Detection of Dataflow Anomalies in Business Process An Overview of Modeling Approaches

Najat Chadli
Mohammadia Engineering School
Mohammed V University, Rabat, Morocco
najatchadli@research.emi.ac.ma

Mohamed Issam Kabbaj
Mohammadia Engineering School
Mohammed V University, Rabat, Morocco
kabbaj@emi.ac.ma

Zohra Bakkoury
Mohammadia Engineering School
Mohammed V University, Rabat, Morocco
bakkoury@emi.ac.ma

ABSTRACT

Most research focus on control flow modeling when modeling and analyzing business process models, but less attention is paid to data flow. However, data flow and control flow are both essential in process modeling. Thus, the data flow modeling and verification have a great importance in detecting anomalies. In this study, some recent approaches for anomaly detection has reviewed. The first is an analytical approach for detecting and eliminating three types of data-flow errors that formally establish the correctness criteria for data-flow modeling. The second formulates the data-flow modeling and verification using a Petri Net based approach. The third one presents an ad hoc approach to detect data modelling errors in business process models by applying for an active help using a DataRecord concept. We explain for each approach its proper method and tools. We then compare and analyze each one of them to discover the added-value of each approach.

CCS CONCEPTS

• **Applied computing** → **Business process management**; *Business process model*

KEYWORDS

Data-flow modeling; Verification; Data-validation; Data anomalies.

ACM Reference format:

N.Chadli, M.I.Kabbaj and Z.Bakkoury.2018. Detection of Dataflow Anomalies in Business Process An Overview of Modeling Approaches. In *12th International Conference on Intelligent Systems: Theories and Applications, Rabat, Morocco, October 2018 (SITA'18)*

DOI:10.1145/3289402.3289537

SITA'18, October 24–25, 2018, Rabat, Morocco

ACM ISBN 978-1-4503-6462-1/18/10…$15.00

DOI:10.1145/3289402.3289537

1 INTRODUCTION

A business process consists of a set of activities structured to collaboratively achieve common business goals using a clearly identified input, output, a beginning and an end of a process. Therefore, the business process has considered as a specific ordering of work activities across time and place [1-2]. So, in Business Process Management, the standard paradigm for process modeling is the workflow concept. However, the workflow specification requires to characterize the other perspectives, including time, data, etc... [3]. Consequently, for workflow modeling, analysis and verification of the time management is an important aspect [3-5]. Similarly, for data-flow modeling, data and routing choices is also important in a process that is typically determined by certain data items [2]. While the business process becomes increasingly complex, the workflow technology constitutes a standard solution to manage it. In workflow design, most efforts are dedicated to control-flow to identify errors. As the workflow system gets more complicated, data flow is becoming more and more prominent in modeling, as for verification. Indeed, the activity sequencing has controlled by some of the operational constraints determined by the relations among data elements. Consequently, data flow perspective has an important role in workflow management [4]. The consumed and produced data, with respect to each activity in a business process, is defined by the "information perspective", i.e. How technological changes interact with individual behaviors, organizations and society to affect the availability and use of information in governance processes [4]. Therefore, to detect the errors of data, it is being necessary to include data flow analysis with control flow into a structured workflow. In this concept, several approaches have been proposed for detecting the errors of data flow modeling in a workflow analysis. To model and analyze the data requirements in workflow systems, some informal and formal modeling tools have been developed [6]. In business process modeling literature, the data flow modeling anomalies are tackled by different approaches. Although their methods to analyze the issues are different from one another, these methods are verified by rules and lemmas and theorems of the correctness criteria and an algorithmic verification. Indeed, there are many error types in data flow modeling. The three basic types are missing data, conflicting data and redundant data. Consequently, the choice of these three types of data flow

anomalies seems to be sufficient to analyze the data flow requirements at the conceptual level [2-3-4]. Indeed, other error type might be represented as a combination of these basic types of data flow errors, e.g. "Mismatched data arise when the structure of an output data item is incompatible with the structure required by the activity that uses the data item as input. Mismatched data can be regarded as the occurrence of both redundant data and missing data".

