Trends in Academic and Industrial Research on Business Process Management - A Computational Literature Analysis

Fabian Muff
Universtity of Fribourg
Switzerland
fabian.muff@unifr.ch

Felix Härer
Universtity of Fribourg
Switzerland
felix.haerer@unifr.ch

Hans-Georg Fill
Universtity of Fribourg
Switzerland
hans-georg.fill@unifr.ch

Abstract

An important aspect of enterprise information systems is the management and execution of business processes. For exploring the evolution of topics in business process management in academia and industry, we present the findings from a computational literature analysis. For this purpose, we revert to the full texts and metadata of the proceedings of the International Conference on Business Process Management and its workshops as a sample. In addition, the data has been enriched with data on the academic or industrial provenance of the authors. For identifying the most important topics in business process management, we performed a content-based analysis of over 1,200 papers using Latent Dirichlet Allocation. This analysis gives insights into the development of topics over time and identifies recently emerging topics.

1. Introduction

The design, execution, and management of business processes has traditionally been recognized as an important component of enterprise information systems for ensuring the consistency of decisions, for describing data exchanges, and for the semantic alignment of business requirements and IT services in general [1, 2]. In today's fast-paced digital world with often changing requirements and ephemeral technologies, also the domain of business process management has to react and take these developments into account for providing optimal support for enterprises [3, 4]. Novel technologies and research trends have to be identified and incorporated into industrial applications.

For investigating the trends and future prospects in the domain of business process management we review in the following the evolution of trends and topics in this area. The method we use for this purpose, is a computational literature analysis [5]. It is based on the papers from the International Conference on Business Process Management and its workshops as the top outlet in this field [6]. Although BPM research is scattered over many different outlets in information systems research, we decided for the BPM conference and its workshops because the data is available from one single publisher, the conference is dedicated only to BPM and it is the widely-accepted top conference in this field with a highly competitive review process. Therefore, it is estimated that the investigated papers will show a representative sample for the overall BPM community. Furthermore, we will conduct comparisons with previous analyses.

The research questions that we will tackle in this study are the following:

- RQ 1: In which geographical regions is BPM research conducted? Are there dominating regions in terms of research output?
- RQ 2: How many authors are active in the BPM conference and do they work in academia or industry?
- RQ 3: What are the major topics in BPM in academia and industry, and how did they evolve over time?

The remainder of the paper is structured as follows. In Section 2 we briefly review related literature analyses in the area of BPM. This is followed by a description of the research methodology in Section 3 and the descriptive and content-based analysis of publications in Sections 4 and 5. Thereafter we discuss the results and describe limitations of the study in Section 6. The paper is concluded with an outlook on future research in Section 7.

2. Related Work

The investigation of literature sources for assessing the state-of-the-art of a discipline is a core part of any research endeavor and requires rigorous documentation [7, 8]. This process has and is still frequently conducted manually and with large effort.

URI: https://hdl.handle.net/10125/80215
978-0-9981331-5-7

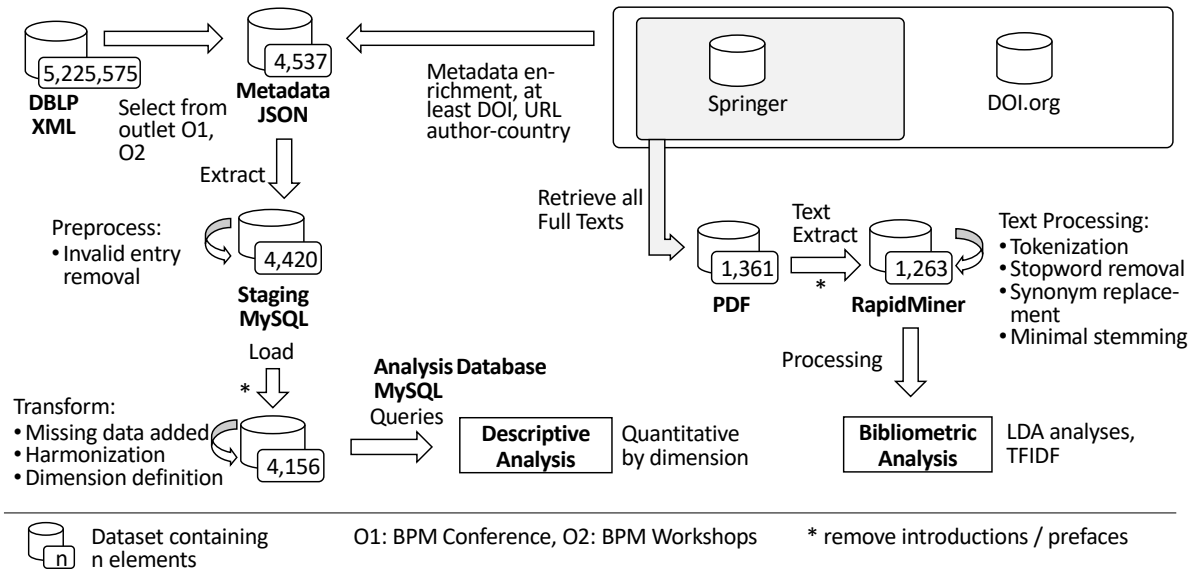

Figure 1. Data collection and analysis process based on the DBLP XML dataset

As a consequence, such analyses are often limited to a subset of available sources, e.g., by considering only particular outlets. For circumventing this limitation and enhancing the analysis, automated approaches have been proposed in addition [9, 10, 11]. In the following we review the most recent manual and automated literature analyses of the BPM discipline.

