Clustering of ERP Business Process Fragments

Riyanarto Sarno[1], Hari Ginardi[2], Endang Wahyu Pamungkas[3] , Dwi Sunaryono[4]
Informatics Department, Faculty of Information Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
[1]riyanarto@if.its.ac.id, [2]hari@its.ac.id, [3]endang.wahyu.pamungkas10@mhs.if.its.ac.id, [4]dwi@its.-sby.edu

***Abstract*— Business process management technology at present has been developed and applied both in small and in large scale. Many companies and organizations use, for instance, Enterprise Resource Planning (ERP) or other business process-oriented system. In this paper, a clustering method in business process model based on its similarity is proposed. This clustering aims to group some similar business processes to form a common business process. A new business process, as a result, can be composed based on similar common business process in order to increase reusability. It is done according to similarity value among business processes that is by calculating the similarity based upon structural and behavioral similarity method. Meanwhile, the clustering process uses a graph partition approach. This research then shows that the clustering result of business process is precise at certain threshold value.**

Keywords—business process, similarity, clustering

I. INTRODUCTION

In recent years, business process management technology is increasingly being used in line with many other existing business processes. Using this technology, a company can build and update any information in the process of rapid business model including repository in order to make every service provided can change quickly either due to changes in policy or due to changes in market conditions [7]. The increasing use of business process management technology can be seen from many companies using automation company performance. Current ERP (Enterprise Resource Planning) becomes the model of the use of business process management technology utilizing Service Oriented Architecture (SOA), which has now been developed into a Service Resource Oriented Architecture (SROA) [5]. For example, a number of Chinese companies listed on Haier have more than 3000 models of business processes and 600 EPC in the SAP reference model. This technology is also applied by one of China's largest companies, CNR Corporation Limited, a combination of 20 companies in China. As each firm, before joining, has nearly 200000 respectively, the final process models that have to be integrated process model are in a significant number [7].

A large corporation, as mentioned above, must have hundreds - even thousands of business processes. For this, discovering and analyzing the similarity of a business process collection owned will be very useful to the companies concerned. First, some business processes that have a high similarity can be formed to increase efficiency. Secondly, some similar business processes can be used as a foundation for the manufacture of new business processes. Third, conclusion from a set of similar business processes can be drawn about the opportunities in the manufacturing business process standardization. To illustrate, for a merger of several companies that have different business processes, the first step that should be done is to group some similar business processes. It is then followed by analyzing the grouped process to obtain a combined model [10]. In addition, the combined model from some business processes with a high similarity can be used as a base in the manufacture of flexible business processes. An appropriate research has been done by [1] on semantic web service aimed to produce a configurable and scalable service. This study was an initial step in the process towards the establishment of a flexible business process.

In the structure of a business process, a model can be regarded as a graph for containing a set of nodes connected by the edge [4]. Many notations can be used as a model for representing business processes such as Business Process Execution Language (BPEL), Business Process Modeling Notation (BPMN), Event-driven Process Chains (EPC), Yet Another Workflow Language (YAWL), Petri Net Markup Language (PNML), and many others. Of these models, Petri net is a notation that is easy to understand and analyze. [12] has also been conducting research on the use of models in ERP Petri net on small and medium-scale enterprise. The results then show that the model can meet the needs of Petri Net model variations in ERP. In addition, many researchers have found a way to change the shape of another notation to Petri Net. Otherwise, it is feasible to convert other forms of the model into a Petri Net [7].

This paper will discuss a clustering method based on the similarity of business process models. Here, a metric integration between structural and behavioural similarity was used to measure the similarity value. The integration was done by weighting the values that is by more focusing on the behavioral similarity in consideration to that most business processes in ERP have different structures but similar transition sequence. By so doing, the behavioural similarity can cover the lack of structural similarity. It is subsequently followed by the process of clustering using the graph partitioning approach to the production of several cluster containing the similar business process models. In this

978-1-4799-1078-6/13/$31.00

process, the threshold was varied to obtain an optimal threshold value.

This paper is organized as follows: the second section is to describe some researches that have been done concerned with this topic. The discussion about the similarity calculation of business process models is then presented in the third section. Following this, the fourth section is designed to present a discussion about the clustering process based on the similarity values that have been calculated. After that, discuss the validation technique from the clustering result. Finally,draw a conclusion from the entire paper.

