Evaluation of Process Model Matching Techniques

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von

Elena Kuß
aus Schweinfurt

Mannheim, 2018

Dekan: Dr. Bernd Lübcke, Universität Mannheim
Referent: Prof. Dr. Heiner Stuckenschmidt, Universität Mannheim
Korreferent: Prof. Dr. Henrik Leopold, Kühne Logistics University Hamburg

Tag der mündlichen Prüfung: 15. März 2019

Abstract

Business process models are commonly used to document a company's operations. They describe internal processes in a chronological and logical order. Business process model matching refers to the automatic detection of semantically similar correspondences in process models. The output of those matching techniques is the basis for many applications. Currently, most research effort has been undertaken to improve the performance of such matching techniques. However, to support the improvement of process model matching techniques further, efficient and fair evaluation strategies are required. Moreover, information about the matching task, regarding the complexity of a data set have to be gathered. In the current literature, complexity is mostly associated with different level of granularity, thus $1 : m$ or $n : m$ correspondences. However, the evaluation should also account for different complexity aspects of the matching task, for example syntactical overlap of correspondences. Moreover, the evaluation of matching results actually strongly depends on the application. In this thesis, we therefore propose an application dependent evaluation. On the one hand, we introduce a non-binary evaluation, which better reflects the uncertainty of a gold standard and propose evaluation metrics, based on this non-binary gold standard which take different application scenarios into account. On the other hand, we propose a conceptually novel evaluation procedure, which offers detailed information about strength and weaknesses of matchers without manually processing the matcher output. It therefore helps to find optimal application scenarios for specific matching techniques. It can further serve as a basis for a prediction for future matching tasks. We conduct experiments to show the insights gained by the introduced evaluation metrics and methods. Moreover, we apply the metrics at the OAEI 2016 and 2017.

Zusammenfassung

Geschäftsprozessmodelle sind in fast jedem größeren Unternehmen zu finden. Sie beschreiben firmeninterne Prozesse in einer chronologischen bzw. logischen Abfolge. Solche Prozessmodelle können sehr große Datenmengen beinhalten, teilweise mit mehreren hunderttausend Prozessmodellen. Solche Datenmengen lassen sich händisch kaum bewältigen, daher müssen sie beispielsweise automatisch bearbeitet werden. Hierzu werden häufig "Matching Technologien" verwendet. Um zu erkennen wie gut solche Technologien in der Praxis funktionieren, bedarf es effizienter Evaluierungstechniken. Aktuell werden dafür hauptsächlich Metriken aus dem Information Retrieval herangezogen, die auf einem binären Goldstandard basieren. In der Praxis zeigt sich jedoch, dass solche binären Goldstandards die tatsächliche Komplexität eines Datensatzes nicht korrekt widerspiegeln. Die Erstellung eines solchen Goldstandards ist einerseits sehr subjektiv und andererseits gibt es nicht immer eine richtige oder falsche Lösung. Um diese "Unsicherheiten" zu berücksichtigen, schlagen wir einen nicht-binären Goldstandard vor, welcher (alle) mögliche Korrespondenzen eines Datensatzes enthält. Darüber hinaus entwickeln wir Evaluierungsmetriken, welche nicht-binäre Werte erlauben und eine Evaluierung je nach Anwendungsfall zulassen. Somit wird die Performance der Matcher aus unterschiedlichen Blickwinkeln betrachtet. Darüber hinaus präsentieren wir eine konzeptionell neuartige Evaluierungsmethode, welche detailierte Informationen über die Performance der Matcher bietet. Dabei wird der Datensatz und Matcher-Output automatisiert in verschiedene Komplexitätsstufen eingeteilt. Die Ergebnisse erlauben zudem optimale Matching-Szenarien für spezifische Matcher abzuleiten. Die eingeführten Metriken und Evaluierungsmethoden wurden bereits bei der OAEI 2016 und 2017 angewendet.

Contents

List of Figures

List of Tables

1

Introduction

In this chapter, we provide an introduction to the field of business process modeling. We focus our review on the main challenges of the evaluation techniques of process model matchers. In this context, we formulate research questions in the area of evaluating process matching techniques, which are addressed in this thesis. We then summarize the contributions of the thesis.

1.1 Business Process Modeling

Business process modeling is a growing discipline in many companies. Conceptual models, like business process models, are commonly used to document a company's operations. It aims in documenting the business processes within a company or institution. Business process modeling is widely used within a company for many reasons: For example to achieve transparency of the business processes or to make the processes comprehensible.

Weske (2012) describes a business process as follows:

"*A business process consists of a set of activities that are performed in coordination in an organizational and technical environment. These activities jointly realize a business goal. Each business process is enacted by a single organization, but it may interact with business processes performed by other organizations.*"

Consequently, business processes are targeted controlling instruments of operations within a company or institution. Therefore the economical point of view of the process is transformed into a technical process. Within such processes, the *activities* of the process models describe an event or task. Examples for notations which are used to document business process models are *Business Process Modeling and Notation* (BPMN) (Owen and Raj, 2003), *Event-driven Process Chain* (EPC) (Van der Aalst, 1999), *Unified Modeling Language* (UML) (Eriksson and Penker, 2000) or *Petri-Nets* (Murata, 1989; Van der Aalst, 1998).

Concrete applications are for instance the automation of manufacturing processes, the improvement of manufacturing processes, to generally increase the quality or to save costs. Therefore *business process modeling* is part of *business process management*. Business Process Management contains the techniques, management and tools which support the design and analysis of business processes. Business Process Models contain one or more Business Processes. In particular, applications are in the context of "Industry 4.0", "Internet of things" and "Smart factory". In this context, process models can be used, e.g., to fully automate the production process in factories. Another application of the development of business processes models is the quality management and best practices.

In some cases, the amount of data is too huge for manual processing business processes. The China railway company, for instance, stores more than 200 000 business process models (Ekanayake et al., 2011). This amount is too huge for manual processing, therefore in such cases automatic processing is required. One prominent example is the automatic matching of process models. In the next section we introduce to process model matching and provide an overview of current challenges in this field.

1.2 Business Process Model Matching

Business process model matching is concerned with the detection of similarities in business process models. On the one hand, the control flow of the process models is an important feature, on the other hand, semantic similarities of the labels are compared. The matching of the activities of the process models are called *correspondences*. Generally spoken, a matcher is a tool which automatically detects correspondences in business process models, or which supports an individual in generating correspondences, in order to save effort and time.

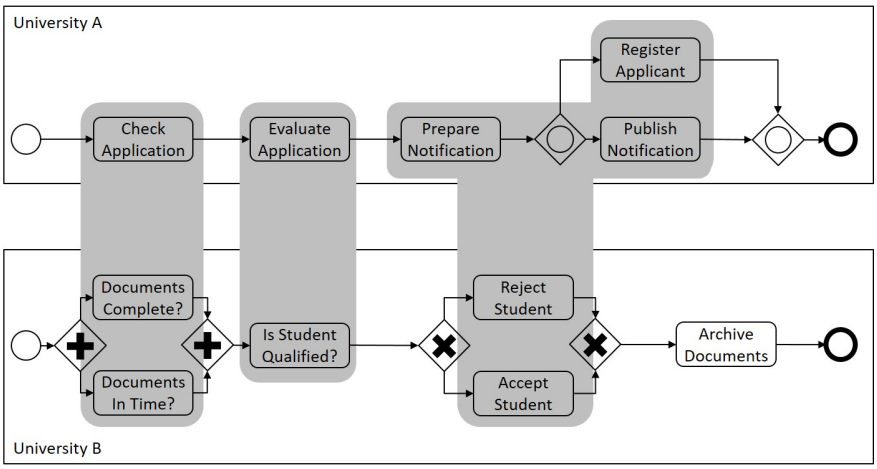


Figure 1.1: Two process models and possible correspondences as shown in (Kuss et al., 2016)

To highlight the challenges associated with such a matching task, we illustrate such specific difficulties in the example depicted in Figure 1.1. It shows two process models describing the steps students have to undertake to be admitted for the graduate programs of two different universities; in this case University A and University B. Possible correspondences are illustrated by gray shades.

Consider, for instance, the correspondence between "*Check application*" from University A and "*Documents complete?*" as well as "*Documents in Time?*" in the process of University B. A matching technique must be able to detect that in the process of University A a pre-check of the application is made, which is described at University B with the question if the documents are complete and in time. Such correspondences are $1 : m$ correspondences (in this case 1:2 correspondences). Similarly, a matching system has to detect the presence of complex correspondence

between "*Prepare Notification*", "*Register Applicant*" as well as "*Publish Notification*" with "*Reject Student*" and "*Accept Student*". To automatically recognize that the latter three activities relate to a stream of action that can be referred to "*Accept or Reject Student*", requires the recognition of complex semantic relationships. (This is an example for an $n : m$ correspondence, here 3:2). Another complex matching is the correspondence between the activity "*Evaluate*" and "*Is student Qualified?*". Here, a matching technique must be capable to automatically recognize that both activities evaluate if a student is suitable. This is especially challenging because on the one hand the words have no syntactical overlap and on the other hand one activity is a verb while the other activity is a sentence. Another challenge is the fact that not all activities from one process model are matched in the corresponding model (like in this example the activity "*Archive Documents*").

Systems that automatically perform such matching tasks are called *matching techniques* or shortly *matchers*. To address such challenges associated with process model matching, many different matching techniques have been proposed in recent years. Typically, these techniques combine different measures to quantify the structural as well as the textual similarity between the considered process models. The first matching techniques that have been defined combined structural measures such as the graph edit distance with syntactic text similarity measures such as the Levenshtein distance (Dijkman et al., 2009; Weidlich et al., 2010a). More recent techniques also consider semantic relationships between words, most commonly by building on the lexical database WordNet (Cayoglu et al., 2013b; Klinkmüller et al., 2013; Leopold et al., 2012). A few techniques also employ alternative strategies. Examples include matching techniques incorporating human feedback (Klinkmüller et al., 2014), techniques selecting the most promising similarity measures based on prediction (Weidlich et al., 2013a), techniques selecting the best correspondences based on voting (Meilicke et al., 2017), and techniques that employ machine learning (Sonntag et al., 2016). Weidlich et al. (2010a) presents a method for $n : m$ node matching.

Process matching techniques are relatively rarely used in practice compared to their many application scenarios. In order to fully exploit the potential in practice, the performance of process matching techniques needs to be improved. Considering the variety of matching techniques that have been defined in previous work, a key question is how to *evaluate* the performance of these techniques. While the specific technologies or model-related aspects exploited by the matching technique

do not change how a matching technique needs to be assessed, the question is how to fairly quantify to what extent the generated correspondences are correct. Overall, it can be observed that most research efforts are spend in advancing the process matching techniques compared to the advancement of their evaluation. However, the improvement process of process model matching techniques is closely linked to an efficient and fair evaluation. Current evaluation mostly relies on a ranking of the evaluated matching techniques. It provides a quantitative analysis of the matching results. However, it does not provide a detailed qualitative analysis. This research gap is filled in this thesis.

The Process Model Matching Contests (PMMCs) (Antunes et al., 2015; Cayoglu et al., 2013a) are the leading forum for the evaluation of process model matching techniques. However, these evaluation experiments assess the performance of matching techniques through a ranking. Consequently, they only provide limited information about the performance of matching techniques, e.g., detailed information about the strength and weaknesses of a matching technique are missing. However, the progress of process model matching techniques is strongly influenced by the available evaluation techniques. Important is that the evaluation techniques are "fair" and that they can be computed efficiently, i.e., without (intensive) manual labor.

In the next section, we discuss research questions which we address in this thesis.

1.3 Research Questions

In this section, we motivate and introduce the main research questions, which are addressed in this thesis.

1.3.1 Gold Standard Creation

The Oxford Dictionary[1] defines a *gold standard* as follows:

> *"A thing of superior quality which serves as a point of reference against which other things of its type may be compared."*

Translated into the domain of process model matching, a gold standard is defined as the optimal result of a "perfect" matcher. But which result is literally optimal is

[1] https://en.oxforddictionaries.com/

not that clear or conclusive as it may appear; in fact, this is highly subjective. We will elaborate on this in details below.

The gold standard, as of the state-of-the-art, has the following main weaknesses:

- In evaluation experiments at least three experts are required to define a gold standard, who then discuss about the correspondences which have to be included into the gold standard. The correspondences which are considered by only one annotator, or a minority of annotators, is not considered in the gold standard, and thus in the evaluation treated as a wrong correspondence. If three different domain experts are asked to create a gold standard, each expert will identify a different set of correspondences as correct. The organizers of the two Process Model Matching Contests (PMMC) 2013 and 2015 stated that there was not even a single pair of process models where two annotators fully agreed on (Antunes et al., 2015; Cayoglu et al., 2013a). This illustrates that it comes with high risk to define only a single set of correct correspondences. In other words, creating a gold standard is a highly subjective task. This is ignored currently. For example, if three experts congregate to discuss about a gold standard, then the resulting gold standard strongly depends on those three experts.

- The procedure to yield a gold standard is very time consuming, since the experts need to discuss and agree on each of the correspondences.

- Generally, the definition of a gold standard has a high effect on the evaluation of matching techniques. All correspondences which are not part of the gold standard are considered as wrong, thus negatively affecting the performance of matchers. A binary evaluation does not differentiate if a correspondence is totally unrelated (thus wrong) or if a correspondence is arguable but related. This leads to unclear evaluation results.

- There is a considerable loss of information when forcing the experts to agree on one single set of correspondences. All other correspondences are excluded from the gold standard. This information is lost in the gold standard.

Therefore, there is a need for a more fine-grained evaluation, which takes the arguability of correspondences into account. In the next section, we illustrate this in more detail with an example.

1.3.2 Absence of a "Perfect Match"

State-of-the-art evaluation procedures for process model matching techniques aim in assessing which of the correspondences identified by a matching technique are correct. While there seems to be no way to circumvent this basic assessment, there are nevertheless several problems attached to it. To illustrate these problems, consider the example depicted in Figure 1.2. It shows two simplified process models from the Process Model Matching Contest (PMMC) 2015 (Antunes et al., 2015). Possible correspondences are shown by gray markings.

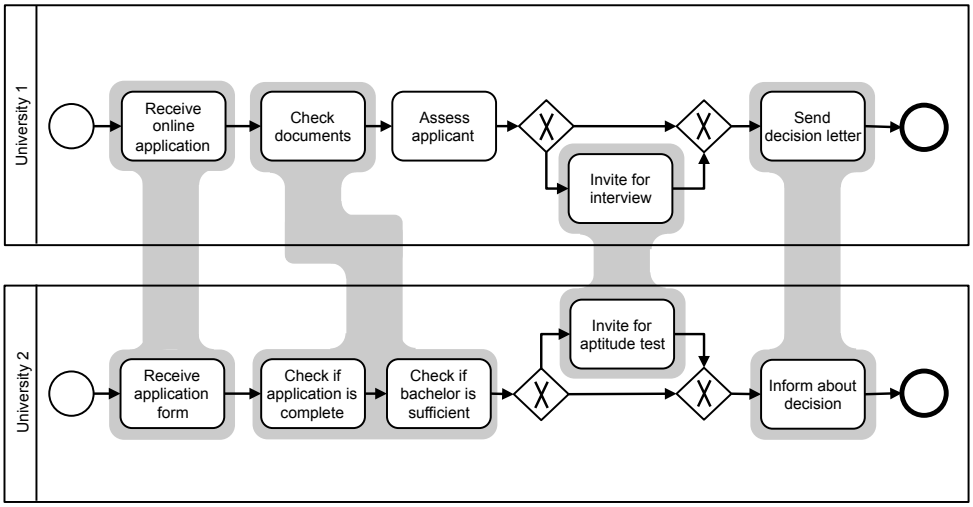


Figure 1.2: Two process models and possible correspondences as shown in Kuss et al. (2016)

Upon closer inspection of the correspondences shown in Figure 1.2, it becomes clear that, it is not that obvious to classify which correspondences are actually correct. While there are arguments in favor of many of the correspondences, there are also counter arguments in many cases. Consider, for instance, the correspondence between "*Receive online application*" from the first University and "*Receive application form*" in the process of the second University. On the one hand, we can argue that these activities do not correspond to each other because the former relates to an online procedure, whereas the second refers to a paper-based application. On the other hand, we can argue in favor of this correspondence because both activities deal with the receipt of an application document. Moreover, for the overall process, the concrete implementation of an activity is not important as long as the result of the activity is the same. Thus, it is disputable whether or not to accept

this correspondence as correct. There are similar arguments for matching "*Invite for interview*" on "*Invite for aptitude test*". An interview is clearly a different assessment instrument than an aptitude test, which makes the correspondence disputable. However, we can argue again that the result of the activity is the same with respect to the overall process. By performing one of these activities, relevant knowledge is acquired that helps to decide upon the suitability of the applicant. Similar arguments can be given in favor or against the other correspondences depicted in Figure 1.2. Consider, for example, the correspondence "*Send decision letter*" and "*Inform about decision*". On the one hand it can be argued in favor of this correspondence, because the activities share the same purpose. However, on the other hand, the activity "*Inform about decision*" does not specify how this is conducted. And therefore one can argue against this correspondence.

