Activity Duration Prediction of Workflows by using a Data Science Approach: Unveiling the Advantage of Semantics

Tobias Weller and Maria Maleshkova

AIFB Institute (KIT)
Kaiserstr. 89, 76131 Karlsruhe, Germany
{tobias.weller,maria.maleshkova}@kit.edu
http://www.aifb.kit.edu.

Abstract. Organizations often have to face a dynamic market environment. Processes must be frequently adapted in order to stay competitive and allow an efficient workflow. Data Science approaches are currently often used in analysis methods to identify influential indicators on processes and learn predictive models to estimate the duration of an activity. However, current methods do not or only partially make use of semantic information in process analysis. The results are unprecise or incomplete, because not all influential indicators have been unveiled and therefore used in the predictive models.
We want to make use of the semantics and show the advantage by applying them on existing data science methods for predicting the duration of an activity in a process. Therefore, we 1) enrich process data with meta-information and background knowledge 2) extend existing data science methods so that they include semantic information in their analysis and 3) apply data science methods for predicting values and compare the results with methods, which do not use semantics.

Keywords: Data Science, Workflow Analysis, Semantic Annotations, Activity Duration Prediction

1 Introduction

Processes and data are key elements for value creation in companies. A large number of heterogeneous processes contribute to the added value. Large amount of data is created during the execution of processes. These process-generated data is increasingly an essential part of the respective business models. Therefore, Information Systems like ERPs (Enterprise Resource Systems), WFM (Workflow Management Systems) and SCM (Supply Chain Management Systems) are recording events occurring in workflows in a structured way as they take place in order to comprehend past workflows but also using the information in possible analysis. In practice, however, it is often not possible for the employees, involved in the processes, to have all the relevant processes within the organization in mind. This hampers to include all the resulting data in possible analysis for

the detection and optimization of weaknesses in a process, as well as using it in predictive models. However, customers and logistics are interested in knowing the duration of an activity. To improve processes and satisfy the requirements of customers and logistics is process analysis a continuous and important task of organizational development. The purpose of the process analysis is to evaluate organization-specific processes, to identify errors and to improve the possibilities, and to identify deviations from predefined standards, guidelines and existing processes in a system [11].

Process Mining [17] is a technique in the rapidly growing data science discipline for analyzing complex processes. It combines traditional model-based process analysis and data mining techniques. It is used in various domains such as e.g. health-care [17, 18], and industry [1]. This well-known technique is often used because it allows for making implicit, and thus hidden, knowledge about a process transparent. Thus techniques are used to estimate the time of a task by using regression models [6] and descriptive statistical methods [2]. At the moment, however, no semantics are used, and no background and expert knowledge is taken into consideration. Previous research has shown that the inclusion of background knowledge on data improves analysis in clustering [29] and similarity analysis of processes [8]. We want to strengthen this fact by including semantics in process data and apply well-established Process Mining methods on process data, which is not enriched with semantic information, and on process data that is enriched with semantics. We draw a comparison between the results of both approaches. Our goal in this work is to unveil the advantages of including semantics in Process Mining methods. We believe that by clearly demonstrating these, the trend of using semantic information in data science, as well as in process mining, approaches will grow, due to the key fact that the analysis results are improved.

In order to achieve our goals, based on previous work we i) represent a perioperative process in sBPMN and extend our annotation tool to ii) capture meta-information, as well as background knowledge about the process, iii) apply data science methods on the process for analyzing influencing factors on performance indicators of the process and iv) predict the duration of an activity by using different approaches, among others based on semantics and the analyzes of influencing factors. The used materials and methods are shown in section 2. We explain the approach, as well as the metrics used in the experiment section for comparing the results. The concrete implementation and evaluation of our approach is shown in section 3. We model a perioperative process and enrich it with meta-information and background knowledge. The evaluation of our approach includes showing the application, and comparing the results of the different approaches by using well-known metrics, including Root Mean Squared Error and Accuracy. Especially we go into detail in comparing other approaches with our approach, which exploits semantic information. The semantic information is modeled according to the Linked Data principles [3]. Related work is described in section 4. A short discussion, outlook and lessons learned is given in section 5.