In 2013, van der Aalst [12] conducted a comprehensive survey on BPM. His work presents twenty different use cases for BPM and how the main concerns of the BPM community are covered by them. For this purpose, 289 papers from the BPM conferences of 2003-2011 and of a previously published book were assigned to one or more use cases. In addition, each paper was manually tagged with one or more key concerns in BPM from a set of 342 tags. This permitted insights into the evolution of key concerns over time. In contrast to our approach that builds on an automated analysis, the analyses in this paper are based on a subjective assessment and assignment.

In 2016, Recker and Mendling published an analysis of 347 papers presented at the BPM conference [6]. For their analysis, they focused on the identity and progress of the BPM conference. This included a classification study and an analysis of citation data. For this purpose, all papers from the conference proceedings from 2003 to 2014 were included. The classification of the papers followed a specifically-developed coding scheme that was applied by the authors in a manual process by reading the full-texts of the papers. The categories for the classification comprised the focus and intent of each paper, the research components, the research method, the positioning within the BPM lifecycle, the empirical evidence, and the type of implementation. For measuring impact, citation data was extracted for each paper using Google Scholar.

An automated literature analysis method has been described by Houy et al. [10] and applied to 905 papers in the Business Process Management Journal and the BPM conference proceedings for the years 2005-2011. Their work focuses primarily on past trends in the community. However, only the abstracts of the individual papers were analyzed and a thesaurus had to be derived by a domain expert in addition. Overall, the focus is put on individual terms and term groups, and the trends were derived over the entire period for individual terms.

More recently, Neder et al. conducted an automated trend analysis in business process management [13]. They reverted to a set of 661 papers with metadata extracted from the Web of Science database on the topic of business process management in the time frame 1995-2018. The papers were divided into those focused on business and management and those associated with information technology for comparing these two directions. With this dataset, a semantic network analysis based on TF-IDF metrics was conducted for identifying the evolution of concepts over different time frames.

In summary, we can state that several profound analyses of BPM as a discipline have been conducted in the past based on the available literature. Apart from manual analyses, some sources also report on automated, computational approaches. What is however

missing in former research is a computational approach joining descriptive (RQ1 / RQ2) and content-based (RQ3) metrics that are verifiable. Furthermore, geographical distributions (RQ1), analysis of the active authors, as well as the potential differences between academic and industrial research on business process management have not been analyzed so far.

3. Research Methodology

Unlike most previous studies and surveys in the BPM community, the analysis method in this work is two-fold. After automated data collection and semi-automated cleaning steps, an ETL workflow with multi-dimensional analysis [14] is applied at first for descriptive statistics. Secondly, existing data is enriched with full texts for applying bibliometric topic identification methods. For classifying the process as whole, it follows the well-known data mining and data analysis approach KDD [15] in its process of data selection, preprocessing, transformation, data mining, and interpretation and evaluation. The method is illustrated in Fig. 1 and further explained in Section 3.2.

3.1. Aims and Scope of the Study

The aim of this work is to analyze recent topics and trends in the discipline of BPM as well as the community itself using automatic data collection and bibliometric analyses based on a sample of top publications. As a first step, a quantitative analysis of the metadata of the BPM Conference and the BPM Workshops between the years 2005 and 2019 has been performed. By conducting and interpreting these analyses, we will highlight several aspects that characterize the BPM community. These include information on the geographical regions in which BPM research is conducted, the development of the quantity of papers and authors in the community, or the academic or industrial background of the participating institutions. Furthermore, we will show the different topics in relation to the conference and the workshops and analyze their evolution over time. Given the high visibility and scientific reputation of the BPM conference and its workshops, it is estimated that this gives valuable insights into the trends and the community in BPM.

3.2. Data Collection

A dump of the DBLP computer science bibliography[1] database from 2020-11-19 was used as the basis for data collection. The BPM conference and its workshops were chosen as the starting point. The XML file from the DBLP contained a total of 1,361 entries on BPM and BPM workshops respectively. In the vast majority of cases, these entries contained title, authors, year, outlet, URL, and DOI and were combined into a JSON file. In addition, this data was enriched using DOI.org and the publisher websites from Springer, adding possibly the DOI, affiliation and the country to the metadata. The data on publications and queries have been made publicly available [16].