These examples illustrate that it may be hard and, in some cases, even impossible to agree on a single *correct* set of correspondences. Despite this, the evaluation of process model matching techniques currently depends on the definition of such a set of correct correspondences referred to as *gold standard*. This alignment is used in an evaluation context to distinguish between correct and incorrect correspondences given an alignment generated by a matching system. It finally is used for the computation of Precision, Recall, and F-Measure, which are the traditional measures used to evaluate process model matching techniques (cf. Antunes et al. (2015); Cayoglu et al. (2013a); Leopold et al. (2012); Weidlich et al. (2010a, 2013b)). Those binary evaluation measures are not always a suitable measure, since it does not fully account for the complexity of the matching task. The binary evaluation clearly states which correspondences are correct and which are not. But as we explained above, sometimes this is blurred. A binary gold standard, however, implies that any correspondence that is not part of the gold standard is incorrect and, thus, negatively affects the above mentioned metrics. This raises the question of why the performance of process model matching techniques is determined by referring to a single correct solution when human annotators may not even agree on what this correct solution is. To take the uncertainty in matching tasks into account, different evaluation metrics are required, which examine the performance of matchers from different perspectives.

1.3.3 How to Evaluate Process Model Matching Techniques?

In the previous section, we stated the problems associated with the definition of one single "perfect match". Another important question is the evaluation technique. Currently the evaluation of process model matching techniques almost exclusively relies on Precision, Recall and F-Measure. Those measures rely on such a binary gold standard, as described above. Moreover, the metrics do not provide information about strength and weaknesses of the matchers. However, an important feature of an efficient evaluation technique is to identify potential for improvement. Therefore, it is necessary to provide a detailed overview about specific strength and weaknesses of the matchers. Besides the limitation of a fair assessment of the matching techniques, in the Process Model Matching Contests the evaluation experiments only provide a grading and ranking of the participating matching techniques. The evaluation experiments do not aim in providing a detailed analysis about specific strength and weaknesses of matching techniques and, therefore, do not aim in providing a feedback about possibilities for improvement. To provide such feedback, currently it is necessary to manually process and interpret the matcher output. One central research question in this thesis is therefore how to automatize this expensive task.

Moreover, one central research question to achieve deeper insights about the performance of matchers is, how to determine what defines the complexity of a matching task. This is necessary to determine which matchers are able to perform well on complex data sets. In the current literature, complexity of process model matching tasks is associated with the different level of granularity of the process models, thus resulting in $1 : m$ or $n : m$ correspondences (cf. Antunes et al. (2015); Makni et al. (2015); Weidlich et al. (2010b)). Although this is one aspect which makes a matching task complex, this is not the only one. In fact, complexity is for instance associated with the extent of syntactic overlap of the activities to be matched. If we look into details, we can observe that $1 : m$ correspondences mostly have a low syntactic overlap, since the same activity is expressed in a different level of granularity, thus different ways.

Therefore, we propose an evaluation procedure, which classifies the matching task and matching results into different levels of complexity.

1.4 Contribution

This thesis contributes to the body of knowledge in several ways. The most significant contributions of this thesis can be summarized as follows:

- We propose a new approach towards a gold standard. We name it non-binary gold standard. To obtain this non-binary gold standard, experts individually develop their own gold standard. Each assignment among the experts' choices is then treated as a vote for the individual correspondence. In this way, each correspondence is assigned with a specific so-called support value. As a result a non-binary gold standard is derived. The proposed non-binary gold standard has the following three main advantages: First, the introduced non-binary gold standard incorporates the uncertainty of the correspondences, in contrast to the state-of-the-art gold-standard. In hardly any practical case "the true" gold standard is achievable. To define a single set of correct correspondences strongly depends on the point of view of the annotators, but even more on the application. Therefore, relying on a binary gold standard does not account for the true complexity of a matching task. Moreover, it also does not account for the subjectivity, which is associated with such a task. Second, it is not necessary to fully agree on one single gold standard. This avoids the sophisticated task to agree on each correspondence, which is actually only feasible for a small group of annotators. Third, because it is no longer necessary to discuss about each correspondence among the annotators, a higher number of annotators can contribute to the definition of a non-binary gold standard. This higher number of annotators additionally may increase the quality of the non-binary gold standard. The approach is presented in Section 5.1.

- We introduce a new evaluation procedure which fairer assesses the performance of matchers, since it takes the arguability of correspondences into account. We adapt Precision, Recall and F-Measure (ProP, ProR and ProFM), to allow non-binary values, derived from the non-binary gold standard. Furthermore, we introduce a new distance-based measure (ReD), which complements the metrics from Information Retrieval. These metrics are presented in Section 5.2 and Section 5.3.

- We develop Bounded versions of ProP, ProR and ProFM, which are adapted to exclude values below a specific threshold in the non-binary gold standard. The non-binary evaluation identifies characteristics of a matcher to derive optimal matching scenarios for the matchers. In this way, we can identify matchers which focus on finding correspondences with a high support value in the data set. On the one hand such high-support correspondences are the most "sure" correspondences, however on the other hand such correspondences are also often rather obvious. For instance, "trivial" correspondences mostly have a high support value in the non-binary gold standard. These Bounded versions of ProP, ProR and ProFM are presented in Section 5.2.

- We introduce a fully non-binary evaluation procedure. In this evaluation procedure, we consider the confidence values of the matchers, but not with their absolute values. Instead, the matcher output as well as the non-binary gold standard are transformed into a ranked collection of correspondences. Then, the confidence and support values are compared with respect to this ranking through a ranking-based correlation. In this way, it can also be assessed if the confidence values of the matchers reflect a realistic confidence of the correspondences. This evaluation method is presented in Chapter 6.

- We propose a conceptually novel evaluation method, which is a category-dependent evaluation via matching patterns. The idea is to automatically divide the matching task as well as the matcher output into categories with different complexity levels. Then standard metrics, like Precision, Recall and F-Measure, can be applied to each of the categories separately. We further compute the false-positive and false-negative alignments for each of the categories. In this way, we obtain an in-depth evaluation providing detailed information about the computed correspondences, where no manually processing of the matcher output is required. This category-dependent evaluation better reveals strength and weaknesses of a matcher. The evaluation procedure further allows to tune matchers to specific applications. The evaluation via matching patterns further allows for an assessment of the data set: it informs about the complexity of the matching task through the identification of the complexity and fraction of correspondences of a data set. Moreover, the quality of the gold standard can be assessed indirectly, e.g., quality and quantity of manual annotations. This is introduced in Chapter 7.

- We provide synthetic test cases, which complement the above described matching patterns, with attributes which cannot be assigned automatically. We furthermore provide an evaluation platform, where all metrics, introduced in this thesis, can be accessed. The synthetic data set and "Evaluation Portal" are described in Section 8.2.1 and Section 8.2.2.
- We apply the concepts and metrics introduced in this thesis at the OAEI 2016 and 2017 (Achichi et al., 2016, 2017).

Some of the work presented in this thesis has already been published:

- Kuss, Leopold, Van der Aa, Stuckenschmidt and Reijers: A probabilistic evaluation procedure for process model matching techniques. Data & Knowledge Engineering, 2018
- Kuss, Leopold, Meilicke and Stuckenschmidt: Ranking-based evaluation of process model matching. On the Move to Meaningful Internet Systems. OTM 2017 Conferences: Confederated International Conferences: CoopIS 2017
- Kuss and Stuckenschmidt: Automatic classification to matching patterns for process model matching evaluation. ER-Forum-Demos 2017
- Achichi et al.: Results of the Ontology Alignment Evaluation Initiative 2016. International Workshop on Ontology Matching co-located with the 16th International Semantic Web Conference (ISWC 2017)
- Kuss, Leopold, Van der Aa, Stuckenschmidt and Reijers: Probabilistic evaluation of process model matching techniques. ER 2016
- Achichi et al.: Results of the Ontology Alignment Evaluation Initiative 2016. International Workshop on Ontology Matching co-located with the 15th International Semantic Web Conference (ISWC 2016)
- Antunes et al.: The Process Model Matching Contest 2015. GI-Edition: Lecture Notes in Informatics, 2015

Publication which is not subject of the thesis:

- Meilicke, Leopold, Kuss, Stuckenschmidt and Reijers: Overcoming individual process model matcher weaknesses using ensemble matching. Decision Support Systems, 2017

1.5 Thesis Outline

The remainder of the thesis is organized as follows:

- Chapter 2 *Background and Basic Definitions*: introduces to basic notions and definitions, which we use as basis in this thesis. Moreover, we introduce the most common metrics from Information Retrieval, which we refer to in our experiments in this thesis.
- Chapter 3 *Process Model Matching Contests*: in this chapter we present the results of the Process Model Matching Contest 2015 and compare those results to the results of the first Process Model Matching Contest 2013. Furthermore, in this chapter the data sets, which we refer to in this thesis, are described.
- Chapter 4 *Related Work*: discusses related work in the field of process model matching evaluation and the evaluation of related fields like schema matching and ontology matching.
- Chapter 5 *Probabilistic Evaluation*: in this chapter, the non-binary gold standard is introduced. Moreover, we present the novel evaluation metrics, which are based on the non-binary gold standard.
- Chapter 6 *Ranking-based Evaluation*: in this chapter, we introduce the completely non-binary evaluation procedure, where we use the confidence values, given by some matchers, to calculate the Spearman's rank correlation.
- Chapter 7 *Evaluation by Automatic Classification to Matching Patterns*: we present our conceptually new evaluation procedure, where all correspondences of the data set as well as matcher output are classified into different categories, depending on the complexity level.
- Chapter 8 *Summary, Conclusions and Outlook*: in this chapter, we summarize the main results of the thesis. Moreover, we give a comprehensive outlook of future research and present first results. Furthermore, we introduce an evaluation portal, where the metrics can be accessed.

2

Background and Basic Definitions

In this chapter, we provide an overview of the basic notions and definitions, which we refer to in this thesis. These definitions serve as a basis of our own definitions, which we state in the corresponding chapters. Moreover, we provide a brief introduction to the most common metrics in the field of Information Retrieval, which are commonly used in state-of-the-art evaluation experiments.

The chapter is organized as follows. Section 2.1 discusses the basic definitions in the context of business process modeling and process model matching, to lay the foundation for business model matching evaluation. Section 2.2 discusses basic notions in the field of Information Retrieval. Those notions are widely used in evaluation experiments of process model matching techniques and related fields like ontology matching and schema matching.

2.1 Basic Notions and Definitions

In the first chapter, we introduced the field of business process modeling and the matching of such process models. In the following, we provide a formal definition of the basic terms which we refer to in this thesis.

Based on the definition by Klinkmüller et al. (2014), we define a process model, and the corresponding set of activities, as follows:

Definition 1 (Process model, set of activities)**.** *Let $\mathcal{L}$ be a set of labels and $\mathcal{T}$ be a set of events. Then a process model $\mathcal{P}$ is a tuple (N, E, λ, τ), in which:*

- *N is the set of nodes;*
- *$E \subseteq N \times N$ is the set of edges;*
- *$\lambda : N \to \mathcal{L}$ is a function that maps nodes to labels; and*
- *$\tau : N \to \mathcal{T}$ is a function that assigns types to nodes,*

and which satisfies $\forall a \in act(\mathcal{P}) = \{a \,|\, a \in N \wedge \tau(a) = \text{ activity}\}$

$$|\{n \,|\, n \in N,\, (a,n) \in E\}| \;\leq\; 1 \qquad \textit{and} \tag{2.1}$$

$$|\{n \,|\, n \in N,\, (n,a) \in E\}| \;\leq\; 1. \tag{2.2}$$

Set $act(\mathcal{P})$, also denoted by A, is called the set of activities for process model $\mathcal{P}$.

The definition of a process model involves the set of events $\mathcal{T}$. Possible events in a process model depend on the notation/format of the process models. Examples are “and”, “or”, “xor”. In our case, mainly the activities are relevant types of events for our considerations. Currently, the matching of process models is mostly based on a comparison of the label strings of the activities in the process models to be matched. Examples for such labels are “Check application” or “Register child”.

For a process model $\mathcal{P}$, we require that each node a has at most one control flow edge originating from a, as ensured by condition (2.1). Similarly, (2.2) ensures that each activity node a has at most one control flow edge into the node a.

Definition 2 (Process Model Matching)**.** *Given two Process Models $\mathcal{P}_1$ and $\mathcal{P}_2$, the goal of process model matching is to automatically identify pairs of equivalent or similar activities from $act(\mathcal{P}_1) \times act(\mathcal{P}_2)$.*

By definition, process model matching aims at automatically identifying which activities in the process model describe an equal or similar behavior/task. Hence, activities from $\mathcal{P}_1$ and $\mathcal{P}_2$, which relate to one of these tasks, are matched with activities which describe the same or a similar task. Such pairs of activities are also called *correspondences* or *alignments*.

Matching techniques, which automatically identify such correspondences in process models, are also called matchers. Mostly the matching of process models is based on the labels strings of the activities. However, some matchers also take structural or background information into account. The result of a matching is called matcher output, which we formally define in the following:

Definition 3 (Matcher output)**.** *For two process models P_1 and P_2 with their activity sets A_1 and A_2, a matcher output O is a subset of all possible alignments, i.e., $O \subseteq A_1 \times A_2$.*

A matcher output is not just any random subset of possible alignments. Rather, the goal of a matcher is to identify activities which describe the same or a similar task. This is reflected in the matcher output O.

In this thesis, we are not only interested in these activity pairs, but also in the confidence that the two activities are matched correctly. For that reason, we define a confidence of an alignment or correspondence between $\mathcal{P}_1$ and $\mathcal{P}_2$ as a function

$$\mathcal{A} : act(\mathcal{P}_1) \times act(\mathcal{P}_2) \rightarrow [0, 1].$$

We refer to $\mathcal{A}(a_1, a_2)$ as the confidence of correspondence (a_1, a_2) from the set $act(\mathcal{P}_1) \times act(\mathcal{P}_2)$.

In the following, we distinguish between two types of alignments:

1. **A binary alignment** is an alignment that uses only two different confidence values, i.e., $\mathcal{A}(a_1, a_2) \in \{0, \alpha\}$ for all $(a_1, a_2) \in act(\mathcal{P}_1) \times act(\mathcal{P}_2)$ and some $\alpha > 0$. Typically α is set to 1. Therefore, a binary alignment only distinguishes between (probably) correct and (probably) incorrect correspondences.

2. **A non-binary alignment** is an alignment that uses more than two values from the range $[0, 1]$. It can thus be used to assign confidence scores for ordering the correspondences on an ordinal scale instead of distinguishing

only between correct and incorrect. We define such an assignment in Section 5.1.

Binary as well as non-binary alignments can be created by human experts, by matching techniques, or by a combination of manual effort and automated matching techniques. Therefore, this distinction holds for correspondences, created by matching techniques (matcher output) as well as manually generated human assessments (gold standard). In all of these cases, confidence scores can be used to express in how far one should trust in the correctness of a generated correspondence. However, most of these approaches do not associate a clear probabilistic meaning to a specific value within a non-binary correspondence. This means, for example, that we cannot assume that a correspondence with a confidence of 1.0 will be correct for sure nor can we assume that a confidence score of 0.5 means that the probability that the correspondence is correct is exactly 50%. Nevertheless, all approaches have in common that a higher confidence value is intended to refer to a higher probability for being correct.

