A Semantic Approach for Process Annotation and Similarity Analysis

Tobias Weller

Institute AIFB,
Englerstr. 11, 76128 Karlsruhe, Germany
tobias.weller@kit.edu,
http://www.aifb.kit.edu

Abstract. Research in the area of process modelling and analysis has a long-established tradition. There a quite few formalisms for capturing processes, which are also accompanied by number of optimisation approaches. We introduce a novel approach, which employs semantics, for process annotation and analysis. In particular, we distinguish between target processes and current processes. Target processes describe how a process should ideally run and define a framework for current processes, which in contrast, capture how processes actually run in real-life use cases. In some cases, current processes do not match the traget processes and can even overhaul target processes. Therefore, one is interested in the similarity of the defined target process and current processes. The comparisons can consider different characteristics of processes such as service quality measures and dimensions. Current solutions try to convert processes to specific ontologies and perform process mining methods to discover hidden structures or try to support the managing of the execution. However, comparing target processes with current processes has not been addressed yet. To this end, we propose an new approach that is based on annotating processes with semantic information. We perform similarity analysis of target processes and current processes that exploits the semantics to show the added value of our approach. As part of the similarity analysis, we consider different service qualities and dimensions in order to determine how they influence both target and current process.

Keywords: Process Annotation, Similarity Analysis, Process Analysis, Quality of Service, Semantic Process Modelling

1 Introduction

Process modelling and analysis has multiple application domains. In the healthcare area one prominent example of processes is clinical pathways. Clinical pathways are an evidence-based response to specific problems and care needs in clinics. They support physicians by providing recommendations of the sequence and timing of actions necessary to achieve an efficient treatment of patients [1, 2]. Each clinic has its own pathways based on its individual evidence and experience. Therefore, there are multiple pathways that target different problems and care needs, as well as same problems and care needs [3–5]. However, physicians are not strictly restricted to the published pathways. Therefore, the process, defined in the pathway (target process), can differ from the actually performed workflow (current process). As a result, there might be discrepancies between the published clinical pathways and the actually performed workflow, which is based on the decisions of physician on how to treat the patient.

This situation is aggravated by the fact that there is a lot of data that is generated during the treatment of patients and that needs to be managed and interpreted. In order to ease this task, the information can be captured semantically and used for enhancing comparisons and processing. For this purpose, there are already many ontologies in the medical domain, which can be used to structure the semantic information – the Disease Ontology[1] that provides descriptions and related medical terms about human diseases and the Foundational Model of Anatomy (FMA)[2], which describes classes, structures and relationships of all parts of the human body. Once the patient data is semantically described, we can exploit it by performing enhanced comparisons between clinical pathways and processes. In addition, processes can be compared based on different service qualities and dimensions, such as complexity, runtime, outcome or costs.

The problem of having current processes diverge from the defined target process does not occur only in the medical domain. The same difficulties arise also in enterprises and in the domain of Internet of Things (IoT) applications, in which the actual communication flow between devices can diverge from a defined target process. This is precisely the topics that we want to explore. One aspect is that it is debatable whether the current process performs better, in terms of certain service qualities and dimensions, than the defined target process and it is, therefore, more advantageous to perform the process by deviating from the target process. Given a set of current processes, we are interested in calculating the similarity between them and the target process, in order to be able to quantify the variety and see how different processes behave in terms of different service quality aspects and dimensions.

2 State of the Art

An important aspect, in order to have a common point of view on processes, is to define the term *process*. We use the process definition from ISO 9000:2015 [6], which is given in the following.

Definition 1. *ISO 9000:2015 Process: Set of interrelated or interacting activities that use inputs to deliver an intended result.*

Note 1 to entry: Whether the "intended result"of a process is called output, product or service depends on the context of the reference.

Note 2 to entry: Inputs to a process are generally the outputs of other processes and outputs of a process are generally the inputs to other processes.