The main motivation of this work is to discuss several approaches using the error types, previously mentioned, to extract advantages of each approach. This will allow later to build an effective new approach based on the hybridization of the discussed methods. The first method [4] used in this study, formally establishes the correctness criteria besides data-flow specification by proposing an additional component for data flow analysis. The second one is a Petri Net based approach using the Petri-net incidence matrix with a polynomial complexity algorithm to formulate the Data-Flow Modeling and Verification [3]. The last one [2] is an ad-hoc approach using an active help method for detecting data modelling errors in workflow and a “DataRecord” concept to manage datasets and activities. The purpose of this paper is to perform a demonstrative description of modeling methods and used tools to analyze and detect the above basic error types. Afterwards, we aim to compare the different strong points and weaknesses of the three approaches to provide new insights and study the added value of each approach.

This paper is organized as follows. Section 2 presents data flow modeling and detection of anomalies in business process and make a critical analysis of each approach. In Section 3, A general discussion and comparisons are reported. Conclusion and perspectives are given Section 4.

2 DETECTION OF DATAFLOW ANOMALIES IN BUSINESS PROCESS

2.1 Introduction

2.1.1 Data Flow modeling. In business process, Modeling involves a progression from a conceptual model to a logical model then to physical schema. As stated before, Data flow modeling is one of the basics of the Structured System Analysis and Design Methods. Data-flow modeling verification concentrates on identifying data-flow errors by means of a set of well-defined correctness criteria. Consequently, verification aims to define the problems caused by incorrect data flow modeling.

2.2.2 The anomalies of data flow modeling. The basic data flow anomalies are missing data, conflicting data and redundant data [7]. However, there are several anomalies of data flow modeling which can be viewed as a combination of these basic types such as Mismatched data, Inconsistent data, Misdirected data or Insufficient data [4].

Missing Data. when data has never been created before or accessed without being initialized during the modeling process, in this case a missing data error occurs.

Conflicting Data. Conflicting data occurs if some data elements are written by an activity, however, activities cannot confirm an update due to the existence of several versions of the same data elements, which causes a conflict of which version should be updated.

Redundant Data. If a data element is written by an activity, but has never been read in all possible continuations, then, this data element is a redundant data.

2.2 Anomalies Detection Approaches

2.2.1 Formulating the data flow perspective for business process modeling. This approach provides a data-flow framework for detecting data-flow anomalies. This framework includes two basic components: data flow specification and data flow analysis. With these components, the business process management has a more analytical stiffness, and an interesting aspect towards a formal methodology for data flow modeling. A data-flow verification, which is a theoretical foundation criterion, is also proposed to eliminate systematic and automatic data-flow anomalies [4].

Explanation of the approach. This approach can be tested in real-world applications; it needs a prototype data flow modeling to develop the formal manner of correcting data flow anomalies. To reform these errors, this approach requires to modify not only the data flow but also some cases of the control flow. Consequently, a new workflow conception based on the data flow analysis is elaborated containing a new operational perspective that specifies methods used in the workflow system such as data flow operations and data flow matrix. So, they have illustrated the concepts by introducing a property loan approval process shown as a UML activity diagram. the correctness criteria for data flow modeling is formally established as a theoretical foundation for the data flow verification. These criteria enable systematic and automatic elimination of data flow errors as in the Figure 1 below [4].

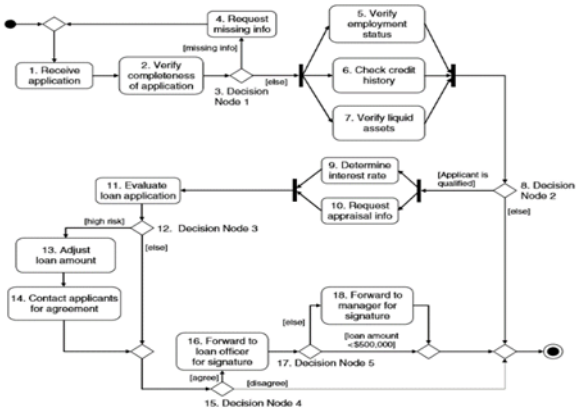

Figure 1:Property Loan Approval Process

Data flow operations. Operations called data flow operations performed in the activity need a data item. Additionally, data items are produced, accessed and modified by means of these operations. Data-flow operations are classified into six operations notably initializing, approving, updating, referring and verifying according to the semantic meanings of data flow operations in the business process [4].

Data-flow matrices. The concept of the data-flow matrix is a two-dimensional table specifying data flow in workflow applications. Indeed, the matrix records the data-flow operations each activity performs on data items in a workflow. Moreover, the decision nodes must be included in the data-flow matrix as activities since they require input data. Finally, with such dataflow matrices, it is easier to find out how each data item is processed in the workflow [4].

