

## Chapter 1

# Unveiling Management Research's Thematic Evolution: An Unsupervised Machine Learning - Latent Dirichlet Allocation Perspective

### Abstract

In the dynamic realm of organizational practices and strategies, navigating complexity demands robust methodologies unveiling intricate themes within management research. This study employs advanced topic modeling, particularly Latent Dirichlet Allocation (LDA), to address: "What thematic patterns emerge in management research?" Unearthing latent themes holds scholarly and pragmatic value. Scholars and practitioners seek insights into trends shaping decisions and strategies. This inquiry delves into concealed research architecture, guiding as they navigate management research. While existing wisdom contributes invaluable frameworks and insights, understanding themes across diverse articles remains challenging. This study leverages LDA's power, employing unsupervised learning for dynamic thematic landscapes, revealing insights into paramount themes. At its core, this study triggers scholarly transformation. Harnessing LDA, it transcends content analysis, heralding data-driven insights into latent themes and their arcs. The narrative is amplified through post-LDA visualization, illuminating enigmatic theme substrata. Innovation continues with multiple correspondence analysis (MCA), harmonizing keywords and topics. This strengthens thematic connections, adding quantitative support to qualitative exploration. By analyzing influential authors and papers through citations, a constellation

---

emerges, shedding light on contributors and milestones sculpting management's evolution.

In summary, this study establishes a comprehensive framework for topic modeling in management research. Empowering scholars and practitioners, it fosters the traversal of management literature, shaping scholarship and strategies, enhancing innovation in theory and practice. These insights navigate the intricate tapestry of management research, providing a roadmap for understanding themes, trends, and their influence. An additional field of application pertaining to the design for a serial entrepreneurship in cyber-physical realities - deciphering visionary management models for marketing has been provided to underpin the diagnostic power of the developed methodology of meta level research.

**Keywords:**

Management research, thematic evolution, Latent Dirichlet Allocation, topic modelling, trends analysis, scholarly discourse

**Introduction**

Management research plays a pivotal role in shaping organizational practices and strategies across various industries. The continuous evolution of this field necessitates robust methodologies that can unveil key themes and trends in the vast corpus of management literature. In this study, we employ advanced topic modelling techniques, specifically Latent Dirichlet Allocation (LDA), to address the pressing research question: "What are the dominant thematic patterns and how have they evolved in management research over time?"

Understanding the underlying themes in management research is crucial both in theory and practice. Academics and practitioners seek to stay abreast of emerging trends and focus on areas that influence decision-making and business strategies. By unravelling the latent structures within the corpus, we aim to provide valuable insights that can aid scholars, practitioners, and journal editors in navigating the dynamic landscape of management research.

(1) What do we know?

Prior research in management has provided valuable theoretical perspectives and empirical findings, which have shaped our understanding of various subfields. However, the complexity and volume of management literature have made it challenging to comprehensively grasp the evolving themes and trends. While existing studies have shed light on specific topics, a major unaddressed puzzle remains: How do we

holistically capture the thematic patterns and shifts across a vast corpus of management research articles?

This study addresses this crucial knowledge gap by employing LDA, a powerful topic modelling algorithm, to uncover key themes and their temporal evolution. By delving into the unsupervised learning approach, we provide a novel and comprehensive perspective on the dynamics of management research themes. This work holds great significance, as it facilitates a deeper understanding of the overarching themes that influence scholarly discourse and practical applications.

(2) What will we learn?

Through our study, we fundamentally change and advance scholars' understanding of management research. By applying LDA, we go beyond traditional content analysis and provide a data-driven approach that enables the identification of latent themes and their transitions over time. The post-LDA processing and visualization techniques enhance the interpretability of results, fostering a deeper understanding of the underlying thematic structures.

Furthermore, our application of multiple correspondence analysis (MCA) offers a quantitative validation of the alignment between article keywords and topic assignments. This innovative approach contributes to a more robust assessment of the thematic associations.

By identifying the most influential authors and papers based on citation counts, we shed light on key contributors and landmark publications that have significantly impacted the development of the management field.

Overall, our study offers a comprehensive and systematic framework for topic modelling and analysis in management research, empowering scholars and practitioners to explore, interpret, and navigate the ever-changing landscape of management literature. The insights gained from this research have the potential to shape future scholarly work, influence strategic decisions in organizations, and foster innovation in the management field.

## **Method**

The method proposed within this research lays out numerous essential aspects and necessary dimensions to implement itself as an important contribution for future researchers, who wish either to cope with a large amount of data or who would like to deliver impactful and solidified research contributions within their research field. The aim is to avoid bias in conducting research. Therefore, all keywords used in research

articles have been put through an unsupervised machine learning algorithm to establish the historical and in-depth content analysis. Typically, conceptual articles cannot entirely renounce themselves from conceptual bias and therefore the analysis through a machine learning algorithm and content analysis delivers the desired results for researchers.

For this research, the total amount of textual data of 21 journals has been analysed with the help of a topic extraction algorithm to compose and provide an objective representation of the historical narrative. Therefore, the scholarly analysed "*lay of the land*" in management has been combined with MLA techniques and tools to allow for more precision and objectivity by establishing unbiased labelling of the "*topic-key-word relations*". To be precise, a "Latent Dirichlet allocation" (LDA) has been used (BLE 03). It is a statistical model and collections of documents (BLE 12) and raises the assumption that each topic is associated with multiple terms and reversed, terms can also be associated with more than one topic but have a different impact and the overall topic distribution. According to Tufts (TUF 17) LDA can be interpreted as follows:

Topic structures in a document are latent, implying hidden structures within the text.

The Dirichlet distribution is assumed to determine the mixture proportions of the topics in the documents and the words in each topic.

Assignment of words to a specific topic (TUF 17).

Based on the developments in machine learning capacity, this methodology provides a solid and robust foundation for research in which large amounts of textual data need to be analysed to create a more holistic grounded design theory. Therefore, LDA illustrates a probabilistic generative model, which takes the assumption that each document is a distribution over topics and each topic is a distribution over words (ARU 10). A dynamic approach is being delivered to emphasise and highlight the evaluation of the topics over time.

A content analysis can first be performed of all papers published in the selected journals, whereas editorial notes, errata, or commentaries should be excluded, to thoroughly examine the content of management literature, trace its evolution and identify main streams and subfields. A subsequent content analysis examines the relation of topics over varying time horizons in an. It reflects the evolution of the respective field and provides journal editors, reviewers and authors with an interpretation of the direction of the respective journal (FUR 08). This analysis is being used for an objective, systematic and quantitative consideration of all selected articles while at the same time

allowing for an interpretation of the shifting priorities of all parties involved, namely editors, reviewers, and authors, who shaped the evolution of the management field.

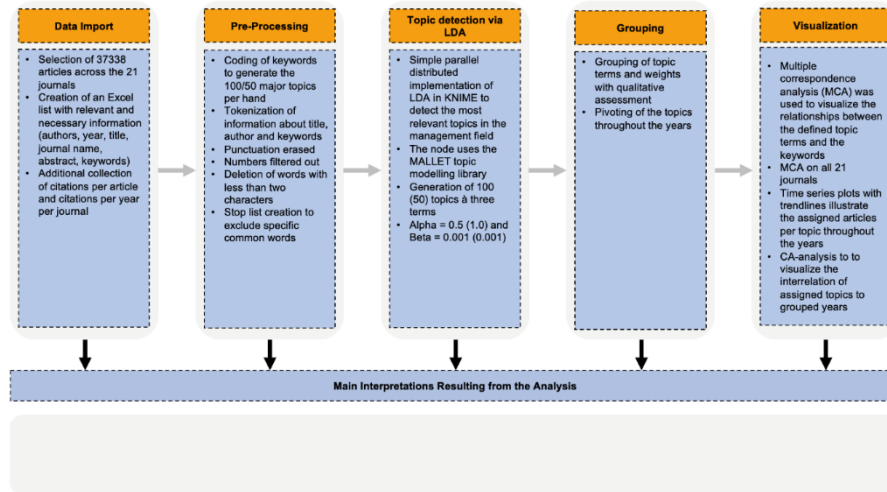


Figure 1: Exemplary workflow of an LDA topic extraction with subsequent content analysis.

As a first step, the article universe has to be selected. Meta-information such as title, author, keywords and abstracts of every article in journals in Journals with highest impact factor should serve as a base. Note, that only a large number of total articles (in our opinion  $n \gg 1000$  should serve as a threshold) justify the use of MLA. In Figure 1 we show an exemplary research project reading out  $n = 37,331$  management articles over 21 journals. We investigated two different topic models, the first with  $K = 50$  topics and the second with  $K = 100$  topics that were automatically extracted by the LDA algorithm. For this, proper pre-processing has to be installed. Further hyper-parameters such as  $\alpha$  and  $\beta$  have to be estimated a-priori.  $\alpha$  is an indicator for the document-topic distribution, that is assumed to follow a special probability distribution, that is called *Dirichlet distribution*. The higher the parameter  $\alpha$  the more topics are assumed to be prevalent in the research articles. Accordingly,  $\beta$  measures the a-priori belief of a topic-term distribution, which is also assumed to follow a Dirichlet distribution. Usually,  $\beta$  is small, i.e. close to zero, indicating that there is only a few number of terms (or keywords) that fully identify a research topic, i.e. the topic content is quite specific, which is usually the case for research articles since academics tend to use highly specialized vocabulary in their respective fields. In a second phase, we conduct a content analysis by pivoting the topic-labelled articles over different year horizons (usually five years) and by counting them. With this we can identify patterns of different topics being prominently discussed in certain years. Finally, we

highlight our results in a bubble diagram showing year group bubbles versus extracted topics. We can see topics treated more frequently in the past and new emerging topics such as digitalization and corporate social responsibility (CSR). Moreover, evergreen topics, being discussed almost every decade, such as pricing, performance and other operation research topics are found in the center of the diagram (cf. Figure 10).

### Pre-processing

For the execution of the analyses the machine learning software “KNIME” can be used. KNIME is an API commercial software with an ML drag-and-drop solution. The KNIME workflow, which has been developed for this methodology consists of 47 nodes in total, illustrated in Figure 2. A node is a pre-implemented snippet of algorithms that can be used in a drag-and-drop environment for building the overall analysis workflow. These nodes are divided into nine “meta-nodes” (combining single nodes) after the data has been send through the LDA. Nevertheless, before generating a reliable and robust output, the previously collected data must be optimized for the LDA. Therefore, eight nodes are necessary to reach the ideal data set for the LDA.

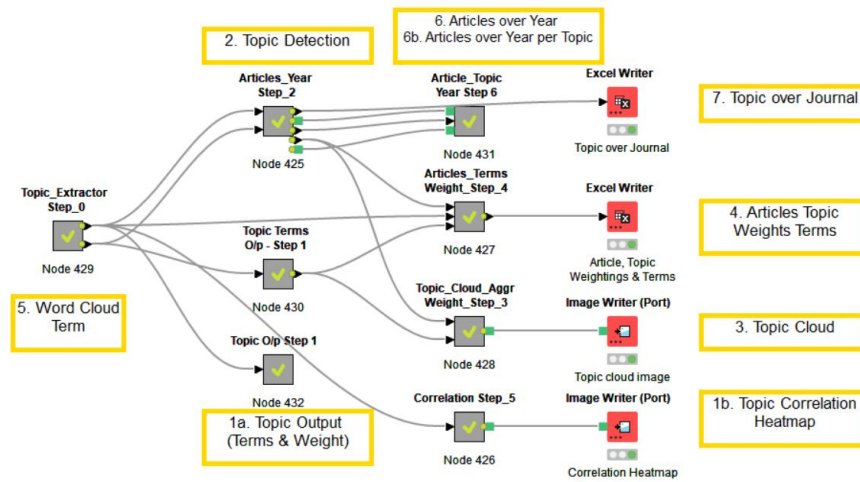


Figure 2: Overall KNIME workflow subsumed in meta-nodes.