[1] http://disease-ontology.org

[2] http://sig.biostr.washington.edu/projects/fm/index.html

This definition is specifically related to quality management systems, but we aim to use it in a broader way. We do not focus on quality management systems in particular but rather on processes in general. In addition, we want to recognise that processes do not necessarily always have to transform inputs into outputs. In some cases, inputs become outputs without transformation.

There are already widely used ontologies, such as Dublin Core Schema[3] that provide a set of metadata that can be used to annotate resources. The advantage of such schemata is that they can be integrated easily in order to annotate resources and provide interoperability with further datasets. In addition to data annotation ontologies, there are also ontologies available to describe the components of a process and the relationships between them such as SUPER [11, 12] and the Process Specification Language[4], which has been approved as an international standard [13]. There are also some ontology-based annotations for process models [14, 15]. The process models are semi-automatically annotated according to process and domain ontology.

Existing approaches describes how service qualities and dimensions can be captured [19]. Thereby, frameworks like SERVQUAL can be used to measure the quality of processes [21]. Service qualities from e-services [20] or other process performance indicators [22] can also be used as metrics to measure the performance of processes.

There are a number of different process similarity measure methods for comparing processes. Some uses node similarity, structural similarity and behavioural similarity. However, most focus on business processes [16, 17] and do not distinguish between target and current processes. The similarity of processes is, among others, used to cluster processes [18]. Other approaches e.g. Process Mining try to reveal hidden structures and create a target process by using current process data [23, 24]. However, these approaches reveal hidden structures but not the influence of processes on different service qualities and dimensions.

3 Problem Statement and Contributions

We focus on performing similarity analysis between target and current processes by exploiting the semantics of process data. The semantics that we use to compare processes consist, among others, of the hierarchical structures of the performed tasks and process flows, the user roles (for example, only specific users are allowed to perform a task or a decision) and the rules that define the workflow of processes. Based on the presented motivation and the current state-of-the-art, we formulate the following research question and its subquestions:

How do we benefit from the combination of process data with semantics in order to improve processes by performing similarity analysis?

RQ1 How can we formally specify process data with semantics?

RQ2 Which service qualities and dimensions can we use to compare processes?

[3] `http://dublincore.org`

[4] `http://www.mel.nist.gov/psl/`

RQ3 Which methods can we use to perform similarity analysis of target processes and current process data?

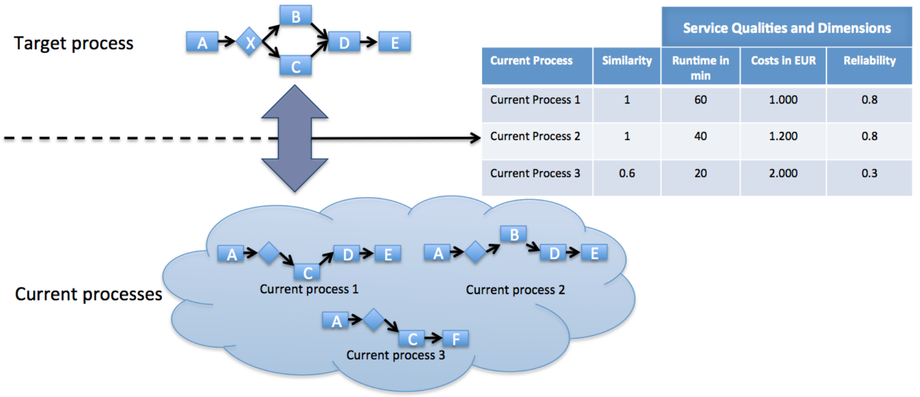
During the PhD we will develop an approach to annotate process data with semantic information and perform similarity analysis of target process and current processes. This approach will be modelled in a common way, so it is generally applicable. In the following, we discuss the subquestions in more detail.

(RQ1) How can we formally specify process data with semantics?
There are already established formal representations for modelling languages e.g. for BPMN 2.0, the standard language BPMN 2.0 XML published by OMG[5] or the Petri Net Markup Language [7] for representing petri nets. However, these standards do not consider formal semantics. Therefore, we will show how to combine formally specified process data with semantics that can be queried and processed. The enriched process data can be used to compare and analyse processes based on the semantic information.