Integration of dataflow in a workflow model. Input and output for an activity can be described by using the object flows in a UML activity diagram. Indeed, to indicate the relationship action-object, each object is connected to one or more activities. In workflow management, the data flow modeling by an object flow is insufficient since that, it offers no details on how different data items associated with one object are processed differently. Hence, the necessity of UML activity diagram [4].

2.2.2 Formulating the Data-Flow Modeling and Verification for Workflow. Developing a systemic formal methodology for characterizing the data-flow modeling and verification requires to formulate the data-flow modeling and verification [3].

Explanation of the approach. The Petri-net [8-9] based approach is proposed to detect data flow modeling anomalies in a workflow-net which can only be represented as the logical relation of workflow that is the control flow. Indeed, extending each activity with its input and output data sets in workflow-net to model each element of data flow is called WFIO-net. Furthermore, the approach comes with the necessity of using the polynomial complexity algorithm that exploits data-incidence matrix to verify the anomalies [3].

Firing rules for classic Petri nets. The firing rules for classic Petri-nets have some basic notions that need to explain. Let N is an arbitrary Petri-net with a set of places P, a set of transitions T and a flow relation F. All places which are connected to a transition by an arc form the set of pre-places and post-places of the specific transition [9]. Accordingly, the rules describing possible changes from one marking to the next are called firing rules. [10].

From WF-Net to WFIO-Net. "Modeling a workflow process definition in terms of a Petri net is rather straightforward, i.e. tasks are modeled by transitions, conditions are modeled by places, and cases are modeled by tokens". A Petri-net which models a workflow process is called a workflow net (WF-net), defined as follows [11-12]:

A Petri net $PN, PN = (P, T; F)$ is a WF-net if and only if

1) PN has two special places: i *and* o, place i is a source place and o is a sink place: $\cdot i = 0$ *place o is a sink place*: $o \cdot = 0$.

2) A transition $t*$ is added to PN which connects place o with place i,then the resulting Petri net is firmly attached. The workflow modeled by WF-net is a sub-set of a Petri Net which is used to characterize formal languages [9-12]. Indeed, the WFIO-net signifies the workflow input/output net that is prospective for modeling the data flow elements of a workflow that is a WF-net while extending each activity with its input and output data sets. As in [3], the figure 2 and figure 3 represent a model by WF-net and WFIO-net.

TA is an activity transition set.

PL is logic place set.

TL is a logic transition set.

Source place ps.

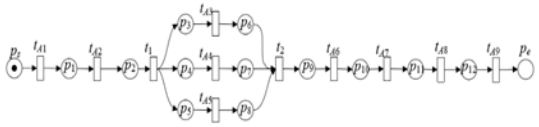

Figure 2:WF-net of the Property Loan Approval BP.

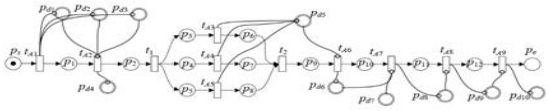

Figure 3:WFIO-net of the Property Loan Approval BP.

2.2.3 Towards an active help on detecting data-flow errors in business process models. In this study, an ad-hoc approach intends to discover the data flow modeling anomalies proposed by the modeler. Indeed, to achieve the issues of data analysis, an active help apply to verify permanently for each fragment in the model when the system is locked. Although, used a concept DataRecord to store the last set of each data and the activities that have read, update or destroy this data [2].

Explanation of the approach. The business process modeler first designs a model and verify it to correct the control or data flow errors. However, repairing these errors doesn't signify that the model is correct. In this case, the revalidation of the entire model is essential to ensure that there is no error that causes a loss of time and cost. Therefore, this technique applies an active help tool for real-time analysis to anticipate the error. Whenever the fragment has an error outbreak, the process of verification is then triggered. This gives us for each time an error-free independent fragment of the model. Consequently, the verification process is done at the modeling time. Additionally, a concept "DataRecord" introduced is an $(n \times m)$ matrix where n is the number of data and m is the number of *Xor* branches in the model. The latest set of data in the model and the activities that have read, update or destroy this data is stored in this DataRecord. Indeed, DataRecord is initially empty, as and when the business process model draws, data is inserted according to the rules of data flow anomalies notably missing data, conflicting data and redundant data errors. Consequently, the various data-items in the workflow are incrementally recorded in the matrix by passing from an activity to the other sequentially [2].