Based on the previous discussion, we define a binary gold standard as follows:

Definition 4 (Binary Gold standard)**.** *Let A_1 and A_2 be the sets of activities of two process models $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively. Then, a binary human assessment can be captured by the subset $H \subseteq A_1 \times A_2$ and the confidence function $\mathcal{A} : act(\mathcal{P}_1) \times act(\mathcal{P}_2) \to \{0, 1\}$ with $\mathcal{A}(a_1, a_2) = 1$ for all $(a_1, a_2) \in H$ and 0 otherwise. Each element $(a_1, a_2) \in H$ specifies that a human assessor considers the activity a_1 to correspond to the activity a_2. Such a binary human assessment is also called gold standard or reference alignment.*

Note two specific details related to this definition. First, Definition 4 also allows for one-to-many $(1 : m)$ and many-to-many $(n : m)$ relationships. If, for instance, the elements (a_1, a_2) and (a_1, a_3) are both part of H, then there exists a one-to-many relationship between the activity a_1 and the two activities a_2 and a_3. The advantage of capturing a complex correspondence based on several elementary correspondences is that the matching technique is not required to identify the entire complex correspondence. If it, for instance, identifies (a_1, a_2) but not (a_1, a_3), it would at least get credit for having identified (a_1, a_2). Second, the information that is available for deciding about a possible correspondence may vary from model to model. In general, we assume that the decision will be mainly based on the labels. If available, however, also data objects can provide valuable input.

Such a binary gold standard is derived from an undefined number of annotators, resulting in one single gold standard, which the annotators have to agree on. Based on the definition of a binary gold standard, in Section 5.1, we define a non-binary gold standard in which we allow the assignment of non-binary values to the correspondences of the gold standard. This is an important generalization of the definition of a binary gold standard, which lies the necessary foundation for our proposed evaluation procedures.

In the evaluation experiments in this thesis, the task of a matcher is to match the process models pair-wise. The result of the matchers is then compared to a manually generated gold standard. The task of a matcher is to identify semantically similar alignments, i.e., to identify the correspondences of the gold standard. Because it is rarely possible to reach a perfect matching, i.e., $O = H$, the matcher output needs to be evaluated. In the next section, we briefly introduce a selection of evaluation metrics, from the field of Information Retrieval, which we refer to in the evaluation experiments in this thesis.

2.2 Introduction to Metrics from Information Retrieval

In process model matching the metrics from Information Retrieval are widely used for evaluation experiments (Van Rijsbergen, 1979). Researchers in this field compare their matchers to the state-of-the-art by such metrics. This is conducted similarly in related fields like schema matching and ontology matching (Do et al., 2002; Euzenat et al., 2011). The advantage of those measures is that they are easy to compute and the results can be intuitively interpreted.

For the calculations of Precision, Recall and F-Measure, the matcher output is compared to a manually generated reference alignment, also called gold standard. Comparing the correspondences computed by a matcher to a manually generated gold standard, then each activity is classified into one of the following four attributes, with respect to the specific gold standard:

1. true-positive (TP), which are correctly computed correspondences;
2. true-negative (TN), which are correctly not-computed alignments;
3. false-positive (FP), which are correspondences which are computed but not correct;

4. false-negative (FN), which are correspondences which are correct, but not computed by a matcher.

By definition, those classifications depend on the choice of the particular gold standard. Then, the following formulas define Precision, Recall, F-Measure and Accuracy:

$$\begin{aligned}
\text{Precision:} \quad & \frac{TP}{TP + FP} \\
\text{Recall:} \quad & \frac{TP}{TP + FN} \\
\text{F-Measure:} \quad & 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \\
\text{Accuracy:} \quad & \frac{TP + TN}{TP + TN + FP + FN}\,.
\end{aligned}$$

Therefore, Precision states the fraction of alignments, which are correct on the whole matcher output. Recall states the fraction of correct alignments of the gold standard, which are computed by the matcher. The F-Measure, is the harmonic mean of Precision and Recall. The Accuracy states the fraction of correct classified correspondences, compared to all possible correspondences. The significance of the information content of the Accuracy measure is rather small for process model matching evaluation, since the Accuracy is a function of the number of true-negative (TN) correspondences. Typically, in big data sets there is a huge number of true-negative alignments. (This leads to a high absolute number of true-negative alignments). Therefore, the fraction of the relevant alignments on the total data set is low, leading to small differences between the matchers. Consequently, the information content of the Accuracy measure is very limited for the comparison of process model matching techniques. Even weak matchers typically achieve an accuracy of more than 90%.

The above described metrics can be either computed at the micro or the macro average. This distinction has its source in the structure of the matching task. Some data sets consist of a collection of process models, i.e., they contain different test cases. At the macro average, the metrics are computed pair-wise within a test case, then the results are averaged. In contrast, for the micro average, the metrics are calculated for the union of the correspondences of the entire data set.

Macro average may result in inconsistencies, in particular if a test case is small. In this case it is possible that empty test cases exist, which can result in an inaccurate measurement. To make this clearer, see the following example, taken from the gold standard of the Process Model Matching Contest 2015 (Antunes et al., 2015) of the University Admission data set of the touple "*University Frankfurt - University Hohenheim*":

The gold standard of this test case consists of two correspondences:

- c_1: *Wait for results – Waiting for the response*
- c_2: *Rejected – Rejection*

There are only two correct correspondences. If a matcher does not detect any of these two correspondences, then the calculation of Precision is undefined. In this small example, where a matcher computes an empty test case, this means for the Precision: $\frac{0}{0+0}$. The result for this is undefined. That means actually there is no result for this test case. One can argue in favor to treat such cases as Precision of 1.0 since the matcher did not compute any correspondence at all, thus no incorrect correspondence is computed. This is very intuitive and was conducted this way in the Process Model Matching Contest 2015 (Antunes et al., 2015). However, on the other hand one can argue that a Precision of 1.0 is not valid since the matcher did not compute any correspondence at all. This shows that such empty alignments lead to unclear results which can be interpreted differently. In the data sets which we use for our evaluations, empty alignments for specific test cases of matchers and gold standards do occur. To avoid such inaccuracies and inconsistencies we always refer to the micro values for the rankings in all evaluation experiments of this thesis. We only use macro values to compare the values to the Process Model Matching Contests (Antunes et al., 2015; Cayoglu et al., 2013a).

In Information Retrieval, additional measures are introduced, e.g., a variety of different F-Measures with a differing weight of Precision and Recall. For more information, we refer the interested reader to Fawcett (2006) and Manning et al. (2008).

We discuss in Chapter 4 a selection of further measures for evaluation experiments.

In the next chapter, we introduce the Process Model Matching Contests where those metrics are used to conduct evaluation experiments.

3

Process Model Matching Contests (PMMCs)

In this chapter, we describe the results of the Process Model Matching Contest 2015 (Antunes et al., 2015). We aim to draw conclusions for future evaluation experiments. Moreover, we compare the results of the Process Model Matching Contest 2015 to the results of the contest 2013 (Cayoglu et al., 2013a), to deduce the improvement of process model matching techniques in this time interval.

Moreover, we describe the data sets which are used in this context and which we always take as basis for our evaluation experiments, which we conduct in this thesis.

The chapter is organized as follows, Section 3.1 introduces the data sets of the Process Model Matching Contest 2015 and compares them to the setting of the 2013 contest edition. In Section 3.2 the results of the matchers are presented. While

Section 3.3 compares the results of 2015 to the contest of 2013, Section 3.4 states conclusions which we can draw for future evaluation strategies.

3.1 Data Sets of the Process Model Matching Contest

In 2013, the first Process Model Matching contest was conducted. The idea was to deliver a common basis for evaluation and to indicate the improvement process of the matching techniques. The 2015 contest was part of the EMISA Workshop (Kolb et al., 2015) and included three different data sets (Antunes et al., 2015). Table 3.1 provides a comparison of the data sets which were used in the 2013 and 2015 edition of the contest. As we can see, the two contests share only one data set. (The differences of both University Admission data sets are explained below.)

	PMMC 2013	**PMMC 2015**
University Admission (UA)	x	x (modified)
University Admission Sub (UAS)	–	x
Birth Registration (BR)	x	x
Asset Management (AM)	–	x

Table 3.1: Data sets of the PMMC 2015 (Antunes et al., 2015) and 2013 (Cayoglu et al., 2013a)

The data sets differ notably, on the one hand, due to the different formats. On the other hand due to their content. The models cover different issues of process modeling.

In the following, we give an overview of the three data sets and present an example for each data set.

3.1.1 University Admission Data Set (UA)

The first data set is the University Admission data set (Figure 3.1). This data set consists of nine admission processes of different German universities. The business process models are in BPMN-format with English text and were created by graduate students at the Humboldt University Berlin. In the Process Model Matching Contest 2013, the data set was in Petri-Nets; later it was transformed into BPMN. Moreover, the gold standard was improved compared to the 2013 version. In 2015, two gold standards were used. The first gold standard contains only activity equiv-

alence, where the activities are classified as equivalent. The second gold standard also included activity subsumptions. This means that one activity is a subsumption of the corresponding activity. The data set contains abbreviations which are specific for the described processes. For instance, the abbreviation “GPA” stands for “Grade Point Average”. This is a grading scale to rank students according to their qualifications. A similar abbreviation is used in the second data set, the Birth Registration data set. However, in the context of the Birth Registration data set the abbreviation “GBA” has a different meaning. This illustrates one difficulty associated with process model matching, where abbreviations may be used and are context-dependent in their meaning.

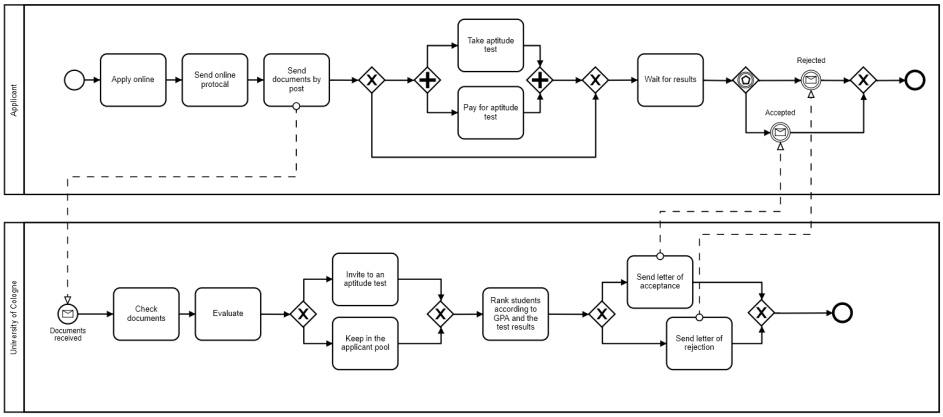


Figure 3.1: Example of a (small) process model of the University Admission data set

3.1.2 Birth Registration Data Set (BR)

The Birth Registration data set contains nine birth registration processes from Germany, Russia, South Africa, and the Netherlands. The business process models are in Petri-Nets and contain Englisch text. The models were created by graduate students at the Humboldt University Berlin and in the context of a process analysis in Dutch municipalities. It can be observed that the data set contains inexact language combinations, as well as dutch abbreviations. One example is again the abbreviation “GBA”, which is again used in a process model label. However, in this context “GBA” stands for “Gemeentelijke Basisadministratie Persoonsgegevens”, which is a local residents registration office. Abbreviations like this are especially difficult to detect for matching techniques, since the models are supposed to be in

English, but have to detect Dutch abbreviations. Moreover, the transitions contain label like “t3” and “t9”. This is also the case for some “places”, like for example “p6” or “p13” in our example (Figure 3.2). Therefore, a suitable match of such labels can only be performed if matchers take structural dependencies into account.

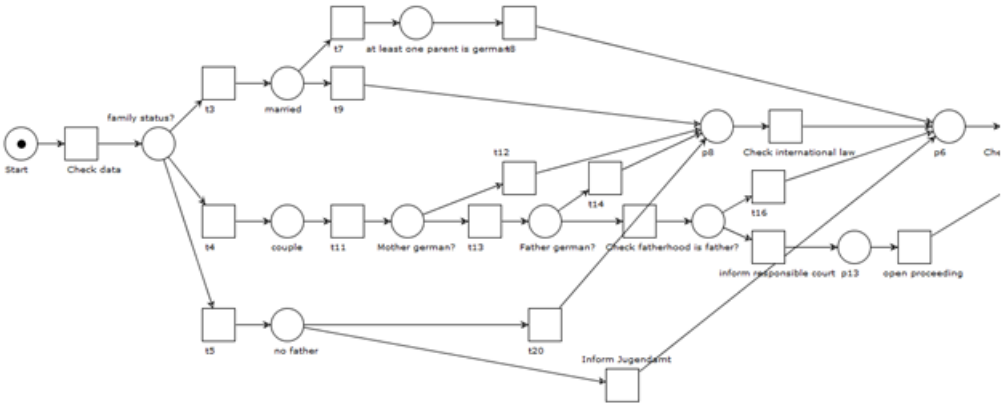


Figure 3.2: Example of a cutout of a process model of the Birth Registration data set

3.1.3 Asset Management Data Set (AM)

The Asset Management data set was firstly introduced in the 2015 process matching contest. The data set consists of 36 model pairs, derived from 72 models from the SAP Reference Model Collection. The process models cover different aspects from the area of finance and accounting. They are EPC models in EPML format, with English text. The data set is very specific due to the high amount of technical terms. The matchers need to have knowledge about those special technical terms. Some examples of such terms and specific abbreviations are covered in the example in Figure 3.3.

3.1.4 Characteristics of the Three Data Sets

Table 3.2 summarizes some characteristics of the three data sets with its corresponding four gold standards. It states the size of the process models as well as the different level of granularity, by stating the minimal and maximal number of activities and the number of $1 : m$ correspondences. Such correspondences are difficult

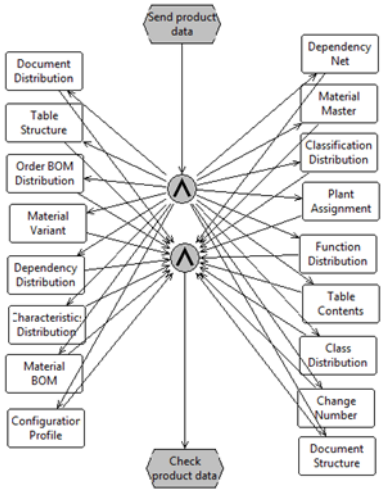


Figure 3.3: Example process model of the Asset Management data set

Characteristic	**UA**	**UA$_S$**	**BR**	**AM**
No. of Activities (min)	12	12	9	1
No. of Activities (max)	45	45	25	43
No. of Activities (avg)	24.2	24.2	17.9	18.6
No. of 1:1 Correspondences (total)	202	268	156	140
No. of 1:1 Correspondences (avg)	5.6	7.4	4.3	3.8
No. of 1:m Correspondences (total)	30	360	427	82
No. of 1:m Correspondences (avg)	0.8	10	11.9	2.3

Table 3.2: Characteristics of the test data sets of the PMMC 2015 (Antunes et al., 2015)

to detect, since they describe either a subset or a subsumption of the corresponding activity. We can see from the numbers in Table 3.2, that the Birth Registration data set consists of a high number of $1:m$ correspondences. This additionally increases the complexity of the data set, besides the characteristics which we state above. As we can see in Table 3.2, the number of $1:m$ correspondences increases for the gold standard which includes subsumptions. However, in some process model pairs the $1:m$ correspondences are actually no subsumptions, but due to different level of granularity.

In the following, we give one example for such an $1:m$ (1:2) correspondence for the pair "*University Cologne*" and "*IIS Erlangen*":

- "*Check Application*" – "*Check application in time*"
- "*Check Application*" – "*Check application complete*"

This example illustrates an equivalence correspondence, because the same activity is described in different level of detail, but describes the same process. Therefore, the 1:2 correspondence results from different level of granularity in the two process models. In fact, it is not an actual subsumption.