2 Material And Methods

The main focus of this work is supporting process analysis by predicting the duration of an activity. We include background knowledge and exploit semantic descriptions of the process in order to make the predictions even more precise, compared to existing approaches. As an additional benefit of this work, the advantage of semantics is revealed. In the following we provide details on our approach. First we state our basic assumptions, as well as the modeling approach for the workflows. After the workflows are modeled, we explain the enrichment with additional information, including background information from experts. In the end we explain the used methods to predict the duration of an activity and describe the evaluation metrics, that are used to compare the results. There, we are keen to make our approach as comparable as possible and reuse well-known data science methods, as well as using leading evaluation metrics for comparing the results. We depict our approach in figure 1.

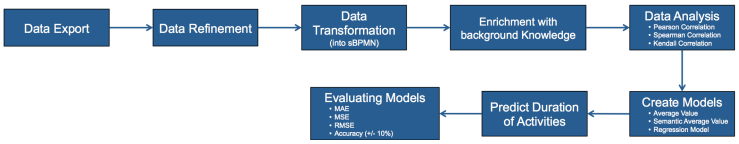

Fig. 1. Performed approach in this work to predict the duration of activities.

We assume a given set of process instances, stored in log files. This assumption is depicted as Data Export in figure 1. Usually the data of a workflow is recorded in an ERP system, which allows to export it in a flat file. Due to incorrect and contradictory entries in the datasets, we also assume a data refinement step. Within this step, we delete false entries or replace them with the correct entry, if the entry can be computed. For analyzing and enriching the process instances and each of the corresponding activities with background knowledge, we first have to model the processes in an appropriate way to have it available in a structured format, which can be queried.

A key fact in Knowledge Management is sharing and creating information collaboratively. Various users might bring different kinds of background knowledge. Thus, we rely on a collaborative design and usage of our methods. Therefore, we apply a collaborative system that allows for a controlled environment and for tracing the provenance of information. Semantic Media Wiki[1] (SMW) [28] serves as collaborative platform to capture and store annotations on documents. We have already tackled the capturing of processes in a collaborative environment such as Semantic MediaWiki as part of previous work [31, 32]. This work builds upon this previous results. We used a BPMN Ontology from the Data & Knowledge Management (DKM) research unit [23] to structure processes. This

[1] `https://semantic-mediawiki.org`, accessed: 2017-07-27

ontology has a very detailed formalization in OWL 2 DL of the BPMN 2.0 specification. Further ontologies that can be applied, in order to describe the data in more detail, are foaf[2], Dublin Core[3] and SKOS[4]. These ontologies are used to structure and describe the process data in detail. Background knowledge from experts can now be applied to have their knowledge available in a structure and reusable format. The captured knowledge can be used for analysis, as well as for predicting the duration of an activity.

In addition to background knowledge from experts, meta-information about the workflow of a process is added to the activities. The captured meta-information is linked to the activities, which are influenced by or influence the information. For instance, if the information about the death of a patient during a surgical process is available, it is attached to the activity that mostly causes this circumstance. The advantage of this approach is to have the information in a structured and standardized format that allows for querying and using it in analysis. Other meta-information includes the persons involved in the task, the runtime of an activity, and several parameters about the patient like if he is a smoker and the number of stays in the hospital. The background knowledge and meta-information are used for correlation analysis to identify and quantify the influence of variables on the process [33]. We used, depending on the level of measurements of the variables, different correlation analysis (Pearson Correlation [21], Spearman Correlation [26], Kendall Correlation [12] and Chi-squared test [22]). We build on these results in our models for predicting the duration of an activity.

The main contribution of this work is the prediction of the duration of an activity. Therefore, we use various approaches, among others one which exploits the semantics in order to show the advantage of semantic information and a well-considered Data Science approach. We are keen to present our methods in a comparable way and reuse well-known methods and evaluation metrics. Therefore, we apply methods on the process instances which do and do not exploit the semantics. Previous methods [2] use the average value (and thus the expected value) of the duration of an activity as prediction. For comparison purposes we apply this approach in our scenario as well and call it in the following *Average Value Approach*. However, this approach is very basic. In our opinion, a process is more complex and has multiple key factors that influence the duration of an activity. Therefore, we analyzed process data according to certain dependencies. Thus we detected coherences between certain variables. The analysis exploits semantic background knowledge [33]. Based on the correlations, we can use the knowledge to select variables that influence the duration of an activity. Instead of using the average of all duration of an activity, as estimator, we shrink the set of activities to compute the average duration, based on similar variables which

[2] `http://www.foaf-project.org/`, accessed: 2017-07-27
[3] `http://dublincore.org`, accessed: 2017-07-27
[4] `https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html`, accessed: 2017-07-27

influence the duration. We call this approach in the following *Semantic Average Value Approach*.