The first node is the “Excel Reader”, which reads Excel files (either single or multiple files at the same time), but is only able to read one sheet per time. The Excel types which are supported are string, number, Boolean, formulas, date, and time. Pictures and diagrams cannot be read. The data is then read in and converted into KNIME

types string, integer, long, double, Boolean, local date, local time and local date and time. The node will scan the input file and determine the number and types of the columns. The output then represents a table with the auto-guessed structure and KNIME-types. Before starting the node, it is important to notice that the presets of the node will only cover 15,000 rows. In the advanced settings of the node it is possible to set this number to the desired number (the total number of articles) (hub.knime.com ---a).

The second node converts the specified strings into documents. A document will be created for each row and then attached to that specific row. The strings of the columns will be used as title, authors, and full text, while additionally the defined category, source type and data will be set. Since this research aims at determining the correlation of articles and keywords to a specific topic, the text section of the node will be assigned to the keywords, which are stated in the input document. The keywords can be retrieved from the databases *Web of Science* and through the websites of the publishers of the respective journals (hub.knime.com ---b).

The information about the title, author and keywords are transformed into a document data type and then adjusted by using a word tokenizer (*OpenNLP English WordTokenizer*). The so-called tokenization is the process of chopping the given sentence into smaller parts, the tokens, which is used in tasks such as spell-checking, processing searches, identifying parts of speech, sentence detection and document classification (TUT 21). Nodes three to six remove all punctuation characters in the input document, filters out all digits, including decimal separators ",", or "." and "+" or "-", and terms in the input document with less than  $N$  characters and converts all terms in the input document into lower- or upper-case letters, respectively. All terms with less than three characters in the input document have to be deleted, all letters have to be turned into lower case letters, and the raw text has to be tokenized based on a set of delimiters, e.g., whitespaces and punctuation.

The next steps include the integration of a built-in “stop list” into KNIME, containing specific words that are very common in the English language, and a manual one with additional very common words in the field of study. Those words are excluded in the analysis due to the likelihood of the distortion of the results given their weight and impact. Figure 3 visualizes the first nine nodes in the KNIME workflow used for this research.

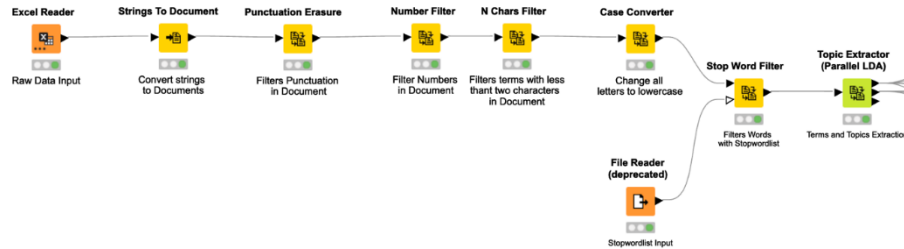


Figure 3: Pre-processing steps in KNIME, visualization of the first nine nodes including the LDA

Once all these eight nodes have been successfully performed, the LDA can be configured and started.

### Topic Detection via LDA

To analyze and detect the most prominent topics in the field of study during the chosen time span, a simple parallel distributed implementation of LDA in accordance with Newman et al. (NEW 09) is used with the addition of the sparse LDA sampling scheme and data structure according to Yao et al. (Yao 09).<sup>1</sup> This node is the second node in the workflow and represents the core of the model. Therefore, LDA needs an a priori defined number of topics ( $K$ ), the number of words, or terms ( $T$ ) as commonly used in LDA literature, associated with the topic, as well as additional distribution parameters alpha ( $\alpha$ ) and beta ( $\beta$ ) as described above. This serves the purpose of strongly favoring sparse word distributions. Again, a high alpha is used for documents that cover many topics and has therefore a lower impact of topic sparsity. A low alpha is being used in cases where only few topics are covered, but with a higher impact of topic sparsity. Usually, alpha is set to a fraction of the number of  $K$  (KON 18). Figure 4 below illustrates the topic sparsity.

<sup>1</sup> From the authors' experience, increasing the number of terms per topic leads to a blurred interpretations of the overall topic detection and is inevitably linked to an increased effort in preprocessing (e.g., customization of the stop word list). The intention was to reduce the preprocessing amount to as few steps as possible.

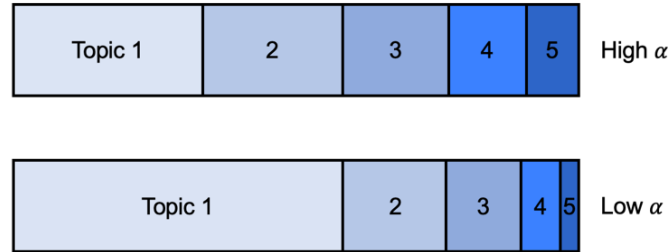


Figure 4: Topic sparsity for high and low Alpha

If a high beta is being used each topic consists of many words, having a low impact of word sparsity, and therefore the output will cover more general topics. However, when a low beta is being used, each topic consists of a few words, hence a higher impact of word sparsity, thus more specified topics. Beta is being used for the granularity of term-topic distribution, indicating that a high beta allows for fewer more general topics, while a low beta allows for more specific topics (KON 18). Figure 5 below illustrates the word sparsity.

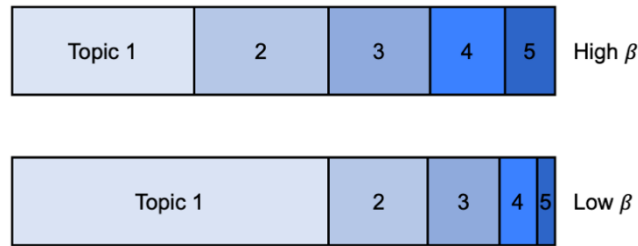


Figure 5: Word sparsity for high and low beta

Nevertheless, to determine the right alpha and beta values three sensitivity analyses, for coarse, fine, and medium granularity, can be performed as a preliminary analysis. Hereby, it has been decided to use the example of the coarse granularity. Figure 6 illustrates a foregoing sensitivity analysis for a coarse granularity of topics, i.e. by using higher alpha and beta values and by evaluating various “goodness of fit” measures for LDA topic modelling. For example, the “coherence”, which was predominantly used in our studies, measures how well the extracted keywords characterize each topic. The higher the coherence, the better the fit. There are also other evaluation measures, such as the so-called “perplexity” that are optimal for our research if they are minimized, i.e. the lower the perplexity scores the better the fit. However, recent studies show the perplexity is often counter-intuitive to human language (CHA

09). Unfortunately, in real-life topic extraction projects not all evaluation measures tend to equally point into the same evaluation direction. As we can see in Figure 6 once again, the perplexity score (to be minimized) would speak for a very small number of topics to be extracted whereas the measure proposed by Cao et al. (CAO 09) indicates a very high number of topics. Arun et al (ARU 10) even has a global minima for  $K = 50$  topics. Similarly for the coherence measures (to be maximized). The standard (textbook method) coherence measure as pre-implemented in the Python package *Gensim* speaks for a higher number of topics, whereas the coherence measure proposed by Mimno et al. (MIM 11) indicates a smaller number of topics. As already indicated, in our research projects realized so far, usually the Gensim coherence correlates to our human judgment when we qualitatively evaluate the proposed research topics with respect to their associated keywords. For other research projects it might be valuable to also hold in mind the other evaluation measures. Therefore, we would like to highlight every single evaluation measure in due brevity.

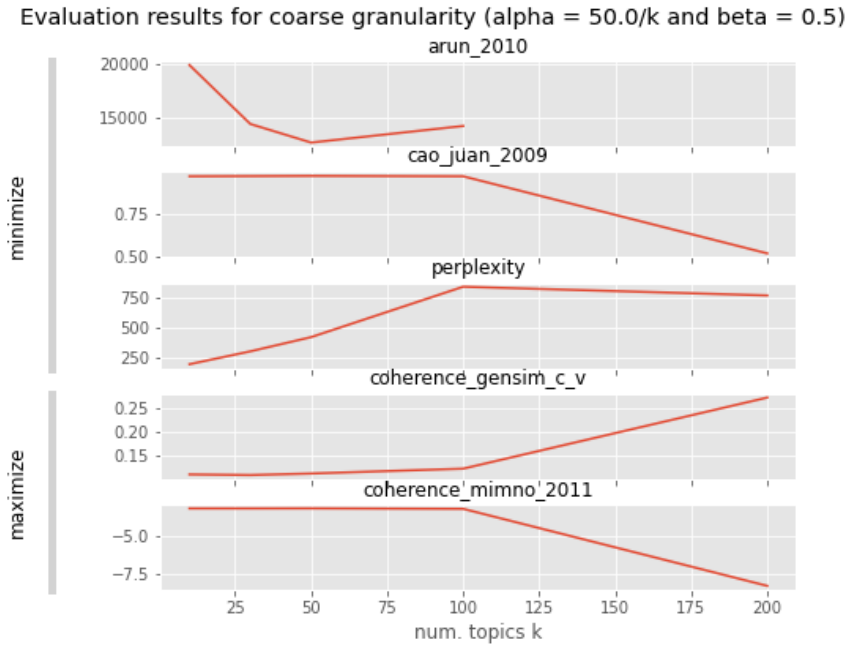


Figure 6: Evaluation results for coarse granularity

Arun et al (ARU 10) present empirical results indicating that the “graph dips down and hits a low” for the “right” number of topics and increases as the number of topics also increases. Therefore, the right number of topics is any number in a small range giving the best accuracy out of the respective dataset. In this case in accordance with Arun et al. (ARU 10) 50 topics seem to be the right number of topics. Nevertheless, the value is still low even for 100 topics.<sup>2</sup>

Cao et al. (CAO 09) propose that the LDA model performs best when the average cosine distance of topics reaches a minimum, hence this graph should be minimized. They find the best number of topics based on the topic density. In this research the graph starts minimizing once it has reached 100 topics. Therefore, this method would suggest using more than 100 topics, with reaching its lowest point around 200 topics.

The third measure deals with the perplexity. Perplexity is a commonly used indicator, where a lower perplexity indicates a better prediction (BLE 03; JAC 16). The graph shows for this case how the perplexity increases as the number of topics increases. Once the graph has reached 100 topics it starts decreasing. Therefore, the perplexity method also suggests using as many topics as possible or as few topics as possible.

The fourth measure deals with the coherence. Röder et al. (RÖD 15) imply that coherence measures have gained importance in text mining and unsupervised learning methods such as topic modeling (proposed in this research) since these methods give no guarantees on the interpretability of the output. Here the graph is being maximized the more topics are being selected and increases once it reaches 100 topics.

Additionally, Mimno et al.(MIM 11) propose a fifth measure, also dealing with the coherence, that generally models with larger numbers of topics are being preferred, as these models offer a higher resolution and the possibility of supporting finer-grained distinctions. Nevertheless, as the number of topics increases, the smallest topics can be of poor quality (MIM 11). The last graph in Figure 6 above clearly indicates that the topic quality after 100 topics decreases; hence, 100 topics should be chosen.

Regardless of the five aforementioned measures, in order to choose the “right” number of topics a certain degree of human judgement is necessary. Hereby it is important to notice that predictive accuracy and human judgement on the quality of topics are often not correlated (CHA 09), hence e.g. the measure of perplexity cannot be observed. After considering all the aforementioned measures it has been decided to use two models, which are explained in the following.

---

<sup>2</sup> The Arun et al. measure was not defined for a topic number higher than 100 due to internal algorithmic constraints.

In our example workflow, as indicated by Figure 1, our first model works with  $K = 100$  topics corresponding with three keywords per topic. Alpha was defined with 0.5. Beta was defined with 0.001, which was also the predefined setting of KNIME. The second model, with  $K = 50$  topics, corresponds to three associated keywords per topic. Alpha was defined with 1.0 and beta was defined with 0.001 once again. Therefore 100, or respectively 50, major topics were identified by coding the keywords of the selected articles. Additionally it is noteworthy that previous studies in Management literature, such as Furrer et al. (FUR 08) have been limited to only 26 keywords (see table 1 in (FUR 08)). Table 1 below once more illustrates the different parameters chosen.