However, the data sets also contain subsumptions. An example for such a subsumption, in this case a $1:3$ subsumption, are the correspondences:

- "*Rank with applicants*" – "*Sum scores*"
- "*Rank with applicants*" – "*Reject application*"
- "*Rank with applicants*" – "*Accept application*"

In this case, the subsumption correspondence do not describe an equal activity of the application process, but one activity is a subsumption of the other activity. Both kind of $1:m$ correspondences can be found in the data sets.

3.2 Results of the PMMC 2015

In the experiments of the contest the gold standards of the evaluation experiments were publicly available, except of the gold standard of the Asset Management data set. In the results (Tables 3.3 – 3.6), we state the micro- as well as the macro-values of Precision, Recall and F-Measure as described in Section 2.2. The reason is that in the 2013 edition of the contest only macro-values were computed. Therefore, to compare the results, we need to compare the macro-values of Precision, Recall and F-Measure. The best results for each metric in each data set are always highlighted in bold.

The results for the University Admission data set (Table 3.3) illustrate a high diversity of the quality of the matching results. The best F-Measure (micro-average) results are obtained by the RMM/NHCM (0.668), RMM/NLM (0.636) and MSSS (0.608).

For the University Admission data set, a second gold standard was used which included subsumptions to the gold standard. The results are shown in Table 3.4. Again the matcher RMM/NHCM achieves the best F-Measure (micro average of

Approach	Precision			Recall			F-Measure		
	Ø-mic	Ø-mac	SD	Ø-mic	Ø-mac	SD	Ø-mic	Ø-mac	SD
RMM/NHCM	.686	.597	.248	.651	.61	.277	**.668**	.566	.224
RMM/NLM	.768	.673	.261	.543	.466	.279	.636	.509	.236
MSSS	**.807**	**.855**	.232	.487	.343	.353	.608	.378	.343
OPBOT	.598	.636	.335	.603	.623	.312	.601	**.603**	.3
KMSSS	.513	.386	.32	.578	.402	.357	.544	.374	.305
RMM/SMSL	.511	.445	.239	.578	.578	.336	.543	.477	.253
TripleS	.487	.685	.329	.483	.297	.361	.485	.249	.278
BPLangMatch	.365	.291	.229	.435	.314	.265	.397	.295	.236
KnoMa-Proc	.337	.223	.282	.474	.292	.329	.394	.243	.285
AML-PM	.269	.25	.205	**.672**	**.626**	.319	.385	.341	.236
RMM/VM2	.214	.186	.227	.466	.332	.283	.293	.227	.246
pPalm-DS	.162	.125	.157	.578	.381	.38	.253	.18	.209

Table 3.3: Results of University Admission data set

Approach	Precision			Recall			F-Measure		
	Ø-mic	Ø-mac	SD	Ø-mic	Ø-mac	SD	Ø-mic	Ø-mac	SD
RMM/NHCM	**.855**	**.82**	.194	.308	.326	.282	**.452**	**.424**	.253
OPBOT	.744	.776	.249	.285	.3	.254	.412	.389	.239
RMM/SMSL	.645	.713	.263	.277	.283	.217	.387	.36	.205
KMSSS	.64	.667	.252	.273	.289	.299	.383	.336	.235
AML-PM	.385	.403	.2	**.365**	**.378**	.273	.375	.363	.22
KnoMa-Proc	.528	.517	.296	.282	.281	.278	.367	.319	.25
BPLangMatch	.545	.495	.21	.247	.256	.228	.34	.316	.209
RMM/NLM	.787	.68	.267	.211	.229	.308	.333	.286	.299
MSSS	.829	.862	.233	.19	.212	.312	.309	.255	.318
TripleS	.543	.716	.307	.205	.224	.336	.297	.217	.284
RMM/VM2	.327	.317	.209	.27	.278	.248	.296	.284	.226
pPalm-DS	.233	.273	.163	.316	.328	.302	.268	.25	.184

Table 3.4: Results of University Admission data set with subsumption

0.452), however the results decrease considerably, due to a strong decrease of Recall, which is halved, and at the same time a weak increase of Precision.

The results of the Birth Registration data set are given in Table 3.5. The matchers results are not as good as the results obtained for the University Admission data set. The best matcher (OPBOT) achieves a micro F-Measure of 0.565. The reason may be a higher complexity level of the Birth Registration data set. However, it may be also an issue of the quality of the gold standard. For the Asset Management data set (Table 3.6), the best results are achieved by AML-PM (micro F-Measure of 0.677). The results show that no matching technique has a high performance on all tested data sets.

	Precision			Recall			F-Measure		
Approach	∅-mic	∅-mac	SD	∅-mic	∅-mac	SD	∅-mic	∅-mac	SD
OPBOT	.713	.679	.184	**.468**	**.474**	.239	**.565**	**.54**	.216
pPalm-DS	.502	.499	.172	.422	.429	.245	.459	.426	.187
RMM/NHCM	.727	.715	.197	.333	.325	.189	.456	.416	.175
RMM/VM2	.474	.44	.2	.4	.397	.241	.433	.404	.21
BPLangMatch	.645	.558	.205	.309	.297	.22	.418	.369	.221
AML-PM	.423	.402	.168	.365	.366	.186	.392	.367	.164
KMSSS	.8	.768	.238	.254	.237	.238	.385	.313	.254
RMM/SMSL	.508	.499	.151	.309	.305	.233	.384	.342	.178
TripleS	.613	.553	.26	.28	.265	.264	.384	.306	.237
MSSS	**.922**	**.972**	.057	.202	.177	.223	.332	.244	.261
RMM/NLM	.859	.948	.096	.189	.164	.211	.309	.225	.244
KnoMa-Proc	.234	.217	.188	.297	.278	.234	.262	.237	.205

Table 3.5: Results of Birth Registration data set

	Precision			Recall			F-Measure		
Approach	∅-mic	∅-mac	SD	∅-mic	∅-mac	SD	∅-mic	∅-mac	SD
AML-PM	.786	.664	.408	.595	**.635**	.407	**.677**	.48	.422
RMM/NHCM	.957	.887	.314	.505	.521	.422	.661	.485	.426
RMM/NLM	**.991**	**.998**	.012	.486	.492	.436	.653	**.531**	.438
BPLangMatch	.758	.567	.436	.563	.612	.389	.646	.475	.402
OPBOT	.662	.695	.379	**.617**	.634	.409	.639	.514	.403
MSSS	.897	.979	.079	.473	.486	.432	.619	.519	.429
RMM/VM2	.676	.621	.376	.545	.6	.386	.603	.454	.384
KMSSS	.643	.834	.282	.527	.532	.417	.579	.482	.382
TripleS	.614	.814	.261	.545	.546	.434	.578	.481	.389
pPalm-DS	.394	.724	.348	.595	.615	.431	.474	.451	.376
KnoMa-Proc	.271	.421	.383	.514	.556	.42	.355	.268	.279
RMM/SMSL	.722	.84	.307	.234	.37	.366	.354	.333	.327

Table 3.6: Results of Asset Management data set

3.3 Comparison of the Results of the PMMC 2015 to the Results of the PMMC 2013

In the following, we summarize the results of the 2015 contest and compare it with the results of the 2013 contest. The 2013 contest took place as part of the "4th International Workshop on Process Model Collections: Management and Reuse" (Cayoglu et al., 2013a). We want to learn in how far we can observe a progress of the matching techniques, in average but also compared to the best performances in 2013 and 2015.

To directly compare the results of 2013 to 2015, the setting has to be the same. This is not the case for those two PMMCs. The only data set which is unmodified

compared to the 2013 edition is the Birth Registration data set. Therefore, we use this data set for our comparisons.

In 2013, the best results were achieved by RefMod-Mine/NHCM with an macro average F-Measure of 0.45. (Note that we compare the macro values, because in the 2013 edition, only the macro values were computed.) Three matchers could not outperform those results from 2013: pPalm-DS (0.426), RMM/NHCM (0.416), and RMM/VM2 (0.402). In 2015, the best matcher on this data set is the matcher OPBOT with macro average F-Measure of 0.54. This is a significant improvement compared to 2013. However, the OPBOT did not participate in 2013. Therefore, it might be more telling to compare the average results of the participating matchers of 2013 to the average participating matchers of 2015 (which is ≈0.35) in 2015 and average approach in 2013 (≈0.29). This indicates a small progress. However, as we indicate in the previous section, the Birth Registration data set is rather special in its characteristics. Therefore, it is difficult to draw a final conclusion about the progress of the matching systems only from this data set.

To compare the results even though the data sets differ, we can compare the average and best F-Measure for the matching techniques. Table 3.7 provides those information. We compare the two data sets which are used in the 2013 edition of the contest, even though the University Admission data set has been modified, compared to 2013. Thus, the comparison is only a hint about the improvement over the past two years. Again, we compare the macro-values of F-Measure, due to the missing micro-values in the contest 2013. As we can see, the results indicate a limited progress from 2013 to 2015 with regard to the average results. However, for the results of the best matchers, we observe a stronger increase of the macro F-Measure.

	UA	**BR**
Average Result 2013 (FM)	.30	.29
Best Result 2013 (FM)	.41	.45
Average Result 2015 (FM)	.37	.35
Best Result 2015 (FM)	.57	.54

Table 3.7: Avg and max results of the PMMC 2015 (Antunes et al., 2015) compared to 2013 (Cayoglu et al., 2013a)

3.4 Conclusions

In this chapter, we provided an introduction to the data sets of the Process Model Matching Contest 2015, which are a running example for our experiments in this thesis.

As we can observe from the results, most matchers aim in a high Precision and therefore miss a considerable amount of correspondences. To understand which correspondences are especially challenging for the matchers, it would be necessary to manually process the matcher output. The experiments from the PMMCs do not provide further information about the individual performance of matchers.

Only OPBOT and RMM/NHCM have a balanced Precision and Recall. Moreover, only OPBOT and RMM/NHCM achieve rather good results for all data sets in the 2015 edition. To include the subsumptions into the gold standard leads to a strong decrease of Recall, with a small increase of Precision for some matchers, since most of the subsumption correspondences are not computed by the matchers. However, in the gold standard of the matching contest they do not always resemble real subsumptions, partially they are actually equivalence correspondences which results from different level of granularity of the process models. Therefore, the information content of this test data set is questionable, since it is a mixture of subsumptions and equivalence correspondences, which should be actually part of the "main" gold standard of the data set. This also explains the small increase of Precision of some matchers.

Moreover, we compared the results of the Process Model Matching Contest 2013 with the results of 2015, to measure the improvement in those two years. We can observe from the comparison that the progress to 2015 is limited, even though the different settings of the PMMCs make a comparison difficult and can only be considered as a hint.

Furthermore, we can state that the evaluation experiments in the two contests do not provide further information about the performance of matchers. It is much more a grading of the tested matching techniques.

Moreover, it does not provide detailed information about the gold standards. In fact, the gold standard has a high effect on the evaluation results. As we argued already in Section 1.3.2, the "perfect match" (gold standard) which is used for the evaluation experiments of the process model matching contest is in fact highly questionable. The organizers of the contest 2015 improved the gold standard com-

pared to 2013. However, the gold standard is obtained by a small group of persons and reflects their point of view.

Furthermore, the evaluation experiments do not provide specific information about the data set and its complexity. The estimation of the complexity of the matching task is only based on the number of $1 : m$ correspondences and the different level of granularity, due to the variations in size of the applied process models. However, there are additional characteristics which make a data set more or less complex. Furthermore, we do not obtain any detailed information about specific strength and weaknesses of matching techniques. To support the improvement of matching techniques and to offer an efficient evaluation procedure, different considerations are required.

In the next section, we provide an overview about related work in the evaluation of process model matching techniques and the evaluation in related fields like schema-matching and ontology-matching.

4

Related Work

In this chapter, we discuss related work in the field of process model matching evaluation and related fields like ontology matching and schema matching. We will take this as a basis to motivate the evaluation procedures, described in this thesis.

State-of-the-art evaluation techniques mostly rely on Precision, Recall and F-Measure for evaluation of process model matching techniques (Baeza-Yates et al., 1999). These are standard metrics from the Information Retrieval field that can be used to quantify the performance of matchers alongside different dimensions. The reliance on these metrics applies to process model matching techniques (cf. Antunes et al. (2015); Cayoglu et al. (2013a); Leopold et al. (2012); Weidlich et al. (2010a, 2013b)) as well as to the related fields of schema matching and ontology matching techniques (cf. Rahm and Bernstein (2001); Shvaiko and Euzenat (2013)).

In general, research about the evaluation of process model matching techniques is rare. One important forum for evaluation experiments are the Process Model

Matching Contests (Antunes et al., 2015; Cayoglu et al., 2013a). In the Process Model Matching Contest 2015, the organizers found that there is no single matcher which has a high performance on all data sets. Moreover, it could be observed that the improvement compared to the PMMC 2013 was limited (cf. Section 3.2). The question rises why the matching techniques have not improved significantly over the past years. There has been limited research effort to answer this question. Recently, in (Jabeen et al., 2017) the authors analyze the most used similarity metrics to answer this question. Moreover, the authors in Weidlich et al. (2013a) propose a prediction of the matching results, in the way that they provide information in how far one should trust in the confidence of a matching result.

To support the improvement of matching techniques, the evaluation needs to offer more detailed insights about the performance of matchers than the current experiments can deliver. Currently, the evaluation of process model matching techniques does not fairly assess the performance of matchers. On the one hand, the generation of a gold standard strongly depends on the experts which generate a gold standard. On the other hand, the evaluation does not take the true complexity of a matching task into account. Moreover, the evaluation only provides a ranking and grading of the matching techniques. The experiments do *not* provide information about strength and weaknesses of matchers.

In ontology matching, an important evaluation forum is the "Ontology Alignment Evaluation Initiative" (OAEI), which has a longer tradition than the Process Model Matching Contests. The OAEI is conducted each year. The OAEI provides a basis of synthetic scenarios to test matchers on those data sets. In addition to that, more research has been dedicated to evaluation strategies in ontology matching (cf. (Ehrig and Euzenat, 2005; Euzenat, 2007; Sfar et al., 2016; Shvaiko and Euzenat, 2013; Zavitsanos et al., 2011)). For example, Euzenat (2007) introduced semantic Precision and Recall. This extension for ontology matching techniques allows to differentiate if the computed correspondences are related, by taking the ontological structure into account. As a consequence, deductible alignments are evaluated. Similarly, Ehrig and Euzenat (2005) propose alternative notions for these measures that take the closeness of results in ontology matching into account. Closeness can, for example, exploit the tree structure of ontologies, where the distance between elements in the tree can be computed to determine if a result is close or remote to the expected result. Therefore, those evaluation techniques mainly focus on relaxing the strict notion of Precision and Recall. Although it better reflects the closeness

of the computed alignments, it does not take the arguability of correspondences into account. Moreover, the evaluation method does not provide information about the kind of correspondences a matcher can identify and thus does not provide information about specific strength and weaknesses of matching techniques.

In the related field of schema matching, Bellahsene et al. (2011) summarize the evaluation experiments in this field. For example, Lee et al. (2007) propose synthetic scenarios to tune a schema matcher to specific applications. Similarly, the annual Ontology Alignment Evaluation Initiatives apply synthetic data sets which allow to test matching systems on specific characteristics, e.g., Achichi et al. (2017). However, these synthetic data sets are artificially generated test cases and therefore cannot always provide a realistic setting. In Euzenat et al. (2011), the authors explain that it is not suitable to apply matchers on synthetic scenarios, since these scenarios are too artificial. It is not clear if matchers have a similar performance on real-world data. Moreover, the automatic generation of the test cases often relies on the same resources as most matchers rely on, e.g., WordNet (Miller, 1995). Therefore, experts generated synonyms manually, but the manual synonym generation comes with high efforts. Moreover, the experiments take the runtime of the schema- and ontology-matcher into account. For process model matching evaluation this is not an interesting property, since the process models are rather small.