Besides these two approaches, we also use a regression model. Usually are regression models used to quantify the coherence between values [10], however the model can also be used to predict the outcome, based on input variables. Thus we learn a regression model, based on the training data, and try to predict the test data, based on the input parameter x. However, as a drawback to the previous approaches, the regression models usually requires a lot of available data. Otherwise, the detected coherence between variables are not statistically verified. We call this last approach *Regression Model Approach*.

Same as the used methods, we use existing and well-known metrics to evaluate the predictions of the duration of activities. Well-known metrics to quantify the quality of predictions are *Mean Absolute Error (MAE)*, *Mean Squared Error (MSE)* and *Root Mean Squared Error (RMSE)*. Especially the last one is commonly used as metric in forecasting models to quantify the result of predictions. RMSE computes the standard deviation between the actually observed value and the predicted value. MSE and RMSE are very similar, however, RMSE punishes outlier more or rather is more sensitive to outliers. Another metric that we use in our evaluation is the accuracy. Therefore, we compute the number of correct predictions divided by the total number of predictions. However, due to a ratio scale of the duration of activities, we allow a deviation of +/- 10%.

An overview of the metrics and their formula is given in table 1.

Name	**Formula**
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - x_i\rvert$
Mean Squared Error (MSE)	$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$
Root Mean Squared Error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}$

Table 1. Metrics for evaluating predictions.

For evaluating our model, we use k-fold cross-validation [27]. This validation technique is commonly used in statistical analysis. It splits the available data in k folds and trains for each fold one estimator and evaluates the estimators on the other $k-1$ folds. This allows for testing the estimators on an independent data set. k-fold cross-validation is exemplary depicted in figure 2 with $k = 4$.

For each estimator the metrics given in table 1 are computed. In addition, we compute the standard deviation of the RMSE. A value close to zero of the standard deviation of the RMSE indicates a stable estimator over each independent fold. Thus, an equally distribution of the duration can be assumed. The advantage of the cross-validation is the usage of each observation for both, training and validation, and each observation is used for training exactly once.

By using well-known metrics and validation techniques, we are able to compare our results with other approaches and allow for an enhanced comprehensibility of the results. The approach, as well as the used metrics, abstract from

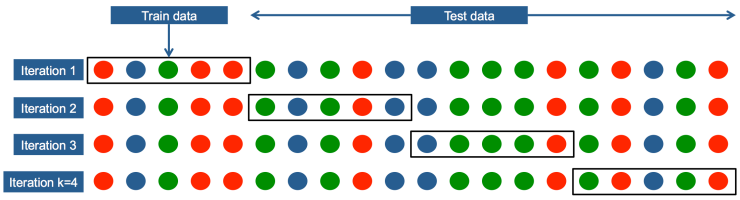

Fig. 2. k-fold cross-validation example by using $k = 4$.

the use-case scenario and the domain. Thus, the approach can be applied to any domain. In the next section we implement our approach in a health-care use-case scenario and evaluate the different approaches. We compare the results by using the above metrics.

3 Experiments

We evaluate our approach by using real-life data from the University Hospital Heidelberg. The process we considered is a perioperative process which describes the workflow of preparing the operating room, bringing a patient into the operating room, the incision and suture of the surgery, and bringing the patient out of the operating room. The considered process is depicted in figure 3 by using BPMN. We only had rectum resection surgeries records available, so we did not look at other performed surgeries. Therefore, we did not include the type of surgery, because it is the same for all considered data sets.

