# Understanding Topic Models in Context: A Mixed-Methods Approach to the Meaningful Analysis of Large Document Collections

Matthias Eickhoff
University of Goettingen
meickho@uni-goettingen.de

Runhild Wieneke
University of Goettingen
rwienek@uni-goettingen.de

## Abstract

*In recent years, we have witnessed an unprecedented proliferation of large document collections. This development has spawned the need for appropriate analytical means. In particular, to seize the thematic composition of large document collections, researchers increasingly draw on quantitative topic models. Among their most prominent representatives is the Latent Dirichlet Allocation (LDA). Yet, these models have significant drawbacks, e.g. the generated topics lack context and thus meaningfulness. Prior research has rarely addressed this limitation through the lens of mixed-methods research. We position our paper towards this gap by proposing a structured mixed-methods approach to the meaningful analysis of large document collections. Particularly, we draw on qualitative coding and quantitative hierarchical clustering to validate and enhance topic models through re-contextualization. To illustrate the proposed approach, we conduct a case study of the thematic composition of the AIS Senior Scholars' Basket of Journals.*

## 1. Introduction

In recent years, scientific awareness of reconciling the historically suggested dichotomy of quantitative and qualitative research approaches has gained momentum [11]. The IS community has intensified its contribution to this discussion. Scholars increasingly draw on mixed-methods research (MMR), i.e. the combination of elements of both quantitative and qualitative research approaches within a single study [49] to examine IS phenomena [1]. Calls for methodological pluralism had already been put forth even before the IS community began to refer to this alternative as MMR [41]. Yet, even though proponents of MMR have gained traction, the long-standing research tradition to rely on quantitative approaches and on the associated positivist paradigm is still considered dominant within the IS discipline [17].

In particular, scholars have mirrored this thread in studies on the analysis of large document collections, i.e. large quantities of qualitative data.

The digitalization of society at large, and particularly the rise of data-generating technologies, e.g., mobile devices and social media, have sparked the rapid, unprecedented proliferation of large unstructured collections of text corpora [40]. Qualitative data analysis approaches, especially qualitative coding, are approved means of drawing meaning from small sample sizes [35]. Yet, due to the continuing increase in data volumes, they reach their limits. When it comes to large sample sizes, qualitative analysis becomes a labor-intensive and time-consuming endeavor. Even if manual coding is replaced by computer-assisted coding, qualitative data analysis often remains virtually impossible [14]. The alternative to qualitative coding for textual analysis is given by quantitative text mining methods, such as sentiment analysis and topic modeling [7, 14]. Yet, quantitative methods do not offer a similar depth of contextual understanding as their qualitative counterparts. Topic modeling, in particular, includes further method-specific challenges, e.g., choosing the "correct" number of topics or validating the estimated model, which are difficult to address in a purely quantitative manner [6]. The limitations of both qualitative and quantitative approaches to the analysis of large qualitative data sets have caused scholars to call for appropriate MMR alternatives [49]. Interest has been expressed in how to synergistically combine the strengths of text mining methods in general and topic models in particular with the strengths of qualitative approaches [40]. We position this paper towards this interest. Prior MMR in this field has already enriched our knowledge of how to provide quantitative assistance to qualitative analysis [25, 30]. Yet, to the best of our knowledge, there is a lack of methodological clarity on the reverse scenario, i.e. how to address the limitations of quantitative topic models by qualitative means. In the context of social media data, Murthy [38] has put forth the idea of advancing LDA [7], a topic model class, by qualitative coding. Yet, the author left the operationalization of his idea unspecified.

In this paper, we address the stated research gap by proposing a structured MMR approach to the analysis of large document collections. It serves the purposes of validating and enhancing [12, 22] quantitative topic models. The aim is to increase the meaningfulness of single topics, and hence to allow researchers to better fathom the meaning of the analyzed document collection at large. We also showcase the proposed approach by analyzing the full content of all articles published in the AIS Senior Scholars' Basket of Journals until the second quarter of 2017. To this end, we address the following **research question**: How can a structured mixed-methods research approach render the analysis of large document collections more meaningful than a purely quantitative topic modeling approach?

## 2. Analysis of qualitative data

In contrast to quantitative numerical data, qualitative data usually assumes the form of texts respectively documents, and features distinct characteristics [35]. In particular, this data type is meaningful [31]. If treated individually, however, words are less meaningful than if considered in their textual context [35]. Hence, qualitative data is contextual, i.e. its context contributes to its content [15]. Considering that *"ideas can become independent of their authors and of the context in which they were originally created and shared"* [16, p. 1229], scholars can process qualitative data according to either pre-defined or emerging schemes. For analytical purposes, data can become subject to de-contextualization and re-contextualization, i.e. detached from its original context as well as re-conceptualized, respectively [48]. Thus in the end, data relates to two contexts. Its first context is the one it belongs to by nature. Its second context, i.e. the re-contextualized output, results from applying (new) organizing principles to the data [32]. To seize the natural and analytical richness of qualitative data [35], scholars can draw on qualitative, quantitative, and mixed-methods threads. Subsequently, we focus on qualitative coding, quantitative topic models, and MMR approaches [46].

**Qualitative coding:** Research has spawned a great variety of qualitative approaches to the analysis of qualitative data, such as grounded theory and hermeneutics [31, 48]. Though differing in detail, these approaches share a key feature. They are less driven by the assumption *"that we do not know all the answers to our problems but rather from an appreciation of the fact that we do not know all the questions"* [15, p. 14]. Thus, qualitative means aim to identify and elaborate key themes from the data [11], and ask scholars to *"decide what things mean"* [35, p. 11]. To do so, scholars usually draw on coding [11].