(RQ2) Which service qualities and dimensions can we use to compare processes?
Processes can be compared based on different service qualities and dimensions such as runtime, outcome, costs or reliability. Capturing these service qualities and dimensions is a first step towards being able to compare the defined target process and the current processes.

(RQ3) Which methods can we use to perform similarity analysis of target processes and current process data?
We will show which methods can be used to compare target process with a set of current process. During the use of different similarity methods, we will exploit semantics such as the hierarchical arrangements of activities and process flows, and user roles, linked to tasks.

Figure 1 shows the comparisons of target process with a current process data.



		Service Qualities and Dimensions		
Current Process	Similarity	Runtime in min	Costs in EUR	Reliability
Current Process 1	1	60	1.000	0.8
Current Process 2	1	40	1.200	0.8
Current Process 3	0.6	20	2.000	0.3

Fig. 1. Determining the similarity between target process and current processes

[5] http://www.omg.org/spec/BPMN/2.0/

The research questions aim to result in multiple contributions. The first contribution is the introduction of an approach that integrates processes with semantic information that can be queried and processed. We would like to integrate as much semantic information as possible to allow, in a later step, enhanced similarity analysis that considers all these aspects. Another contribution is a set of service qualities and dimensions that can be used to compare processes. We will show different metrics and how they can be used. The last contribution is the similarity analysis between target process and current processes. Thereby, we will use methods that exploit the semantics, captured in the previous step, such as the hierarchy of activities and process flows, to quantify the similarity.

4 Research Methodology and Approach

The structure of the research methodology and approach is directly derived based on the research question (Section 3). Research methodologies can be classified as quantitative, qualitative and mixed research methodologies. Quantitative research methods collect numerical data and use it to analyse and explain a circumstance [8]. We will apply quantitative methodologies to plan and approach the research problems. In particular we will collect a large sample of data, process the data, and compare the results to other existing similarity approaches. We will investigate how semantics effect the analysis and comparisons of processes, and test different methods to compare processes.

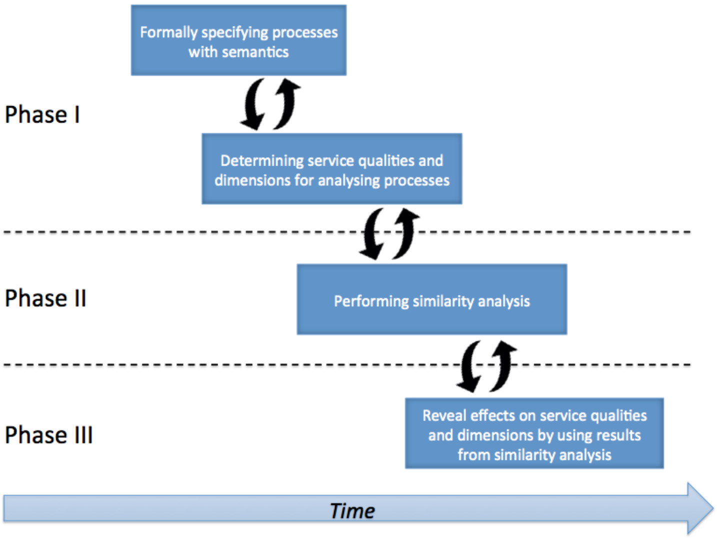

Formally specifying processes with semantics
Phase I
Determining service qualities and dimensions for analysing processes
Phase II
Performing similarity analysis
Phase III
Reveal effects on service qualities and dimensions by using results from similarity analysis
Time

Fig. 2. Research approach – divided into three phases. Each tackles another aspect and influences tasks in other phases.

Figure 2 shows the planned thesis approach, divided into three main phases. Each phase tackles a specific part of the thesis, consists of performed activities and influences or is influenced by other activities. In the following, we will explain each phase in more detail.