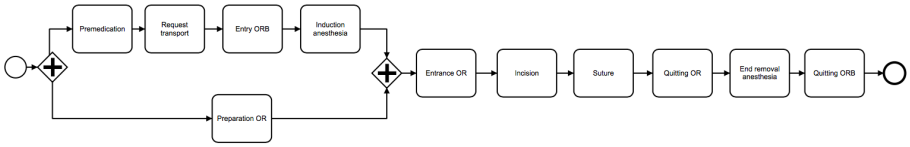

Fig. 3. Considered perioperative process in the evaluation.

The data we received were stored in spreadsheet format. It contained information about the timestamps of every activity, basic information about the patient, like e.g. the age, height and weight, as well as information about former diseases and progress of the surgery. In total, 65 attributes were available. Not all attributes were for each data set available. Some were missing or contained false entries, like e.g. a negative loss of blood. Therefore, we first refined the data. The refinement of the data was done manually. We calculated the duration of each activity according to the timestamps. However, some duration were negative, which we ignored in our analysis, because negative durations are not

possible. We had in total 1,690 data sets available so that the refinement of the data was a huge workload.

In the following, we consider three approaches: The first approach is the basic approach by using the average value of the historic data as prediction of the duration for an activity [2] (*Average Value Approach*). The second approach is by using our semantic analysis in combination with the existing approach to show the advantage of semantics (*Semantic Average Value Approach*). We performed the correlation analysis on the semantically enriched data in a previous paper [33], so that we can build upon those results in order to use them for predicting the duration of an activity. The third approach is by using regression analysis to predict the duration of an activity (*Regression Model Approach*). We use $k = 10$ as cross-validation, because it is commonly used and size of the folds are appropriate. We use MAE, MSE and RMSE as evaluation metrics of each fold and compute the overall accuracy and the deviation of RMSE to evaluate the results.

Average Value Approach: We split the data set in 10 uniform folds and computed on each the average value as predictive value for the other folds. We computed for each fold the MAE, MSE and RMSE. Afterwards, we computed for the RMSE the standard deviation to show how the value vary. A small deviation would show that the root man square error is on every fold uniformly. We computed for each activity the duration. Thus, in total, we performed our approach on 11 activities. The standard deviation of the RMSE for the duration of surgery is 15.2793. This shows that the predicted value of each fold is close to each other and does not vary. An average accuracy of 19.2683% is accomplished by using this method. Similar underperforming results are achieved by predicting other durations like e.g. the preparation of the operating room (Accuracy: 13.4146%) and the operating room time (Accuracy: 22.0122%), which is the time between incision and suture.

Semantic Average Value Approach: There were 65 attributes available. However, due to incomplete and inconsistent data, we could not use every data point in our analysis [33]. Due to the performed correlation tests, we analyzed the data and found out which attributes correlate to each other, among others to the duration of activities. The facts if a patient has a relapse, the amount of red cell concentrate, received Fresh frozen plasma, and if a patient received an intraoperative radiotherapy influence the duration of a surgery. We focused on meta-information, which was known for an activity in advance. We calculated the average value of similar initial positions of the considered process, according to the correlated attributes, and used this as predictive value. The standard deviation of the RMSE for the duration of surgery is 16.3042. Thus it is a bit higher than the RMSE of the previous approach. However, this is lead by the adapted predictions due to the correlated attributes. The average accuracy of predicting the duration of a surgery is 24.6104% and thus better than the previous approach, were no semantics and the correlations were considered. Predicting the

other duration were better than the previous approach, too. The preparation of the operating room achieved an overall accuracy of 19.5092% and the operating room time had an accuracy of 25.6173%. This clearly shows that our approach performs better than the basic procedure in all the three presented predictions of the duration of an activity. We also achieved improved results in the other nine cases of duration prediction. Nevertheless, we want to point out, that the results of $20\% - 25\%$ is still not satisfying and needs to be improved.