Table 1: Model description with parameter selection

Parameters	Model 1	Model 2
Topics ( $K$ )	100	50
Terms ( $T$ )	3	3
Alpha ( $\alpha$ )	0.5	1.0
Beta ( $\beta$ )	0.001	0.001
Sensitivity	Coarse Granularity	Coarse Granularity

The topic relevance is firstly assessed through the topic weightings given as an output measure by the LDA algorithm and then by a quantitative assessment based on the proximity of the originally published article keywords to the topic context in terms of a decadal and thematic (topic) assessment by an additional content analysis. The LDA algorithm makes use of the “*MALLET: A Machine Learning for Language Toolkit*” topic modeling library. MALLET is a Java-based package designed to deliver statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text (MCC 02).

Probabilistic topic modeling can be assigned to the branch of unsupervised machine learning algorithms aiming to annotate large achievements of documents with thematic information (BLE 12), i.e. to discover abstract topics in collections of unlabeled documents (KON 18). Since the model is unsupervised it does not require any a priori labeling of the documents. The topics then emerge from the analysis of the original texts in form of a dimension reduction which is similar to the numeric counterpart method of a principal component analysis (PCA), which is a simple, non-parametric method for extracting relevant information from cluttered datasets (PAU 13). Nevertheless, LDA differentiates itself as it aims to reduce the information of textual (non-metric) data with the use of Bayesian statistics. LDA assumes that all collected documents share the same set of topics, but each document is exhibiting these topics in different proportions (BLE 12). Speaking from a model-theoretic-based approach,

a topic is the distribution over a fixed number of words, and it is assumed that these topics exist even before the generation of the data. LDA is a three-level hierarchical Bayesian model. Here each word of a collection is being modeled as a finite mixture over a pre-existing set of topics. Each topic is modeled as an infinite mixture over a set of topic probabilities (BLE 03) . The described three-level process can be further explained as follows:

During step one, each document exhibits all topics but in different proportions. Step two randomly selects a topic which is chosen from the per-document distribution over topics. Afterwards, during step three, every word in each document is being drawn from the previously chosen topic. To solidly ground the models methodology, the probabilistic model is going to be described, based on the preceding literature such as Blei (BLE 12)

The stochastic problem can be attributed to the calculation of the posterior distribution, or the conditional distribution of the hidden (or latent) variables for the documents provided. Therefore, the LDA can be described through the following parameters. The topics are described through  $\beta_{1:K}$  while each  $\beta_k$  is describing the distribution over words. The topic proportions for the  $d$ -th document are  $\theta_d$ , whereas  $\theta_{d,k}$  describes the topic proportion for topic  $k$  in document  $d$ . The topic assignments documents are  $z_d$  , whereas  $z_{d,n}$  describes the topic assignment for the  $n$ -th word in document  $d$ . Lastly, the observed words for each document  $d$  are  $w_d$  , whereas  $w_{d,n}$  describes the  $n$ -th word in document  $d$ , which is a component from the fixed vocabulary that originates from the automatic read-out bag-of-words representation, which is a way of text data representation when text is being modeled through machine learning algorithms (BLE 12). Therefore, the LDA process can be described and the joined distribution of the hidden and the observed variables can be stated as follows(BLE 12):

Equation 1: Joined distribution of hidden and observed variables

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^K p(\beta_i) \prod_{d=1}^D p(\theta_d) \left( \prod_{n=1}^K p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

Through this formula, numerous dependencies are being uncovered. According to this the topic assignment  $z_d$ , is depending on the per-document proportions  $\theta_d$ . Moreover, the observed word  $w_{d,n}$  is dependent on both, the topic assignment  $z_{d,n}$  as well as on all topics  $\beta_{1:K}$  (BLE 12).

These dependencies define the LDA and are visualized in figure 8. This figure illustrates, how each node displays a random variable to its role in the generative process. Words within a document are characterized by shaded nodes. The rectangles denote replication. The  $N$  plate denotes the collection words within the documents,

while the  $D$  plate denotes the collection of documents within the collection (BLE 03). Hidden variables are displayed by unshaded nodes whereas observable variables are depicted by shaded notes. Rectangular boxes indicate multiple iterations.

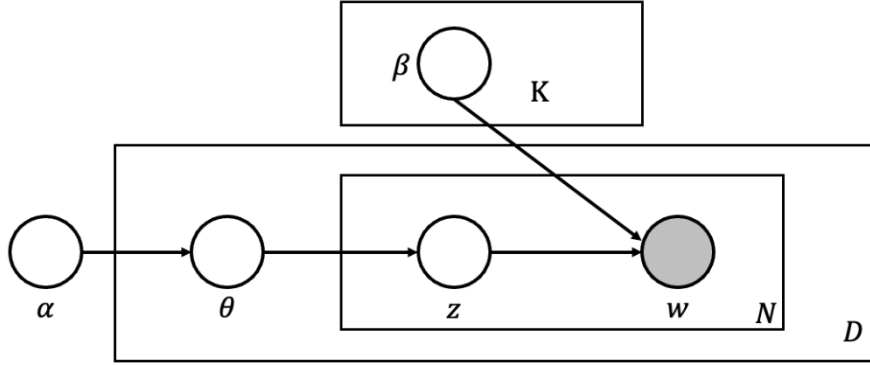


Figure 7: The graphical model for LDA according to (BLE 12).

The posterior probabilities can be calculated as follows:

Equation 2: Posterior probabilities

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:D}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

In this formula, the numerator is describing the joint variable of all random variables, precisely the topic distributions, document-specific topic proportions, topic assignment, and the word collection. The denominator is describing the marginal probability of the words. Theoretically they are computed through the aggregation of the joint distribution over every possible instance of the underlying hidden topic structure. Practically speaking this distribution is difficult to calculate (DIC 83). Therefore, a variational Bayesian framework as means of an approximation method for the posterior probabilities is being proposed. This approximation method is assuming a Dirichlet distribution for each topic  $\beta k$  (ATT 01; KAM 21). Additionally also Blei et al. (2003) suggest to use another Dirichlet approximation. This approximation estimates the per-document topic proportions  $\theta d$  based on a fixed parameter called  $\alpha$  (see Figure 8 above).

Since for this research two exemplary models have been created the constant parameter differ as previously described in this chapter. The constant parameters for

model one (model two) are  $\alpha = 0.5$  ( $\alpha = 1.0$ ) and  $\beta = \beta k = 0.001$  ( $= \beta k = 0.001$ ) as well as  $K = 100$  ( $K = 50$ ) been chosen. Each topic then is being associated with  $T = 3$  major terms, whereas these terms describe the keywords indicating the highest probability that certain terms are associated with that one specific topic. As already mentioned, the alpha parameter is defining the Dirichlet prior on the per-document topic distributions to the prior weight of the topics in a document. The KNIME library recommends using an alpha parameter of usually less than 1 for all topics to prefer sparse topic distribution, indicating few topics per document. The beta parameter defines the prior on per-topic multinomial distribution over words. Here the KNIME library recommends using a number much smaller than 1, hence the default value is set to 0.001. This serves the purpose to strongly prefer sparse word distributions. To further validate the robustness, a grid search optimization over the hyper-parameters  $\alpha$ ,  $\beta$ ,  $K$  and  $T$  could be performed. Nevertheless, the grid search optimization presents itself as computationally intense, hence a smaller variation analysis is proposed. This analysis aims at illustrating a first impression of how a successful grid search optimization could look like (KAM 21).

### ***Post LDA***

After the first nine nodes in the KNIME workflow, including the topic extraction through the LDA have been completed, the next four nodes play a vital part for the processing and evaluation of the input data.

The first node (GroupBy) groups the rows of a table by the unique values in the selected group columns. For each unique set of values, a row is created within the selected group column. The columns which remain are aggregated and are based on the specified aggregation settings. Finally, the output table contains one row for each unique value combination of the selected group columns. The next node is a “JavaSnippet” which allows to execute arbitrary java code to create new columns, respectively replace already existing ones. In this case it renames the topics, being named “topic\_0” to “topic\_00”, for reasons of simplicity while dealing with the output. The next node, the “Sorter”, sorts the rows according to the predefined criteria, meaning it sorts the topics starting from topic\_00 to topic\_99. Once these three nodes have been processed the last node of this process is an “Excel writer”. This node transfers the input data table into an Excel file, which directly downloads once it is being created by KNIME. Table 2 illustrates the output of the Excel writer, for this research exemplary for the first 10 topics of the first model with 100 topics of this research.

*Table 2: Exemplary first 10 topics of the 100 topic model*

Topic id	Concatenate(Term)	Sum(Weight)	Detection
topic_00	customer, satisfaction, relationship	2584	Customer Satisfaction
topic_01	health, care, healthcare	892	Healthcare
topic_02	sharing, contracts, coordination	888	Contracting
topic_03	marketing, science, management	774	Marketing Science
topic_04	intertemporal, replication, formation	243	Intertemporal Choice
topic_05	capacity, management, project	1273	Project Management
topic_06	modeling, structural, model	998	Structural Modeling
topic_07	capital, human, venture	999	Human Capital
topic_08	simulation, estimation, variance	826	Simulation & Estimation
topic_09	goal, motivation, outsourcing	526	Outsourcing
topic_10	queue, systems, queueing	719	Service Optimization

At last, topic terms and weights have to be interpreted, and each topic is named by a qualitative assessment of the authors plus an additional of two or more academics before they can be pivoted over the years. Once this list has been created, an extra column, named “Detection” has to be added next to the column illustrating the weight. The detection column specifies the topic for each row, e.g., topic\_00 consists of the term’s customer, satisfaction, and relationship, therefore the assigned topic in this case can be interpreted as “Customer Satisfaction”. Once this process has been completed for both models, 100 and 50 topics, the excel list is once again put into an excel reader (previously described in section 0) within the KNIME workflow. Once that list has been put into the KNIME workflow, several analyses can be performed, as to be described in section 0.

Further, various robustness checks can be conducted to evaluate the quality of the resulting topic-term distributions, as already indicated in Figure 6. The association of terms to keywords and topics is analyzed qualitatively for an independently chosen subset of articles suggesting the economical soundness of the LDA clusters. In the following section 0, we further present a MCA analysis to quantitatively validate the concordance of article keywords and topic-term associations. Therefore, a two-component based MCA can be used to project the underlying variation of data into two main factors. As our experience shows, an enlargement in the number of principal components, does not generally lead to an increase in explained variation (regarding the maximum Eigenvalue criteria). Therefore, it can be assumed that the validation analysis with two main components is sufficient to test the suitability of the LDA projections (KAM 21).

Nevertheless, before turning to the results of the MCA and the content analysis, several limitations in regard to the proposed method can be depicted.

### ***Limitation of the LDA Methodology and Comments***

LDA makes three assumptions which are going to be commented in the following.

The first assumption of LDA is the “bag of words” representation, i.e., the order of the words in the document does not matter. This is problematic from a semantic viewpoint, especially when it comes to language generation. Nevertheless, this assumption does not limit analysis mainly consisting of stand-alone keywords.

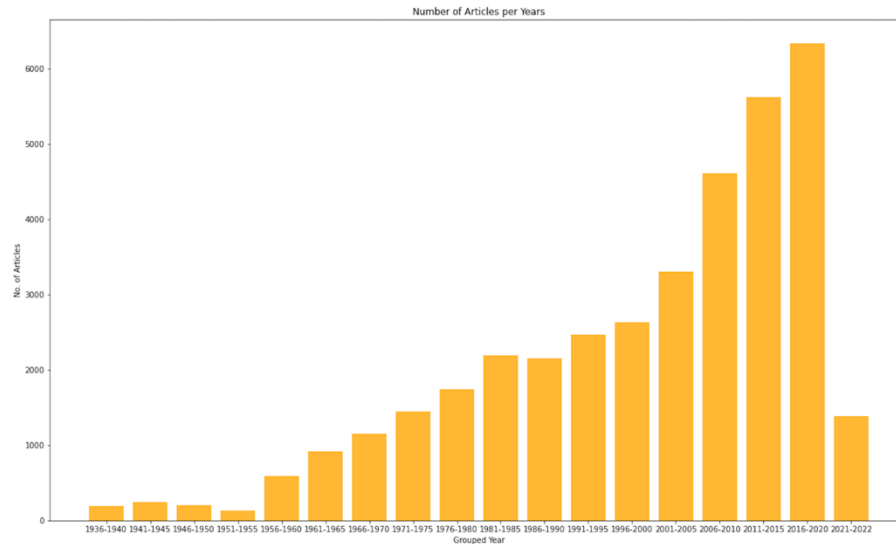
The second assumption is that the order of the documents does not matter. As stated in the subsequent below the focus in certain topics can shift among time. Thus, it may be reasonable to estimate multiple LDAs over time. By using this approach, a topic would be a sequence of distributions over words. To capture these dynamics, we performed a content analysis that highlights the relevance of each topic over decades.