Phase I: We assume that semantics provide a huge potential for similarity analysis and revealing the effects on service qualities and process performance indicators. In order to exploit semantics in processes, we first have to annotate the processes. In this phase we will consider different approaches for capturing and annotating target processes with semantic information.

The annotation of processes with semantics influences the selection of service qualities and dimensions for comparing processes. Dimensions, which cannot be captured with semantic annotations, cannot be used for comparing processes. However, if we want to use specific dimensions for comparisons, we have to find the corresponding information in process data. The determining of dimensions partially overlaps with the formal specification of processes with semantics. Both activities are performed in phase I.

Phase II: This phase focuses on performing similarity analysis of target process and current processes. We will use different similarity methods e.g. node similarity, structural similarity and behavioural similarity. Among others, we will also use methods that do not exploit semantics and compare them to methods that exploit semantics in order to show the advantages of having semantic annotations. We will also consider combining different methods for similarity analysis, resulting in a hybrid approach.

Phase III: The last phase uses the similarity analysis as an input in order to reveal the effect on service qualities and dimensions. We will evaluate whether individual types of process tasks influence the outcome of the process. In addition, during this phase, we can also discover new insights that motivate to capture additional service qualities and dimensions. Therefore, this activity influences in turn phase II.

Capturing, annotating, performing similarity analysis and revealing the effect on service qualities and dimensions of processes, independently from a specific domain, is the benefit of this work. We will show that the methods are not constrained to a single domain by applying them to different domains (Section 6).

5 Preliminary Results

Currently, we are facing the first phase (see Section 4), which is about formally specify process data with semantics. To this end, we analysed different tools that allows to model processes. However, existing tools do not allow to enrich process data with semantics. In addition, we aim to follow the Linked Data Principles[6] for publishing data.

In order to combine processes with semantic information, we created a tool to capture process and allow to enrich them with semantic information. We

[6] http://www.w3.org/DesignIssues/LinkedData.html

used bpmn-io as web modeller and extended it with further functionality. bpmn-io[7] is a JavaScript renderer that allows to model and checking the syntax of BPMN networks. It is part of Camunda BPM[8], which is an open source platform for workflows. We embedded our developed tool into a Semantic MediaWiki[9]. We used Semantic Forms[10], which is an extension to Semantic MediaWiki, to allow the creation of forms to capture semantic information of processes. Thus, Semantic MediaWiki in combination with our developed tool serves as platform to capture, annotate, query and process the information in a structured way.

With this tool, we can integrate processes, stored in the standard format BPMN 2.0 XML[11] into Semantic Media Wiki and enrich them with semantics. Furthermore, the functionality of Semantic MediaWiki allows the collaborative modelling and sharing of processes and information. The integrated and semantically enriched processes can in turn be exported into BPMN 2.0 XML format, allowing for exchange and reuse of the modelled processes.

As the next step, we will determine service quality measures and dimensions for comparing processes but also to measure the efficiency of a target process such as runtime, outcome or costs. Therefore, we will investigate different dimensions, which are used to measure the efficiency of processes such as runtime and outcome. In addition, we will consider different similarity methods to quantify the similarity between target process and current processes and necessary information that will improve the calculation of similarity. These consideration will influence the enrichment of semantic information, since we have to capture it during the annotation of the processes. Following, we will compare target processes with current processes, according to the considered service quality measures and dimensions with different similarity methods.

6 Evaluation Plan

For validating our solution, we will implement the designed approach and methods in different use-case scenarios. This ensures on the one hand that our approach and methods abstract from the used domain and on the other hand the capturing of two independent results that can be evaluated.