After the collection of the raw data, the metadata was extracted into a staging database consisting of 4,420 entries. Then, a manual harmonization of all names, countries, institutions, cities and outlets, as well as the elimination of invalid entries, including non-paper posts, such as editorials and introductions and placeholders with missing authors, was conducted on this dataset. The remaining 4,156 entries were converted into the star scheme with partially normalized dimensions or ”snowflaking” as shown in Fig. 2 for the analysis of multiple dimensions.

Fortunately, we were able to rely on a single publisher for the full texts of the papers. The full texts could thus be downloaded automatically using a Node.js scraping script. Subsequently, the reduced set of 1,263 full texts, resulting from removing introductions and prefaces, was converted into text files and the titles of the respective papers were filtered out of the full texts in order to prevent a possible bias inherent in the data. After that, the documents were loaded into RapidMiner Studio 9.8. Within RapidMiner, the normalization of plurals and frequent inflected forms, as well as further NLP activities such as tokenization, stop word elimination, synonym substitution, and limited stemming were applied.

3.3. Data Analysis

For the descriptive analysis, the database served as a direct source. Through a quantitative analysis, detailed results were retrieved by executing queries over the dimensions, e.g., for the frequencies of publications. For multi-dimensional queries, the database design proved useful, e.g., authors associated with organizations from countries with varying granularity levels, such as individual countries, continents or all countries. Second, a bibliometric data analysis was conducted on the full texts of the documents. For this purpose, the Latent Dirichlet Allocation (LDA) method was used, which is a statistical tool for identifying topics in documents. Last, a TF-IDF analysis for unique terms occurring in the different topics was performed to get an overview of

[1] https://dblp.org/

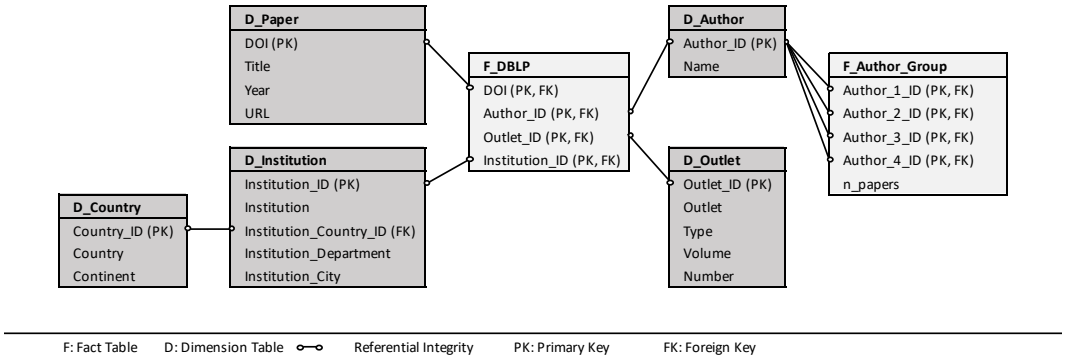

Figure 2. Schema of the analysis database. Fact tables (prefix F) store DBLP publications and author groups according to dimension tables (prefix D) [11]. Note: for all author groups metrics are calculated; author IDs are only stored for up to the first four authors.

the evolution of these terms over time.

4. Descriptive Analysis of BPM Publications

For the following analyses, the time period between 2005 and 2019 is considered. This is due to the fact that the BPM workshops have only been held regularly since 2005 and that the papers from 2020 were not yet available at the time of the analysis. In order to achieve a well-comparable data basis, the total period has additionally been divided into three intervals of five years each. The first descriptive analysis of the dataset targets the number of published papers in each time period – this is depicted in Table 1. On the first axis the different time periods are listed and on a second axis, the different continents are shown as a further dimension.

Table 1. Evolution of the number of published papers (n papers) over three time periods on different continents with sum and distinct total for showing overlapping values – see SQL queries Q1 - Q4 [16].

Continent	Year			
	05 - 09	10 - 14	15 - 19	05 - 19
	n papers			
Africa	2	11	5	18
Asia	32	53	36	121
Europe	316	365	310	991
N. America	46	53	23	122
Oceania	63	38	35	136
S. America	10	19	36	65
Σ	469	539	445	1,453
Distinct Tot.	417	469	377	1,263

Concerning the geographical regions of BPM research (RQ 1), it could be found that authors from 35 countries authored papers at the conference and from 49 countries at the workshops. Thereby, Europe plays a major role, with 78% of all articles having at least one author who was affiliated with a European institution at the time of publication. This applies to all three time periods. This insight can be further substantiated by regarding the number of papers and authors per country. We can observe, that among the top ten countries in terms of published papers, eight are European countries and that these top ten countries contributed to 47% of all published papers. This means, that 47% of all authorships are from authors affiliated to an institution in one of these ten countries.

In RQ 2 we considered how many authors are active in the BPM conference, their typical number of papers and their institutional affiliation. As shown in Table 2, the number of distinct authors between 2005 and 2019 is 2,174. By regarding the average number of authors per paper (avg auth./pap.), we note that this value does increase clearly over the three periods. Regarding the development of the average number of papers per author (avg pap./auth.) over the different years, we see that this number seems to increase steadily, as shown by the last row in Table 2.

