each subset of articles

(A) Subsets of articles based on the main search word (Number of articles it appears in Title)	Most frequent terms (as % of A)			(E) Two most frequent terms in title after the main search word (as % of A)
	(B) In abstract	(C) In authors' keywords	(D) Database's keywords	
Decline (684)	Decline (62%)	Decline (11%)	Management (8%)	Organizational (10%); Rise (8%)
Failure (2,267)	Failure (71%)	Failure (36%)	Failure (6%)	Success (12%); Corporate (9%)
Bankruptcy (1,361)	Bankruptcy (86%)	Bankruptcy (67%)	Financial (5%)	Corporate (15%); Prediction (13%)
Survival (1,424)	Survival (71%)	Survival (41%)	Survival (8%)	Firm (26%); Analysis (10%)
Growth (5,666)	Growth (77%)	Growth (40%)	Growth (7%)	Firm (22%); Model (9%)
Success (3,478)	Success (66%)	Management (19%)	Management (6%)	Business (10%); Management (7%)
Failing (156)	Failing (35%)	Business (12%)	Management (6%)	Firm (20%); Organization (11%)
Mortality (400)	Mortality (91%)	Mortality (50%)	Mortality (17%)	Risk (11%); Stochastic (9%)
Growing (654)	Growing (33%)	Management (15%)	Management (4%)	Firm (12%); Management (6%)
Death (406)	Death (58%)	Death (26%)	Analysis (4%)	Life (12%); Organizational (6%)

Absence of shared standards among different database providers

We came across lack of standards throughout the entire process of sample selection. This is especially critical concerning how database providers organize information on the items they carry. For instance, finding ways to match duplicates both intra and inter databases is a major concern.

Scopus presented the higher level of duplicates intra database. After handling them, 10,355 articles were retained – out of the 11,090 originally extracted. In other words, that database had 6.6% of duplicates. In turn, Ebsco and Web of Science had 1.1% and 0.2% duplicates, respectively. The matching of articles inter databases was also essential to combining databases. As illustrated in figure 3, there is a relevant number of intersections between the bibliographic databases, and the lack of standard across them adds an additional layer of difficulty as one seeks to combine their information accurately.

Finally, the lack of standards concerning cited references makes impossible any citation or coupling analysis, since 710 out of the 15,848 items of the comprehensive database have no cited reference information, and the remaining items present 562,302 cited references, almost 37 references per article. In addition, one same reference may be cited in different formats as already mentioned in the "Altman (1968)" case.

CONCLUSIONS

Needless to mention the importance of connecting to the right literature to adequately make sense of a topic's different perspectives. This paper brings to attention the increasing complexity of going through that process of sense-making and sense-giving (Cronin and George, 2020) due to challenges the identification of academic work pose. Furthermore, these challenges are embedded in the choices one makes, and directly affect the quality of one's academic work. Both deciding what search words refer to the investigated literature and carrying out the search for items in the available databases constitute an exponential concern because of the staggering amount of published work and the absence of shared standards among different databases providers. One wonders how faster scientific work might advance should institutionalized standards be available. Moreover, one also wonders whether the time is ripe for setting in motion a combined effort by the academic community to address this issue.

REFERENCES

Aguinis, H., Ramani, R.S., Alabduljader, N., 2020. Best-Practice Recommendations for Producers, Evaluators, and Users of Methodological Literature Reviews. Organizational Research Methods 1094428120943281. https://doi.org/10.1177/1094428120943281

Altman, E.I., 1968. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. The Journal of Finance 23, 589–609. https://doi.org/10.2307/2978933

Antons, D., Breidbach, C.F., Joshi, A.M., Salge, T.O., 2021. Computational Literature Reviews: Method, Algorithms, and Roadmap. Organizational Research Methods 1094428121991230. https://doi.org/10.1177/1094428121991230

Aria, M., Cuccurullo, C., 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics 11, 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Breslin, D., Gatrell, C., 2020. Theorizing Through Literature Reviews: The Miner-Prospector Continuum. Organizational Research Methods 1094428120943288. https://doi.org/10.1177/1094428120943288

Cobo, M. J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F., 2011a. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. J. Informetrics 5, 146–166. https://doi.org/10.1016/j.joi.2010.10.002

Cobo, M. J., López-Herrera, A.G., Herrera-Viedma, E., Herrera, F., 2011b. Science mapping software tools: Review, analysis, and cooperative study among tools. Journal of the

American Society for Information Science and Technology 62, 1382–1402. https://doi.org/10.1002/asi.21525

Creswell, J.W., 2014. Research design: qualitative, quantitative, and mixed methods approaches, 4th Edition. ed. SAGE Publications, Thousand Oaks, California.

Cronin, M.A., George, E., 2020. The Why and How of the Integrative Review. Organizational Research Methods 1094428120935507. https://doi.org/10.1177/1094428120935507

Davis, M.S., 1971. That's Interesting!: Towards a Phenomenology of Sociology and a Sociology of Phenomenology. Philosophy of the Social Sciences 1, 309–344. https://doi.org/10.1177/004839317100100211

Hiebl, M.R.W., 2021. Sample Selection in Systematic Literature Reviews of Management Research. Organizational Research Methods 1094428120986851. https://doi.org/10.1177/1094428120986851

Jesson, J., Matheson, L., Lacey, F.M., 2011. Doing your literature Review: traditional and systematic techniques.

Mellahi, K., Wilkinson, A., 2004. Organizational failure: a critique of recent research and a proposed integrative framework. International Journal of Management Reviews 5–6, 21–41. https://doi.org/10.1111/j.1460-8545.2004.00095.x

Pretorius, M., 2008. Critical variables of business failure: A review and classification framework. South African Journal of Economic and Management Sciences 11, 408–430.

Serra, F.A.R., Pinto, R., Guerrazzi, L., Ferreira, M.P., 2017. Organizational Decline Research Review: Challenges and Issues for a Future Research Agenda. BAR, Braz. Adm. Rev. 14. https://doi.org/10.1590/1807-7692bar2017160110

Snyder, H., 2019. Literature review as a research methodology: An overview and guidelines. Journal of Business Research 104, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Waltman, L., 2016. A review of the literature on citation impact indicators. Journal of Informetrics 10, 365–391. https://doi.org/10.1016/j.joi.2016.02.007

Wanyama, S.B., McQuaid, R.W., Kittler, M., 2021. Where you search determines what you find: the effects of bibliographic databases on systematic reviews. International Journal of Social Research Methodology 0, 1–13. https://doi.org/10.1080/13645579.2021.1892378

Weitzel, W., Jonsson, E., 1989. Decline in Organizations: A Literature Integration and Extension. Administrative Science Quarterly 34, 91–109. https://doi.org/10.2307/2392987

Whetten, D.A., 1980. Organizational Decline: A Neglected Topic in Organizational Science. The Academy of Management Review 5, 577–588. https://doi.org/10.2307/257463

Zupic, I., Čater, T., 2013. Bibliometric Methods in Management and Organization: A Review. Academy of Management Proceedings 2013, 13426–13426. https://doi.org/10.5465/AMBPP.2013.13426abstract