Coding is the structured analytical process of organizing qualitative data, primarily text data, through reduction and complication [11, 48]. Reducing means segmenting and hence de-contextualizing data into a manageable number of common denominators. Complicating means modifying, e.g., synthesizing and classifying, and hence re-contextualizing (derived) rationales based on (new) organizing principles. To this end, scholars attach codes to data chunks of varying size, e.g., words, phrases and passages. Codes are *"labels for assigning units of meaning to the descriptive or inferential information compiled during a study"* [35, p. 56]. While descriptive codes conceive the apparent occurrences that surface at the literal level, inferential codes allow for deciphering the dispositional and implicit contents not directly observable yet underlying a phenomenon at the interpretative level [29]. Codes are hence useful to the analysis of both the manifest and latent meaning of textual data [8]. Thus, it is distinguished between *"data directly accessible to the investigator (manifest), and parameters (latent) which in some way must be inferred from the manifest data"* [29, p. 48]. In either case, codes need to feature both internal homogeneity and external heterogeneity [3].

To enhance rigor and reliability, a research team is advised to conduct an inter-coder agreement test [35]. Initially, individual team members separately code (a sample of) the given data set. Afterwards, they compare and discuss the assigned codes and applied coding rules, and recode the data according to the agreed solution. In doing so, the research team can correct for discrepancies in individual judgment, and for joint mistakes that become apparent during analysis. To further challenge and also structure the coding process, researchers should rely on memo writing [42]. Memos are graphical and textual notes that come to the analyst´s mind during data collection and analysis [20]. Both coding and memo writing can be accomplished in a manual and in a computer-assisted manner [42].

Coding aside, yet closely related to it, qualitative data analyses span three chains of reasoning, namely deductive, inductive, and combined logics [3]. While deductive reasoning entails arguing from the general to the particular, inductive reasoning involves the reverse scenario [49]. In particular, the type of reasoning used in a study is reflected in the assignment of codes to textual data [45]. Deductive coding is associated with quasi-statistical methods. It starts from a predetermined set of codes that is derived from established or anticipated theories, frameworks and hypotheses. The codes are then assigned to the studied data [8, 45]. By contrast, inductive coding is data-driven. Researchers immerse into the given data sample, and let the coding scheme emerge in a bottom-up fashion through data analysis [8, 20]. Finally, in combined coding processes,

scholars combine the two generic modes of reasoning stated above [28], or rely on abductive inferences along with deductive and inductive logics [49].

In this paper, given our interest in making sense of qualitative data in due consideration of its particular context, we ascribe our analytic inquiry to inductive coding [35]. Most often, this approach involves several coding cycles. First-cycle coding is the process of developing and continuously evolving first-level codes from the studied data. In absence of preconceived ideas, researchers open-mindedly attach these codes to data segments, and thus gradually discover salient data traits at a fine-grained analytical level [12]. Researchers move to higher levels of inference during subsequent coding cycles [42]. By elaborating, grouping, and abstracting similar lower-level codes, researchers develop them into higher-level codes, often referred to as pattern codes [35]. As hierarchical coding schemes reveal regularities at different analytical levels [31], they are comparable to cluster-analytic methods in quantitative inquiries [35]. In addition to establishing hierarchical linkages across lower and higher levels of analysis, codes can also be linked on the horizontal line at a single analytical level [20, 28]. Overall, multiple coding cycles unfold in an iterative process, during which codes at all analytical levels possibly undergo several alterations [35]. Among others, codes can be re-labeled, eliminated, re-classified by subsuming them under different higher-order codes, and assigned to a different analytical level, if applicable [28, 42].

**Quantitative topic models:** Topic models deal with the challenge of retrieving thematically similar documents from text corpora. With over 18 thousand cites and counting, LDA-type models [7] constitute a popular answer to this task. LDA models primarily result in two pieces of information usable for further analysis. First, a set number of topics is estimated, for each of which every word in the document collection is ranked regarding its likelihood to appear in a document given that a topic is present. Second, the likelihood for each topic in each document is calculated. Based on these two pieces of information, i.e. the word to topic and topic to document assignments, LDA topics can be used to study the thematic composition of large document collections. However, as noted by Blei et al. [7, p. 994], these topics are not necessarily akin to what humans consider to be such, but rather constitute a representation of *"intra-document statistical structure"*. Still, when inspecting these topics, many researchers have found them to be informative regarding the actual topics discussed in the documents under study. A number of LDA variants have been developed, including the Hierarchical Dirichlet Process [47]. It can provide some guidance on the number of topics needed to model a given text collection, and on implementations of LDA that enable the analysis of much larger corpora and streamed documents [51]. Also, dynamic [5] and document influence models [19] have been developed, the latter of which enable the modeling of topic evolution over time, and of the impact of documents regarding the future topic evolution in a document collection. As the approach outlined in this paper relies on the two basic pieces of information all LDA-type models have in common, i.e. word to topic assignments and topic to document assignments, it is applicable to these variants as well as to basic LDA models. Indeed, we use a dynamic model ourselves for the case study presented in section 4 because the given document collection spans many years, and this model type allows for changes in wording over time.

Mimno et al. [36] note that topic models typically produce low-dimensional topics that appear flawed to human domain experts. In this paper, we consider this low-dimensionality a result of the de-contextualization of data that results from quantitative analysis. The pre-processing of documents treats them as a bag of words [50]. Thus, it ignores word orders, and removes the words from their immediate contexts. For the purpose of training the model, words are treated as independent features. Some approaches towards the incorporation of immediate word contexts in bag of words models exist, such as n-grams [50]. Yet, they tend to inflate the size of the term-document matrix, which makes the training of topic models significantly more time consuming. From a qualitative point of view, this implies a major disadvantage, i.e. the bag of words approach disregards substantial information content [33]. Concentrating *"solely on numbers shifts attention from substance to arithmetic"* [35, p. 56]. Hence, the decision whether certain aspects of the phenomenon under study are of superior or inferior importance cannot be exclusively made based on quantifiable measures. Instead, it is also important to account for the meaning and context of the studied data. For instance, some words or compound terms may arouse an ambiguous sense, e.g., owing to negations, amplifications and dilutions. Consequently, to decipher their actual substance, an analytic inquiry should not only suit the general research interest, but also allow for analyzing data in due consideration of its context [33]. Besides finding n-grams in a given data set itself, another approach is to find n-grams in the topics estimated on the basis of a unigram model [18]. Still, regardless of any improvements made towards keeping some context in the bag of words or the topic model, simplification is the very nature of modeling. Also, it is indeed desirable during the quantitative analysis of large document collections. After all, gaining a content-related overview is the point of this method. This means that word context will be lost during the pre-processing and modeling of documents. Yet, the interpretation of

topic models can be difficult in absence of context. The lack in context is a byproduct of quantitative analysis.