Table 2. Number of papers published at the BPM in the three time periods with the total (n aut.) and average number of authors per paper (avg auth./pap.), as well as the average number of papers per author (avg pap./auth.) – see SQL queries Q5 - Q7 [16].

	Year			
	05 - 09	10 - 14	15 - 19	05 - 19
n auth.	820	958	782	2,174
avg auth./pap.	3.0	3.3	3.6	3.3
avg pap./auth.	1.5	1.6	1.7	1.9

In Table 3, the number of authors and institutions

Table 3. Overview over the institution categories and the number of authors (n authors) affiliated in these categories, as well as the number of papers (n papers) with at least one participating institution in the respective category. The table contains sum and distinct total to show overlaps. The categories are Higher Level Education Institutions, Research Institutions, Industry and Other – see SQL queries Q8 - Q9 [16].

Category	Year							
	2005 - 2009		2010 - 2014		2015 - 2019		2005 - 2019	
	n authors	n papers	n authors	n papers	n authors	n papers	n authors	n papers
Higher Level Educ.	700	355	671	368	623	341	1,595	1,082
Research Institution	228	94	216	120	121	87	433	301
Industry	94	55	93	52	40	26	210	133
Other	5	4	3	3	12	6	20	13
Σ	1,027	508	983	561	796	460	2,258	1,529
Distinct Total	820	417	958	469	782	377	2,174	1,263

according to affiliation categories is shown. The institutions were divided into four categories. These comprise purely academic ones such as universities and universities of applied sciences (*Higher Level Education Institutions*), institutions with academic and industrial relations such as Fraunhofer in Germany or Unité Mixte de Recherche (UMR) in France (*Research Institutions*), purely industrial institutions, i.e., companies (*Industry*), to organizations that could not be assigned to one of these categories (*Other*), e.g., municipalities. The institutions were assigned to one of these categories manually [16]. The classification has been derived and cross-checked over multiple iterations by the authors.

As shown in Table 3, an author is assigned to a given category if they co-authored at least one paper representing an institution that was assigned to this category. It is evident that most authors come from *Higher Level Education Institutions*. 1,082 out of 1,529 (71%) of all authorships observed in the dataset have been made in connection with an institution assigned to this category. Another considerable proportion of authors work in *Research Institutions* or in *Industry*. It should be noted that the number of authors in the field of *Research Institutions* is twice as high as in *Industry*, with roughly the same number of institutions in each category. Furthermore, it must also be considered here that an author can be affiliated to multiple institutions. Therefore, a sum row and a distinct total row have been added to Table 3. This is to show these overlaps.

5. Content-based Analysis of BPM Publications

We examined the contents of the papers found in our dataset in the second phase. For this, we used the MALLET (**MA**chine **L**earning for **L**anguag**E** **T**oolkit) and the LDA implementation that is part of RapidMiner 9.8. LDA is a topic modeling methodology operating at the level of documents in order to classify their topics. In comparison to simpler approaches such as word frequency, TF-IDF, and n-gram analysis, LDA constructs a probabilistic model allowing for multiple topics per document. Given a set of documents, any document d is represented by a statistical distribution θ_d over its topics. That is, each subject has a specific probability or weight for d, and for any topic k a distribution of words $\theta_{d,k}$ [17]. The hidden variables of the distributions are computed with parallel processing by the Gibbs sampling scheme, where per-word weights are determined so that their probability of occurring in a specific topic is maximized [18]. For all conducted LDA studies we present the top five terms according to their weight (cf. [17, 19]). The topics are sorted by cumulative weight, with the weight of a topic k and word w as occurrence measure over w assigned to k. Note that it is only possible to include identified weights of the top five words. This procedure replicates closely the approach established before in [11].

The bibliometric analysis with LDA was conducted on the full-texts of documents between 2005 and 2019 and the different sub-periods of all papers, over the workshop and conference category, over all continents, over a subset of institution categories, and over the most involved countries – see [16] for the details of the configuration. Due to space limitations, we present in the following only a subset of the results, concentrating mainly on the differences between academic and industrial institutions between 2010-2014 and 2015-2019.

The LDA results are shown in a standardized format. Per analysis, 8 topics are listed in a table, sorted by cumulative weight of the 5 included terms. After testing different numbers of topics, 8 seemed to be an appropriate number. Thus, a table always refers to an independent analysis and the weights of the individual terms can only be compared within a table.