We plan to use the following two domains to evaluate our appraoch:

1.) Medical Domain: Actual processes in clinics differ from target processes. This is caused by latest insights and developments in the medical domain and the slow adaption of clinical pathways. In addition, there are many ontologies i.e. Foundational Model of Anatomy ontology (FMA)[12] or Gene Ontology[13] that can be used to structure process data with semantic information. Therefore,

[7] `https://github.com/bpmn-io/bpmn-js`

[8] `https://camunda.org`

[9] `https://semantic-mediawiki.org`

[10] `https://www.mediawiki.org/wiki/Extension:Semantic_Forms`

[11] `http://www.omg.org/spec/BPMN/2.0/`

[12] `http://sig.biostr.washington.edu/projects/fm/`

[13] `http://geneontology.org`

we will use our approach to calculate the similarity between target and process data and show the influences of processes on different service qualities and dimensions.

2.) Internet of Things: Another field of application is the domain of Internet of Things. In this domain the communication and data flow between devices is not strictly given. Hence, there are more ad-hoc processes, which makes it hard to get an overview of the processes in general. Although this domain is rather new, there are already some ontologies available [9, 10]. We will use data from devices (i.e. communication data and process data) and annotate the tasks with semantic information. This allows enhanced analysis of communication workflows, and to see the deviation of current processes from target processes.

We benefit from implementing the used similarity methods in different domains in order to compare how well they perform on different data sets. We will compare the outcome of methods that do not exploit semantics with methods that use semantics and describe the differences in outcomes.

For evaluating the first research question, we will validate the formalised process data, enriched with semantics, by comparing the usability of the provided methods with different approaches and the expressiveness of the formally specified processes according to the defined service qualities and dimensions. The formalisation of data should not be focused on a specific scenario or domain. Therefore, we will a apply our approach and methods in multiple scenarios and domains. In addition to the application in multiple domains, we will also show how well the defined service qualities and dimensions can be queried and processed. We will query the data and show which issues can be answered by the formalised process data. We will compare the data before enriching it with semantics and show which issues can be answered by using semantics. With this evaluation, we will outline the advances of semantics in process data.

To evaluate the second and third research question, we will start with comparing very simple process and gradually extend the process with further details and expressiveness. Hence, we will start performing similarity analysis and reveal the effect of different service qualities and dimensions in each applied domain with a sequential process and then successively extend the expressiveness of the process and the used service qualities and dimensions.

7 Conclusions

We aim to develop an approach to annotate and perform similarity analysis between target and current processes. We have taken a first step towards process annotation. This approach shows how domain experts can enrich process data with semantics.

The process annotation is an important aspect to perform similarity analysis. These annotations are used to compare target process with current processes. Thereby, we will show how different approaches can be improved by using semantics.

We will consider the similarity in relation to service qualities and dimensions in order to 1.) describe a framework, in which the process runs and 2.) reveal weak spots, which has influence on different service qualities and dimensions. This knowledge can be used to enhance the execution of the processes and improve the target process.

In conclusion, we have taken a first step towards process annotation and similarity analysis. The approach will be evaluated in different domains in order to get multi-faceted results. The results of the similarity analysis and the impact of the service qualities and dimensions can be used for processes improvement and optimisation.