**Mixed-methods approaches:** MMR combines elements of both quantitative and qualitative research approaches within a single study [49]. Mixed methodologists take advantage of the strengths of either approach while compensating for their respective constraints [12]. In doing so, they build upon the assumption that *"there is no fundamental clash between the purposes and capacities of qualitative and quantitative methods or data"* [20, p. 17]. This assumption revises the outdated view that scientists largely held in the mid-20th century, namely the mutual exclusiveness of quantitative and qualitative inquiries mostly grounded in the underlying paradigms. While constructivist paradigms are usually associated with qualitative research, positivist paradigms dominate quantitative research [46]. As third research thread, MMR draws on other paradigms, such as pragmatism, that abandon the traditionally suggested dichotomy of quantitative and qualitative paradigms [11]. Miles and Huberman [35] advise researchers to rely on both numbers and words, i.e. quantities and qualities, to be able to make holistic sense of a phenomenon. In a similar vein, Kaplan [26, p. 207] remarks that *"whether something is identified as a quality or as a quantity depends on how we choose to represent it in our symbolism"*. The author further notes that any *"transformation of quantity into quality, or conversely, is a semantic or logical process, not a matter of ontology"*. While qualities are quantifiable through the introduction of scales, researchers render quantitative measures qualitative by the assignment of labels. Assuming this methodological compatibility, MMR serves important purposes, such as the validation and enhancement of research results [12, 22]. Compared to purely qualitative or purely quantitative studies, MMR aims at stronger inferences, greater depths and breadths of understanding, and more insightful results [21].

To serve a chosen research purpose, mixed methodologists have to select a suitable MMR design. To do so, they can draw on different design typologies [12]. In this paper, we rely on the classification scheme proposed by Greene and Caracelli [21]. The authors differentiate between two design classes that differ in the level of methods' integration. First, *component designs* implement different methods as discrete aspects. The methods used are rarely combined during data collection and analysis, but rather at the level of drawing conclusions. Second, *integrated designs* place emphasis on the integration of different methods, and thus likely generate more insightful results. In this paper, given our interest in advancing topic models, we opt for an integrated design. In particular, we build on a *nested design*. The main characteristic of this design
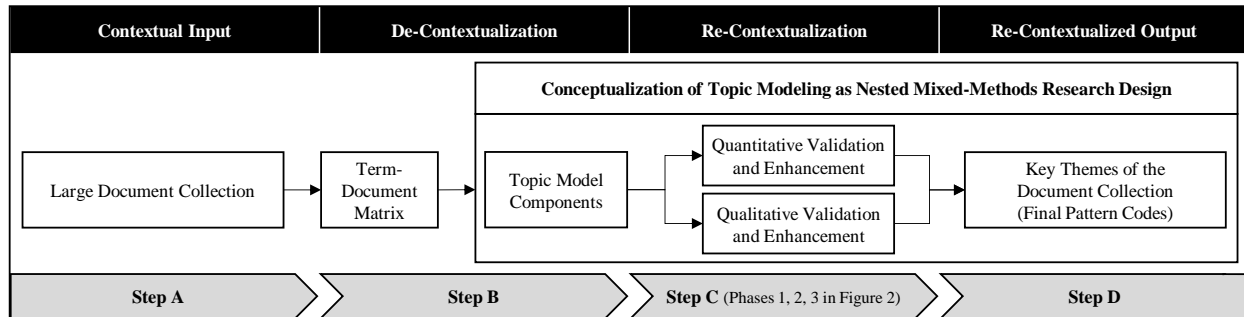
type is the positioning of at least one method within another primary method. As detailed in section 3, we propose a MMR approach to the analysis of large textual data sets that shows how researchers can advance quantitative topic models by methodological pluralism. Extant MMR endeavors have elaborated on quantitative assistance to qualitative analyses. For instance, Lemke et al. [30] have proposed a blended reading approach. It builds on and integrates two approaches to analyzing textual data known from research in the humanities, namely close reading (i.e. in-depth qualitative analysis) and distant reading (i.e. quantitative analysis with computational programs). Towards the goal of blended reading, a document exploration tool named Leipzig Corpus Miner (LCM), which supports a number of topic models (LDA, HDP, Online-LDA), has been developed. Similarly, Janasik et al. [25] have proposed the use of self-organizing maps (SOM) to advance Grounded Theory analysis. Both LCM and SOM offer quantitative assistance for qualitative work. However, we approach the reverse case to address the lack of methodological clarity on how to inform text mining methods, especially topic models, by qualitative means [38, 40]. Thus, in the next section, we address this research gap in detail.

## 3. A mixed-methods approach to the analysis of large document collections

We propose a structured MMR approach to the analysis of large document collections. It is conducive to validating and enhancing the meaningfulness of individual topics generated by topic models, and hence to fathoming the meaning of the analyzed document collection at large. The research process consists of four steps, which are shown in Figure 1, and detailed below.

**Step A:** Each topic model-based analysis begins by creating a large textual qualitative data set, i.e. a large document collection to be analyzed. Subsequently, we do not elaborate on the corpus creation itself. For the purpose of this paper, we rather consider the document collection the predetermined input of analysis. This input is meaningful and contextual [15, 31].