The last assumption about LDA is that the number of topics is to be assumed known and fixed. The Bayesian non-parametric topic model provides a solution for this (TEH 06). Hereby the number of topics is determined through the collection during posterior inference analysis and new documents can exhibit previously unseen topics.

## **Analyses**

To begin with, it is worth knowing how many articles have been published in each year. If there is a large number of years to investigate, it may be beneficial to define different timespans. Figure 8 exemplifies the number of chosen articles over a 4-year timespan, for our example workflow. As it can be seen, the number of research articles in Management Science (comparably to other disciplines) tend to rise in the recent decade. Note that the last time span declines since the collected articles were selected in the beginning of the year 2022. Thus, the overall collection pre-dominantly consists of articles from 2021 compared to a full 4-year timespan of articles in previous periods.

For understanding, readability, and simplicity purposes, only one model of the previously mentioned two is going to be used for explaining the ongoing analyses as additional part of our research methodology. A multiple correspondence analysis (MCA) is used to examine and analyze the relationship between the topic terms and keywords to gain a deeper insight of the relationships between the topics. Additionally, the MCA can be used to evaluate the relationships between the topics over time (FUR 08).



*Figure 8: Number of articles per timespan*

MCA describes an exploratory data analysis technique used for graphical representation of multivariate categorical data (BEN 84; HOF 86; LEB 84). This technique aims at explaining the interdependence between a number of categorical variables, in this case the topics (themes) and the keywords. It is similar to the principal component analysis introduced by Hoffmann and Leeuw (LEE 92). Through this technique the researcher is being given the opportunity to examine and analyze multipath tables and to determine the structure in the relationship between the nominal variables (FUR 08). Through this analysis similar counting patterns between the different rows and columns of a frequency table can be highlighted.

Figure 9 below shows that the collected keywords from all articles (blue dots) of the example data overlap perfectly with the associated terms (green dots) and the subordinate topics (orange dots). This outcome supports the use of the LDA topic extractor, which has been previously described.

For the purpose, a two-dimensional factor has been determined as the most suitable option for the graphical representation. The corresponding eigenvalues were both around one with an explained inertia of 0.02% for each component. As Kaciak and Louviere (KAC 90) noted, the proportion of total variance explained by the dimensions is often very small. This is closely related to the (necessarily) binary nature of the transformed nominal data (LEB 84). Additional sensitivity analyses can be carried

out for higher orders of components ( $n = 2 \dots 10$ ), which did not lead to any improvement in the maximum explained inertia per component in this example.

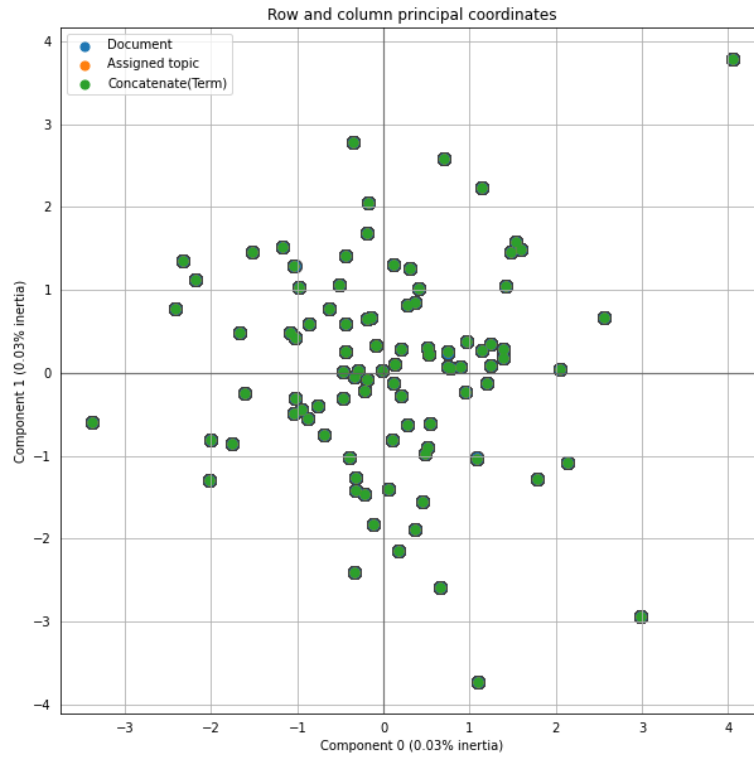


Figure 9: Exemplary MCA Analysis illustrating the proximity of observable research article keywords (blue dots), topic assignments (orange dots), and (concatenated) terms assignment (green dots), with the dot size indicating the number of articles associated with the categories.

In general, the MCA results illustrated above proof a high match between the observed literature articles and the topic-term distributions. In subsequent analyses, researchers can evaluate advanced visualization methods which are included in the Python library LDAvis developed by Sievert et al. (SIE 14). The authors propose a new relevance score in order to assess the degree to which a term belongs to a topic by using a fraction of log-likelihood of the term  $w_{d,n}$  and the approximated Bayesian

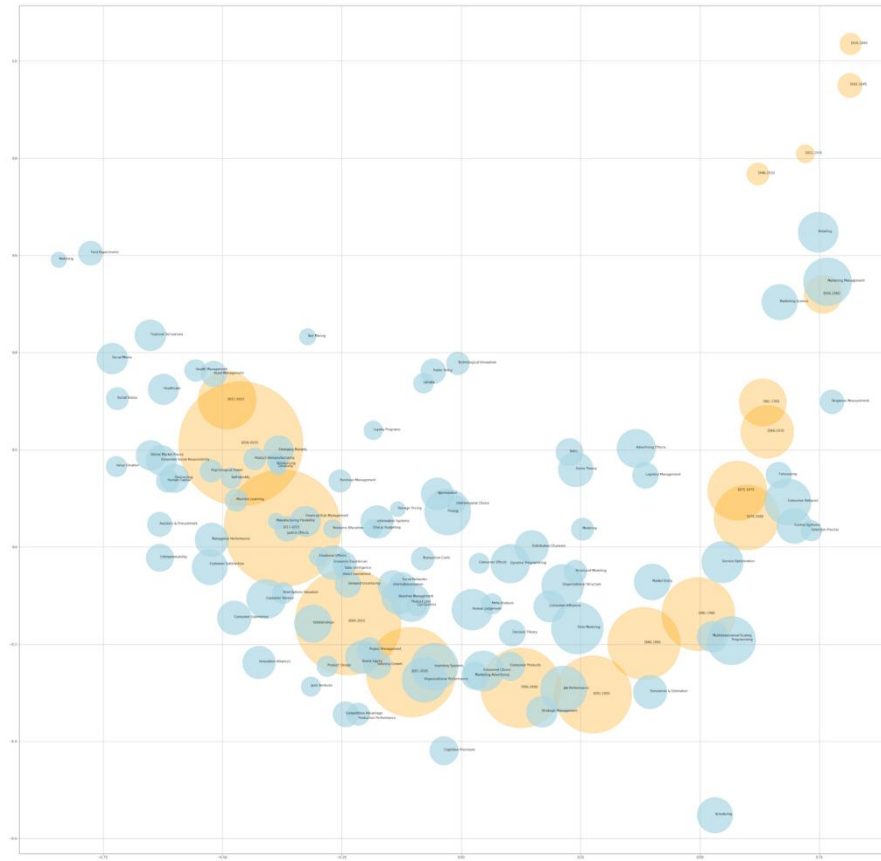
distribution over the empirical term frequency (as introduced in Equation 1 of section 0).

Previous studies such as Bischof and Airolidi (AIR 12) simply ranked the terms by the probability under a topic. This inadequacy is being overcome by the proposed measure in this methodology. In the previous section a different visualization, comparable to Sievert et al.'s (SIE 14) method, by using an MCA between topic term distributions, has been used. Additionally, to highlight the changes of topics over time, a content analysis can be performed, distinguishing between topic and year associations.

### **Results of the Content Analysis**

The exemplary content analysis as depicted in figure 10 was executed to illustrate the relationships between the assigned topics and the grouped years in our example. Through a pivot table, which was created after the LDA extraction from the assigned topics over years, it was determined how frequent a published article appeared within each topic over each year group. Once again, to simplify the graphical representation, a two-factor model has been chosen as the most suitable option. The blue bubbles describe the topics, while the orange bubbles describe the time-period. The size of the orange bubbles indicates the number of articles associated with the respective timespan, while the size of the blue bubbles indicates the number of articles associated with each topic. The closer the year and topic bubbles are together the closer their association in sense of thematic proximity. This is also the case from a geometrical viewpoint, in accordance with the Euclidean distance, which is the length of a line segment between the two points. The size of the topic bubbles describes the number of articles published on that topic; hence larger topic bubbles have more articles associated with. The larger the size of the year bubble, more articles have been assigned to that specific time-period, and hence more articles have been published during that time-period.

A pair of coordinates in a two-dimensional space is provided by the analysis for all articles in the exemplary input dataset. Nevertheless, the map would be impossible to interpretate if a dot for each article would be presented, and therefore only the position of the topics is being indicated. Moreover, the MCA described earlier implicates that the article keywords and the assigned topics are closely correlated.



*Figure 10: Exemplary content analysis indicating the association of topics (blue bubbles) with time decades (orange bubbles)*

Further, numerous time series analyses can be performed to marginally assess the evolution of a single topic over time, a small excerpt of examples is presented in Figure 11. A positive, respectively negative slope of the trend line (red line) is indicating an increase or decrease in the average amount of articles which can be associated with each topic throughout the timespan of analysis. The time series plots underline the relevance of the topics over the years proving a constant increase or a high number of publications constantly throughout the years or respectively the opposite.

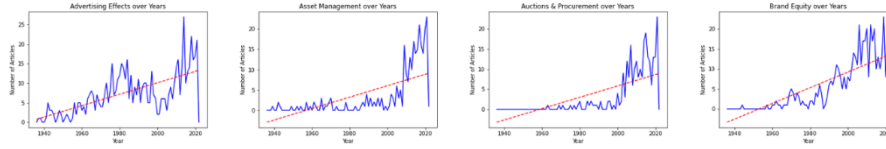


Figure 11: Exemplary time series plots (blue) with trendlines (red) showing the evolution of articles per topic over time

Once again, it has to be emphasized that the underlying topic modeling methodology belongs to the category of unsupervised learning algorithms which were not fed by a-priori labeled topic or term names, as can be seen in previous literature studies such as Furrer et al. (FUR 08). The algorithm simply groups keywords that are close in topic, which, as the previous robustness checks have shown, seem reasonable from a management perspective. This overcomes subjective topic labeling by researchers embedded in a particular research environment, which could lead to biased topic evaluations.

In addition to the assignment of topics to a specific timespan, the analysis can also output the following Figure 12. Here the topics are being assigned to the journals analyzed, indicating which focus each journal has. This offers a particularly interesting insight especially for researchers regarding new publications and where they could best submit their newest publications based on their topic. For example, researchers focusing on “Logistics Management”, “Programming” or “Scheduling” have a high probability of being published with one of those topics in the *Journal of Operations Research*, while researchers focusing on “Strategic Management”, “Innovation Alliances” or “Internationalization” will likely be rejected by the *Journal of Operations Research* but have a better chance of being published in the *Strategic Management Journal*. Similar to Figure 11, which assigned the topics to a specific timespan, this illustration also presents topics, which cannot be precisely assigned to one journal since their overlap with the topics represented in the journals is too marginal. Namely, for example these topics are, “Game Theory”, “Gender”, “Technological Innovation” or “Purchase Management” to only name a few in the example data set. These topics are additionally represented in relatively small bubbles simply because not enough articles could be associated to these topics and therefore the assumption arises that they could not be assigned to a specific journal.