Regression Model Approach: We compared our method with a regression analysis as last evaluation part. Regression Analysis is used to predict the duration of an activity in the past (see [6]). Therefore, for comparison reasons, we applied a regression analysis, too. We used a Polynomial Regression Analysis [14]. The best results were achieved by using a degree of 6 and using the age of the patient as input parameter for the duration of a surgery. The standard deviation of the RMSE over all folds is 7.0497. This clearly points out the stability of the estimators. However, the results are not as good as the previous approach, in which we exploited the semantics. The accuracy of the *Regression Model Approach* for predicting the operating time is 18.7662%. Similar underperforming results are achieved by predicting other duration like e.g. the Preparation of the operating room (Accuracy: 14.1830%). Likewise the previous results, the operating room time had the highest accuracy of the three considered ones of 23.0263%. The coefficient of determination of the regression analysis is $R^2 = 0.0168$, which shows that this model cannot explain the data sufficiently to predict values. Table 2 summarizes the results of the overall accuracy of the evaluated approaches and shows that exploiting meta-information and semantics achieves best results. Exploiting these information, if available, and using it in data science methods is essential for achieving better results.

Approach	**Operating Time**	**Preparation of the Operating Room**	**Operating Room Time**
Average Value Approach	19.2683%	13.4146%	22.0122%
Semantic Average Value Approach	24.6104%	19.5092%	25.6173%
Regression Model Approach	18.7662%	14.1830%	23.0263%

Table 2. Overview of Overall Accuracy of the different approaches.

4 Related Work

Our approach is addressed by roughly three kinds of work: 1) Match BPMN process to sBPMN, 2) annotating business processes with meta-information and 4) using Data Science methods for predicting the activity duration.

BPMN [20] is a de-facto standard for representing processes in a very expressive graphical. BPMN defines semantics partly, which means that the symbols

have meanings. However, the semantics just have little weights and there is no much attention paid to the formal definitions of the symbols. One advantage of BPMN is its executability [5]. Since 2011, it's current version is BPMN 2.0 [20]. So far, existing work already addressed the transformation of BPMN into other languages like e.g. BPEL [7] and Petri-Net [16]. sBPMN (Semantic BPMN) was developed to allow for tackling the disadvantage of the lack of formal definitions and provide an unambiguous and consistent semantics [13]. sBPMN extends BPMN elements with additional information and background knowledge to enhance analysis [**?**].

The second aspect that is tackled in this paper is the annotation of business processes with meta-information. This issue had been addressed among other by us in previous works by stating out a meta-model for processes [34], as well as approaches for annotating decision trees and processes with meta-information [31]. Existing works has also pointed out the advantage of using meta-information in process mining [24]. A published survey summarizes existing approaches for business process annotations [15]. One aspect, considered in the survey, was the possibility to capture semantic annotations. The meta-information in processes are used for reasoning purposes, which is used to support analytics and optimization of processes [19].

The last addressed issue in our work is using Data Science methods for predicting activity duration. Previous approaches has mostly focused on control-flow discovery [30, 9]. However, when event logs contain time information, the discovered models can be extended with timing information. Most related to our work is the time prediction by van der Aalst [2]. Van der Aalst took the expected value in order to estimate the time of a future activity [2]. They only took the structure of the workflow into account but no other meta-information. However, there are factors, other than the structure of the workflow, which influence the duration of an activity. In health-care these factors might be the age of a patient, the condition and drugs that had been used by a patient. Therefore, this work is not sufficient for a more precise prediction. Another related work is the estimation of workflow execution time [4]. In this work, a more advanced method to estimated the execution time of workflows were used. The execution time of workflows was computed based on stochastic estimates of tasks' execution time. For calculating the tasks' execution time, they split up its execution time in multiple variables like e.g. resource preparation time, queuing time and data transfer time. We do not go into this detail, because we even do not have this detail of information in our time logs. For calculating the executions time, they performed, like we did, statistical hypothesis test for checking dependent variables. Based on the results of statistical tests, they proposed a combination of Chebyshev-like distribution-free inequalities and distribution-based approaches to computer a tasks' runtime.

5 Conclusions

We showed an approach of exploiting semantic information in process data. Therefore, we first captured the process data and refined in the case of inconsistent and incomplete data. Afterwards we transformed the data into sBPMN by using a BPMN Ontology from the Data & Knowledge Management (DKM) research unit [23]. This enabled us to enrich it with further information in a structured way and to use the semantic information about the process. We used the results from the enrichenment with meta-informatio in our prediction models and compared it to existing standards. We used an existing approach [2], based on the average value of historic data as estimator. We compared this approach with ours, which exploits the semantics, by computing the average value of similar data sets, according to performed correlation tests, from historic data as estimator. Thus, we did not used all attributes for determining similar data sets, but selected ones, which were determined by using correlation tests. As a last approach, we used a polynomial regression analysis to predict the duration of an activity.