**Step B:** The given document collection is subject to two subsequent processes of de-contextualization. First, the documents are pre-processed into a term-document matrix. This pre-processing results in a loss of word context as it treats documents as bags of words. Second, the topic model itself further detaches words from their original context in order to arrive at topics suitable to model the entire document collection. At this stage, quantitative topic modeling would end in most cases, and the resulting topics may be given ad-hoc labels. Potentially, the generated topics would be processed further, e.g., by selecting relevant topics for regression

| Contextual Input | De-Contextualization | Re-Contextualization | Re-Contextualized Output |
|---|---|---|---|

**Conceptualization of Topic Modeling as Nested Mixed-Methods Research Design**

Large Document Collection → Term-Document Matrix → Topic Model Components → Quantitative Validation and Enhancement / Qualitative Validation and Enhancement → Key Themes of the Document Collection (Final Pattern Codes)

| Step A | Step B | Step C (Phases 1, 2, 3 in Figure 2) | Step D |

**Figure 1: Mixed-methods research process**

using econometric methods [14]. In contrast to a purely quantitative content analysis approach, our proposed MMR approach considers the generated topic model components, i.e. word to topic assignments and topic to document assignments, as preliminary results.

**Step C:** The topic model components form the basis of our nested MMR design. Using both qualitative and quantitative methods, we propose an approach to the validation and enhancement of individual topics found in the document collection. In particular, this approach targets the re-contextualization of the previously de-contextualized data. Compared to a purely quantitative topic modeling approach, it allows researchers to draw more meaningful inferences. Given its novelty, a detailed explanation of the nested research design is provided in Figure 2, and detailed below.

**Step D:** The research process finally yields re-contextualized output, namely the key themes of the document collection. As detailed below, these themes are presented by the developed final pattern codes.

For the sake of convenience and clarity, Figure 2 provides a detailed account of how we conceptualize topic modeling as nested MMR design. To avoid repetitions, we subsequently delineate the nested design by focusing on its core, i.e. step C consisting of the following three phases:

**Step C - Phase 1:** The word to topic assignments, i.e. the first topic model component, provide the very basis for inductive *topic coding* [42], likely associated with ad-hoc labeling [14]. Topic coding is a thorough qualitative coding approach suitable for summarizing the essence of the word to topic assignments, through which we ensure inter-coder reliability [35]. Individual members of the research team initially develop and assign descriptive codes to each word to topic assignment separately. Then, they discuss and resolve discrepancies in individual judgement, e.g., definitional ambiguities. This inter-coder reliability check results in a preliminary collection of initial topic codes. Topic coding is important because it enables researchers to reason about the data collection concisely. Yet, initial topic codes are derived from de-contextualized data. To achieve re-contextualization, researchers have to

validate them through an in-depth analysis of the topic to document assignments. As further insights about the manifest and latent meaning of the individual topics arise, initial topic codes are retained, eliminated, or re-labeled, if applicable. Given the dynamic nature of this analytical process, they become evolving topic codes.

**Step C - Phase 2:** Subsequent to first-cycle coding, researchers engage in second-cycle coding [42]. To draw inferences at a higher level of abstraction, evolving topic codes are grouped into initial pattern codes. Again, an inter-coder reliability test is inevitable.

**Step C - Phase 3:** We propose a validation of the initial pattern codes based on quantitative hierarchical clusters [44] derived from the word to topic assignments (Figure 3). This allows researchers to assess the quantitative hierarchical relation between individual topics, while making use of the previously developed topic codes. To assess the quality of individual pattern codes, we propose a relative metric comparing the quantitative closeness of individual topic codes included in a pattern code by seeking their closest parent cluster. We count the steps necessary to join all topic codes in a pattern code. We use the mean of these step counts over all pattern codes as a metric of the average internal homogeneity of topic code to pattern code assignments. If a pattern code exhibits more than one standard deviation of the individual step counts than this average, we revisit its topic to pattern code assignments, i.e. the step count from the topic codes within a single pattern code to their common parent cluster must not be larger than $mean(steps) + \sigma$, where $mean(steps)$ refers to the average step count to the parent clusters across all pattern codes. The reliance on quantitative hierarchical clusters serves the purpose of compensating for the drawbacks regularly associated with qualitative coding, e.g., shortcomings in terms of validity and reliability [14]. The resulting adjustments in the initial pattern codes are reflected in evolving pattern codes, which likely imply corresponding alterations in the evolving topic codes. Thus, first-cycle and second-cycle coding unfolds in an iterative fashion. Ultimately, the iterations will reach saturation and further analysis does not yield substantial new insights. At this point, researchers
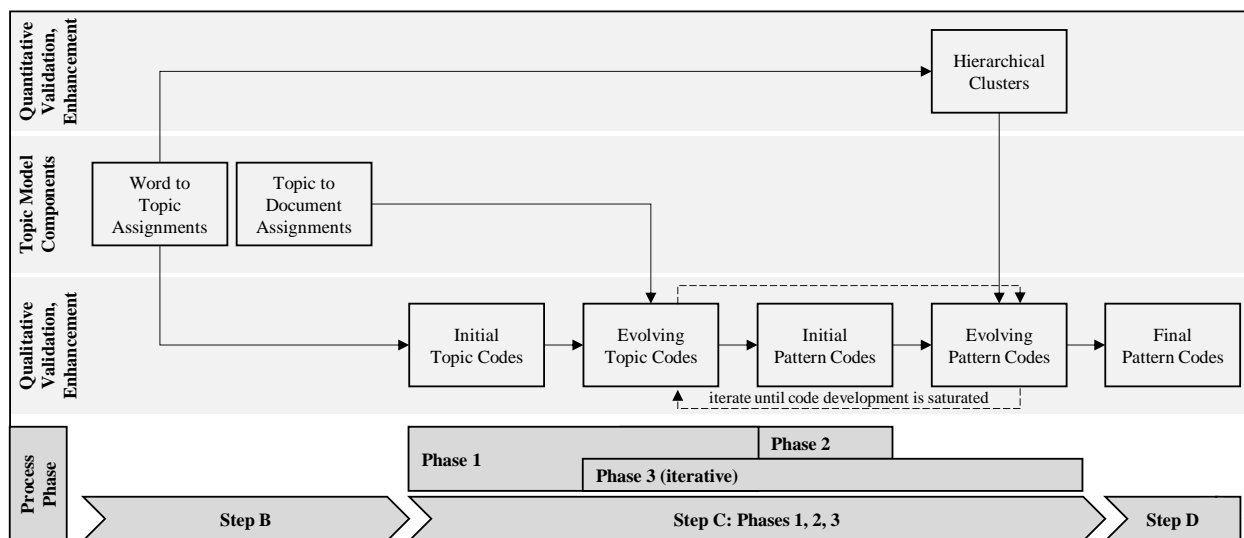
**Figure 2: Nested mixed-methods research design**

terminate the validation and enhancement of the topic model. The final pattern codes and the subsumed topic codes represent the key themes of the analyzed document collection. After all, even though not detailed, we recommend researchers to engage in memo writing to complement the overall coding approach.