Sagi and Gal (2012) adapt Precision and Recall to evaluate non-binary confidence values which are computed by schema matching techniques. In this paper, the authors introduced an extension of a similarity matrix to evaluate schema matching techniques. Despite the existence of these different measures, what they all have in common is that they rely on the existence of a binary gold standard, i.e., on a single set of correct correspondences. Another limitation of such approaches is to use the confidence values which are computed by a matcher for the metrics: As we show in Section 6.1.1, the confidence values of the matchers are not comparable, since they have different ranges. The huge range differences also avoid a normalization. For example, some matchers compute confidence values between 0.95 and 1.0. For the normalization it would be necessary to stretch the confidence values range from the current 0.95 to 1.0. to the range of the gold standard (which is 0.125 to 1.0). However, such a normalization does not lead to stable results. We will explain this in greater detail in Section 5.3.

Recently, the authors extended their work in Sagi and Gal (2018). In their paper, the authors state:

"In particular, most data integration evaluation methods use a binary (match/no match) approach, while the variety and veracity of big data requires to broaden evaluation to a full scope in between."

However, they do not propose a non-binary gold standard for their evaluations. Similarly, in Thaler et al. (2014) the authors call for a more adequate evaluation of process model matching techniques. The authors state that it is sometimes impossible to agree on one single gold standard and that there are sometimes many possibilities which correspondences are actually correct. However, they do not propose an alternative evaluation procedure, which solves the subjectivity of reference alignments. Instead, the authors provide guidelines on how to build a gold standard.

Another research direction in ontology matching is the use of crowd-sourcing (Kittur et al., 2008; Paolacci et al., 2010). This research direction focuses on the problem how to establish a gold standard, which is very time-consuming and requires experts in the domain. In (Cheatham and Hitzler, 2014), the authors propose crowd-sourcing to establish a gold standard for ontology matching via Amazon Mechanical Turk (Turk, 2012). The authors call this "Wisdom of the Crowds". Since individuals who are not familiar with ontology matching establish a gold standard via crowd-sourcing, the authors introduce a metric which takes the "fuzzyness" (as they state) into account. However, in their paper, the authors compare the confidence values, computed by the matchers, directly to the values of the gold standard, e.g., the values of the matcher and the gold standard are multiplied or the difference between both values is calculated. In this proposed calculation, depending on the absolute number of the confidence value, the correspondence is considered as false-positive or false-negative. Since the range of the confidence values of the matcher output and the confidence values of the gold standard are not normalized, this leads to unwanted effects. Hence, the difference between the values of the gold standard and the confidence values of the matchers, are calculated without any normalization. As we will explain in Chapter 5, the confidence values of the matchers have no common range. (Each matcher uses a different threshold.) Therefore, the results cannot be compared and it may lead to unfair results, when simply using those values for the calculations. We will illustrate this in more detail in Section 6.

The Spearman's rank correlation, which we utilize in this thesis, already contains some kind of normalization, since the confidence values are only considered to establish a rank of the correspondences.

As an alternative measure to the rank correlation coefficient by Spearman, the Kullback Leibler divergence measures the divergence of two probability distributions. In statistics it is commonly used to measure the loss of information in approximations of probability distributions (Kullback and Leibler, 1951). In the following, we state the formula and explain why it is not an appropriate measure for process model matching evaluation.

Definition 5. *Let P and Q be two discrete probability distributions with $P(i) > 0$ and $Q(i) > 0$ for all i. Then, the Kullback-Leibler divergence is calculated by*

$$D_{KL}\left(P|Q\right) := \sum_i P(i) \cdot \ln\left(\frac{P(i)}{Q(i)}\right), \tag{4.1}$$

with the natural logarithm $\ln(\cdot)$.

To apply the Kullback-Leibler divergence towards process model matching evaluation, we have to make two changes to the matcher output and the non-binary gold standard. Therefore, let P be the confidence values of the matcher output and Q be the support values of the non-binary gold standard. Then, first, to avoid taking the logarithm of 0, $P(i) > 0$ for all correspondences i including those not contained in the matcher output but present in the non-binary gold standard. To achieve this, we assign a small constant ε as confidence value for such correspondences. Similarly, small constants ε have to be added to all correspondences i with $Q(i) = 0$ to avoid the division by 0. Second, to yield a probability distribution for P and Q, we have to divide the confidence values and the support values, respectively, by the sum of the entries. For example, the matcher output P then yields the following normalized values

$$\begin{aligned} \text{norm}_P &= \sum_i P(i) \\ P(i) &\leftarrow \frac{P(i)}{\text{norm}_P}. \end{aligned}$$

The Kullback-Leibler divergence has the following main disadvantage for process model matching evaluation. First, the Kullback-Leibler divergence measure is not symmetric, i.e., when exchanging the role of P and Q, then the resulting

Kullback-Leibler divergence typically does yield different values. This is not a desired property. Second, the choice of ε is arbitrary. The particular choice of ε can result in unwanted effects, even if it is a very small value. For example, assume $P(i) = 0.5$ and $Q(i) = \varepsilon$, i.e., correspondence i is not included in the gold standard but computed by the matcher with confidence value of 0.5. Now, when choosing $\varepsilon = 0.00001$, we obtain $0.5 \cdot \ln(\frac{0.5}{0.00001}) \approx 5.4099$. If we lower the value of ε to 0.000001, then the fraction inside the logarithm increases by factor 10 to yield a larger value of $0.5 \cdot \ln(\frac{0.5}{0.000001}) \approx 6.5612$. This illustrates the strong effect which an arbitrary choice of ε might have on the value of the Kullback-Leibler divergence. Because of the nature of the process model matching evaluation, we expect many entries with $Q(i) = \varepsilon$, i.e., FPs. Third, the normalization of both the matcher output and the non-binary gold standard changes their absolute values. For example, if a correspondence has both the same confidence and support value (for example 0.5 or 1.0), then the Kullback-Leibler divergence should be 0 for this correspondence (because $\ln(1) = 0$). Unfortunately, the normalization does not preserve the property of equal confidence and support value, i.e., after normalization these values are no longer the same which causes also the Kullback-Leibler divergence to be non-zero for this particular correspondence.

Consequently, the Kullback-Leibler divergence seems not to be a good choice for the evaluation of process model matching techniques.

A smoothed version of the Kullback–Leibler divergence is presented by the Jensen–Shannon divergence (Lin, 1991). As such, it also measures the difference between two probability distributions. It is computed as follows.

Definition 6. *Let P and Q be two discrete probability distributions with $P(i) > 0$ and $Q(i) > 0$ for all i. Then, the Jensen–Shannon divergence is calculated by*

$$D_{JSD}(P, Q) := \frac{1}{2} D_{KL}(P|M) + \frac{1}{2} D_{KL}(Q|M), \tag{4.2}$$

with $M = \frac{1}{2}(P + Q)$.

It is noteworthy that the Jensen–Shannon divergence is symmetric, i.e., by definition $D_{JSD}(P, Q) = D_{JSD}(Q, P)$. As such, the Jensen–Shannon divergence overcomes the first drawback of the Kullback–Leibler divergence, as stated above. The Jensen–Shannon divergence also requires the assignment of ε values for both P and Q. However, the undesired effect described above does not occur because the logarithm decreases smaller than linearly for small values. For example, $P(i) = \epsilon$

and $Q(i) = 0.5$ yields $\epsilon \cdot \ln(\frac{\epsilon}{\frac{1}{2}(\epsilon+0.5)})$ which is ≈ -0.0001 for $\epsilon = 0.0001$ and ≈ -0.00001 for $\epsilon = 0.00001$ for the first term in (4.2) for correspondence i. Therefore, the Jensen–Shannon divergence also overcomes the second drawback of the Kullback–Leibler divergence for our application. Unfortunately, the Jensen–Shannon divergence also requires a normalization, just like the Kullback–Leibler divergence, because both metrics compare probability distributions. Thus, the third drawback of the Kullback–Leibler divergence also remains for the Jensen–Shannon divergence. Therefore, the rank-correlation is a more appropriate measure for process model matching evaluation, since it does not have the described drawbacks. The rank-correlation already implies a normalization for the different values of the confidence values of the matchers and the support values of the gold standard, since only the rank is considered. We will explain this in more detail in Chapter 6.

All in all, it can be summarized that current evaluation experiments for process model matching as well as related fields do not provide detailed information about the performance of matchers, without manually processing the matcher output. Moreover, the evaluation experiments do not adequately take the uncertainty of a reference alignment into account.

In the next chapter, we therefore introduce a non-binary gold standard, which avoids the problem to agree on one single gold standard. Instead, questionable alignments are included but are assigned with a specific weight. This also avoids a loss of information, since "possible" correspondences are included. We will further show that there is a very low fraction of correspondences where all the annotators agreed on.

5

Probabilistic Evaluation

In the previous chapters, we discussed state-of-the-art evaluation and gave an introduction to the field of Information Retrieval. We further argued that a binary gold standard does not take the true complexity of a matching task into account. In this chapter, we introduce a probabilistic evaluation procedure, which takes a non-binary gold standard as basis for the evaluation. We believe that the probabilistic evaluations can be used for a more fine-grained evaluation of matchers and, thus, can help to improve the matchers itself. The evaluation can be assessed without the need for additional information, thus can be used to evaluate existing matching systems.

In this chapter, we introduce a new, so-called non-binary gold standard, which does not exhibit the weaknesses described above. The idea is to move away from a 0 or 1 measure of false or correct correspondences to a non-binary value between 0 and 1. A correspondence with value 0 is still regarded as a wrong correspondence; value 1 remains a correct correspondence. All values in between 0 and 1 are in-

tended to measure the strength of correctness of the correspondence, reflecting the expert opinion of the annotators. For example, if two out of three experts define a correspondence as correct, then the non-binary gold standard contains that correspondence with weight 0.67. Therefore, instead of discarding correspondences from single annotators, they are considered as correspondences with a specific support value. To include all such correspondences in the non-binary gold standard avoids a loss of information.

This chapter is organized as follows. In Section 5.1, we provide a formal definition of our non-binary gold standard. Our definition of the non-binary gold standard allows for the adjustment of the well-known and intuitive metrics Precision, Recall and F-Measure. These definitions are adjusted to deal with the non-binary values in Section 5.2. Those measures are metrics which can be intuitively interpreted. It further indicates if matchers focus on a high Precision or on a high Recall or if the results are balanced. Moreover, we introduce new performance measures, which take the non-binary gold standard as basis for the evaluation in Section 5.3. The relative distance measures a correspondence with low support values closer to 0 than to 1. Finally in Section 5.4, we conduct experiments for all metrics and show the insights which we gain by applying the described metrics to two data sets and matchers of the Process Model Matching Contest 2015 as well as the OAEI 2016 and 2017. We will further demonstrate the robustness of our metrics through our experiments in Section 5.4.6.

Some of the work in this chapter has already been published in Kuss et al. (2016, 2018).

5.1 Definition of a Non-binary Gold Standard

In Section 2.1, we defined a binary gold standard. Based on this definition we define the non-binary gold standard which we refer to.

Definition 7 (Non-Binary Gold Standard)**.** *A non-binary gold standard is a tuple* $\mathcal{GS} = (A_1, A_2, \mathcal{H}, \sigma)$ *where*

- A_1 *and* A_2 *are the sets of activities of two process models,*
- $\mathcal{H} = \{H_1, \ldots, H_n\}$ *is a set of independently created binary human assessments, and*

- $\sigma : A_1 \times A_2 \to \mathbb{R}$ *is a function assigning to each* $(a_1, a_2) \in A_1 \times A_2$ *a* support value*, which is the number of binary human assessments in* $\mathcal{H}$ *that contain the correspondence* (a_1, a_2) *divided by the total number of binary human assessments* $|\mathcal{H}|$*.*

The overall rationale of the non-binary gold standard from Definition 7 is to count the individual opinions from the binary human assessments as votes. Such a binary human assessment according to Definition 4 should be created independently and solely reflect the opinion of a single assessor. Based on a number of such independently created binary human assessments, we can then define a non-binary gold standard. In this way, we obtain a *support value* σ for each correspondence according to the number of votes in favor of this correspondence. We can understand this support value as *confidence values*. In this way, any correspondence with a support value $0.0 < \sigma < 1.0$ can be regarded as an uncertain correspondence. For these correspondences, there is no unanimous vote about whether or not it is a correct correspondence.

We can observe that the non-binary gold standard covers a broad range of correspondences.

To take the uncertainty of correspondences into account, a non-binary gold standard is required. To obtain a gold standard with confidence values we collected assessments created by individual human annotators. Each of these *binary human assessments* captures the correspondences that a single annotator identifies between two given process models. We asked 8 individuals to identify the correspondences for the 36 model pairs from the University Admission data set. We prepared respective templates for each model pair and asked the annotators to complete this task model pair by model pair. We instructed them to not spend more than two hours in a row on this task to avoid low quality results caused by depletion. The group of involved annotators was heterogeneous and included 4 researchers being familiar with process model matching and 4 student assistants from the University of Mannheim in Germany. Some student assistants were already familiar with process model matching. The remaining student assistants were introduced to the problem of process model matching and of creating a gold standard. They were told that the gold standard expresses their point of view and therefore were not influenced in the way they identified correspondences. Then all correspondences from each annotator were collected and each annotator's choice was counted as a vote for the corresponding annotation. The result of this step, was a non-binary

gold standard based on 8 binary assessments. On average, the annotators spent around one hour per model pair (i.e., approximately 36 hours per annotator). Note that we did not apply any changes to the individual assessments. We included them in their original form into the non-binary gold standard. Similarly, the procedure was conducted for the Birth Registration data set.[1]

5.2 Probabilistic Precision, Recall, and F-Measure

Based on the support values provided by the non-binary gold standard, we define probabilistic versions of Precision, Recall, and F-Measure, which take the uncertainty of correspondences into account. For notational convenience, we introduce $\mathcal{C}$ to refer to the set of all correspondences that have a support value above 0.0.

Definition 8 (Probabilistic Precision, Recall, and F-Measure)**.** *Let A_1 and A_2 be the sets of activities of two process models, $M : A_1 \times A_2$ the correspondences identified by a matching technique, and $\mathcal{GS} = (A_1, A_2, \mathcal{H}, \sigma)$ a non-binary gold standard. Then, we define probabilistic Precision, Recall, and F-Measure as follows:*

$$\text{Probabilistic Precision (ProP)} = \frac{\sum_{m \in M} \sigma(m)}{\sum_{m \in M} \sigma(m) + |M \setminus \mathcal{C}|} \tag{5.1}$$

$$\text{Probabilistic Recall (ProR)} = \frac{\sum_{m \in M} \sigma(m)}{\sum_{c \in \mathcal{C}} \sigma(c)} \tag{5.2}$$

$$\text{Probabilistic F-Measure (ProFM)} = 2 \times \frac{\text{ProP} \times \text{ProR}}{\text{ProP} + \text{ProR}} \tag{5.3}$$

Probabilistic Precision and Recall are adaptations of the traditional notions of Precision and Recall that incorporate the support values from a non-binary gold standard $\mathcal{GS}$. We define *probabilistic Precision* (ProP) as the sum of the support values of the correspondences identified by the matching technique (M) divided by the same value plus the number of correspondences that are not part of the non-binary gold standard ($|M \setminus \mathcal{C}|$). This definition gives those correspondences that have been identified by many annotators a higher weight than those that have only been identified by a few. Therefore, it accounts for the uncertainty associated with correspondences in the non-binary gold standard. The metric further rewards

[1] We received three out of eight individual gold standards of the Birth Registration data set from researchers of the Karlsruhe Institute of Technology (KIT) in Germany.

matchers which are also able to detect correspondences with low support values in the non-binary gold standard.

The correspondences, identified by the matchers are considered as *binary* values, thus they have a confidence value of 1.0. We use only binary values for the correspondences of the matcher, because of the following two reasons:

1. Most matchers do *not* provide confidence values for the computed correspondences, i.e. the confidence values are already binary.

2. The matchers which provide confidence values use a different range of confidence values, e.g., they use different thresholds. Therefore, the confidence values cannot be compared directly, without any normalization. However, a normalization does not lead to stable results, due to the very small interval of the confidence values, which some matchers compute. In Section 6.1 we will explain this in more detail.