For the following analyses we divided the dataset into institutions with category 0 or 1, i.e, *Higher Level Education* or *Research Institution*, and institutions with category 2 and 3, i.e., *Industry* and *Other*. This division was chosen since, according to our interpretation, because the first two categories likely represent the academic and the second two categories the industrial sector. Table 4 represents the results of the LDA analysis across all papers between 2010-2014 with an authorship of an author affiliated to an institution of category 1 or 2. Table 5 shows the results of the LDA analysis of the same institution category, but in the period between 2014-2019. We can observe, for example, over these two periods, the term *business* is prominent in the topics 1 and 3 between 2010-2014 and in the topics 3, 4 and 7 in the period of 2015-2019. Topic 1 is the most weighted with the terms *model*, *business*, *information*, *management* and *case* for the period 2010-2014. Topic 2 follows, including terms on *log*, *model*, *event*, in addition to *mining* and *trace*. Further, we note in topic 3 the terms *event*, *rule* and *data*, with topic 4 involving *query* and *match*. Excerpts for the LDA analysis results for all papers with authorships of authors affiliated to institutions of the categories *Industry* or *Other* for the periods 2010-2014 and 2015-2019 are shown in Table 6, resp. 7. In the period of 2010-2014, for example, topic 1 contains the terms *model*, *task*, *business*, *information* and *user*.

If we look at the occurrence of the different words from the LDA analysis, in the categories 0 and 1 between 2010-2019, we note the occurrences of the terms *model* (10), *event* (5), *business* (5), *log* (4), *data* (4) and *task* (4) – see Tables 4 and 5. The number in the brackets behind the different words represents the number of occurrences in the according tables. In the categories 2 and 3 between 2010-2019, the terms *model* (6), *data* (5), *event* (4) and *business* (4) occurred the most, followed by *task* (3) and *log* (3) – see Tables 6 and 7.

Table 4. LDA topics for all papers from 2010 to 2014 with institution category Higher Level Education or Research Institution ordered by cumulative topic weight.

Topic 1	
Word	Weight
model	7884
business	4602
information	2252
management	2249
case	2073

Topic 2	
Word	Weight
log	5012
model	4261
event	4211
mining	2466
trace	2376

Topic 3	
Word	Weight
event	5980
business	3305
data	2890
rule	2339
model	1985

Topic 4	
Word	Weight
model	5138
node	1375
similarity	1332
query	1203
match	1118

Topic 5	
Word	Weight
model	2713
net	2510
transition	1911
petri	1210
state	1166

Topic 6	
Word	Weight
task	1902
service	1858
user	1660
workflow	1547
social	1488

Topic 7	
Word	Weight
model	2038
compliance	1510
task	1462
rule	1014
patient	910

Topic 8	
Word	Weight
service	1789
value	1069
time	1006
customer	994
performance	759

6. Discussion and Limitations

This section summarizes our main findings for the descriptive and content-based analyses and discusses possible interpretations in the context of business process management. Finally, we reflect on the identified research topics in light of prior work that used comparable methods by pointing out the topics supported by and differing from this study.

The research questions set out initially fall into two categories. RQ 1 and 2 can be answered directly in terms of plain data and descriptive statistics (Section 4), related to geographical distribution (RQ 1) and the number of active authors with their affiliations and their according categories (RQ 2). Secondly, RQ 3 involves the data and interpretation of the content-based analysis (Section 5).

For RQ 1, the ample geographical diversity of BPM becomes obvious through the fact that people from all continents authored papers at the BPM conference in the past. However, we can also note a strong European influence – see Table 1. It must be mentioned at this point, that we did analyze the BPM conference and its workshops as representative for the whole BPM community. Due to the limitation to one conference, however, only an indication for the overall community can be given. When adjusting for relative changes over time, the parameter of geographical origin reveals a remarkably stable participation share from Europe between 68% and 70% of all papers over 15 years, while

Table 5. LDA topics for all papers from 2015 to 2019 with institution category Higher Level Education or Research Institution ordered by cumulative topic weight.

Topic 1	
Word	Weight
model	5301
log	4707
event	3655
trace	3240
algorithm	1608

Topic 2	
Word	Weight
event	2845
data	2825
mining	2035
log	1554
patient	1391

Topic 3	
Word	Weight
model	2942
data	2514
business	2294
case	1742
goal	1049

Topic 4	
Word	Weight
model	7390
business	954
pattern	768
information	722
result	694

Topic 5	
Word	Weight
decision	2566
model	2412
constraint	1790
rule	1135
task	1115

Topic 6	
Word	Weight
time	1951
event	1866
case	1664
performance	1199
log	1182

Topic 7	
Word	Weight
business	2509
management	1653
research	1276
organization	1109
resource	1074

Topic 8	
Word	Weight
task	1190
data	1073
service	695
blockchain	551
smart	544

South America is the only continent with a clear upward trend – see Table 1.

In terms of the number of authors who published papers at the BPM and their work for either academia or industry (RQ 2), the total number of authors did not change notably over time – see Table 2. At the same time, the average number of authors per paper published at the BPM grew from 3.0 to 3.3 and 3.6, indicating a greater relevance of collaborations – see Table 2. When regarding the development of the average number of papers per author over the different years, we see that this number increases steadily, as shown by the last row in Table 2. This may be an indication that is reflected upon only in the BPM conference. The data on affiliations for the conference and workshops suggests considerably fewer publications of authors from *Research Institutions* and *Industry* in recent time

Table 6. LDA topics for all papers from 2010 to 2014 with institution category Industry or Other ordered by cumulative topic weight.