For evaluating the three approaches, we used k-fold cross-validation, with $k = 10$. Therefore, every data set was used as test and training data. Besides the fact that every data set is used as test and training data, we used it because it provides more accurate results of the models [25]. We computed for every fold the MAE, MSE and RMSE and provided the number of hits by allowing a deviation of +/- 10%. These well-known metrics are usually used in Data Science to indicate the results of a prediction model and allows for comparing the results. We showed that our approach performs best, compared to the other ones and achieved a better overall accuracy. In this work we focused on three out of 11 activity duration due to a better clarity. However, similarly to the presented results, the *Semantic Average Value Approach* performed best for predicting the other activity duration. The standard deviation of the RMSE over all folds were in the *Semantic Average Value Approach* higher, compared to the other ones. However, this is due to the individual predictions, based on the correlated attributes.

Even if our approach outperforms the other two approaches, there is still a lot of work to be done in this area. Computing the duration of an activity is still a challenging task and the results also shows that, although we improved the accuracy, they are still not satisfying. Therefore, we have to improve our model. One way we would like to improve it is to include more attributes and data sets. Due to the increasing amount of data that is nowadays stored, we see this as a chance to collect more data and use it in our analysis and prediction models. Another future work that we would like to tackle is to transfer our approach to other domains and show its applicability. We built the approach in such a way that it abstracts from the underlying data and domain. Therefore, we expect that the approach is also usable in other domains and not limited to a certain scenario.

In conclusion, we have taken a first step towards exploiting semantic information of processes, stored according to the Linked Data principles for predicting

the duration of activities. We showed that our approach outperforms existing ones. The semantics in the data can be used in order to improve existing methods.

References

1. van der Aalst, W.M.P., Reijers, H.A., Weijters, A.J.M.M., van Dongen, B.F., Alves de Medeiros, A.K., Song, M., Verbeek, H.M.W.: Business process mining: An industrial application. Inf. Syst. 32(5), 713–732 (Jul 2007)
2. van der Aalst, W., Schonenberg, M., Song, M.: Time prediction based on process mining. Information Systems 36(2), 450 – 475 (2011)
3. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. Semantic services, interoperability and web applications: emerging concepts pp. 205–227 (2009)
4. Chirkin, A.M., Kovalchuk, S.V.: Towards better workflow execution time estimation. IERI Procedia 10, 216 – 223 (2014)
5. Dijkman, R., Van Gorp, P.: BPMN 2.0 Execution Semantics Formalized as Graph Rewrite Rules, pp. 16–30. Springer Berlin (2010)
6. Dongen, B.F., Crooy, R.A., Aalst, W.M.: Cycle time prediction: When will this case finally be finished? In: Proceedings of the OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008. pp. 319–336. Springer-Verlag (2008)
7. Doux, G., Jouault, F., Bzivin, J.: Transforming bpmn process models to bpel process definitions with atl. In: In GraBaTs 2009 : 5th International Workshop on Graph- Based Tools (2009)
8. Ehrig, M., Koschmider, A., Oberweis, A.: Measuring similarity between semantic business process models. In: Proceedings of the Fourth Asia-Pacific Conference on Comceptual Modelling - Volume 67. pp. 71–80. Australian Computer Society, Inc. (2007)
9. Günther, C.W., Van Der Aalst, W.M.P.: Fuzzy mining: Adaptive process simplification based on multi-perspective metrics. In: Proceedings of the 5th International Conference on Business Process Management. pp. 328–343. Springer-Verlag (2007)
10. Hosmer, D.W., Lemesbow, S.: Goodness of fit tests for the multiple logistic regression model. Communications in Statistics - Theory and Methods 9(10), 1043–1069 (1980)
11. Huang, S.M., Yen, D.C., Hung, Y.C., Zhou, Y.J., Hua, J.S.: A business process gap detecting mechanism between information system process flow and internal control flow. Decision Support Systems 47(4), 436 – 454 (2009)
12. Kendall, M.G.: A new measure of rank correlation. Biometrika 30(1/2), 81–93 (1938)
13. Kossak, F., Illibauer, C., Geist, V., Kubovy, J., Natschläger, C., Ziebermayr, T., Kopetzky, T., Freudenthaler, B., Schewe, K.D.: A rigorous semantics for bpmn 2.0 process diagrams. In: A Rigorous Semantics for BPMN 2.0 Process Diagrams, pp. 29–152. Springer (2014)
14. Lai, T., Robbins, H., Wei, C.: Strong consistency of least squares estimates in multiple regression ii. Journal of Multivariate Analysis 9(3), 343 – 361 (1979)
15. Lautenbacher, F., Bauer, B.: A survey on workflow annotation & composition approaches. In: Proceedings of the Workshop on Semantic Business Process and Product Lifecycle Management (SemBPM) in the context of the European Semantic Web Conference (ESWC). pp. 12–23 (2007)