II. RELATED WORK

Clustering is based on the value of similarity between business process models. [3] has performed a similarity search approach by comparing the value of 4 different methods: A-star algorithm, greedy, heuristic and exhaustive approach. Of these 4 methods, A-star is found to have the best performance. Meanwhile, [4] re-creates a notation approach using Event Driven Process Chains (EPC) as a model of the process. 3 types of similarity are involved: label match similarity, structural similarity, and behavior similarity. The result of precision and recall suggest that structural similarity is slightly better than the other two methods. [3] represents a pertinent as a graph structure to compute similarity based on the maximum common edge sub graph (MCES) algorithm and [4] calculates a behavior similarity using a reach ability graph approach. Here, both aforementioned studies are implemented using a tool called Beehivez.

In some previous studies, [8] have tried to cluster based on similarity by using 2 vectors - activity and transition. The calculation of the two vectors is used as a distance between the business processes. The distance, afterward, is used as the benchmarks in doing clustering, in this case using the technique of hierarchical clustering. Then [15] clustering is applied with parameter names of web-service operations and then leveraged to quantify degree similarity of web service. Another study [10] uses a two-level clustering in which clustering at the first level is done by topic or labels that have made the most appearances. The next level furthermore is based on structural similarity using a specific structure matching algorithm. This study is also equipped with a retrieval technique from a existing process model.

In this paper we propose clustering method that use graph partition approaches and Petri Net as models notation. As we know, sequence of execution in ERP is more important than structure of business process. We more explore at the calculation of similarity value between models using the integration of structural value and behaviour similarity as distance between models. Unlike in [10] , they just use structural similarity as distance. We also explore the provision of threshold value in the clustering process using graph partition approaches. Furthermore the clustering results will be evaluated by using silhouette index. Thus, the result of clustering is expected to be maximum.

III. SIMILARITY BUSINESS PROCESS

A. Structural Similarity

In our research, Beehivez tools have been used to calculate the similarity value..Here, an algorithm is used to follow what is implemented in these tools. In Beehivez a structural similarity is computed by considering the dependency graph as discussed in [14]. The paper explained that the matrix incidence of Petri net is used to build process matrix. Process matrix is build by dependency graph. It is then followed by calculating the difference dependency matrix of the two models as the different distance (d). After that, the similarity value is obtained by the formula $1/(1+d)$. The following are the steps to finding the value of distance [2]:

- Forming dependency graph. In [2] the dependency graph (DG) is interpreted by a tuple <DN, DE> where DN represents a set of activity nodes and DE represents a collection of edges connecting nodes activity.
- Filtering comparability. In this section, comparisons between dependency graph and δ-Comparability are made. In this process, the threshold value δ is required. Two dependency graphs DG_1 and DG_2 are said δ-Comparable if fulfilling the conditions.

$$\frac{|DN_1 \cap DN_2|}{|DN_1 \cup DN_2|} \geq \delta \text{ dimana } 0 < \delta \leq 1 \quad (1)$$

- Forming process matrix. Process similar to adjacency matrix M is a *n x n* matrix, which contains information on whether one node to another node is interconnected.

$$M(i,j) = \begin{cases} 1 & \text{if there is edge connecting else} \\ 0 & \end{cases} \quad (2)$$

- Normalizing the matrix process. Normalization is intended to equalize the number of row and column in the second matrix process – for example in model 1 one node that does not exist in the other models; the row and column in the matrix process is added with value 0. The other models do either. Suppose 2 dependency graphs of $DG_1 = (DN_1, DE_1)$ and$DG_2 = (DN_2, DE_2)$ and NM_1 and NM_2 are matrices that have been normalized, and $\cup DN_2 = \{a_1, a_2, \ldots a_3\}$ is obtained then as formulated as follows:

$$NM_1(i,j) = \begin{cases} 1 & if(a_i, a_j) \in DE_1 \\ 0 & otherwise \end{cases} \quad (3)$$

$$NM_2(i,j) = \begin{cases} 1 & if(a_i, a_j) \in DE_2 \\ 0 & otherwise \end{cases} \quad (4)$$

- Getting a distance value. Distance value can be found by performing matrix subtractions $NM_1 - NM_2$. The matrix will subsequently be obtained in which there is a value of 1 or -1. Distance is obtained by summing the values but the value -1 must be valued as positive first.