Example of Model with a XOR split.

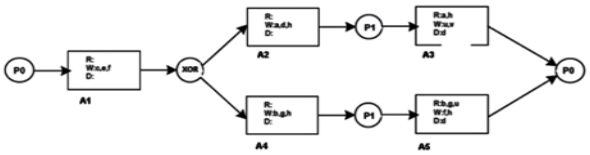

Figure 4:Model with an XOR split.

For the above example in figure 4, the approach results are reported in table 1. The column "state" of DataRecord reflects the latest state of each data item while the next columns represent the last state of each *Xor* branch. Indeed, by applying "Rule 1", data item u is detected as a missing data in activity *A5* since this data was created only in the first *Xor* branch of activity *A3*, but not in the second *Xor* branch. After locking the system to correct the errors, the modeler chooses the state to write or destroy the element u. In this case, the modeler decides to destroy u. Similarly, "Rule 2" is applied to detect elements f and h as conflicting data, and applying "Rule 4" also detects redundant data elements v, c and e.

Table 1: The Last State of DataRecord of the model with an XOR split.

Data	State	State branche1	State branche2
c	$(0, W_{A1}, 0)$		
e	$(0, W_{A1}, 0)$		
f	$(0, W_{A1}, 0)$		$XOR2(0, W_{A5}, 0)$
a	$(0, W_{A2}, 0)$	$XOR1(R_{A3}, W_{A2}, 0)$	
d	$(0, W_{A2}, 0)$	$XOR1(0, W_{A2}, D_{A4})$	$XOR2(0, W_{A2}, D_{A5})$
h		$XOR1(R_{A3}, W_{A2}, 0)$	
u		$XOR1(0, W_{A3}, 0)$	$XOR2(0, W_{A5}, 0)$
v		$XOR1(0, W_{A3}, 0)$	
b			$XOR2(R_{A5}, W_{A4}, 0)$
g			$XOR2(R_{A5}, W_{A4}, 0)$

3 GENERAL DISCUSSION OF THE THREE APPROACHES

In this section, we aim to explain and compare these approaches by analyzing the similar and the opposed ideas in each approach.

3.1 Case Study of the Three Approaches

Following our study, we choose to select some cases of each approach that authors analysed. Three matrices represented in the tables below are constructed: Data-flow-matrix table 2, 3, Activity-Data Incidence-matrix table 4,5 and DataRecord-matrix table 6. Indeed, these approaches use $(n \times m)$ matrices (n rows, m columns). In the first approach, n is a number of data and m is a number of activities [4]. In the second approach, n is a number of transition(activities) and m is a number of data places [3]. In the last approach, n is the number of data item and m the number of possible XOR branch's [2]. For each element (i, j) such as $i, j \in \mathbb{N}$ in a matrix is a state indicator either the data has been read (r)or written(w). Then, a missing data error check is applied to the three approaches.

3.1.1 Approach of formulating the data flow perspective for business process modeling.

Table 1:Symbols Used in the Property Loan Approval Process

Data items		
d1 Applicant name	d8 Account balance	d15 Risk
d2 Loan amount	ĺ9 Account balance verified	d16 Amount adjusted
d3 Annual income	d10 Applicant qualified	d17 Agreed by applicants
d4 Application complete	d11 Interest rate	d18 Property insured
d5 Application summary	d12 Property address	d19 Signed by applicants
d6 Employment status verified	d13 Current owner of property	d20 Signed by loan officer
d7 Credit score	d14 Appraised value of property	d21 Signed by manager
Activities		
v1 Receive application	v8 Decision Node 2	v15 Decision Node 4
v2Verify completeness of application	v9 Determine interest rate	v16 Forward to loan officer for signature
v3 Decision Node 1	v10 Request appraisal info	v17 Decision Node 5
v4 Request missing info	v11 Evaluate loan application	v18 Forward to manager for signature
v5 Verify employment status	v12 Decision Node 3	
v6 Check credit history	v13 Adjust loan amount	s Start node
v7 Verify liquid assets	v14 Contact applicants for agreement	e End node
Operation		
r Read		
w Write		