Topic 1	
Word	Weight
model	1179
task	684
business	679
information	412
user	367

Topic 2	
Word	Weight
event	830
mining	372
log	357
data	344
system	237

Topic 3	
Word	Weight
case	438
project	353
use	267
organization	246
management	221

Topic 4	
Word	Weight
application	277
wsn	212
service	204
web	178
component	161

Topic 5	
Word	Weight
test	277
data	245
access	208
security	131
control	129

Topic 6	
Word	Weight
state	266
quality	246
model	147
place	111
set	108

Topic 7	
Word	Weight
component	217
product	163
business	106
compliance	100
service	98

Topic 8	
Word	Weight
graph	127
event	121
email	107
voting	103
dcr	99

frames, whereas the number of authors from *Higher Level Education Institutions* has only slightly decreased – cf. Table 3. In summary, we can observe concentration effects shown by the decreasing number of affiliations and geography. Further studies would be required to analyze possibly hidden relationships in these areas and reveal the causes for these developments, as well as to draw definitive conclusions for the entire community.

The third research question RQ3 concerned the content-based analysis of publications and the differences in academia and industry. LDA describes topics by individual terms objectively; however, the discussion of topics has a subjective component because of semantic ambiguities.

Due to space limitations we restrict in the following the analysis of the LDA results to the two time periods 2010-2014 and 2015-2019. Further, we only regard the results for papers originating from authors coming

Table 7. LDA topics for all papers from 2015 to 2019 with institution category Industry or Other ordered by cumulative topic weight.

Topic 1	
Word	Weight
model	476
business	419
goal	367
case	250
task	226

Topic 2	
Word	Weight
event	295
data	227
patient	208
mining	196
log	145

Topic 3	
Word	Weight
data	264
business	207
model	200
support	173
management	166

Topic 4	
Word	Weight
log	265
search	197
graph	195
event	166
model	165

Topic 5	
Word	Weight
model	353
resource	209
time	141
ontology	139
tenant	133

Topic 6	
Word	Weight
organization	202
workaround	187
activitie	130
improvement	116
factor	91

Topic 7	
Word	Weight
constraint	196
data	126
time	103
privacy	91
task	86

Topic 8	
Word	Weight
service	166
language	107
problem	102
text	99
ballerina	97

from *Higher Level Education* or *Research Institution* on the one hand and to authors coming from *Industry* or *Other* on the other hand. As papers may have been jointly authored by people from academia and industry, overlaps may occur.

Despite of the limitations of the dataset, which only presents a sample of the overall research on BPM, some interesting insights can be gained. First, topics related to *process mining* are increasingly prominent in academic research as the terms related to this area increased relative to the other topics of the time frames - Topic 2 in 2010-2014 compared to Topics 1 and 2 in 2015-2019 – see Tables 4 and 5. In the papers with authors who have an industry affiliation, process mining also takes a prominent role, although it did not make it to the first topic in both time spans - see Tables 6 and 7. Rather, there is an indication that 'traditional' business process modeling tasks are still of primary interest in industrial research.

Further, comparatively novel approaches such as *decision modeling* and *blockchains* recently appeared in academic papers at the BPM conference - see Topics 5 and 8 in Table 5, whereas they are so far absent in industrial research papers on BPM. One interpretation may be that these topics are still at the fundamental research stage and have yet not been investigated broadly in industrial research, despite several potential applications [20, 21]. An interesting observation can be made in Topic 5 of the industrial papers between 2015-2019 in Table 7 where the term *ontology* appears. Semantic business process management and the use of ontologies has long been studied in academic and applied BPM research - see e.g., [22]. Whereas this term has recently not appeared prominently in academic research papers at the BPM conference, there seems to be interest from the side of industry in this field.

On the other hand, the terms *workaround*, *improvement*, *data*, and *privacy* stand out in the topics of recent industrial research papers - see Topics 6 and 7 in Table 7. These terms cannot be found among the primary topics inferred for academic research papers. This may be due - on the one hand - to the importance of these topics in practice, as can be confirmed for example for the topic of business process improvement by the recent BPTrends survey [23]. On the other hand, academic approaches to business process improvement are also discussed in the context of quality management and thus may not appear prominently at the BPM conference [24, 25].

The observations made through the LDA analysis suggest that the topic of process mining has been important both in academic research on BPM and in industry. Whereas the focus of academic publications at the BPM conference seems to be more oriented towards the use of novel technologies such as decision models or blockchains, it is obvious that industry papers have a stronger focus on business-related topics, such as business goals, improvement, or workarounds as these topics are slightly higher weighted in the analysis. Thus, it could be an opportunity for future academic research to focus more on aspects of process improvement and combine them with technological solutions, as already pursued in recently-funded research projects in this area[2].