B. Behavioural Similarity

This similarity is computed by Principle Transition Sequence (PTS) metric [13]. In searching for the value of similarity behavior in Petri net, the principle of *coverability* *tree* is very important [9]. PTS is considered to have a good behavior in the calculation of similarity for being able to handle multiple cases at Petri net such as complex structure, non-free choice, or arbitrary loop. Hence, it can be expected that the behavior similarity values can be obtained in a good approximation [14]. The formula in finding the value of similarity using PTS can be described as follows:

Suppose LP and LP' is labeled Petri net 2, then $L(\sigma)$ and $L(\sigma')$ are a transition sequence of the second Petri net. Similarity of 2 PTS values is obtained based on the longest common subsequence (LCS) as presented in the following formula:

$$Sim\ (\sigma,\sigma') = \frac{length(lcs(L(\sigma),L(\sigma')))}{\max\ (length(L(\sigma)),length(L(\sigma')))} \qquad (5)$$

Having obtained the formula on how to find the value of 2 sequencess similarity, it is afterwards developed by searching for similarity of two sets of transition sequence. Suppose P and Q are two sets that contain a collection of transitions:

$$Sim(P,Q) = \frac{\sum_{\sigma\in P\ \sigma'\in Q}\max\ Sim\ (\sigma,\sigma') + \sum_{\sigma'\in Q\ \sigma\in P}\max\ Sim\ (\sigma,\sigma')}{|P|+|Q|} \qquad (6)$$

Then, we can obtain a new formula to find similarity between two labeled Petri net as follow.

$$Sim(LP,LP') = \sum_{i=1}^{3}\lambda_i x\ Sim(P_i(LP),P_i(LP')),\ \ \lambda_i = \frac{|P_i(LP)|+|P_i(LP')|}{|pts(LP)|+|pts(LP')|} \quad (7)$$

where P_i and $pts(LP)$ follow the definition of Principal Transition Sequence (PTS) as follows:

E.g.: LP is labeled as Petri net, C_{LP} is the incidence matrix of the Petri net, CT_{LP} is coverable from the tree with root labeled Petri net v_r and V_d is a collection of dead-end node, an old V_o node. Then $pts(LP)$ can be determined with the following provisions:

- If $v_d \in V_d, ts(v_r,v_d,CT_{LP}) \in pts(LP)$
- If $v_o \in V_o\ , ts(v_r, anchor(v_o), CT_{LP}) \in pts(LP)$
- If $v_o \in V_o\ , ts(anchor(v_o), v_o, CT_{LP}) \in pts(LP)$

$ts(v_1,v_2,CT_{LP})$ means a transition node from v_1 and v_2 at *coverability tree* CT_{LP}. From the above definition, $pts(LP)$ is obtained. On the other hand, P_i is obtained by model of Petri net, the category into which the PTS is. The following categories of PTS are in question:

- P_1= PTS that do not contain repetitive transition.
- P_2= PTS repetitive but is finite.
- P_3= PTS is recurrent and is infinite.

IV. CLUSTERING WITH GRAPH PARTITION APPROACH

Clustering method based on the distance between the 2 entities similar to the graph partition. This method is usually used when the object in the cluster is difficult to be represented in a mathematical form [10]. A distance used in the clustering process is based on the similarity metric from similarity calculation. The set of business processes that have a high similarity will be in 1 cluster. Then clustering algorithms are used:

- Determining the threshold.
- Calculating all values of similarity between the models.
- Repeating for each model compared with the threshold value of similarity.
- For the 2 models similarity value above the threshold given edge that connects the two.
- For models not connected with our models do not go to any means of any cluster.
- A graph that shows the clusters formed. Amount equal to the number of clusters formed graph.