16. Lohmann, N., Verbeek, E., Dijkman, R.: PetriNet Transformations for BusinessProcesses – ASurvey, pp. 46–63. Springer Berlin Heidelberg (2009)
17. Mans, R.S., van der Aalst, W.M.P., Vanwersch, R.J.B.: Process Mining, pp. 17–26. Springer International Publishing (2015)
18. Mans, R.S., Schonenberg, M., Song, M., Aalst, W., Bakker, P.J.: Application of process mining in healthcare–a case study in a dutch hospital. Biomedical engineering systems and technologies pp. 425–438 (2009)
19. Niedermann, F., Radeschütz, S., Mitschang, B.: Business Process Optimization Using Formalized Optimization Patterns, pp. 123–135 (2011)
20. OMG: Business process model and notation (bpmn), version 2.0 (January 2011)
21. Pearson, K.: Note on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London 58(347-352), 240–242 (1895)
22. Pearson, K.: On the criterion that a given system of derivations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine 50(302), 157–175 (1900)
23. Rospocher, M., Ghidini, C., Serafini, L.: An ontology for the business process modelling notation formal ontology. In: Information Systems – Proceedings of the Eighth International Conference. pp. 133–146. IOS PRess BV (Sep 2014)
24. Saylam, R., Sahingoz, O.K.: Process mining in business process management: Concepts and challenges. In: 2013 International Conference on Electronics, Computer and Computation (ICECCO). pp. 131–134 (Nov 2013)
25. Seni, G., Elder, J.: Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Morgan and Claypool Publishers (2010)
26. Spearman, C.: The proof and measurement of association between two things. The American Journal of Psychology 15(1), 72–101 (1904)
27. Stone, M.: Cross-validatory choice and assessment of statistical predictions. Journal of the royal statistical society. Series B (Methodological) pp. 111–147 (1974)
28. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H., Studer, R.: Semantic wikipedia. In: Proceedings of the 15th International Conference on World Wide Web. pp. 585–594. ACM (2006)
29. Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S.: Constrained k-means clustering with background knowledge. In: Proceedings of the Eighteenth International Conference on Machine Learning. pp. 577–584. Morgan Kaufmann Publishers Inc. (2001)
30. Weijters, A.J.M.M., van der Aalst, W.M.P.: Rediscovering workflow models from event-based data using little thumb. Integr. Comput.-Aided Eng. 10(2), 151–162 (Apr 2003)
31. Weller, T., Maleshkova, M.: Capturing and annotating processes using a collaborative platform. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 283–284. International World Wide Web Conferences Steering Committee (2016)
32. Weller, T., Maleshkova, M.: Towards a collaborative process platform: Publishing processes according to the linked data principles. In: Proceedings of the Workshop on Linked Data on the Web, LDOW 2016 (2016)
33. Weller, T., Maleshkova, M., Wagner, M., Ternes, L.M., Kenngott, H.: Analysis of semantically enriched process data for identifying process-biomarkers. In: Proceedings INTELLI. p. 6. IARIA XPS Press (November 2016)
34. Weller, T., Maleshkovaa, M.: Towards a process meta-model. In: Proceedings of the second Karlsruhe Service Summit Workshop-Advances in Service Research, Karlsruhe, Germany (2016)