## 4. Illustration of the proposed mixed-methods approach by the example of the AIS Senior Scholars' Basket of Journals

To illustrate the MMR approach proposed in section 3, we conducted an analysis of the thematic composition of the AIS Senior Scholars' Basket of Journals. We chose this literature sample for our case study because it ensures working with a subject matter likely familiar to our readers. Our goal is not to provide novel insights into the thematic structure of the IS discipline but to showcase that our structured MMR approach arrives at high level themes that are familiar to IS scholars based on the re-contextualized final results. We analyze all research articles published since the foundation of each journal, namely *EJIS, ISJ, ISR, JAIS, JIT, JMIS, JSIS,* and *MISQ.*

**Applying Step A:** As some of our sample PDFs did not contain textual content, we performed optical character recognition where needed before extracting the full text of each article for analysis. This left us with 7,356 articles spanning 40 years (1977 – Qr. 2 2017) in total. Indeed, this particular document collection could have also been analyzed by other means than looking at the full content of each article, e.g., a keyword-based analysis. However, as noted by Moffitt et al. [37], keywords often poorly represent the actual document content. Besides, large document collections do usually not feature keywords or similar metadata at all. Thus, they do not enable a comparably comfortable analysis. For these reasons, we subsequently further delineate the MMR approach proposed in section 3.

**Applying Step B**: First, we pre-processed the document collection by removing stop-words as well as domain-specific noise, e.g., journal names. Besides, we lemmatized the documents [13]. The pre-processing resulted in a term-document matrix. Second, we applied a dynamic topic model [5] to the pre-processed corpus. In contrast to conventional LDA models, this model type ensures a better model fit in document collections that include texts written over many years as it allows for changes in per-topic word assignments over time. The dynamic topic model consisted of two key components, namely 40 word to topic assignments and the resulting topic to document assignments. These two components formed the basis of step C.

**Applying Step C - Phase 1:** Based on the 50 most likely words per topic, two members of our research team inductively assigned topic codes to each of the 40 topics independently. Thus, we achieved an inter-coder reliability of 65%, measured as the number of equal codes developed divided by the total number of codes developed per researcher, i.e. 40. The inter-coder reliability check continued by in-depth discussions of the discrepancies in individual judgement based on the memos developed during individual topic coding. Ultimately, we agreed upon a final list of 40 initial topic codes. From these codes, we gained a tentative idea of the themes contained in the data set. To validate this preliminary result, we went back to the document level on the basis of the topic to document assignments. Precisely, we carefully read the titles and abstracts of the

**Table 1: Exemplary illustration of topic code re-labelling**

| Initial topic code | Evolving topic code (re-labeled) | Exemplary presentation of ten most likely words per topic (NN = noun, JJ = adjective, VB = verb) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Organizational Learning | Cognition and Behavior | mode l NN | construct VB | user NN | factor NN | behavior NN | item NN | effect NN | social JJ | usage NN | team NN |
| Design Science | User Involvement in Software Development | user NN | project NN | development NN | organization NN | manager NN | group NN | job NN | support NN | decision NN | team NN |

15 most representative articles per topic determined by the dynamic topic model. Based on the emerging contextual insights, we revised the initial topic codes. If the topics did not show internal homogeneity, we excluded them from further analysis. Thus, we excluded five topics. Also, given a lack of external heterogeneity, we joined three topics. In 17 cases, we assigned more meaningful labels to the topics. Table 1 exemplifies this re-labelling of topic codes for two topics. We retained 16 initial topic codes. In sum, the revision process yielded a list of 33 evolving topic codes.

**Applying Step C - Phase 2:** To draw inferences at a higher level of abstraction, each individual researcher engaged in pattern coding. Thus, we achieved an inter-coder agreement of 60%. Having resolved individual biases, we arrived at ten initial pattern codes.

**Applying Step C - Phase 3:** To validate the initial pattern codes, we relied on quantitative hierarchical clustering. The 40 clusters were derived on the basis of the 40 word to topic assignments. The result is illustrated in Figure 3, and annotated with the initial topic codes. Our evaluation of the average step distance from topic codes to parent clusters revealed two initial pattern codes which included topic codes that violated the metric described in section 3 (Step C - Phase 3), i.e. the steps required in these pattern codes were larger than $mean(steps) + \sigma$. This rule implied an iterative analytical process, during which went back and forth between the evolving pattern codes and evolving topic codes. The iterations let us to re-assign two topic codes under a different pattern code. The worst offender was

given by the *User Satisfaction* topic code (Table 2), which we initially assigned to the *Design Science* pattern code. We re-assigned it to the *(IT-enabled) Organizational Change* pattern code, which reduced the step count to the parent cluster from nine to six. Also, we re-labeled one pattern code, i.e. the *Business Value of IT* pattern code was initially labeled *Performance*. When the iterations did not yield substantially new insights, we terminated the revision of the topic model.

**Applying Step D and Results:** The final pattern codes and their respective final topic codes represent the thematic composition of the AIS Senior Scholars' Basket of Journals which we derived by means of our MMR approach. We present these results in Table 2. In this table, we also include clear operational descriptions [35] of our final pattern codes labeled as: *Decision Support Systems; Strategic Management; (IT-enabled) Organizational Change; Design Science; Behavior, Cognition, Affect; Risk Management; Business Value of IT; Electronic Markets; Knowledge Management; Theoretical and Methodological Foundations*.