*Figure 12: Content analysis indicating the association of topics (blue bubbles) with journals (orange bubbles)*

Once again, the size of the orange bubbles indicates the number of articles associated with, in this case, each journal, while the size of the blue bubbles indicates the number of articles associated with each topic. The closer the bubbles are together the higher their correlation. Therefore, this figure allows researchers, aiming to publish an article, to pick the right journal precisely fitted to their research field.

Figure 13 illustrates the number of articles associated to each extracted topic of our exemplary workflow (with  $K = 100$  topics). In this example, there was a large group of research articles associated to the topics of “Data Modeling”, “Programming”, and “Inventory Systems”, underpinning once again the impact of digitalization and automation issues in Management Science.

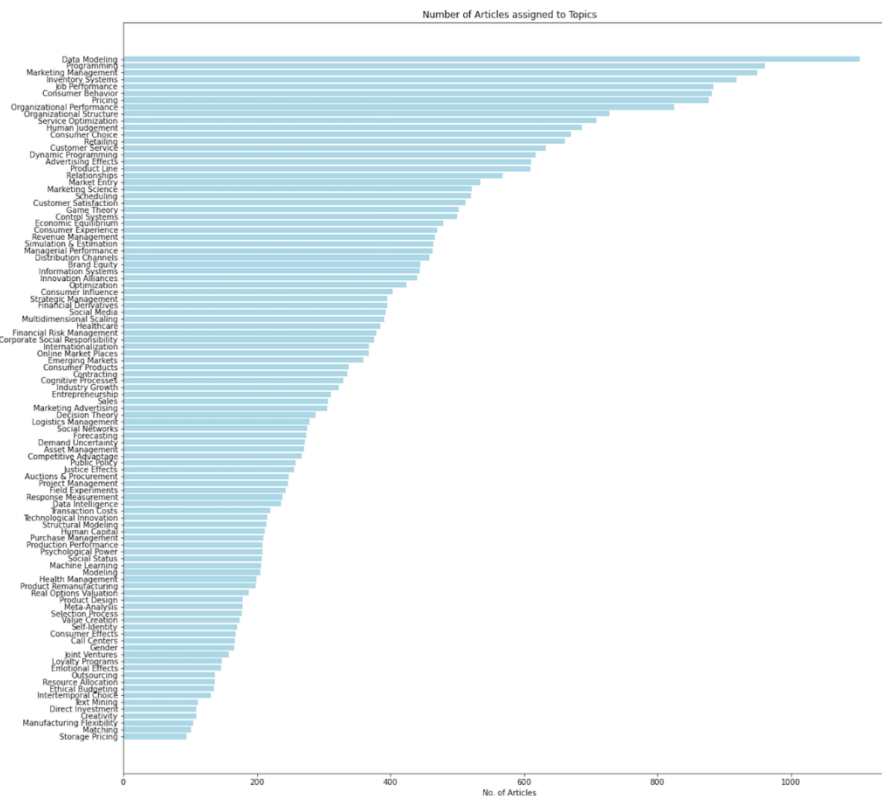


Figure 13: Number of articles assigned to topics

## Contributing Authors

Bergh et al. (BER 06) concluded that the characteristics of each author have the greatest explanatory power regarding the impact of the article. They implied that those researchers, who published the most within a certain period, subsequently had the strongest influence on topics researched in the subsequent period. For the purpose of this research, the authors with the most publications throughout the time span of publication of all articles are proposed to be analyzed to fully understand the history of the management field and to allow for possible future implications and topics. Moreover, Berry and Parasuraman (BER 93) proposed a model regarding the forces which influence the development of a certain academic field. Hereby key individuals (authors) and key publications act as accelerating factors for each academic field. Therefore the most published authors as well as the most published papers in regard to their impact can be analyzed with the proposed methodology (FUR 08).

Through this analysis, the contributions of researchers over the time period analyzed are highlighted, differentiating between long-established scholars as well as new generations of scholars. New scholars will face the task of determining future research and the direction the field of management is heading. The number of published papers by an author runs proportionally to his or her career length, hence the authors on top of the list have the longest career, and therefore their influence on the development of the management field can be regarded as the highest.

### **Most Influential Papers**

Specific publications can be assigned with a vital and central role regarding the growth of the scientific field they were published in, since their influence accelerated the growth of the discipline. In order to understand the future possibilities and capabilities of the management discipline from a holistic perspective, it is necessary to detect and identify the major papers published in all journals analyzed for the respective time span. Therefore the approach of summed citation counts can be used to detect and assess their impact or influence on research papers (BER 06; RAM 04; TAH 99).

The most cited articles of the respective journals can be determined based on the citation count they have been assigned with in the Web of Science. Since papers, which have been published at an earlier point in time have a greater chance of receiving a higher citation count than papers published at a later stage in time, the articles have to be sorted by the number of citations divided by the number of years they have been published. The procedure proposed by Furrer et al. (FUR 08) can therefore be used.

### **Box plotting**

In addition to the determination of the most published authors and the most influential papers, regarding their citation per year, several boxplots for each topic can be determined. This serves to determine whether a topic, a researcher is writing about, leads to a potentially high citation count of the respective article.

Boxplots are diagrams used to graphically represent the distribution of an at least ordinal scaled characteristic. An ordinal scale sorts variables with expressions between which there is a ranking order. It thereby combines various robust measures of dispersion and position in one representation. A boxplot is intended to quickly give an impression of the range in which the data lies and how they are distributed over this range (KRO 14). Figure 14 below illustrates an exemplary of the structure of a boxplot.

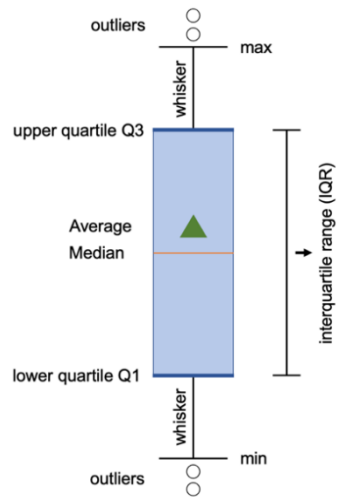


Figure 14: Exemplary Boxplot structure.

The lowest observed score, excluding the outliers, is shown at the end of the bottom whisker. 25% of all scores are illustrated below the first, while 75% of all scores are illustrated below upper quartile, hence 25% of all scores are above that value. The median marks the mid-point of the analyzed data and divides the box into two parts. Similar to the lowest observed score, the highest score, excluding outliers, is shown at the end of the top whisker. The whiskers illustrate the scores which are outside the middle 50%. The interquartile range simply illustrates the middle 50% of all scores. The outliers can be seen as an observation, which is numerically distant from the rest of the observed data. These outliers are located outside the whiskers (MCL 19). The float determines the reach of the whiskers beyond the first and third quartile. In other words, where IQR is the interquartile range ( $Q3 - Q1$ ) the upper whisker extends to the last datum less than  $Q3$  plus the whisker times the IQR (or briefly  $1.5 \cdot IQR$ ). Similarly, the lower whisker will extend the first datum greater than  $Q1$  minus the whisker times the IQR (or briefly  $1.5 \cdot IQR$ ). The outliers are plotted as individual points. In the Python code this value is set unreasonably high to force the whiskers to show the minimum and maximum values.

The higher the boxes the higher is the probability that an article about the respective topic is being cited, hence the lower the box the lower the probability to reach a high citation count with that respective topic. The higher the value on vertical axis the higher is the citation count of the respective topic. If the median and the average are on top of each other or at least close together the more symmetrically distributed the topic.

An additional boxplot analysis of only the most cited articles per journal can be performed to be compared to the results of the total dataset. The two boxplot analyses give a deeper insight into which topics offer the highest opportunity to be cited for researchers.

## **Conclusion**

In this study, we have presented a comprehensive methodology for topic modeling and analysis of management research publications. By applying LDA to extract topics, we successfully revealed latent thematic structures within the corpus. The post-LDA processing facilitated the visualization and interpretation of topic-term distributions, contributing to a better understanding of key research themes.

The application of multiple correspondence analysis (MCA) provided quantitative validation, confirming the alignment between article keywords and topic assignments. Through content analysis and time series plots, we delved into the evolution of topics over time, highlighting persistent areas of interest and emerging trends in management research.

By examining the most influential authors and papers based on citation counts, we acknowledged the significant contributions of key individuals and landmark publications in shaping the management field. Moreover, the boxplot analysis allowed us to identify topics with higher potential for citations, aiding researchers in making informed decisions when selecting research topics.

Overall, our methodology offers valuable insights into the management research landscape, aiding scholars, practitioners, and journal editors in understanding and navigating the diverse and ever-evolving field of management. As the body of management literature grows, the utilization of topic modeling and analysis becomes increasingly important in facilitating knowledge discovery and fostering new avenues of research.

## References

- [ARU 10] Arun, R.; Suresh, V.; Veni Madhavan, C. E.; Narasimha Murthy, M. N. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In: Zaki, M. J.; Yu, J. X.; Ravindran, B.; Pudi, V. (eds.): *Advances in Knowledge Discovery and Data Mining, Part I. 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21-24, 2010, Proceedings*. Vol. 6118. New York: Springer (Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence Ser), 2010, pp. 391–402.
- [ATT 01] Attias, H.; Platt, J. C.; Acero, A.; Deng Li: *Speech Denoising and Dereverberation Using Probabilistic Models*, 2001.
- [BEN 84] Benzécri, J.-P.; Bellier, L.: *L'analyse des données. Leçons sur l'analyse factorielle et la reconnaissance des formes et travaux du laboratoire de statistique de l'Université de Paris VI*. 4. éd. comportant de nouveaux programmes et des compléments théoriques, Paris : Dunod (Leçons sur l'analyse factorielle et la reconnaissance des formes et travaux du laboratoire de statistique de l'Université de Paris VI, vol. 2), 1984.
- [BER 06] Bergh, D. D.; Perry, J.; Hanke, R.: Some predictors of SMJ article impact. In: *Strategic Management Journal*, 27. (2006), No. 1, 2006, pp. 81–100.
- [BER 93] Berry, L. L.; Parasuraman, A.: Building a new academic field—The case of services marketing. In: *Journal of Retailing*, 69. No. 1, 1993, pp. 13–60.
- [BLE 12] Blei, D. M.: Probabilistic topic models. In: *Communications of the ACM*, 55. (2012), No. 4, 2012, pp. 77–84.
- [BLE 03] Blei, D. M.; Ng, A. Y.; Jordan, M. I.; Lafferty, J.: Latent dirichlet allocation, 2003.
- [CAO 09] Cao, J.; Xia, T.; Li, J.; Zhang, Y.; Tang, S. (2009): A density-based method for adaptive LDA model selection. In: *Neurocomputing*, 72. No. 7-9, 2009, pp. 1775–1781. (<https://www.sciencedirect.com/science/article/pii/S092523120800372X>).
- [CHA 09] Chang, J.; Boyd-Graber, J. L.; Gerrish, S.; Wang, C.; Blei, D. M.: Reading Tea Leaves: How Humans Interpret Topic Models. In: *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.*, 32. (2009), pp. 288–296. ([https://www.researchgate.net/publication/221618226\\_Reading\\_Tea\\_Leaves\\_How\\_Humans\\_Interpret\\_Topic\\_Models](https://www.researchgate.net/publication/221618226_Reading_Tea_Leaves_How_Humans_Interpret_Topic_Models)), 2009.
- [DIC 83] Dickey, J. M.: Multiple Hypergeometric Functions: Probabilistic Interpretations and Statistical Uses. In: *Journal of the American Statistical Association*, 78. No. 383, 1983, pp. 628. (<http://www.jstor.org/stable/2288131>).
- [ELL 06] Ellen, P. S.: Building Corporate Associations: Consumer Attributions for Corporate Socially Responsible Programs. In: *Journal of the Academy of Marketing Science*, 34. (2006), No. 2, 2006 pp. 147–157. (<https://link.springer.com/article/10.1177/0092070305284976>).
- [FUR 08] Furrer, O.; Thomas, H.; Goussevskaia, A.: The structure and evolution of the strategic management field: A content analysis of 26 years of strategic management research. In: *International Journal of Management Reviews*, 10. (2008), No. 1, 2008, pp. 1–23.