TABLE I. Similarity Matrix

	M1	M2	M3	M4	M5	M6
M1	1	0.65	0.75	0.42	0.23	0.18
M2	0.65	1	0.72	0.58	0.22	0.19
M3	0.75	0.72	1	0.42	0.19	0.13
M4	0.42	0.58	0.42	1	0.55	0.41
M5	0.23	0.22	0.19	0.55	1	0.83
M6	0.18	0.19	0.13	0.41	0.83	1

TABLE II. Adjacency Matrix

	M1	M2	M3	M4	M5	M6
M1	1	0	1	0	0	0
M2	0	1	1	0	0	0
M3	1	1	1	0	0	0
M4	0	0	0	1	0	0
M5	0	0	0	0	1	1
M6	0	0	0	0	1	1

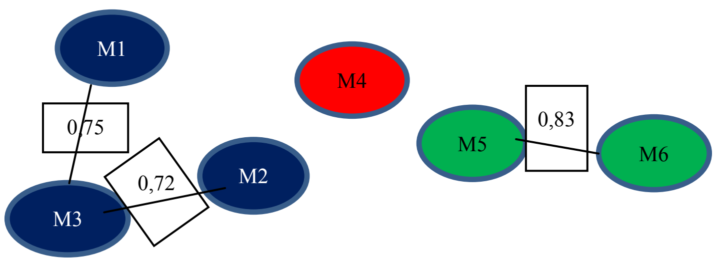

Fig.1. Clustering Method

In the figure above one example of a cluster is formed by a graph approach partition. We determined that the threshold is 0.7. Then, the 2 models with similarity above 0.7 will be connected by an edge, as formed in the model of M1 with M3, M2 with M3, and M5 with M6. Once the edge is added, it will form a graph, each of which represents 1 cluster. In the figure above M1, M2, and M3 are in 1 cluster, while M5 and M6 form another cluster. M4 is an example of a model that does not go into any clusters but forms its own clusters. This occurs because it has a value under threshold similarity with other models.

V. SILHOUETTE INDEX

Validation is very important in a series of clustering process. This validation has two capabilities. The first capability is the usage of this validation as a basis for clustering in order to determine whether quality is good. The second capability is the ability to compare the quality of several clustering algorithms. There are a lots of methods to validate the process of clustering, it depends on the method and the type of data cluster. In the case of a business process model, clustering is based on similarity in which its value percentage is used as the distance value. that would be very suitable to use silhouette index as validation method.

Silhouette index calculates the quantity cluster or clustering based on the coherence between data clusters. Dissimilarity values, meanwhile, are used instead of the similarity value. In accordance with [11], the purpose of this silhouette index is to know how tacky coherence value of a data cluster with other members of the cluster is and how tacky that data with the data in other clusters nearby is. The following picture below presents a better understanding:

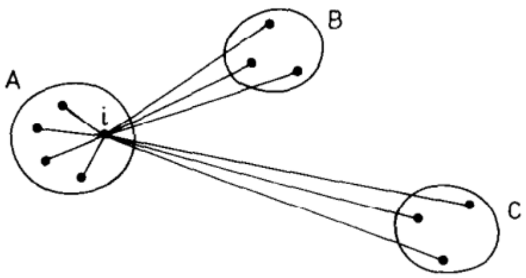

Fig 2. Silhouette Index

In the picture above, there are 3 clusters: A, B, and C. Suppose we intent to find the value of silhouette index node *i*. It calculated as the average dissimilarity value i to other node in the cluster and the dissimilarity value of node i with the nodes in the nearest cluster of node i. The following formula is used to calculate the silhouette index:

$$s(i) = \begin{cases} 1 - a(i)/b(i) & if\ a(i) < b(i) \\ 0 & if\ a(i) = b(i) \\ b(i)/a(i) - 1 & if\ a(i) > b(i) \end{cases} \qquad (8)$$

Where : s(i): value silhouette index point i
a(i): the average dissimilarity value i with the other points of the cluster
b(i): the average dissimilarity value i with dots of other nearby clusters

The normal result value is between 0 and 1. However, it is also possible to get a negative value. The closer we get to 1, the better determination of the cluster. If the result is a negative value, the point is in the wrong cluster. If the value is 0, the point is the dissimilarity between the same clusters. Values above zero indicate that the point has to be in the right cluster. From all the values, the average silhouette value of each cluster or even the entire cluster can be calculated. This value can be used as the basis of whether the overall clustering result is good.