Since we cannot use the confidence values of the matchers, we keep the binary values in the evaluation on matcher side for all matchers. Therefore, only the values of the gold standard are non-binary. As a result, the impact of false-positives, i.e., correspondences that have been identified by the matching technique but are not part of the non-binary gold standard, result in a strong penalty of 1.0. We justify this high penalty by the high coverage of uncertain correspondences included in non-binary gold standards. These gold standards can be expected to contain a broad range of potential correspondences (thus are almost complete), including those identified by only one single annotator. Any correspondence not included in this broad range can be considered to be certainly incorrect, which is reflected in the penalty of 1.0 for false-positives. We will show this in more detail in Section 5.4.6.

Probabilistic Recall (ProR) follows the same principle as the probabilistic Precision. It resembles the traditional definition of recall, but incorporates the support values from the non-binary gold standard respectively. As a result, identifying correspondences with a higher support has a higher influence on the Recall than identifying correspondences with a low support. *The probabilistic F-Measure* (ProFM) presents the harmonic mean of probabilistic Precision and Recall. It is computed in the same way as the traditional F-Measure, though it is here based on ProP and ProR.

To illustrate these metrics, consider the correspondences, their support values, and the output of three matchers depicted in Table 5.1. The support values reveal that 5 out of 6 correspondences are considered to be correct correspondences by one or more binary human assessor. Matcher $\mathcal{M}_1$ identifies exactly these 5 correspondences. Therefore, $\mathcal{M}_1$ achieves ProP and ProR scores of 1. This indicates a "perfect match" since the matcher did not compute any wrong alignments. Moreover, the matcher did not miss any correspondence, which is classified as reasonable. For some applications, this is a wanted outcome of a matcher. For example, if the results of a matcher are used to incorporate human feedback as proposed by Klinkmüller et al. (2014). Therefore, the metric rewards matchers which focus on finding all reasonable correspondences. However, the penalty of not computing correspondences with low support value is very low. By contrast, matcher $\mathcal{M}_2$ identifies only 3 of the 5 correct correspondences. The matcher also includes the incorrect correspondence c_6 in its output. This results in a ProP value of 0.71 and a ProR value of 0.77. Although matcher $\mathcal{M}_3$ correctly identifies 4 correspondences, instead of the 3 identified by $\mathcal{M}_2$, it achieves the exact same ProP and ProR values. This occurs because $\mathcal{M}_3$ identifies c_4 and c_5, which have a combined support value of 0.75, i.e., the same support value as correspondence c_3 that is identified by $\mathcal{M}_2$. This shows that correspondences with a high support value have a greater contribution to the metrics than those with low support.

Table 5.1: Exemplary matcher output and metrics

$\mathcal{C}$	σ	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
c_1	1.00	1	1	1
c_2	0.75	1	1	1
c_3	0.75	1	1	0
c_4	0.50	1	0	1
c_5	0.25	1	0	1
c_6	0.00	0	1	1
ProP		**1**	**0.71**	**0.71**
ProR		**1**	**0.77**	**0.77**
ProFM		**1**	**0.74**	**0.74**

These exemplary calculations show that our non-binary/probabilistic version of Precision and Recall takes the support values explicitly into account. In this way, correspondences that have been identified by many assessors have a more

significant contribution to the metric than those that have only been identified by a few. At the same time, matchers that compute correspondences with low support value are rewarded. This is a desired feature for many applications.

However, the non-binary gold standard also allows us to change this point of view. It allows moreover to obtain more fine-grained insights into the performance of matchers. We can achieve this by computing probabilistic Precision and Recall scores for correspondences with a minimal support level. By adapting the equations from Definition 8. In this way, we can differentiate between matchers that identify correspondences with a broad range of support values and those that focus on the identification of correspondences with high support values. We capture this notion of *Bounded* probabilistic Precision, Recall, and F-Measure in Definition 9.

Definition 9 (Bounded Probabilistic Precision, Recall, and F-Measure)**.** *Let A_1 and A_2 be the sets of activities of two process models, $M : A_1 \times A_2$ the correspondences identified by a matching technique, $\mathcal{GS} = (A_1, A_2, \mathcal{H}, \sigma)$ a non-binary gold standard, and $\mathcal{C}_\tau$ refers to the set of correspondences with a support level $\sigma \geq \tau$. Then, we define Bounded probabilistic Precision, Recall, and F-Measure as follows:*

$$\text{ProP}(\tau) = \frac{\sum\limits_{m \in M} \sigma(m)}{\sum\limits_{m \in M} \sigma(m) + |M \setminus \mathcal{C}_\tau|} \tag{5.4}$$

$$\text{ProR}(\tau) = \frac{\sum\limits_{m \in M} \sigma(m)}{\sum\limits_{c \in \mathcal{C}_\tau} \sigma(c)} \tag{5.5}$$

$$\text{ProFM}(\tau) = 2 \times \frac{\text{ProP}(\tau) \times \text{ProR}(\tau)}{\text{ProP}(\tau) + \text{ProR}(\tau)} \tag{5.6}$$

By computing Bounded Precision and Recall values, we can directly gain insights into the differences between the results obtained by the matchers, consider now Table 5.2. The correspondences c_4 and c_5 are highlighted in red, since those two correspondences turn into false-positives for ProP(0.75) and ProR(0.75). We can see from the values of Table 5.2, that the results for the matchers $\mathcal{M}_1$ – $\mathcal{M}_3$ change considerably (cf. Table 5.1). For instance, $\mathcal{M}_2$ improves, since the matcher focuses on computing correspondences with high support values. In contrast, the results for the matchers $\mathcal{M}_1$ and $\mathcal{M}_3$ decrease, since those matchers also identify the correspondences with low support values. For the Bounded probabilistic evaluation,

Table 5.2: Exemplary matcher output and metrics for Bounded probabilistic FM at $\tau = 0.75$ with the matchers of Table 5.1

$\mathcal{C}$	σ	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$
c_1	1.00	1	1	1
c_2	0.75	1	1	1
c_3	0.75	1	1	0
c_4	~~0.50~~ 0.00	1	0	1
c_5	~~0.25~~ 0.00	1	0	1
c_6	0.00	0	1	1
ProP(0.75)		**0.56**	**0.71**	**0.37**
ProR(0.75)		**1.0**	**1.0**	**0.7**
ProFM(0.75)		**0.71**	**0.83**	**0.48**

matcher $\mathcal{M}_2$ achieves the best results, since this matcher focuses on computing correspondences with a high support value in the non-binary gold standard.

With the Bounded version of the metrics it can be determined, which matchers focus on correspondences with high support values and which identify correspondences with low support values. In particular, with the Bounded version of the metrics, it is possible to determine a threshold; all correspondences under this threshold τ are excluded from the evaluation. Therefore, the Bounded version of the metrics do not penalize any more if low-support correspondences under a selected threshold τ are not computed. This threshold can be selected depending on the application scenario, to assure that the focus of the matcher fits the specific application. For example, if it is desired to build a matcher, which focuses on finding only "sure" correspondences, the Bounded version of the metrics can indicate this.

5.3 Relative Distance

The previously introduced notions of ProP, ProR, and ProFM implicitly build on the premise that matchers should also identify correspondences with low support values. In fact, they reward matchers that identify correspondences with low support values and penalize matchers that fail to identify them. As an illustration, consider a correspondence for which 2 out of 5 human annotators agree that this is a correct correspondence. If identified by a matcher, the ProP, ProR, and ProFM scores of the matcher will increase, because the correspondence has a non-zero

support value. However, it is important to recognize that also 3 out of the 5 annotators agree that this is *not* an actual correspondence, i.e., the majority of the annotators disagree with the correspondence. Generally, the evaluation strongly depends on the applications. For example, for some applications it might be required that only "sure" correspondences are considered. A metric which rewards matchers that also identify uncertain correspondences would not be a reasonable metric for such applications. The previously introduced metrics do not fully take such a majority of disagreements into account. The Bounded Precision and Recall in fact allows to evaluate only correspondences of the non-binary gold-standard with a higher threshold, this circumvents the above described problems. However, those correspondences with a lower threshold are excluded from the evaluation and therefore again this information is lost in the non-binary gold standard. Recognizing such characteristics, we also introduce an alternative performance measure that explicitly considers agreements and disagreements in a non-binary gold standard, without the loss of information. This performance measure builds on the notion of *distance* between the matcher output and the support values from the non-binary gold standard. The overall rationale is to explicitly account for agreements and disagreements with the annotators of the non-binary gold standard. Intuitively, this means that correspondences with low support values are no longer favorable since most annotators disagree with these correspondences. We define the measure *Relative Distance (ReD)* as follows.

Definition 10 (Relative Distance)**.** *Let A_1 and A_2 be the sets of activities of two process models, $M : A_1 \times A_2$ the correspondences identified by a matching technique, $\mu : A_1 \times A_2 \rightarrow \{0, 1\}$ a function that returns 1 if a correspondence $m \in M$ and 0 if a correspondence $m \notin M$, and $\mathcal{GS} = (A_1, A_2, \mathcal{H}, \sigma)$ a non-binary gold standard. Then, we define the Relative Distance as follows:*

$$\text{Relative Distance (ReD)} = \sum_{m \in (M \cup \mathcal{C})} (\mu(m) - \sigma(m))^2 \tag{5.7}$$

The core idea underlying the ReD measure is to compute the distance between the matcher output (which can be 1.0 or 0.0) and the support value σ from the non-binary gold standard. We square the values to obtain a lower penalty for correspondences that have a high support. To illustrate the mechanism of ReD, consider Table 5.3. It shows how the output of the three matchers from Table 5.1 is evaluated by ReD.

		$\mathcal{M}_1$		$\mathcal{M}_2$		$\mathcal{M}_3$	
$\mathcal{C}$	$\sigma(c_n)$	$\mu(\mathbf{c_n})$	$\mathbf{ReD}(\mathbf{c_n})$	$\mu(\mathbf{c_n})$	$\mathbf{ReD}(\mathbf{c_n})$	$\mu(\mathbf{c_n})$	$\mathbf{ReD}(\mathbf{c_n})$
c_1	1.00	1	0	1	0	1	0
c_2	0.75	1	0.063	1	0.063	1	0.063
c_3	0.75	1	0.063	1	0.063	0	0.563
c_4	0.50	1	0.25	0	0.25	1	0.25
c_5	0.25	1	0.563	0	0.063	1	0.563
c_6	0.00	0	0	1	1	1	1
Total			**0.938**		**1.438**		**2.438**

Table 5.3: Illustration of Relative Distance

The example depicted in Table 5.3 illustrates three key characteristics of ReD. First, matchers identifying a correspondence that is not part of the non-binary gold standard or fail identifying a correspondence with a support of 1.0 receive a penalty of 1. Second, it does not matter whether a matcher identifies a correspondence with a support of 0.5 (see c_4). The distance in both cases is identical. This is a reasonable approach taking into account that the matcher agrees/disagrees with half of the annotators. Third, the penalty for identifying a correspondence with a low support is higher than for not identifying it (see c_5). This is again in line with the argument of taking agreements into account. Given a support of 0.25 of c_5, a matcher that does not identify c_5, disagrees with 25% of the annotators. A matcher that does identify c_5, disagrees with 75% of the annotators.

Note that we again did not consider the confidence values of the matchers, for the reasons which we explained above.

In this way, we complement the above introduced evaluation measures; in contrast to ProFM, ReD does not reward matchers which identify a correspondence with low support value in the non-binary gold standard. However, low support correspondences are not treated as incorrect, as this was the case for the Bounded variants of Precision and Recall. Therefore, the information of the uncertain correspondences are preserved and not lost, as this is the case in the Bounded version of ProFM.

In summary, it can be stated that the choice of the particular metric depends on the application of the matchers. If matchers should also indentify uncertain correspondences then ProP, ProR and ProFM is a suitable measure. It further can be used to analyze the focus of the different matchers; with the Bounded variants

of ProP, ProR and ProFM it is possible to understand if matchers focus on identifying high supported correspondences (which are also often obvious), or if they also identify low-support correspondences. In contrast to ProFM, ReD rewards matchers that aim in identifying high-support correspondences, since it is penalized if matchers compute correspondences with a low support value. The metric treats a low-support correspondence closer to a wrong than a correct correspondence. This changes with a support value of ≥ 0.5.

In the next section, we apply our probabilistic evaluation procedure with all introduced metrics to the University Admission data set and the Birth Registration data set introduced in the context of the Process Model Matching Contest 2015 (Antunes et al., 2015).

As described in Section 5.1, we created a non-binary gold standard, based on correspondences identified by 8 individual annotators (for each data set), and compute the probabilistic measures for up to 16 different matchers that solved this matching problem. The overall goal of our experiments is to demonstrate the usefulness of the non-binary perspective and the value of the insights that our evaluation procedures delivers.

5.4 Experiments

To illustrate the insights which we gain by the above introduced evaluation measures, we apply them to the data sets and matchers of the University Admission data set and the Birth Registration data set of the PMMC 2015 (Antunes et al., 2015), which we introduced and described in Section 3.1.

The University Admission data set consists of nine BPMN process models describing the admission processes for graduate study programs of different German universities. The size of the models varies between 10 and 44 activities. This indicates the high level of complexity, since there are big differences in granularity of the matched data sets.

The Birth Registration data set also consists of 36 model pairs that were derived from 9 models representing the birth registration processes of Germany, Russia, South Africa, and the Netherlands. The models were created by graduate students at the HU Berlin and in the context of a process analysis in Dutch municipalities. The data set is in Petri-Nets. This data set has also been used in the PMMC 2013 (Cayoglu et al., 2013a).

The task of the Process Model Matching Contest 2015 was to match these models pair-wise, resulting in a total number of 36 matching pairs.

Based on the non-binary gold standard, we calculated ProP, ProR, ProFM and ReD for a total of 16 matchers. Twelve matchers solved this matching problem in the context of the PMMC 2015 and 4 matchers solved it in the context of a subtrack of the Ontology Alignment Evaluation Initiative (OAEI) 2016 and 2017 (Achichi et al., 2016, 2017). In line with the report from both the PMMC 2015 and OAEI 2016 and 2017, we distinguish between micro and macro average. Macro average is defined as the average Precision, Recall, and F-Measure of all 36 matching pairs. Micro average, by contrast, is computed by considering all 36 pairs as one matching problem. The micro average scores take different sizes of matching pairs (in terms of the correspondences they consist of) into account. As a result, a poor Recall on a small matching pair has only limited impact on the overall micro average Recall score.

5.4.1 Results

This section discusses the results of our experiments. Section 5.4.2 elaborates on the characteristics of the non-binary gold standard of the University Admission data set. Section 5.4.3 presents the results from the evaluation with ProP, ProR, and ProFM and compares them to the results of the non-binary evaluation. Section 5.4.4 discusses the insights from the evaluation with the Bounded versions of ProP, ProR, and ProFM. Section 5.4.5 presents the results from the evaluation with ReD. Section 5.4.6 investigates when the ProFM and ReD metrics become robust.

5.4.2 Attributes of the Non-binary Gold Standard

Figure 5.1 illustrates exemplary the average correspondence support values for the University Admission data set from the PMMC 2015 (Antunes et al., 2015) of the eight experts for each process model pair. The average support values differ notably for the various models. It is characteristic for this applied data set that some process model pairs vary strongly in their size and structure, thus have a significantly different level of granularity. Some process model pairs are similar regarding the structure, size and the syntax of the process model label.