- [HOF 86] Hoffman, D. L.; Franke, G. R.: Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research. In: *Journal of Marketing Research*, 23, No. 3, 1986, pp. 213. (<http://www.jstor.org/stable/3151480>).
- [HOF 92] Hoffman, D. L.; Leeuw, J.: Interpreting multiple correspondence analysis as a multidimensional scaling method. In: *Marketing Letters*, 3, No. 3, 1992, pp. 259–272. ([https://www.academia.edu/2611242/Interpreting\\_multiple\\_correspondence\\_analysis\\_as\\_a\\_multidimensional\\_scaling\\_method](https://www.academia.edu/2611242/Interpreting_multiple_correspondence_analysis_as_a_multidimensional_scaling_method)).
- hub.knime.com (---a): Excel Reader (<https://hub.knime.com/knime/extensions/org.knime.features.ext.poi/latest/org.knime.ext.poi3.node.io.filehandling.excel.reader.ExcelTableReaderNodeFactory>). Accessed on 01.04.2022.
- hub.knime.com (---b): Strings To Document (<https://hub.knime.com/knime/extensions/org.knime.features.ext.textprocessing/latest/org.knime.ext.textprocessing.nodes.transformation.stringstodocument.StringsToDocumentNodeFactory2>). Accessed on 01.04.2022.
- [BIS 12] J.M. Bischof; E.M. Airolidi: Summarizing topical content with word frequency and exclusivity (vol. 1), 2012.
- [JAC 16] Jacobi, C.; van Atteveldt, W.; Welbers, K.: Quantitative analysis of large amounts of journalistic texts using topic modelling. In: *Digital Journalism*, 4, No. 1, 2016, pp. 89–106. ([https://www.researchgate.net/publication/283671339\\_Quantitative\\_analysis\\_of\\_large\\_amounts\\_of\\_journalistic\\_texts\\_using\\_topic\\_modelling](https://www.researchgate.net/publication/283671339_Quantitative_analysis_of_large_amounts_of_journalistic_texts_using_topic_modelling)).
- [KAC 90] Kaciak, E.; Louviere, J.: Multiple Correspondence Analysis of Multiple-Choice Experiment Data. In: *Journal of Marketing Research*, 27, No. 4, 1990, pp. 455–465.
- [KAM 21] Kamran, Q.; Topp, S.; Becker, M.: The Structure and Evolution of the Marketing Field: A Content Analysis of Five Decades of Research within the Academy of Marketing Science Journals. Dortmund : PDF document, 2021.
- [KON 18] Konrad, M.: Probabilistic Topic Modeling with LDA. Practical topic modeling: Preparation, evaluation, visualization, 2018. (<http://dsspace.wzb.eu/pyug/topicmodeling2/slides.html>). Accessed on 01.04.2022.
- [KRO 14] Kronthaler, F.: Statistik angewandt. Datenanalyse ist (k)eine Kunst. Heidelberg : Springer Spektrum (Springer-Lehrbuch), 2014.
- [LEB 84] Lebart, L.; Morineau, A.; Warwick, K. M.; Berry, E. M.: Multivariate descriptive statistical analysis. Correspondence analysis and related techniques for large matrices. New York : Wiley & Sons (Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics), 1984.
- [MAG 04] Maignan, I.; Ferrell, O. C.: Corporate Social Responsibility and Marketing: An Integrative Framework. In: *Journal of the Academy of Marketing Science*, 32, No. 1, 2004, pp. 3–19. (<https://link.springer.com/article/10.1177/0092070303258971>).
- [MAC 02] McCallum, A. K.: MALLET: A Machine Learning for Language Toolkit (<http://mallet.cs.umass.edu>), 2022. Accessed on 01.04.2022.

- [MCL 19] Mcleod, S.: Box plots (also known as box and whister plots) (<https://www.simp-lypsychology.org/boxplots.html>), 2019. Accessed on 01.04.2022.
- [MIM 11] Mimno, D. M.; Wallach, H. M.; Talley, E. M.; Leenders, M.; McCallum, A.: Optimizing Semantic Coherence in Topic Models. In: EMNLP (ed.): Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIG-DAT, a Special Interest Group of the ACL. Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 262–272. ([https://www.researchgate.net/publication/221012637\\_Optimizing\\_Semantic\\_Coherence\\_in\\_Topic\\_Models](https://www.researchgate.net/publication/221012637_Optimizing_Semantic_Coherence_in_Topic_Models)).
- [NEW 09] Newman, D.; Asuncion, A. U.; Smyth, P.; Welling, M.: Distributed Algorithms for Topic Models. In: Journal of Machine Learning Research, 10, 2009, pp. 1801–1828. ([https://www.researchgate.net/publication/220320734\\_Distributed\\_Algorithms\\_for\\_Topic\\_Models](https://www.researchgate.net/publication/220320734_Distributed_Algorithms_for_Topic_Models)).
- [PAU 13] Paul, L. C.; Al Suman, A.; Sultan, N.: Methodological analysis of Principal Component Analysis (PCA) method. In: International Journal of Computational Engineering & Management, 16, No. 2, 2013, pp. 32–38. (<https://researchers.mq.edu.au/en/publications/methodological-analysis-of-principal-component-analysis-pca-metho>).
- [RAM 04] Ramos-Rodríguez, A.-R.; Ruíz-Navarro, J.: Changes in the intellectual structure of strategic management research: a bibliometric study of the Strategic Management Journal, 1980–2000. In: Strategic Management Journal, 25, No. 10, 2004, pp. 981–1004.
- [RÖD] Röder, M.; Both, A.; Hinneburg, A.: Exploring the Space of Topic Coherence Measures, 2015.
- scimagojr.com (---): Website: Scimago Journal & Country Rank (<https://www.scimagojr.com/>). Accessed on 01.04.2022.
- [SEN 06] Sen, S.: The Role of Corporate Social Responsibility in Strengthening Multiple Stakeholder Relationships: A Field Experiment. In: Journal of the Academy of Marketing Science, 34, No. 2, 2006, pp. 158–166. (<https://link.springer.com/article/10.1177/0092070305284978>).
- [SIE 14] Sievert, C.; Shirley, K.: LDAvis: A method for visualizing and interpreting topics. In: Chuang, J.; Green, S.; Hearst, M.; Heer, J.; Koehn, P. (eds.): Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, Maryland, USA Stroudsburg, PA, USA: Association for Computational Linguistics, 2014, pp. 63–70.
- [TAH 99] Tahai, A.; Meyer, M. J.: A Revealed Preference Study of Management Journals' Direct Influences. In: Strategic Management Journal, 20, No. 3, 1999, pp. 279–296. (<http://www.jstor.org/stable/3094106>).

- [THE 06] Teh, Y. W.; Jordan, M. I.; Beal, M. J.; Blei, D. M.: Hierarchical Dirichlet Processes. In: Journal of the American Statistical Association, 101., No. 476, 2006, pp. 1566–1581. ([https://www.researchgate.net/publication/221997021\\_Hierarchical\\_Dirichlet\\_Processes](https://www.researchgate.net/publication/221997021_Hierarchical_Dirichlet_Processes)).
- [TUF 17] Tufts, C.: The Little Book of LDA [work in progress], 2017 ([https://miningthedetails.com/LDA\\_Inference\\_Book/](https://miningthedetails.com/LDA_Inference_Book/)). Accessed on 01.04.2022.
- [TUT 21] tutorialspoint.com: OpenNLP - Tokenization ([https://www.tutorialspoint.com/opennlp/opennlp\\_tokenization.htm](https://www.tutorialspoint.com/opennlp/opennlp_tokenization.htm)), 2021. Accessed on 01.04.2022.
- [YAO 09] Yao, L.; Mimno, D.; McCallum, A.: Efficient methods for topic model inference on streaming document collections. In: Elder, J.; Fogelman, F. S.; Flach, P.; Zaki, M. (eds.): KDD'09. Proceedings of the 15th ACMKDD International Conference on Knowledge Discovery & Data Mining; June 28 - July 1, 2009, Paris, France. the 15th ACM SIGKDD international conference, Paris, France, 6/28/2009 - 7/1/2009New York, NY: ACM, 2009, pp. 937.

## **Appendix 1: Application of the machine learning methodology to investigate the domain of entrepreneurship and marketing**

### **TITLE: DESIGN FOR A SERIAL ENTREPRENEURSHIP IN CYBER-PHYSICAL REALITIES - DECIPHERING VISIONARY MANAGEMENT MODELS FOR MARKETING**

#### **Abstract**

This research is developed as a case to analyze and develop a holistic literature review of 21 leading journals in management science from 1960 to 2022 by using unsupervised machine learning, namely KNIME and Python, to determine the evolutionary history of the fields and to decipher the foundations of integrative models of serial entrepreneurship enhancing entrepreneurially-driven-marketing strategies.

**Keywords:** literature review, entrepreneurially-driven-marketing, serial entrepreneurship, interdisciplinary Weltanschauung of marketing, KNIME

**Description:** This application of methodology provides an overview of entrepreneurship, innovation, leadership, and cybernetics by analyzing 36.348 articles published in 21 leading journals from 1960 until January 2022 to develop a new conceptual framework supporting marketers to operate in more complex and dynamic markets.

#### **Introduction**

Especially in recent years, the world has been marked by unexpected crises, such as the COVID-19 pandemic and the Russian invasion of Ukraine, significantly changing the business environment. While these crises are still lingering and have dramatically increased the range of entrepreneurial and marketing pursuits, the enduring spirit of creativity and perseverance will address these challenges. Therefore, a design for serial entrepreneurship to decipher visionary management models delivering profound recommendations to marketers will be delivered. By applying design thinking as the methodology of this article, it aims first to point out the status quo in entrepreneurship, leadership, innovation, and cybernetics. In a second step, the holistic second-order entrepreneurial Design Weltanschauung model (SOEDWAM) will be developed. Our framework is based on an interdisciplinary Weltanschauung of marketing. It combines the sciences of entrepreneurship, cybernetics, and philosophy into a coherent whole, thus helping marketers find their way in complex marketing situations.