To find the silhouette index values, a dissimilarity table between models is needed. Dissimilarity value is obtained by the formula1-value similarity. With the example in Fig.1, the following table is obtained:

TABLE III. Table dissimilarity

	M1	M2	M3	M4	M5	M6
M1	0	0.35	0.25	0.58	0.77	0.82
M2	0.35	0	0.28	0.42	0.78	0.81
M3	0.25	0.28	0	0.58	0.81	0.87
M4	0.58	0.42	0.58	0	0.45	0.59
M5	0.77	0.78	0.81	0.45	0	0.17
M6	0.82	0.81	0.87	0.59	0.17	0

Having obtained the new dissimilarity table, the silhouette index value is then being calculated. The following table is the calculation of Table 3:

TABLE IV. Silhouette value

	A	B	Silhouette Index
M1	0.3	0.58	0.482758621
M2	0.315	0.42	0.25
M3	0.265	0.58	0.543103448
M4	0	0	0
M5	0.17	0.45	0.622222222
M6	0.17	0.59	0.711864407

The mean value of the silhouette index is searchable by finding the mean value of the silhouette index and by ignoring zero value. In Table 4 average silhouette index value is 0.52.

VI. EXPERIMENTAL RESULT

As experimental materials we used 28 models of ERP in Petri net notation. These 28 models were made by using woped tools [16]. With these tools, it was possible to model business processes that fitted our own business processes. Then, in finding similarity values, we used the help of tools called beehivez [17] used to find the value of structural similarity [14] and behavior similarity [13].

Similarity values are used as the basis for clustering a combination of structural and behavior similarity. The way to incorporate is by using the weighted value of each similarity value. Weighting a little emphasis on the behavior value is to avoid the confusion of same structure but different task. In this experiment, we tested three types of weighting: 40% /60%, 30% /70%, and 20% /80%. As explained before, we have given a bigger value for the behavioral similarity score expectedly to cover the lack of structural similarity that cannot capture the transition sequence as described in the introduction.

Once the similarity values are obtained, the next process is to perform clustering. As explained, the process of clustering is using the graph partitioning approach. In the clustering process we tried a variety of threshold values to identify and to analyze each cluster of its results. The following clusters are generated by each of its values threshold:

TABLE V. Cluster results threshold 0.5

Threshold = 0.5		
cluster 1	Cluster 2	Cluster 3
MTO01	MTO18	MTO19
MTO02		MTO20
MTO03		MTO21
MTO04		MTO22
MTO05		MTO23
MTO06		MTO24
MTO07		MTO25
MTO08		MTO26
MTO09		MTO27
MTO10		
MTO11		
MTO12		
MTO13		
MTO14		
MTO15		
MTO16		
MTO17		

TABLE VI. Cluster results threshold 0.6

Threshold = 0.6							
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
MT O01	MT O9	MT O10	MT O11	MT O11	MT O18	MT O19	MT O25
MT O02			MT O12	MT O12		MT O20	MT O26
MT O03			MT O13			MT O21	
MT O04			MT O14			MT O22	
MT O05						MT O23	
MT O06						MT O24	
MT O07						MT O25	
MT O08						MT O26	
MT O17							

TABLE VII. Cluster results threshold 0.7

Threshold = 0.7							
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
MT O01	MT O9	MT O10	MT O11	MT O11	MT O18	MT O19	MT O25
MT O02			MT O12	MT O12		MT O20	MT O26
MT O03			MT O13			MT O21	
MT O04			MT O14			MT O22	
MT O05						MT O23	
MT O06						MT O24	
MT O07						MT O25	
MT O08						MT O26	
MT O17							

Further, the validation is performed against each cluster variation. Here, it can be seen which variation produces the best cluster result. Following table is the validation calculations using the Silhouette Index:

TABLE VIII. Average Silhouette Index

	40/60	30/70	20/80
0.5	0.534	0.536	0.538
0.6	0.524	0.528	0.531
0.7	0.524	0.528	0.531

The above table shows the average of Silhouette index of each generated cluster. The row indicates the threshold clustering and the column shows the weighting similarity. On the other side, the silhouette index of each cluster is obtained from the average silhouette index of each point.