When comparing the results of our analysis with previous investigations, we can find the following. In Houy et al.'s automated analysis [10], a set of BPM terms and concepts was proposed that occur frequently in the abstracts of BPM papers. For comparison, we

[2]See for example the ERC project by Marlon Dumas: `https://cordis.europa.eu/project/id/834141`

extracted a selection of 49 BPM terms and concepts from their result table, which we matched with the terms of our LDA analysis over all papers. From the three periods 2005-2009, 2010-2014, and 2015-2019, it can be shown that 45% of the terms from the LDA analysis can be mapped directly to the terms found in [10], suggesting a shared understanding of common topics. However, the comparison is limited since TF-IDF tends to have a variety of relatively specific subject matters due to the inverse document frequency method favoring terms differentiating topics well, while LDA identifies distributions of frequent topics.

In [13], 16 BPM themes over time were derived using network centrality measures. Unlike the LDA and TF-IDF, these result from nodes in a semantic network. The individual terms of these network nodes often relate to specific subsets of a topic such as *traditional-BPM*, while our analysis shows individual aspects of the topics. Even though broad areas match with our analysis, a full comparison of the results does not seem applicable.

In comparison to the analysis conducted by van der Aalst in [12], the identified LDA topics of the most recent time frames suggest the support of three out of six key concerns: *Process Mining*, *Process Enactment Infrastructures* (i.e., workflow system and service topics), and *Process Model Analysis* (i.e., business and data topics). There are three key concerns of his study that cannot be directly found in our analysis. These are *Process Modeling Languages*, *Process Flexibility* and *Process Reuse*. When comparing the major topics of our LDA analysis, it however revealed topics that do not seem to be explicitly reflected by the key concerns such as *business, management, research, organization, resource*, or *service, value, time, customer, performance* - however, this may depend on a subjective view on these topics and would need to be discussed between the authors of each study in detail. In addition, there are methodological differences to our work. The study by van der Aalst focused on key concerns and use cases, where use cases primarily concern the functions and qualities, not directly capturing topics. For this reason, we only considered key concerns for the comparison.

In a study by Recker and Mendling [6], the phases of the BPM lifecycle were taken as a foundation. The major topics we identified in our LDA analysis are consistent with the BPM lifecycle phases *Process discovery* that can be related to the topics of process mining, *Process analysis* and *Process monitoring and controlling* which are related to business, event, and data topics, *Process identification* and *Process re-design* that are related to process improvement as well as business and management topics, and *Process implementation and execution* that is related to workflow topics.

The study we conducted is however not without some limitations. First, only data from one outlet on business process management research has been analyzed. Although the BPM conference together with its workshops is one of the most prominent and competitive outlets, there are a large number of outlets available for publishing research on business process management. This includes not only business and management related outlets as well as information systems and business informatics journals and conferences. A large body on BPM research is found in computer science related publications and industry journals. Therefore, it needs to be stressed that the investigated data only represents a sample of total BPM research and may overlook important developments in other outlets. Second, the results of the LDA have only been presented and interpreted here in their original form. In further research, the results could be used as a basis for further empirical studies which is out of scope for the paper at hand.

Furthermore, the application of LDA as we used it for our analysis only focuses on the most highly weighted topics. Recently emerging terms that have not yet found their way into a larger number of papers may not surface. This includes for example the recently emerging topic of *robotic process automation* that is already present in BPM workshops and dedicated forums - e.g., [26] - but has not appeared yet under the major topics identified by LDA. A further limitation of LDA is that it rests solely on the occurrence of terms in the overall dataset but does not consider the occurrence on the level of documents. Thus, terms that occur very frequently in single documents may be over-represented. This can for example be suspected for the term 'ballerina' in Topic 8 of Table 7 that obviously originates from one particular paper [27].

7. Conclusion and Future Research

In this paper we conducted a computational bibliometric study on the metadata and the contents of papers at the BPM conference and its workshops from 2005-2019 as a representative for the entire BPM community. The analysis suggests that the BPM community is an international community with contributions coming from all continents. However, the majority of research activities in this field takes place in Europe and the number of authors who are active in the BPM conference and its workshops recently declined.

The content-based analysis of papers from authors with an academic and those with an industrial background confirmed a rising interest in the topic

of *process mining* throughout the last years. While many topics overlap in academic and industrial research papers, the analysis showed that industrial research puts more emphasis on *process improvement* and *data privacy* topics, while academic research in the BPM community rests more on novel technologies such as *decision modeling* or *blockchains*. In the scope of the BPM conference such emerging topics are typically discussed in affiliated forums and workshops, which may serve as indicators for future emerging topics. This includes for example the Blockchain Forum or the Robotic Process Automation Forum.

We hope that the results stipulate discussions in the community on the future of BPM in the context of enterprise information systems. In the future we plan to systematically reflect the gained insights with other members of the community and extend the study to further outlets for enhancing generalizability.

References

[1] K. Andresen and N. Gronau, "Adaptability concepts for enterprise resource planning systems - A component framework," in *11th Americas Conference on Information Systems*, p. 150, AIS, 2005.