Table 2:Dataflow matrix as in [4]

	v1	v2	v3	v4	v5	v6	v7	v8	v9
d1	w	r			r	r	r		r
d2	w	r			r				r
d3	w	r							
d4		w	r	r					
d5	w								
d_6					w				
d_7						w			r
d_8	w	r							
d_9									
d_{10}					w	w			
d_{11}									w
d_{12}	w	r							
d_{13}									
d_{14}									r

d_{15}									
d_{16}									r
d_{17}									
d_{18}									
d_{19}									
d_{20}									
d_{21}									
	V_{10}	V_{11}	V_{12}	V_{13}	V_{14}	V_{15}	V_{16}	V_{17}	V_{18}
d_1	r	r		r	r		r		r
d_2					r		r	r	r
d_3									
d_4									
d_5									r
d_6		r							
d_7		r							
d_8		r							
d_9		r							
d_{10}		r			r		r		r
d_{11}		r		r			r		r
d_{12}	r								
d_{13}	w								
d_{14}	w	r		r					
d_{15}		w	r	r			r	r	r
d_{16}							r	r	r
d_{17}							r		r
d_{18}									
d_{19}							r		r
d_{20}							w		
d_{21}									w

3.1.2 Approach of Formulating the Data-Flow Modeling and Verification for Workflow.

Table 3:Activity Information of this Property Loan Approval

Activity Name	Meaning	Pre-activities	Write Data	Read Data
A1	Receive application	Ø	$\{D_1, D_2, D_3\}$	Ø
A2	Verify completeness of the application	$\{A1\}$	$\{D_4\}$	$\{D_1, D_2, D_3\}$
A3	Verify employment status	$\{A2\}$	Ø	$\{D_5\}$
A4	Check credit history	$\{A_2\}$	Ø	$\{D_5\}$
A5	Verify liquid asset	$\{A_2\}$	Ø	$\{D_5\}$
A6	Determine interest rate	$\{A_3, A_4, A_5\}$	$\{D_5\}$	$\{D_6\}$
A7	Evaluate loan application	$\{A_6\}$	$\{D_6, D_7\}$	$\{D_8\}$
A8	Contact applicant for agreement	$\{A_7\}$	$\{D_8\}$	$\{D_9\}$
A9	Forward to loan officer for signature	$\{A_8\}$	$\{D_9\}$	$\{D_{10}\}$

Table 4:Activity-Data Incidence Matrix of WFIO-net as in [3]

	P_{d1}	P_{d2}	P_{d3}	P_{d4}	P_{d5}	P_{d6}	P_{d7}	P_{d8}	P_{d9}	P_{d10}
T_{A1}	w	w	w							
T_{A2}	r	r	r	w						
T_{A3}					w					
T_{A4}					w					
T_{A5}					w					
T_{A6}					r	w				
T_{A7}						r	r	w		
T_{A8}								r	w	
T_{A9}									r	w

3.1.3 Approach of Towards an active help on detecting data-flow errors in business process models.

Table 5:DataRecord Matrix for example in figure 4

	A1	A2	A3	A4	A5
c	w				
e	w				
f	w				
a		w			
d		w		d	d
h		w	r		w
u			w		
v			w	w	
b				w	r
g					r

In the first matrix, the errors of missing data are mentioned like d_{18} in V_{14} table 3. In the second matrix, the modeler detects P_{d7} in the transition T_{A7} table 5. Since the approaches apply the passive help, the verification of the process begins until processing the whole model. So, in these matrices, the errors haven't been corrected up till the modeler ended the verification process. Contrariwise, in the third matrix, the data flow anomaly of data item u in activity A_5

table 6 is detected as missing data. So, the modeler chooses that element *u* should not be read in A_5. Certainly, the error doesn't appear in the matrix as the verification is applied on the modelling time using an active help. Consequently, the first two approaches have wasted time to correct the errors in comparison to the third approach.

Table 6:Completeness test result

approaches	Detection Type	Completeness test
Approach1	passive	completely
Approach 2	passive	completely
Approach 3	active	completely

3.2 Similarities of the Three Studies

The authors have pursued the same goals to detect data flow anomalies in business process modeling especially the basic three anomalies: missing data, conflicting data and redundant data. They used keywords e.g; data flow modeling, workflow, data flow verification, data flow anomalies, data flow analysis. They made data-flow analysis more manageable and covered the key issues of ensuring data flow integrity in the workflow at the conceptual level, besides the control flow to address the data flow anomalies. Moreover, workflow systems have become a standard solution for managing complex processes in business domains such as supply chain management. In addition, the successful business process management depends on a workflow management system that is a data-perspective based workflow modeling and analysis.