VII. CONCLUSION AND FUTURE WORKS

The experimental results show that the weighting on the structural and behavioral similarity does not significantly affect the similarity value. This is due to that many business processes models used in the experiment are sequential; thus making structural and behavioral values not really significantly different. Furthermore, the variation of threshold values had an influence on the clusters form. In the experiment, there was a difference in the threshold current of 0.5 and 0.6. From the results of the evaluation with the silhouette index cluster, threshold 0.5 turned out better as when the threshold was at 0.6 many models did not fit into the cluster. So, the value of the silhouette index divisor was low. Overall clustering results were good. The value of the average silhouette index at 0.5 can prove that. It can be deduced that there are any models already on the correct cluster.

In a subsequent study, we will try another similarity method in consideration to that we have already known that the research on the measurement of similarity between business process models continues to evolve. In addition, improving on clustering techniques also need to be done in which the current hierarchical clustering technique now begins to be used. The results of clustering itself can be analyzed and used to build a new business process. This technique aims to acquire some new flexible business processes which can be done by deriving common fragment from each cluster.

ACKNOWLEDGEMENT

We would like to thank the Higher Education Directorate of the Education Ministry of Indonesia, Japan International Cooperation Agency (JICA) and the Indonesian Agency for Agricultural Research and Development for supporting the research.

REFERENCES

[1] Anang,_K. and Sarno, R., Semantic Web Service Composition for ERP Business Process,Journal Kursor, 2013.

[2] Bae, J., Liu, L., Caverlee, J., and Rouse, W. B., Process mining, Discovery, and Intregation using Distance Measure, Proceedings of the IEEE International Conference on Web Services, pp. 479-488, 2006.

[3] Dijkman, R., Dumas, M., Garcia-Banuelos, L. Graph Matching Algorithms for Business Process Model Similarity Search,Proceedings of the 7th International Conference on Business Process Management, pp. 48–63, 2009.

[4] Dijkman R., Dumas M., van Dongen B., Kaarik R., and Mendling J., Similarity of Business Process Model : Metric and Evaluation, Information Systems 36(2), pp. 498-516, 2011.

[5] Hermawan, and Sarno, R., Developing Distributed System with Service Resource Oriented Architecture 10(2), TELKOMNIKA International Journal, 2012.

[6] Jin T., Wang J., and Wen L., Efficient Retrieval of Similar Business Process Model Based on Structure, On the move to meaningful internet systems – volume part 1, pp 56-63, 2011.

[7] Jin T., Wang J., and Wen L., Querying Business Process Model based on Semantic, Proceedings of the 16th international conference on Database systems for advanced applications: Part II, pp.164-178, 2011.

[8] Jung J.Y., Bae J., and Liu L., Hierarchical Clustering of Business Process Models, SCC 2008, IEEE International Conference on Volume2, pp 613 – 616.

[9] Murata, T., Petri nets: properties, analysis and applications. Proceedings of the IEEE, 77, pp.541–580, 1989.

[10] Qiao M., Akkiraju R., and Aubrey J. Rembert, Towards Efficient Business Process Clusteringand Retrieval: Combining Language Modelingand Structure Matching, Proceedings of the 9th international conference on Business process management, pp. 199-214, 2011.

[11] Rousseeuw P. J., "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics,vol. 20, pp. 53-65, 1987.

[12] Sarno, R. et al., Petri Net Model of ERP Business Process Variations for Small and Medium Enterprises, Journal of Theoretical and Applied Information Technology, vol. 54, no. 1, pp.31-38, 2013.

[13] Wang, J., He, T., Wen, L., Wu, N., ter Hofstede, A.H.M and Su, A.J. behavioral similarity measure between labeledPetri nets based on principal transition sequences, Proceedings of the 2010 international conference on On the move to meaningful internet systems - Volume Part I, pp. 394–401.

[14] Wang, J., Wong, R. K., Ding, J., Guo, Q., and Wen, L., On Recommendation of Process Mining Algorithms, Proceedings of the 2012 IEEE 19th International Conference on Web Services, pp. 311-318.

[15] Dong, X., Halevy, A.Y., Madhavan, J., Nemes, E., and Zhang, J. Simlarity Search for Web Services, Proc. of VLDB 2004, pp. 372–383 (2004)

[16] Workflow Petri Net Designer (WOPED). Available : http://woped.dhbw-karlsruhe.de/woped/?page_id=22 [March. 23, 2013].

[17] Business Process Model and Instance Management System (Beehivez). Available : https://code.google.com/p/beehivez/ [March. 18, 2013]