[2] H. Fill, "Design of semantic information systems using a model-based approach," in *AAAI Spring Symposium, Technical Report SS-09-08, Stanford, California, USA, March 23-25, 2009*, pp. 19–24, AAAI, 2009.

[3] B. Bender, C. Bertheau, and N. Gronau, "Future ERP systems: A research agenda," in *23rd International Conference on Enterprise Information Systems*, pp. 776–783, SCITEPRESS, 2021.

[4] H. Fill, "Enterprise modeling: From digital transformation to digital ubiquity," in *2020 Federated Conference on Computer Science and Information Systems*, pp. 1–4, 2020.

[5] M. J. Mortenson and R. Vidgen, "A computational literature review of the technology acceptance model," *International Journal of Information Management*, vol. 36, no. 6, pp. 1248–1259, 2016.

[6] J. Recker and J. Mendling, "The state of the art of business process management research as published in the BPM conference - recommendations for progressing the field," *Bus. Inf. Syst. Eng.*, vol. 58, no. 1, pp. 55–72, 2016.

[7] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MISQ*, vol. 26, no. 2, pp. xiii–xxiii, 2002.

[8] J. v. Brocke, A. Simons, B. Niehaves, B. Niehaves, K. Reimer, R. Plattfaut, and A. Cleven, "Reconstructing the giant: On the importance of rigour in documenting the literature search process," *ECIS'2009*, 2009.

[9] R. T. Watson and J. Webster, "Analysing the past to prepare for the future: Writing a literature review a roadmap for release 2.0," *Journal of Decision Systems*, vol. 29, no. 3, pp. 129–147, 2020.

[10] C. Houy, K. Sainbuyan, P. Fettke, and P. Loos, "Towards automated analysis of fads and trends in information systems research: Concept, implementation and exemplary application in the context of business process management research," in *IEEE RCIS Conference*, pp. 1–11, IEEE, 2013.

[11] F. Härer and H. Fill, "Past trends and future prospects in conceptual modeling - A bibliometric analysis," in *International Conference on Conceptual Modeling (ER'20)*, pp. 34–47, Springer, 2020.

[12] W. Aalst, van der, "Business process management: a comprehensive survey," *ISRN Software Engineering*, no. Article ID 507984, 2013.

[13] R. Neder, P. Ramalho, O. Rabelo, E. Zambra, C. Maciel, and N. Benevides, "Business process management: Terms, trends and models," in *FedCSIS 2018*, vol. 17, pp. 163–170, 2018.

[14] S. M. F. Ali and R. Wrembel, "From conceptual design to performance optimization of ETL workflows: current state of research and open problems," *VLDB J.*, vol. 26, no. 6, pp. 777–801, 2017.

[15] U. M. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview," in *Advances in Knowledge Discovery and Data Mining*, pp. 1–34, AAAI/MIT Press, 1996.

[16] F. Muff, F. Härer, and H. Fill, "Index and Queries for the Bibliometric Data of the HICSS 2022 Submission: Trends in Academic and Industrial Research on Business Process Management - A Computational Literature Analysis," Sep 2021. doi:10.5281/zenodo.5511770.

[17] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, 2012.

[18] D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed algorithms for topic models.," *Journal of Machine Learning Research*, vol. 10, no. 8, 2009.

[19] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," p. 487–494, AUAI Press, 2004.

[20] H. Fill, P. Fettke, and S. Rinderle-Ma, "Catchword: Blockchains and enterprise modeling," *Enterp. Model. Inf. Syst. Archit. Int. J. Concept. Model.*, vol. 15, pp. 16:1–16:8, 2020.

[21] J. Mendling and I. Weber, "Blockchains for business process management - challenges and opportunities," *EMISA Forum*, vol. 38, no. 1, pp. 22–23, 2018.

[22] *Proceedings of the 5th International Workshop on Semantic Business Process Management SBPM 2010*, CEUR-WS.org, 2010.

[23] P. Harmon and J. Garcia, "The state of business process management 2020," *BPTrends*, 2020. `https://www.bptrends.com/bpt/wp-content/uploads/2020-BPM-Survey.pdf`.

[24] F. Johannsen and H. Fill, "Meta modeling for business process improvement," *Bus. Inf. Syst. Eng.*, vol. 59, no. 4, pp. 251–275, 2017.

[25] F. Johannsen and H. Fill, "Codification of knowledge in business process improvement projects," in *22st European Conference on Information Systems*, 2014.

[26] S. Agostinelli, A. Marrella, and M. Mecella, "Research challenges for intelligent robotic process automation," in *Business Process Management Workshops*, pp. 12–18, Springer, 2019.

[27] S. Weerawarana, C. C. Ekanayake, S. Perera, and F. Leymann, "Bringing middleware to everyday programmers with ballerina," in *Business Process Management - 16th International Conference, BPM*, pp. 12–27, Springer, 2018.