In addition, to illustrate how the importance of the key themes within the analyzed AIS Senior Scholars' Basket of Journals has evolved throughout the analyzed timeframe, we include a plot of their relative historical development in Figure 4. This presentation is just one of many possible ways to provide more details on the composition of the document collection. Future studies applying our MMR approach could cover other possible analyses, e.g., how topics differ within different sub-samples of the corpus, or who authored key documents.
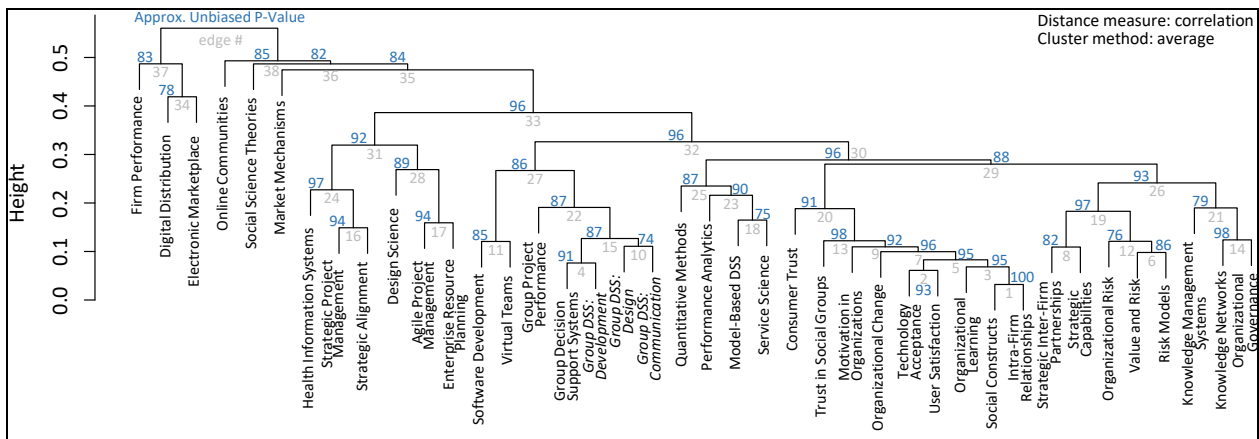


**Figure 3: Hierarchical clusters annotated with initial topic codes**

**Table 2: Final topic code to pattern code assignments**

| Final Topic Codes | Final Pattern Codes | Description of Final Pattern Codes |
|---|---|---|
| Model-Based DSS<br>Group Decision Support Systems<br>Performance Analytics | Decision Support Systems | *"Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision makers who deal with semistructured problems."* [27] |
| IS Management in Public Sector Organizations<br>Strategic Information Systems Planning<br>Organizational Governance<br>Strategic Inter-Firm Partnerships<br>Strategic Alignment<br>Enterprise Resource Planning | Strategic Management | Strategic management is inherently pluralistic. Scholar most often associate the firm's strategic initiatives, internal organization, managers and owners, resources, performance, and business environment with this concept. [39, p. 947] |
| Technology Acceptance and Adoption<br>Innovation Adoption<br>Resistance to Change<br>Information Technology Capabilities<br>User Satisfaction | (IT-enabled)<br>Organizational Change | *"[Organizational change] is a process through which an organization (e.g., a system of routines) is re-aligned with its environment."* Technology is often a major enabler and driver of this process. [4, p. 117] |
| Software Development<br>Web Personalization<br>User Involvement in Software Development<br>Agile Software Development | Design Science | *"The design-science paradigm seeks to extend the boundaries of human and organizational capabilities by creating new and innovative artifacts."* [24, p. 75] |
| Cognition and Behavior<br>Neuro IS<br>Trust in Social Groups<br>Motivation in Organizations | Behavior, Cognition, Affect | IS research concerned with what people do (behavior), what people think (cognition), and/or what people feel (affect). [9, p. 659] |
| Risk Models<br>Security Risks | Risk Management | *"The term risk management [...] is concerned with a phased, systematic approach (possibly implemented through a unique tool or technique) to the analysis and control of the risks occurring within a specific context."* [10, p. 373] |
| IT Impact on Firm Performance | Business Value of IT | *"IT business value research examines the organizational performance impacts of information technology."* [34, p. 284] |
| Technology-Mediated Distribution<br>Pricing Mechanisms<br>Electronic Marketplace | Electronic Markets | *„An electronic marketplace or electronic market system is an interorganizational information system that allows the participating buyers and sellers in some market to exchange information about prices and product offering."* [2, p. 1676] |
| Organizational Knowledge Management<br>Geographically Dispersed Teams<br>Knowledge Sharing | Knowledge Management | *"Knowledge management is the generation, representation, storage, transfer, transformation, application, embedding, and protecting of organizational knowledge."* [43, p. 218] |
| Process Modeling<br>Theories and Research Designs | Theoretical and Methodological Foundations | This topic subsumes articles dedicated to discussions about methods, and theories. [23] |