## Method

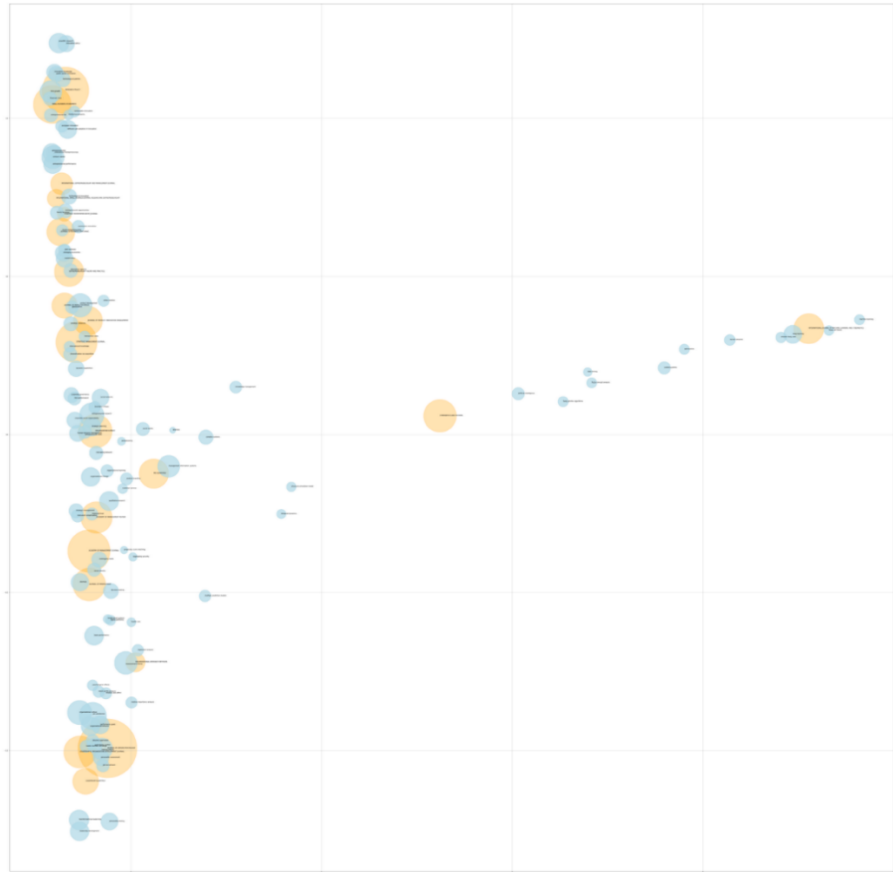
The author used a Latent Dirichlet Allocation (LDA) algorithm for unsupervised topic modeling and Python as an additional tool for the following analyses. Several content analyses (CA) must be examined to connect the existing managerial literature with cybernetic journals upon which the SOEDWAM for serial entrepreneurship in cyber-physical realities can be built. This is because a few latent topics are enough to represent a large corpus effectively. According to [ARU 10], such models have proved very effective. The related multiple correspondence analysis (MCA) was conducted to validate the mapping between the digitized keywords from 36,348 research articles and the topic labels and terms resulting from the performed LDA analysis. In this way, the scientific analysis of the "Lay of the Land" in innovation, entrepreneurship, leadership, and cybernetics with the techniques and tools of artificial intelligence were combined to achieve more precise and objective labeling of the "topic-keyword relationships." LDA assumes that each topic is associated with multiple terms, and conversely, terms may belong to multiple topics but have different effects on the overall topic distribution. Based on current developments in ML capabilities, this methodology provides a solid foundation not only for this research but also for future approaches where a large amount of textual data is to be analyzed for a more holistic grounded theory design. A holistic view presents the most discussed past, present, and future topics within the main research areas. Using unsupervised machine learning to analyze the massive number of articles in combination with the mixed research approach, namely quantitative and qualitative, ensures the robustness and applicability of the model referring to the most essential literature in the field. **Table 1** visualizes the method of the analysis.

Table 3. Method of the analysis

Data Import	Pre-Processing	Topic Detection via LDA	Grouping	Visualization
<ul style="list-style-type: none"> <li>Selection of 35,640 / 38,348 / 33,206 articles (21 top journals in entrepreneurship, innovation, leadership, and cybernetics)</li> <li>Creation of relevant Excel list including all relevant information (authors, title, journal, keywords, abstract, citations, citations per year)</li> </ul>	<ul style="list-style-type: none"> <li>Coding of keywords to generate 100 major topics per hand</li> <li>Tokenization of information about title, authors, and keywords</li> <li>Punctuation erasure</li> <li>Filter out numbers</li> <li>Delete words with less than 3 characters</li> <li>Creation of stop list to exclude specific common words</li> </ul>	<ul style="list-style-type: none"> <li>Single-parallel distributed implementation of LDA in KNIME for detecting the topics most relevant to the research project</li> <li>The node uses the MALLET topic modeling library</li> <li>Generation of 100 topics based on sensitivity analysis of 5 terms each</li> <li><math>\alpha = 0.01</math> and <math>\beta = 0.001 \Rightarrow</math> fine granularity</li> </ul>	<ul style="list-style-type: none"> <li>Grouping of topic terms and weighting with qualitative evaluation</li> <li>Pivoting of topics over years</li> </ul>	<ul style="list-style-type: none"> <li>MCA was applied to identify the relationship between topic terms and keywords</li> <li>MCA on the relevant journals for this research project</li> <li>Time series plots with trendlines show assigned articles per topic over years</li> <li>CA-analyses to present the interrelation of assigned topics to grouped years and journals</li> <li>Box Plots</li> </ul>
Main Interpretations Resulting from the analyses				
<ul style="list-style-type: none"> <li>MCA-Analyses indicate accurate results regarding the applied methodology</li> <li>Entrepreneurial and cybernetical topics are identified as the most discussed topics in present and possible future due to the trend lines</li> <li>Substantial scientific gap between cybernetics and management will be closed by this thesis</li> </ul>				

The data analysis software “KNIME” was used for the analyses, a commercial API software with an AI-based drag-and-drop solution. A simple parallel distributed implementation of LDA by [NEW09] using the sparse LDA sampling scheme and the data structure of [YAO 09] was used to identify the most relevant topics in the existing literature related to entrepreneurship, innovation, leadership, and cybernetics over the last seven decades. Unsupervised machine learning methods assessed the 100 most essential topics in these disciplines, and each article was assigned to one precise topic.

To show the interrelation of the assigned topics to grouped years, various CA-Analyses were performed. The frequency of published articles within a topic was based on a pivoting table drawn from the assigned topics over the years after LDA extraction. The author chose a two-factor model to simplify the graphical representation. Topics (blue bubbles) that are closely related to a period (orange bubbles) are also geometrically (in terms of the Euclidean distance) close together. The bigger the topic bubble, the more articles were published on the topic, and the bigger the journals’ bubble, the more articles were published in the respective journals.

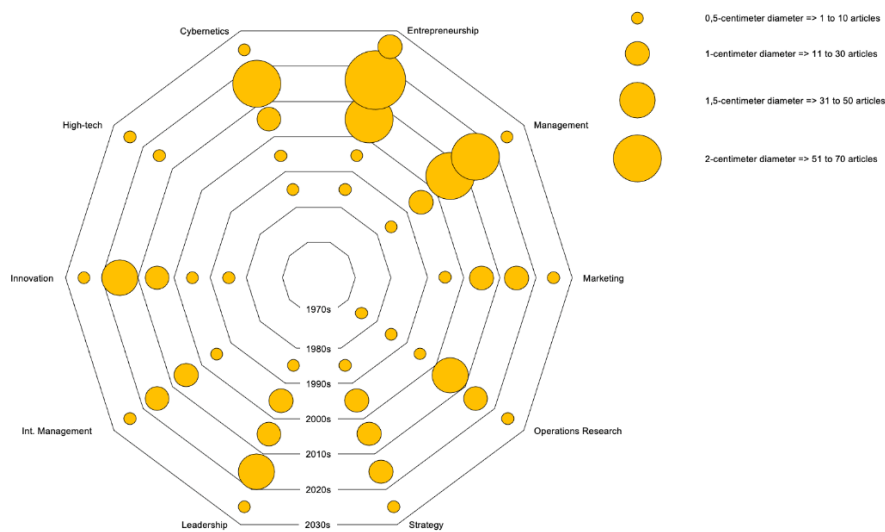


**Fig. 15.** CA-Analysis, including Cybernetics, indicates the association of topics (blue bubbles) with time decades (orange bubbles). The size of the bubbles indicates the number of associated articles to either the topic or year category, respectively. Bubbles closer together indicate thematic proximity induced by the underlying LDA classifier.

### **Content Analysis – The most cited articles per journal per year**

In a second step, the author analyzed which of the 100 labeled topics occurred in the 840 most cited articles. The 40 most cited articles per journal per year, resulting in 840 articles, are categorized into ten distinct topics, which are built by clustering the 100 most essential topics in management science, such as cybernetics, entrepreneurship, innovation, international management, leadership, marketing, operations research, and strategy in consultation with an expert in the managerial field. As the scope of this article is limited, not all 840 articles can be described in-depth. To give

a short overview of the topic distribution regarding timespans and the categories that have been discussed in the essential literature since 1960, Fig. 16 is shown. The bigger the diameter, the more articles were published in the respective category and decade. It indicates the importance of cybernetics and entrepreneurship nowadays and the constant importance of marketing throughout the 62-year period.



**Fig. 16.** 840 most cited articles per year assigned to category and timespan.

### *Deciphering visionary management models via the SOEDWAM*

Design Weltanschauung model.

By working through existing models and literature in Design Science, the Design Weltanschauung model developed by [KAM 21], has been detected, indicating the highest fit for this research to build upon. Operating in cyber-physical realities with global crises such as the COVID-19 pandemic indicates that businesses operate in markets shifting towards the complexity of massive proportions [KAM 20]. The second axis is the time to react to upcoming crises and challenges, which periodically becomes shorter [KAM 21]. Until an adequate strategy is formulated [MIN 20], the crises could become even more severe because the situation could have already changed significantly [BEE 93].

Building upon the DWA model, marketers must ask themselves three key questions: 1. Who are the authors, and what is our legitimacy? 2. What does the author

need to do today? 3. What does the author need to do in the future? [KAM 21]. The essential component of this model for SOEADWAM is the differentiation between the future and Jacques Derrida's analogy of *l'avenir*. The future that will comparably sure come is predicted and foreseeable. It is considered the future of controlled, handled, and programmed timetables. On the opposite, *l'avenir* is totally unexpected and unpredictable, signaling the otherness of the future and the coming arrivant [TRE 19]. *L'avenir* is not an event with a beginning and ending point and cannot be pinned down. *L'avenir* will always escape philosophy and mastery [FRE 22]. However, the situation does not mean that the author is powerless and incapable; the aporia is not the end of the road but the beginning. While the author cannot willfully actualize or make true an event, the author can make himself available to it. The author can be true to and to the events [TRE 19]. *L'avenir* is that which distinguishes or unhinges the movement in question. [MAR 07]. The analogy describes an individuum or event coming without the competitor being able to anticipate the arrival [DIC 05]. The difference between *le futur* and *l'avenir*, therefore, is what the future does or what the author does with the future (*l'avenir*) over what the future is or holds (*le futur*) [MAR 07]. *L'avenir* is a certain impossibility, an aporetic structure, something that is to come, although it will never fully arrive [DER 05]. Furthermore, Derrida argues that *l'avenir* is "the other is what is never inventable and will never have waited for your invention. The call of the other is a call to come, and that happens only in multiple voices" [DER 07]. To enable a way out of the present impasse toward *l'avenir*, Derrida proposed practicing deconstruction, which means not to be enclosed or dominated by the same or majority [WHI 07].

SOEDWAM is introducing *L'avenir* II.

However, a philosophical and cybernetical perspective must be adopted to decipher the foundations of integrative models of serial entrepreneurship enhancing entrepreneurially-driven marketing strategies. Therefore, the Design Weltanschauung model by [KAM 21] needs to be examined via the SOEDWAM, as illustrated in Fig. 17.