3.3 The Differences

Each approach uses a matrix that contains data and activities, but there are major differences between their purposes. In the first approach, data flow matrix is used to record the data-flow operations that each activity performs on various data items. The second approach used Activity-Data Incidence Matrix of WFIO-net because this type of matrices shows the control-flow relation, but no data-flow information is reflected whereas the third approach uses DataRecord matrix to record the various data items and last activities in the workflow. In more details, the first approach aims to present a data perspective including two basic components: data-flow specification and data-flow analysis. A simple predicate logic format is used instead of following the event-role-object-condition-action format. The presented lemmas and theorems gave rise to data-flow verification rules. In this study also, algorithms have been developed to be used as a roadmap for the implementation of the data-flow perspective. A data flow matrix and an extension of the unified modeling language (UML) activity diagram is introduced including a special type of control activity nodes.

Second approach is based on the Petri-net system. A WFIO-net extending the WF-net as its basic concepts. The authors used control flow and data flow to formulate the data-flow modeling verification in workflow. They've also proposed for interpreting, to introduce the activity task incidence matrix of a WFIO-net enriched by the rules of data flow anomalies. However, the authors didn't use a time factor in WFIO-net to give a more accurate verification. There is also a lack of detailed taxonomy for each kind of data flow errors. While the abovementioned approaches provide a passive help to the designer as they need to check the correctness of the model at the end of the modelling phase, repairing detected errors doesn't ensure that the result is a correct model, it is obligatory to revalidate the model. The third study, an ad hoc approach uses an active help method and a "DataRecord" concept to store the last state of each data item in the model and the last activity that has performed read, update and destroy operations. This approach applies verification at the modelling time. The authors tested this approach on a simple linear model and an *Xor*-Split model of two branches. However, they didn't apply a loop modelling. In this case, for using the loop, it is necessary to enrich the approach by special looping rules.

4 CONCLUSIONS

In this paper, an overview of detecting anomalies approaches in a workflow management system is provided. The first approach's goal is to formulate the data flow perspective by means of dependency analysis. The data flow matrix and an extension of the UML activity diagram are proposed to specify the data flow in the business process. The second approach had a goal to formulate the data-flow modeling and verification. A Petri-Net based approach applied a polynomial complexity algorithm and the activity-data incidence matrix of the WFIO-net. The third approach aim is to introduce an ad hoc approach for an active help on detecting data modeling errors. They used a "DataRecord" concept to store the last state of each data set in the model and the last activity. This concept is a key feature of the third approach which allowed the designer to anticipate and correct errors at the modelling time.

This overview of the three approaches is the beginning for the upcoming research as we need to use parts of each approach, to resolve the issue when a loop modeling is used in an active help approach.

REFERENCES

[1] TH Davenport. 1993. Process innovation: reengineering work through information technology. - *books.google.com.*

[2] MI Kabbaj, A Bétari, Z Bakkoury, A Rharbi . 2015.Towards an active help on detecting data flow errors in business process models, - IJCSA, researchgate.net.

[3] LIU Cong, Q ZENG, D Hua. 2014.Formulating the data-flow modeling and verification for workflow: A petri net-based approach. *International Journal of Science and, - ijsea.com.*

[4] SX Sun, JL Zhao, JF Nunamaker .2006. Formulating the data-flow perspective for business process management, - Information Systems, - pubsonline.informs.org.

[5] JQ Li, YS Fan, MC Zhou .2004. Performance modeling and analysis of workflow. Cybernetics-Part A: Systems, -ieeexplore.ieee.org.

[6] GKPLS Rausch, SW Retschitzegger. Workflow Management Based on Objects, Rules, and Roles. *Data Engineering - Citeseer.*

[7] S Sun, L Zhao, O Sheng .2004. Data flow modeling and verification in business process management. AMCIS 2004 Proceedings, aisel.aisnet.org.

[8] JL Peterson .1977. Petri nets. ACM Computing Surveys (CSUR), dl.acm.org.

[9] T Murata .1989. Petri nets: properties, analysis and applications. Proceedings of the IEEE, - ieeexplore.ieee.org.

[10] Popova-Zeugmann, Louchka.2013. Time petri nets. In: *Time and Petri nets*. Springer, Berlin, Heidelberg, p. 31-137.

[11] W Reisig .2013. *Understanding petri nets: modeling techniques, analysis methods, case studies*. Springer

[12] WMP Van der Aalst .1998. *The application of Petri nets to workflow management*. Journal of circuits, systems, and computers, World Scientific.