References

1. Panella, M., Marchisio, S., Di Stanislao, F. Reducing clinical variations with clinical pathways: do pathways work? International Journal for Quality in Health Care (2003), 15(6):509-521
2. Kinsman, L., Rotter, T., James, E., Snow, P., Willis, J. What is a clinical pathway? development of a definition to inform the debate. BMC Medicine (2010), 8(31)
3. Zand, e. K. Integrated care pathways: eleven international trends. International Journal of Care Coordination (2002), 6(3):101-107
4. Vanhaecht, K., Bollmann, M., Bower, K., Gallagher, C., Gardini, A., Guezo, J., Jansen, U., Massoud, R., Moody, K., Sermeus, W., Zelm, R., Whittle, C., Yazbeck, A. M., Zander, K., Panella, M. Prevalence and use of clinical pathways in 23 countries – an international survey by the european pathway association. International Journal of Care Coordination (2006), 10(1):28-34
5. Hindle, D., Yazbeck, A. Clinical pathways in 17 european union countries: a purposive survey. Australian Health Review (2005), 29(1):94-104.
6. European Committee for Standardization, Quality management systems - Fundamentals and vocabulary (ISO 9000:2015), September 2015
7. Billington, J., Søren, C., van Hee, K., Kindler, E. Kummer, O., Petrucci, L., Post, R., Stehno, C., Weber, M. The Petri Net Markup Language: Concepts, Technology, and Tools, Applications and Theory of Petri Nets 2003, 2679:483-505, Springer Berlin
8. Muijs, D. Doing Quantitative Research in Education with SPSS. 2nd edition. London: SAGE Publications, 2010.
9. Hachem, S., Teixeira, T., Issarny, V. Ontologies for the Internet of Things. ACM/IFIP/USENIX 12th International Middleware Conference, Dec 2011, Lisbon, Portugal. Springer, 2011.
10. Wang, W., De, S., Toenjes, R., Reetz, E., Moessner, K. A Comprehensive Ontology for Knowledge Representation in the Internet of Things. Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference, pp.1793-1798, June 2012, doi: 10.1109/TrustCom.2012.20
11. Dimitrov, M., Simov, A., Stein, S., Konstantinov, M. A BPMO Based Semantic Business Process Modelling Environment, Semantic Business Process and Product Lifecycle Management. Proceedings of the Workshop SBPM 2007, Innsbruck, April 2007
12. Hepp, M., Roman, D. An Ontology Framework for Semantic Business Process Management. Wirtschatsinformatik Proceedings 2007
13. European Committee for Standardization, Industrial automation systems and integration – Process specification language (ISO 18629-1:2004), November 2004

14. Lin, Y., Ding, H. Ontology-based Semantic Annotation for Semantic Interoperability of Process Models. Computational Intelligence for Modelling, Control and Automation, 2005 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, 1: 162-167, November 2005, doi: 10.1109/CIMCA.2005.1631259
15. Lin, Y., Strasunskas, D. Ontology-based Semantic Annotation of Process Templates for Reuse, 10th Int. Workshop on Exploring Modeling Methods in System Analysis and Design (EMMSAD05), Porto, Portugal, 2005
16. Dijkman, R., Dumas, M., van Dongen, B., Käärik, R. Mendling, J. Similarity of business process models: Metrics and evaluation, Information Systems, 36(2):498-516, April 2011, ISSN 0306-4379
17. Zhang, Y., Liu, J., Wang, L. Product Manufacturing Process Similarity Measure Based on Attributed Graph Matching, 3rd International Conference on Mechatronics, Robotics and Automation (ICMRA 2015), June 2015
18. Jung, J., Bae, J. Workflow Clustering Method Based on Process Similarity, Computational Science and Its Applications - ICCSA 2006, 3981: 379-389, 2006
19. Bauer, H.H., Falk, T., Hammerschmidt, M. eTransQual: A transaction process-based approach for capturing service quality in online shopping, Journal of Business Research, 59(7): 866-875, July 2006, ISSN 0148-2963
20. Collier, J.E., Bienenstock, C.C. A Conceptual Framework for measuring E-Service Quality, Proceedings of the 2003 Academy of Marketing Science (AMS) Annual Conference, 2015, pp. 158-162, ISSN 2363-6165
21. Gawyar, E.T.H., Ehsani, M., Kozehchian, H. Measuring service quality of state clubs in Lorestan province using SERVQUAL model, International Journal of Sport Studies, 4(2):233-237, 2014, ISSN 2251-7502
22. del-Río-Ortega A., Resinas, M., Ruiz-Cortés, A. Defining Process Performance Indicators: An Ontological Approach, Confederated International Conferences: CoopIS, IS, DOA and ODBASE, pp. 555-572, October 2010, ISSN 0302-9743
23. 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, Information Systems, 32(5):713732, July 2007
24. van Dongen, B.F., de Medeiros, A.K.A., Verbeek, H.M.W., Weijters, A.J.M.M., van der Aalst, W.M.P. The ProM Framework: A New Era in Process Mining Tool Support, Applications and Theory of Petri Nets 2005, 3536, pp. 444-454, 2005