# 5. Conclusion

We address the challenge of analyzing the thematic composition of large document collections using quantitative topic models, which usually implies an erosion of contextual meaning. To address this challenge, we propose a structured mixed-methods research (MMR) approach. Its centerpiece is given by a nested mixed-methods research design building upon the model components resulting from quantitative topic modeling. This design draws on qualitative coding and quantitative hierarchical clustering to validate and enhance topic models through re-contextualization. To highlight the utility and applicability of the proposed approach, we conduct an illustrative study of the AIS Senior Scholars' Basket of Journals. Its resulting thematic composition exhibits a useful outcome of in-depth mixed-methods topic model analysis. As will be apparent to many readers, some of the resulting themes
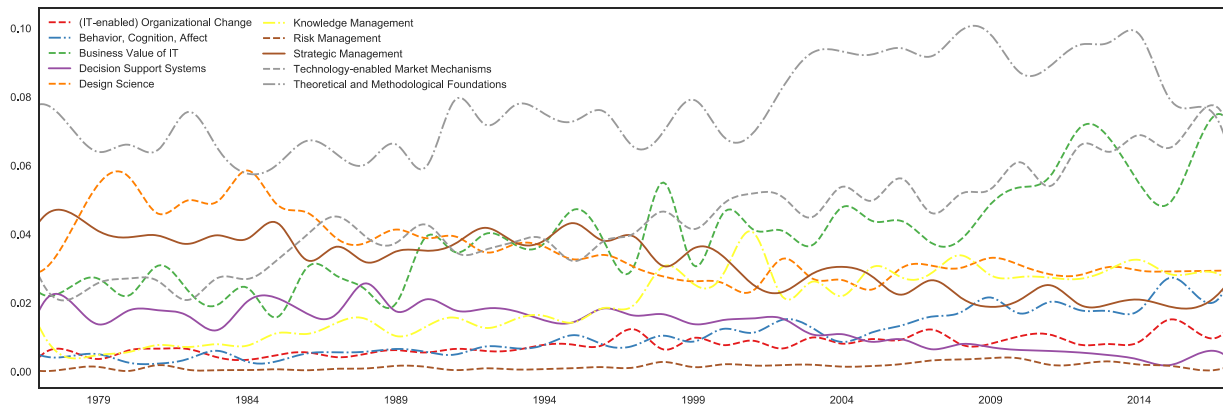
**Figure 4: Development of relative pattern code importance over time**

are not very granular, i.e. it would be sensible to use a topic model with a higher topic count (40 in our case). As choosing the topic count is a crucial parameter selection step, this also highlights the importance of domain knowledge in topic modeling research. Thus, our approach supports researchers in translating their domain knowledge into parameter choices. Based on this retrospective assertion, they can revise parameter choices in a contextually informed way. In sum, the proposed MMR approach mitigates shortcomings of both quantitative research, such as the problematic choice of the "correct" amount of topics included in a model, and de-contextualization of textual data, as well as qualitative research, such as deficiencies regarding validity, reliability, and replicability. Future work may consider validating and refining the proposed MMR approach, as well as creating software environments that support researchers in implementing such an approach.

## 6. References

[1] Ågerfalk, P. J., "Embracing diversity through mixed methods research," *European Journal of Information Systems,* vol. 22, 2013, pp. 251-256.

[2] Bakos, J. Y., "Reducing Buyer Search Costs: Implications for Electronic Marketplaces," *Management Science,* vol. 43, 1997, pp. 1676-1692.

[3] Bandara, W., Furtmueller, E., Gorbacheva, E., Miskon, S., and Beekhuyzen, J., "Achieving Rigor in Literature Reviews: Insights from Qualitative Data Analysis and Tool-Support," *Communications of the AIS,* vol. 37, 2015, pp. 154-204.

[4] Besson, P. and Rowe, F., "Strategizing information systems-enabled organizational transformation: A transdisciplinary review and new directions," *The Journal of Strategic Information Systems,* vol. 21, 2012, pp. 103-124.

[5] Blei, D. M. and Lafferty, J. D., "Dynamic Topic Models," in *23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006, pp. 113-120.

[6] Blei, D. M. and Lafferty, J. D., "Topic models," in *Text mining: Classification, Clustering, and Applications*, A. Srivastava, and Sahami, M., Ed., ed Boca Raton: CRC Press, 2009, pp. 71-93.

[7] Blei, D. M., Ng, A. Y., and Jordan, M. I., "Latent Dirichlet Allocation," *The Journal of Machine Learning Research,* vol. 3, 2003, pp. 993-1022.

[8] Boyatzis, R. E., *Transforming Qualitative Information: Thematic Analysis and Code Development*. Thousand Oaks, CA: SAGE Publications, 1998.

[9] Burton-Jones, A. and Gallivan, M. J., "Toward a Deeper Understanding of System Usage in Organizations: A Multilevel Perspective," *MIS Quarterly,* vol. 31, 2007, pp. 657-679.

[10] Charette, R. N., "The mechanics of managing IT risk*," *Journal of Information Technology,* vol. 11, 1996, pp. 373-378.

[11] Coffey, A. and Atkinson, P., *Making Sense of Qualitative Data: Complementary Research Strategies*. Thousand Oaks, CA: SAGE Publications, 1996.

[12] Creswell, J. W. and Clark, V. L. P., *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage Publications, 2007.

[13] De Smedt, T. and Daelemans, W., "Pattern for Python," *Journal of Machine Learning Research,* vol. 13, 2012, pp. 2031–2035.

[14] Debortoli, S., Müller, O., Junglas, I., and vom Brocke, J., "Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial," *Communications of the AIS,* vol. 39, 2016, pp. 110-135.

[15] Edson, C. H., "Our Past and Present: Historical Inquiry in Education," *Journal of Thought,* vol. 21, 1986, pp. 13-27.

[16] Faraj, S., Jarvenpaa, S. L., and Majchrzak, A., "Knowledge Collaboration in Online Communities," *Organization Science,* vol. 22, 2011, pp. 1224-1239.

[17] Galliers, R. D., "In celebration of diversity in information systems research," *Journal of Information Technology,* vol. 26, 2011, pp. 299-301.

[18] Gerrish, S. and Blei, D. (2009). *Modeling Influence in Text Corpora*. Available: www.umiacs.umd.edu/~jbg/nips_tm_workshop/30.pdf

[19] Gerrish, S. M. and Blei, D. M., "A Language-based Approach to Measuring Scholarly Impact," in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010, pp. 375-382.

[20] Glaser, B. G. and Strauss, A. L., *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine Publishing Company, 1967.