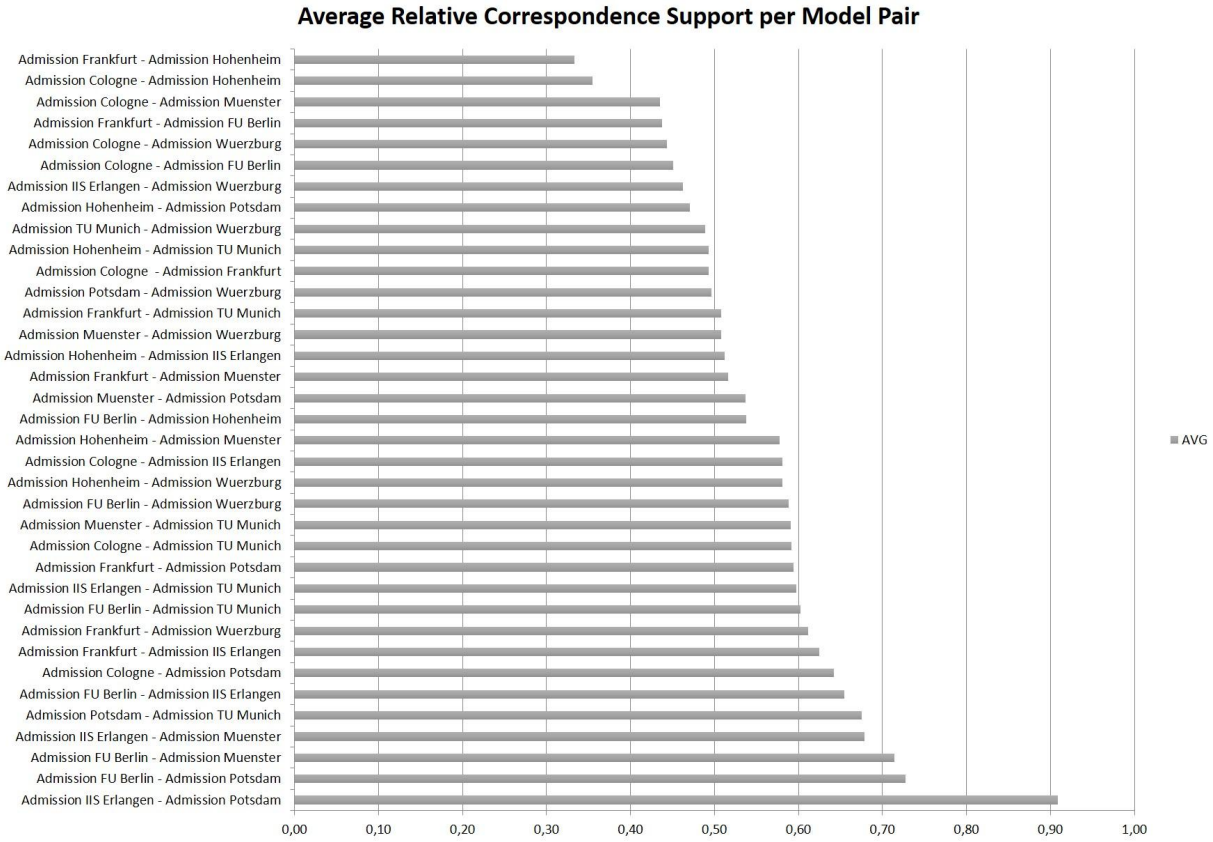


Figure 5.1: Average support values of the test cases

In Figure 5.1, we can observe that process model pairs which have a similar structure and size, get a higher average correspondence support level of the eight annotators, (*e.g.*, IIS Erlangen - Potsdam). Opposed to this, process model pairs which differ significantly in size and structure get a low average correspondence support level, (*e.g.*, Frankfurt - Hohenheim, Cologne - Hohenheim). A different level of granularity in the data set implies a higher complexity of the matching task; for humans as well as matching techniques.

The non-binary gold standard from the University Admission data set resulting from the 8 binary assessments consists of a total of 879 correspondences. The binary gold standard from the PMMC 2015 only consisted of 234 correspondences, which is less than a third. The average support value per model pair ranges from 0.33 to 0.91. This illustrates that the models considerably differ with respect to how obvious the contained correspondences are.

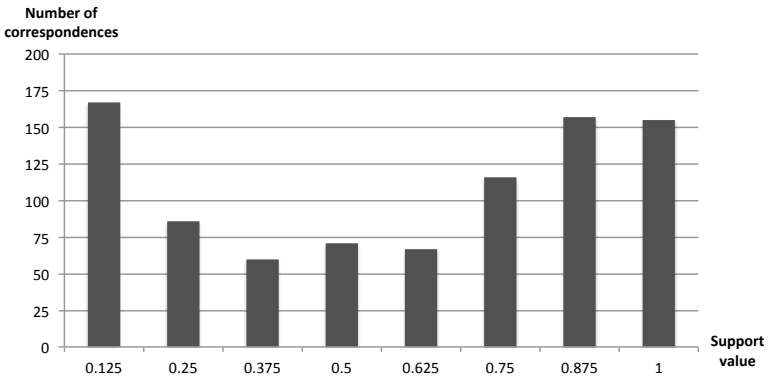


Figure 5.2: Distribution of support values in the non-binary gold standard of the University Admission data set

Figure 5.2 illustrates the distribution of the support values of the University Admission data set. It shows that there are two extremes. On the one hand, there is a high number of correspondences with 6 or more votes (support value ≥ 0.75). On the other hand, there is also a high number of correspondences with three votes or less (support value ≤ 0.375). Overall, the number of correspondences which would be included based on a majority vote (support value ≥ 0.5) amounts to 495, which is only a little more than half of the correspondences from the non-binary gold standard. These numbers illustrate the complexity associated with defining a binary gold standard and highlight the risks of a purely binary evaluation

procedure. Instead of excluding a high number of possible correspondences, we include them with a respective support value. This avoids a loss of information. The broad coverage of the non-binary gold standard implies that all reasonable correspondences are included. Thus, correspondences which are not part of the non-binary gold standard can be considered as wrong. This is one major difference to a binary gold standard. We will show that the probabilistic evaluation is robust and does not profit from further annotators.

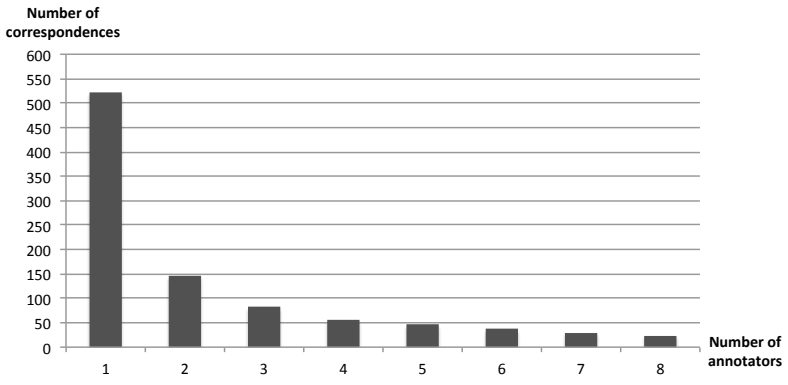


Figure 5.3: Average increase of number of correspondences with additional annotators

Figure 5.3 further illustrates the average number of correspondences that are added to the non-binary gold standard by an additional annotator. The numbers from Figure 5.3 emphasize that the number of correspondences added by an additional annotator decreases very quickly. While the second annotator, on average, adds about 145 new correspondences to the non-binary gold standard, the 8th annotator only adds 24 new correspondences. Note that the correspondences that are newly introduced by the 8th annotator only have a support of 0.125, since none of the previous annotators agreed with these correspondences. Overall, these numbers show that we quickly reach a point where hardly new reasonable correspondences are added. This is in line with the notion of *theoretical saturation* in qualitative research settings (Bowen, 2008). In this context, theoretical saturation describes the point where no new insights can be obtained from analyzing additional data.

5.4.3 Evaluation Using Probabilistic Precision, Recall and F-Measure

Based on the non-binary gold standard, we calculated ProP, ProR, ProFM, and ReD for a total of 16 matchers. Twelve matchers solved this matching problem in the context of the PMMC 2015 and 4 matchers solved it in the context of a subtrack of the Ontology Alignment Evaluation Initiative (OAEI) 2016 and 2017 (Achichi et al., 2016). In line with the report from both the PMMC 2015 and OAEI 2016/2017, we distinguish between micro and macro average. Macro average is defined as the average Precision, Recall, and F-Measure of all 36 matching pairs. Micro average, by contrast, is computed by considering all 36 pairs as one matching problem. The micro average scores take different sizes of matching pairs (in terms of the correspondences they consist of) into account. As a result, a poor Recall on a small matching pair has only limited impact on the overall micro average Recall score.

Table 5.4 presents the probabilistic evaluation results based on the non-binary gold standard. It shows the micro and macro values of probabilistic F-Measure (ProFM), Precision (ProP) and Recall (ProR) for each of the 16 matchers that participated in the PMMC 2015 or the OAEI 2016/2017. The column *Rank - New* indicates the rank the matcher has achieved according to the ProFM micro value. The column *Rank - Old* shows the rank the system has achieved according to the binary evaluation, as conducted in the Process Model Matching Contest 2015. In line with the report of the PMMC 2015 we distinguish between “micro” and “macro” values, which we describe in Section 2.2.

The results in the table illustrate that the probabilistic evaluation has notable effects on the ranking. For instance, for the University Admission data set, the matcher *AML-PM* moves from rank 14 to 5 and the matcher *RMM-NLM* moves from rank 3 to rank 14. A brief analysis of the matchers’ inner workings provides an explanation for this development. The matcher *AML-PM* does not impose strict thresholds on the similarity values it uses for identifying correspondences. As a result, it also identifies correspondences with low support values. In the binary gold standard, however, these correspondences were simply not included and resulted in a decrease of Precision.

Table 5.5 illustrates this effect by showing an excerpt from the correspondences generated by the matcher *AML-PM* and the respective entries from the binary and the non-binary gold standard. We can see that from the 5 correspondences from Table 5.5 only two were included in the binary gold standard. In the context

Rank			Approach	ProFM		ProP		ProR	
New	Old	Δ		mic	mac	mic	mac	mic	mac
1	2	+1	RMM-NHCM	.432	.391	.83	.777	.292	.297
2	11	+9	LogMap	.42	.366	.683	.676	.304	.301
3	1	-2	AML	.419	.376	.795	.728	.284	.289
4	6	+2	Know-Match-SSS	.411	.358	.679	.788	.295	.297
5	14	+9	AML-PM	.408	.395	.411	.46	.406	.408
6	13	+7	KnoMa-Proc	.406	.345	.573	.594	.314	.302
7	5	-2	OPBOT	.369	.318	.669	.676	.254	.248
8	12	+4	BPLangMatch	.361	.327	.559	.505	.267	.265
9	7	-2	RMM-SMSL	.358	.325	.6	.712	.255	.256
10	9	-1	DKP-lite	.347	.284	.895	.911	.215	.219
11	8	-3	DKP	.341	.285	.759	.691	.22	.223
12	15	+3	RMM-VM2	.318	.307	.333	.337	.304	.306
13	4	-9	Match-SSS	.315	.249	.827	.814	.194	.203
14	3	-11	RMM-NLM	.312	.253	.73	.583	.198	.203
15	10	-5	TripleS	.301	.21	.518	.498	.212	.216
16	16	±0	pPalm-DS	.275	.261	.229	.289	.345	.344

Table 5.4: Results of probabilistic evaluation of the University Admission data set with non-binary gold standard

Correspondence (C)		Gold Standard	
Activity 1	Activity 2	Binary	Non-binary
Send documents by post	*Send appl. form and documents*	0	0.750
Evaluate	*Check and evaluate application*	0	0.500
Apply online	*Complete online interview*	0	0.375
Wait for results	*Waiting for response*	1	0.875
Rejected	*Receive rejection*	1	0.625

Table 5.5: Effect of gold standard on assessment of output of matcher *AML-PM*

of an evaluation based on this binary gold standard these three correspondence would therefore reduce the Precision of this matcher. An evaluation based on the non-binary gold standard, however, would come to a different assessment. The non-binary gold standard does not only include the two correspondences from the binary gold standard, but also includes the three other correspondences. It is obvious that this positively affects the ProP of the matcher and improves its overall ProFM respectively. For the matcher *RMM-NLM* we observe the opposite effect. In the context of the evaluation with the non-binary gold standard it misses

a huge range of correspondences. Consequently, the ProR of this matcher decreases considerably.

Rank			Approach	ProP		ProR		ProFM	
New	Old	Δ		mic	mac	mic	mac	mic	mac
1	1	±0	OPBOT	.65	.614	.517	.446	.576	.5
2	3	+1	RMM-NHCM	.781	.718	.443	.364	.565	.458
3	11	+8	LogMap	.834	.78	.411	.308	.551	.39
4	8	+4	Know-Match-SSS	.865	.812	.379	.292	.527	.39
5	6	+1	BPLangMatch	.661	.524	.417	.327	.511	.39
6	10	+4	TripleS	.651	.588	.426	.328	.515	.38
7	7	±0	AML-PM	.513	.458	.505	.44	.509	.439
8	12	+4	I-Match	.812	.644	.366	.267	.504	.345
9	2	-7	pPalm-DS	.469	.462	.521	.442	.493	.425
10	5	-5	AML	.467	.417	.515	.44	.49	.41
11	13	+2	Match-SSS	.974	.991	.315	.23	.476	.323
12	4	-8	RMM-VM2	.454	.419	.48	.41	.466	.402
13	9	-4	RMM-SMSL	.518	.542	.42	.344	.464	.379
14	14	±0	RMM-NLM	.912	.967	.293	.21	.443	.295
15	15	±0	KnoMa-Proc	.224	.207	.437	.342	.296	.248

Table 5.6: Results of probabilistic evaluation of the Birth Registration data set with non-binary gold standard

Different behavior can be observed for the Birth Registration data set in Table 5.6. The matcher RMM-NHCM, for instance, increases its performance not just relatively (compared to the other matchers), but also with the absolute numbers. For instance, the F-Measure of RMM-NHCM increases from .456 in the binary evaluation, to .565 in the non-binary evaluation. (Such an effect can be observed for other matchers as well.) This is surprising, because the non-binary gold standard contains a high number of uncertain alignments, which can be expected to result in a decrease of the Recall of the matchers. The improvement of the absolute values for the F-Measure of some matchers indicates that the reliability of the *binary* gold standard for the Birth Registration data set is highly questionable. This highlights the problems associated with a binary evaluation.

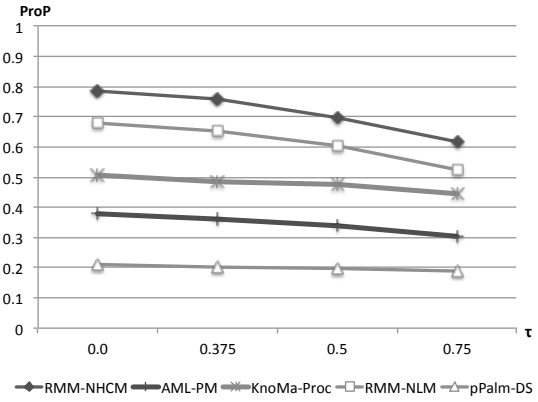
5.4.4 Evaluation Using Bounded Probabilistic Precision, Recall, and F-Measure

The Bounded variants of ProP, ProR, and ProFM provide the possibility to obtain more detailed insights into the performance of the matchers. Figure 5.4 illustrates this by showing the values of ProP, ProR, and ProFM for $\tau = 0.0$, $\tau = 0.375$, $\tau = 0.5$, and $\tau = 0.75$ for 5 selected matchers from the University Admission data set of the PMMC 2015. We selected these matchers to illustrate the different observations by the Bounded probabilistic evaluation.

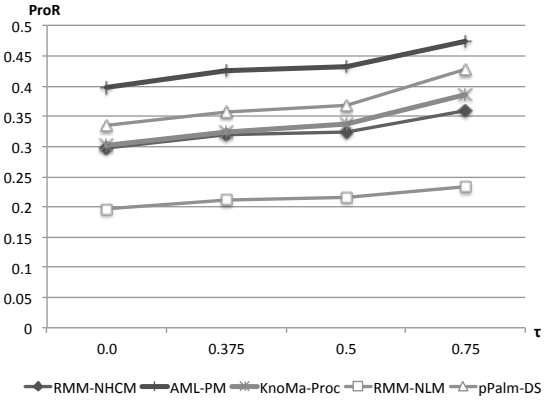
The results from Figure 5.4 show that the effect of a change in the minimum support level τ varies for the different matchers. In general, we observe a decreasing ProP(τ) and an increasing ProR(τ) for higher values of τ. This is intuitive because a higher value of τ results in the consideration of fewer correspondences. However, for some matchers this effect is stronger than for others. For instance, we observe hardly any change in ProP(τ) and a strong increase in ProR(τ) for the matcher *pPalm-DS*. This means that this matcher mainly identifies correspondences with high support. It therefore benefits from a stricter gold standard. The matcher *RMM-NLM* represents a contrasting case. The ProP(τ) of this matcher decreases dramatically with an increase of τ, while its ProR(τ) slightly increases. This reveals that this matcher also identifies a considerable number of correspondences with low support. Since these correspondences turn into false-positives when we increase τ, the ProP(τ) drops respectively.