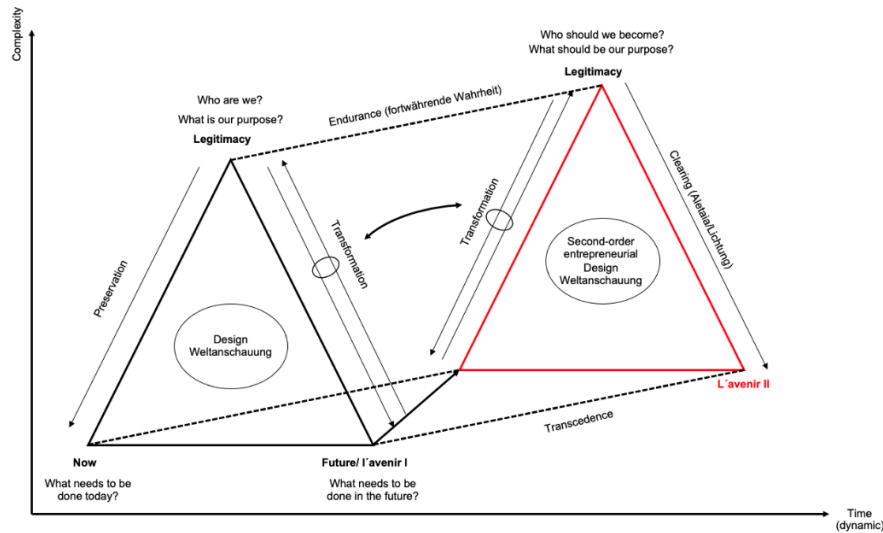


Fig. 17. SOEDWAM introducing l'avenir II

For “usual” companies, the Design Weltanschauung model by [KAM 21] remains the valid model in Design Science to operate successfully in cyber-physical realities. However, the findings in previous chapters require developing an additional layer to enhance entrepreneurially driven marketing strategies. Therefore, a philosophical and cybernetical dimension must be implemented to introduce l'avenir II, enabling marketers to handle even higher complexity and dynamic issues. Therefore, the marketer needs to follow the fortwährenden Wahrheit (Endurance) to maintain its raison d'être following the analogy of l'avenir II. Simultaneously, marketers must be able to transcend current knowledge and use Dasein to base their operations on SOEDWAM. They need to operate from a constructivist perspective. Knowledge does not represent "objective" facts but is a compendium of concepts, conceptual relationships, and rules that have proven helpful in the world [FOE 84]. Foerster introduced the concept of second-order cybernetics and got rid of the control and regulation euphoria that characterized the first-order cybernetics of the time. In second-order cybernetics, the observer and the observed are inextricably intertwined. Second-order cybernetics brings into play the constant obligation to reflect the idiosyncrasies and blind spots of the observer, never to separate assertions from the individual who makes them, but always to evaluate them in a serious sense as the product of an individual. Going a step further, radical constructivism can be seen as an epistemology that offers a pragmatic approach to questions of reality, truth, and human understanding [GLA 84]. Knowledge is categorized according to its realizability in the realm of experience rather than according to the traditional philosophical position that it is constitutive of

truth, i.e., that it corresponds to objective reality (phenomenology). The two basic principles of radical constructivism are:

Knowledge is not passively received through the senses but is actively constructed by the cognizing subject, the learner,

The function of cognition is the organization of the world of experience rather than the discovery of an independent reality [WAL 20].

Now, what about the light of truth covered by the right side of the triangle? Is the clearing also, in this case, a more original dimension in which there can only be something like truth? Yes, and even in an even deeper sense, because the essence of truth as αλήθεια, the coming forth from concealment into unconcealment - i.e., a coming to light that is essentially related to darkness - is the essence of the clearing itself. The essence of truth as αλήθεια, the emergence from the hidden into the unconcealed - that is, an emergence essentially related to darkness - is the essence of the clearing itself. Heidegger explains the Greek word αλήθεια as a “privative expression.” [HEI 79]. On the one hand, the openness of Dasein is being enlightened, and on the other hand, the unconcealment of Being is becoming enlightened, expresses the original essence of truth, which remains veiled in the word “truth.” Heidegger also calls the development of Dasein and the discovery of Being “openness” or “revelation.” [HEI 76]. The openness of Dasein, as letting in on the being, is, according to its essence, a “letting-be” [HEI 76]. Thus, it is not only an ontic but an ontological relation: as letting-be, letting-be does not only refer to the being but only to be, i.e., to “the open and its opennesses, which every being brings with itself, as it were. This open has been conceived by Western thought in its beginnings as τα ἀληθέα, the unconcealed” [HEI 76] in which only something can open: This area of reference is the clearing of being. “The name of this clearing is αλήθεια” [HEI 76].

“Clearing,” even more than “truth,” is the appropriate “translation” of αλήθεια as the realm of unconcealment arising from concealment. In order to translate the word αλήθεια in its fullest sense as “clearing,” the following conditions must be met. If the clearing is the place, the realm of the truth event as a present happening, it must not be thought of as a place already given: Rather, it retains the literal sense of “clearing,” i.e., it is the breaking open of an open. Connected with this is the essential relationship of darkness and light (of concealment and unconcealment), for the clearing points to an emergence from the irreducible, harboring darkness into the light [AMO 83]. All the previously described philosophical and cybernetical dimensions are foundations of integrative models of serial entrepreneurship enhancing entrepreneurially-driven marketing strategies.

Based on the findings, l’avenir II can be defined as follows:

- |   |
|---|
| <p>I) Entrepreneurs need the ability to endure to follow the analogy of l'avenir II. The if-question is replaced by the when-question.</p> <p>II) Knowing the science is a necessary prerequisite to participate in knowledge-exchange on <i>Augenhöhe</i>.</p> <p>III) Risk-taking and focusing on entrepreneurial opportunities instead of boundaries and doubts.</p> <p>IV) Being creative is a key pillar for entrepreneurs applying l'avenir II.</p> <p>V) Co-creation is required for entrepreneurs to follow the analogy of l'avenir II.</p> <p>VI) Referring to Kant, entrepreneurs need to follow the epistemological "a priori" (before experience) instead of "posteriori" (after experience). This means, that basing its activities solely on experience does not allow to follow the analogy of l'avenir II for serial entrepreneurship in cyber-physical realities. To give an example, Musk achieved to develop reusable space vehicles without any experiences observed due to the novelty based on his visions, risk-taking, and ability to endure.</p> |
|---|

**Fig. 18.** Key pillars l'avenir II [KAN 1781]

For example, the foundation of Zip2 in 1995, Elon Musk's first company, followed the analogy of l'avenir II. Nowadays, many revenue models of contemporary location-based services, smart cities, and web maps are based on Musk's company, founded over 20 years before competitors base their operations on this model. Before the foundation of Zip2, the Dasein of such opportunities was not given to represent the unforeseeable future called l'avenir II.

## Conclusion

The author concludes by designing a foundational model derived from the data observed to serve as the meta-framework for establishing effective marketing management models in contemporary disruption and change. For marketers of today and the future, the SOEDWAM would deliver a boost in their capabilities and phenomenologically creative work of creating value. By enhancing the boundaries of serial entrepreneurship and embedding a philosophical and cybernetical dimension, the essence of a competitive advantage can be achieved.

To summarize, entrepreneurs must ask three key questions to themselves building upon the SOEDWAM:

Who should the author become, and what should our purpose be?

What needs to be done in the future?

How to transcend the future?

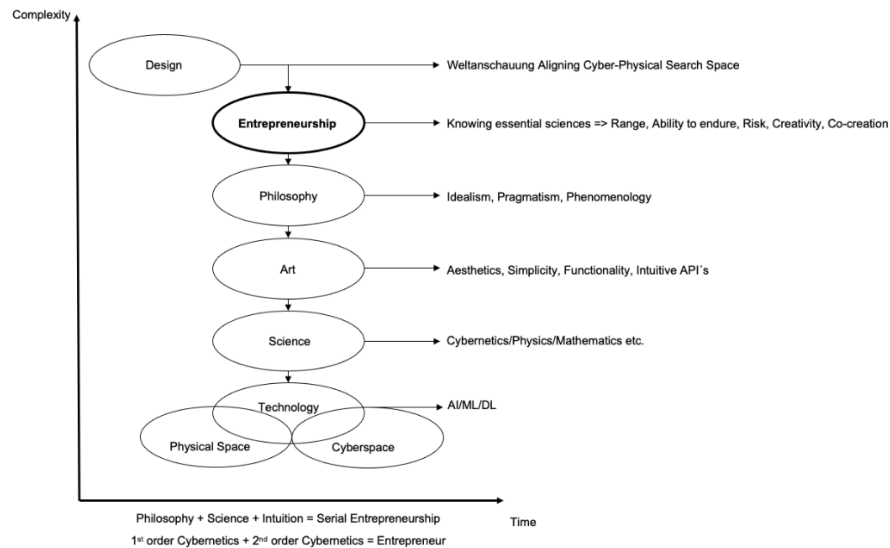


Fig. 19. Describing the SOEDWAM Applied and Aligned Diverse Essential Fields for Serial Entrepreneurship Enhancing Entrepreneurially Driven Marketing

Fig. 19 above displays the foundations of the SOEDWAM, where entrepreneurially driven marketers can navigate the turbulent markets within the dimensions of “transforming” the organizational *raison d’être* by the structural dynamics of an ambidextrous embodiment (adapted by [KAM 21]), while focusing on the dimension of clearing/Lichtung, it is essential to follow the analogy of *l’avenir II*. Therefore, it illustrates how designer-based serial entrepreneurship can be applied because the cyber-physical markets are moving toward increasing complexity and dynamics, leading to a reduced time to respond effectively and innovate. Special attention must be given to the philosophical foundation, and idealism, pragmatism, and phenomenology must be detailed here.

## References

- [ARU 10] Arun R, Suresh V, Veni Madhavan CE, Narasimha Murthy MN.; On finding the natural number of topics with latent Dirichlet allocation: some observations. In: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC et al (eds) *Advances in Knowledge Discovery and Data Mining*. Lecture Notes in Computer Science, vol 6118. Springer, Berlin, 2010, pp 391–402. [https://doi.org/10.1007/978-3-642-13657-3\\_43](https://doi.org/10.1007/978-3-642-13657-3_43)
- [BEE 93] Beer S., *Designing Freedom*. House of Anansi, Toronto, 1993.

- [DER 05] Derrida J., Chapter machine. In: *Paper Machine*. Cultural Memory in the Present. Stanford University Press, Stanford, CA, 2005.
- [DER 05] Derrida J., *Rogues: Two Essays on Reason*. Meridian Crossing Aesthetics Series. Stanford University Press, Stanford, CA, 2005.
- [DER 07] Derrida J., *Psyche: Inventions of the Other*. Meridian Crossing Aesthetics Series. Stanford University Press, Stanford, CA, 2007.
- [FOE 84] Foerster H von., On cybernetics of cybernetics and social theory. In: Roth G, Schwegler H (eds) *Self-Organizing Systems*. Campus Verlag, Frankfurt am Main, 1984, pp 102–105.
- [FRE 22] French MF., Imminence over immanence: A lesson in L’Avenir. Patheos Blog. <https://www.patheos.com/blogs/mariafrancescafrench/2022/03/imminence-over-immanence-a-lesson-in-lavenir/>, 2022. Accessed 29 Mar 2022
- [GLA 84] Glasersfeld E von., An introduction to radical constructivism. In: Watzlawick P (ed) *The Invented Reality: How Do We Know What We Believe We Know?*. Norton, New York, 1984.
- [HEI 72] Heidegger M., *Sein und Zeit*, 12th edn. Max Niemeyer, Tübingen, 1972.
- [HEI 76] Heidegger M., *Vom Wesen der Wahrheit*, 6th edn. Vittorio Klostermann, Frankfurt am Main, 1976.
- [KAM 21] Kamran Q., *Managing Complexity in Marketing: From a Design Weltanschauung*. University of Twente, Enschede, 2021.
- [KAM 21] Kamran Q, Topp S, Becker M., The structure and evolution of the marketing field: A content analysis of five decades of research within the Academy of Marketing Science Journals. *J Acad Mark Sci* 49, 2021, pp. 1021–1047. <https://doi.org/10.1007/s11747-020-00758-x>
- [KANN 1781] Kant I., *Kritik der reinen Vernunft*. Johann Friedrich Hartknoch, Riga, 1781.
- [MAR 07] Martinon J., *On Futurity: Malabou, Nancy, and Derrida*. Palgrave Macmillan, Basingstoke, 2007
- [MIN 20] Mintzberg H, Ahlstrand BW, Lampel J., *Strategy Safari: A Guided Tour Through the Wilds of Strategic Management*, 2nd edn. FT Publishing International, Harlow, 2020.
- [NEW 09] Newman D, Asuncion A, Smyth P, Welling M., Distributed algorithms for topic models. *J Mach Learn Res* 10, 2009, pp. 1801–1828. <https://www.jmlr.org/chapters/v10/newman09a.html>
- [WAL 20] Walshe G., Radical constructivism—von Glasersfeld. In: Akpan B, Kennedy T (eds) *Science Education in Theory and Practice*. Springer, Cham, 2020, pp 359–371. [https://doi.org/10.1007/978-3-030-43620-9\\_24](https://doi.org/10.1007/978-3-030-43620-9_24)
- [WHI 07] White E., A passage toward the other: The legacy of Jacques Derrida (1930–2004). *Eur Legacy* 12(4): 2007, pp. 407–408. <https://doi.org/10.1080/10848770701395926>

[YAO 09] Yao L, Mimno D, McCallum A., Efficient methods for topic model inference on streaming document collections. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, 2009, pp 937–946. <https://doi.org/10.1145/1557019.1557121>