[21] Greene, J. C. and Caracelli, V. J., "Crafting Mixed-Method Evaluation Designs," in *New Directions for Evaluation - Advances in Mixed-Method Evaluation: The Challenges and Benefits of Integrating Diverse Paradigms*, J. C. Greene and V. J. Caracelli, Eds., ed San Francisco, CA: Jossey-Bass, 1997, pp. 19-32.

[22] Greene, J. C., Caracelli, V. J., and Graham, W. F., "Toward a conceptual framework for mixed-method evaluation designs," *Educational evaluation and policy analysis,* vol. 11, 1989, pp. 255-274.

[23] Gregor, S., "The Nature of Theory in Information Systems," *MIS Quarterly,* vol. 30, 2006, pp. 611-642.

[24] Hevner, A. R., March, S. T., Park, J., and Ram, S., "Design Science in Information Systems Research," *MIS Quarterly,* vol. 28, 2004, pp. 75-105.

[25] Janasik, N., Honkela, T., and Bruun, H., "Text Mining in Qualitative Research: Application of an Unsupervised Learning Method," *Organizational Research Methods,* vol. 12, 2009, pp. 436-460.

[26] Kaplan, A., *The Conduct of Inquiry: Methodology for Behavirol Science*. Scranton, Pennsylvania: Chandler Publishing Company, 1964.

[27] Keen, P. G. W. and Morton, M. S. S., *Decision Support Systems: An Organizational Perspective*. Reading, MA: Addison-Wesley, 1978.

[28] King, N. and Brooks, J., *Template Analysis for Business and Management Students*. London: Sage Publications, 2017.

[29] Lazarsfeld, P. F., *Qualitative Analysis: Historical an Critical Essays*. Boston: Allyn and Bacon, 1972.

[30] Lemke, M., Niekler, A., Schaal, G. S., and Wiedemann, G., "Content Analysis between Quality and Quantity: Fulfilling Blended-Reading Requirements for the Social Sciences with a Scalable Text Mining Infrastructure," *Datenbank Spektrum,* vol. 15, 2015, pp. 7-14.

[31] Lincoln, Y. S. and Guba, E. G., *Naturalistic Inquiry*. Beverly Hills, CA: SAGE Publications, 1985.

[32] Marton, F., "Phenomenography - A Research Approach to Investigating Different Understandings of Reality," *Journal of Thought,* vol. 21, 1986, pp. 28-49.

[33] Mayring, P., *Qualitative Inhaltsanalyse: Grundlagen und Techniken*, 11th ed.: Beltz Verlag, 2010.

[34] Melville, N., Kraemer, K., and Gurbaxani, V., "Review: Information Technology and Organizational Performance: An Integrative Model of IT Business Value," *MIS Quarterly,* vol. 28, 2004, pp. 283-322.

[35] Miles, M. B. and Huberman, A. M., *Qualitative Data Analysis: An Expanded Sourcebook*, 2nd ed. Thousand Oaks, CA: SAGE Publications, 1994.

[36] Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A., "Optimizing Semantic Coherence in Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Edinburgh, 2011, pp. 262-272.

[37] Moffitt, K. C., Richardson, V. J., Snow, N. M., Weisner, M. M., and Wood, D. A., "Perspectives on Past and Future AIS Research as the Journal of Information Systems Turns Thirty," *Journal of Information Systems,* vol. 30, 2016, pp. 157-171.

[38] Murthy, D., "The ontology of tweets: Mixed methods approaches to the study of Twitter," in *The SAGE Handbook of Social Media Research*, L. Sloan and A. Quan-Haase, Eds., ed London: Sage Publications, 2017, pp. 559-572.

[39] Nag, R., Hambrick, D. C., and Chen, M. J., "What is Strategic Management, Really? Inductive Derivation of a Consensus Definition of the Field," *Strategic Management Journal,* vol. 28, 2007, pp. 935-955.

[40] Rai, A., "Editor's Comments - Synergies Between Big Data and Theory," *MIS Quarterly,* vol. 40, 2016, pp. iii-ix.

[41] Robey, D., "Research Commentary - Diversity in Information Systems Research: Threat, Promise, and Responsibility," *Information Systems Research,* vol. 7, 1996, pp. 400-408.

[42] Saldaña, J., *The Coding Manual for Qualitative Researchers*. Thousand Oaks, CA: SAGE Publications, 2009.

[43] Schultze, U. and Leidner, D. E., "Studying Knowledge Management in Information Systems Research: Discourses and Theoretical Assumptions," *MIS Quarterly,* vol. 26, 2002, pp. 213-242.

[44] Suzuki, R. and Shimodaira, H., "Pvclust: An R package for assessing the uncertainty in hierarchical clustering," *Bioinformatics,* vol. 22, 2006, pp. 1540-1542.

[45] Tashakkori, A. and Teddlie, C., "Introduction to mixed method and mixed model studies in the social and behavioral sciences," in *Mixed Methodology: Combining Qualitative and Quantitative Approaches*, L. Bickman and D. J. Rog, Eds., ed Thousand Oaks, CA: SAGE Publications, 1998, pp. 3-19.

[46] Tashakkori, A. and Teddlie, C., "Major Issues and Controversies in the Use of Mixed Methods in the Social and Beavioral Sciences," in *Handbook of Mixed Methods in Social and Behavioral Research*, A. Tashakkori and C. Teddlie, Eds., ed Thousand Oaks: SAGE Publications, 2003, pp. 3-50.

[47] Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M., "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association,* vol. 101, 2006, pp. 1566-1581.

[48] Tesch, R., *Qualitative Research: Analysis Types and Software Tools*. New York: Falmer Press, 1990.

[49] Venkatesh, V., Brown, S. A., and Sullivan, Y. W., "Guidelines for Conducting Mixed-methods Research: An Extension and Illustration," *Journal of the AIS,* vol. 17, 2016, pp. 435-495.

[50] Wallach, H. M., "Topic Modeling: Beyond Bag-of-Words," presented at the Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, USA, 2006.

[51] Wang, C., Paisley, J., and Blei, D. M., "Online Variational Inference for the Hierarchical Dirichlet Process," in *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 752-760.