As we observe in Figure 5.4c, the ProFM(τ) of the matcher KnoMaProc increases with rising τ, because this matcher mostly identifies correspondences with high support values in the non-binary gold standard. Contrary, the performance of the matcher AML-PM decreases with rising τ since this matcher identifies a huge number of correspondences with low support values, which turn into false-positives for the Bounded metrics. Therefore, the ProFM of AML-PM decreases compared to KnoMaProc. Hence, KnoMaProc achieves a higher ProFM at $\tau = 0.75$ than AML-PM.

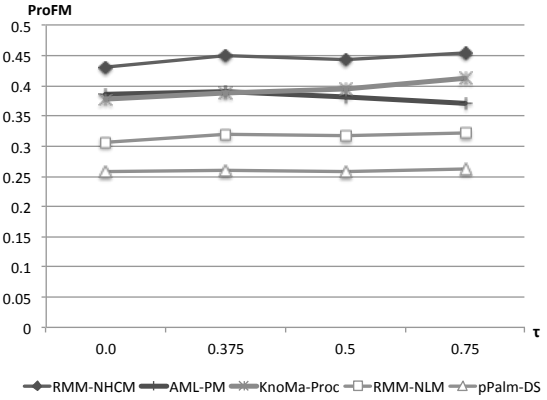
The consideration of the bounded variants of ProP, ProR, and ProFM illustrate that an evaluation based on a non-binary gold standard facilitates a more detailed assessment of specific matchers. It is possible to identify whether a matcher focuses on rather obvious correspondences (with high support) or whether a matcher also identifies less apparent correspondences (with low support). Therefore, it allows for an application dependent evaluation. For instance, it can be considered if



(a) Bounded probabilistic Precision



(b) Bounded probabilistic Recall



(c) Bounded probabilistic F-Measure

Figure 5.4: ProP, ProR, and ProFM for different values of τ

matchers aim at focusing on obvious, thus "sure", correspondences or also focus on computing uncertain correspondences.

5.4.5 Evaluation Using Relative Distance

The relative distance ReD explicitly takes the number of agreements and disagreements with the annotators from the gold standard into account. As a result, matching systems that identify correspondences with low support values are slightly penalized. Table 5.8 gives an overview of the results obtained using this distance measure. It shows for each matcher the ReD value, the ProFM value, the ranks based on the respective measures, and the delta between the ranks.

Rank			**Approach**	**ReD**	**ProFM**	
ReD	ProFM	Δ			mic	mac
1	1	± 0	RMM-NHCM	261.1	.432	.391
2	10	+8	DKP-lite	265.6	.347	.284
3	3	± 0	AML	269.8	.419	.376
4	13	+9	Match-SSS	276.6	.315	.249
5	11	+6	DKP	288.6	.341	.285
6	2	-4	LogMap	295.2	.42	.366
7	14	+7	RMM-NLM	297.6	.312	.253
8	4	-4	Know-Match-SSS	298.8	.411	.358
9	7	-2	OPBOT	313.9	.369	.318
10	9	-1	RMM-SMSL	340.6	.358	.325
11	8	-3	BPLangMatch	343.4	.361	.327
12	6	-6	KnoMa-Proc	344.9	.406	.345
13	15	+2	TripleS	347.4	.301	.21
14	5	-9	AML-PM	510	.408	.395
15	12	-3	RMM-VM2	533.8	.318	.307
16	16	± 0	pPalm-DS	815.7	.275	.261

Table 5.7: Results of probabilistic evaluation with non-binary gold standard for the University Admission data set

The results depicted in Table 5.8 illustrate that the use of ReD has notable effects on the ranking. We can identify several matchers whose rank changed considerably. For instance, the matcher AML-PM went from rank 5 to rank 14 and the matcher DKP-lite went from rank 10 to rank 2. However, it is also interesting to note that the first and the last rank did not change. The matcher RMM-NHCM has both the

lowest ReD value as well as the highest ProFM value. The matcher pPalm-DS has both the highest ReD value as well as the lowest ProFM value. As a result, they remain on the first and the last rank, respectively.

To better understand these results, it is necessary to look into the specific correspondences that the matchers identify. An analysis of the correspondences identified by the matcher AML-PM reveals, for instance, that this matcher establishes a high number of correspondences with low support values. This means that the fairly good ProFM value of AML-PM results from a high number of small rewards for low-support correspondences. Since ReD does not reward but penalizes the identification of such correspondences, ReD is rather high in comparison to other matching systems. For the matcher DKP-lite, which moved 8 ranks up, we observe the opposite effect. This matcher mainly produces correspondences with high support values. While this resulted in a rather moderate ProFM value because of all the unidentified low-support correspondences, the ReD value of this matcher is very low, resulting in a good rank.

The two extreme cases of AML-PM and DKP-lite illustrate that ReD penalizes matchers that identify a high number of correspondences with low support values and rewards matchers that do not. This also reveals the specific characteristics of the matching systems on the first and the last rank. The matcher RMM-NHCM identifies a considerable number of correspondences with high support values. As a result, both ProFM as well as ReD yield good results. The matcher pPalm-DS, by contrast, simply produces a considerable amount of noise. The high number of false positives (and at the same time a low number of true positives) results in a bad performance from the perspective of both measures.

This can also be observed for the Birth Registration data set. The non-binary gold standard contains many alignments with support value 0.125. This effect also leads to the strong ranking differences between ReD and ProFM. (ProFM slightly penalizes matchers which do *not* compute uncertain alignments.)

5.4.6 Robustness of the Results

The advantage of the probabilistic evaluation procedure is that it builds on the individual assessments of a number of annotators. In this way, we circumvent the almost unfeasible task of defining a single set of correct correspondences. However, building on the assessments of annotators also raises the question when the

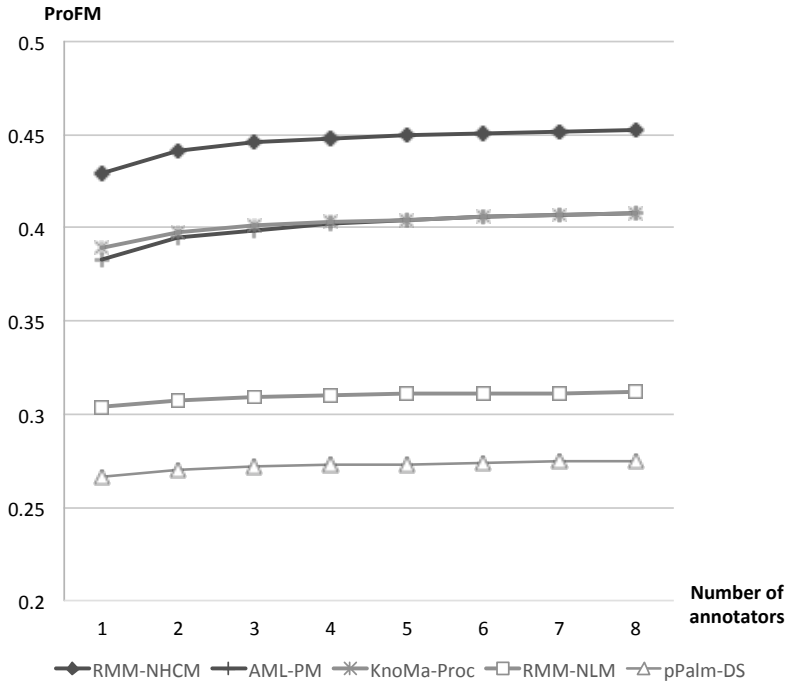
Rank			Approach	ReD	ProFM	
ReD	ProFM	Δ			mic	mac
1	12	+11	Match-SSS	105.4	.476	.323
2	4	+2	Know-Match-SSS	108.0	.527	.390
3	14	+11	RMM-NLM	121.1	.443	.295
4	11	+7	AML	122.0	.490	.410
5	8	+3	I-Match	129.4	.504	.345
6	3	-3	LogMap	148.1	.551	.390
7	2	-5	RMM-NHCM	153.6	.565	.458
8	5	-3	TripleS	164.6	.515	.380
9	6	-3	BPLangMatch	184.1	.511	.390
10	1	-9	OPBOT	216.6	.576	.500
11	10	-1	RMM-SMSL	256.4	.464	.379
12	7	-5	AML-PM	264.1	.509	.439
13	9	-4	pPalm-DS	320.9	.493	.425
14	13	-1	RMM-VM2	352.4	.466	.402
15	15	±0	KnoMa-Proc	630.6	.296	.248

Table 5.8: Results of probabilistic evaluation with non-binary gold standard for the Birth Registration data set

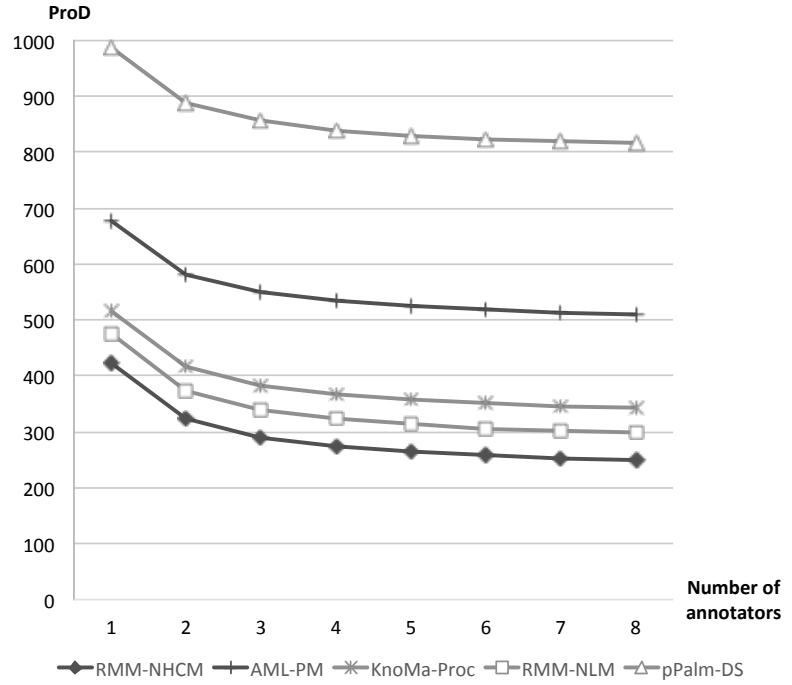
evaluation results actually become robust, i.e., how many annotators are required before the presented performance measures stabilize.

Figure 5.5 illustrates how the ProFM and ReD develop for 5 representative matching systems with an increasing number of annotators for the University Admission data set. To avoid a bias resulting from the order of the annotators (including someone as the 8th annotator who identified a lot of correspondences, would lead to a non-representative movement in the graph), we computed the average values for both evaluation measures based on all possible annotator combinations. For example, the values for 4 annotators are obtained by computing and averaging ProFM and ReD for all possible combinations of 4 annotators in the University Admission data set.

The values in Figure 5.5a show that ProFM converges after only including 4 annotators, i.e., the inclusion of additional annotators has a negligible effect on the results. For instance, the additional correspondences included by the 7th annotator do not even change the third decimal place for most matching systems. For ReD, we observe that more annotators are required. We see that ReD changes quite



(a) Probabilistic F-Measure



(b) relative distance

Figure 5.5: Development of probabilistic evaluation measures with increasing number of annotators

drastically when including additional annotators. This can be explained by the strong effect of low-support correspondences on this measure.

Additional annotators are likely to include more correspondences, which reduces the number of correspondences that are considered as false positives. Despite this rather strong decrease, we still observe that ReD converges. After including 7 annotators, the change is below 2% for all matching systems.

To get insights into the differences between the two annotator groups (student assistants and researchers), we also analyzed the binary assessments from both groups and compared the correspondences they created. We found that the student group came up with more correspondences than the researcher group (825 versus 615). The total number of correspondences where the entire subgroup agreed on a correspondence was, however, slightly higher for the researcher group (242 versus 211). These numbers indicate that the student group had a more diverse view on the correspondences and, as a result, had a higher degree of disagreement. These insights emphasize once again that the idea of consulting several annotators is a promising strategy. The higher the number of annotators, the less individual opinions affect the evaluation.

Altogether, we can state that the presented performance measures stabilize after including 4 to 7 annotators. While we cannot give a general recommendation (independently from the data sets) about the number of annotators that is required, our analysis showed that this number is likely to be below 10. Taking into account that annotators only need to be familiar with the domain and not with process model matching, this is a feasible number.

5.5 Summary, Observations and Findings

In this chapter, we introduced a probabilistic evaluation procedure, which takes a non-binary gold standard as basis for the evaluation. This probabilistic evaluation procedure takes the disputability of correspondences into account. The evaluation experiments in this section illustrate that the presented performance measures have a different focus. The metrics are designed to provide additional insights from different angles. ProFM (together with ProP and ProR) is based on the well-known measure from Information Retrieval and, therefore, might be considered as intuitive by many people. A specific characteristic of this measure is that it rewards matching systems that also recognize correspondences with low support values.

Whether this is a desired outcome, largely depends on the application scenario of the matching system. If the output of a matching system is used as input for humans, i.e., the matching system's task is to suggest possible correspondences, identifying a larger number of correspondences is helpful. A notable advantage of this measure is the low number of annotators that is required for the non-binary gold standard. We found that ProFM already converges after including 4 annotators. The Bounded version of this measure further allows to use a specific threshold, to exclude uncertain correspondences from the evaluation. It further allows a deeper understanding if matchers focus on finding correspondences with low support values or if the matchers aim in identifying the correspondences which have a high support value in the non-binary gold standard. This helps to tune the matcher to a specific application.

In contrast, ReD takes the number of disagreements with the annotators of the non-binary gold standard explicitly into account. As a result, it rather favors matchers that focus on identifying high-support values. In contrast to the ProFM, ReD treats a low probability score closer to 0 than to 1. In other words, correspondences with low support values in the gold standard, computed by a matcher are slightly sanctioned. Therefore, matchers which focus on computing "sure" correspondences achieve better results than matchers which also identify correspondences with low support values in the non-binary gold standard. If this is a desired feature of a matcher, ReD provides a better impression of the performance than ProFM. A small disadvantage of ReD is that it requires more annotators than ProFM to produce stable results. Our analysis showed that ReD converged after including 7 annotators, as opposed to 4 for the ProFM metric.

Our conducted experiments further reveal that the matchers did not fail in identifying correspondences in the Birth Registration data set as it seemed to be the case in the binary experiments of the PMMC 2015. In fact the gold standard of the PMMC 2015 seems to have major shortcomings. This again indicates that a purely binary evaluation does not account for the full complexity of a matching task. This again highlights the risk of a binary evaluation procedure.

In summary, we can state that the choice of the performance measure mainly depends on the application scenario of the evaluated matching system. The evaluation procedure, introduced in this chapter, for conceptual reasons assumes that the output of the matching technique is binary. In fact, some matching techniques compute confidence values that indicate the reliability of the identified correspon-

dences. The transformation of these confidence values into binary values does not only come with the loss of information, but also results in a less accurate assessment of the performance of the matching technique. In the next chapter, we therefore introduce a completely non-binary evaluation measure. In this evaluation procedure the confidence values, computed by the matchers are considered as well. However, the confidence values of the matcher as well as the non-binary gold standard are not considered with their absolute values. The values are only used to transform the matcher output as well as gold standard into a ranked set of correspondences. Then the correlation between both is computed. We will present the rank-correlation in the next chapter.

6

Ranking-based Evaluation

In the previous chapter we introduced a non-binary reference alignment and introduced evaluation metrics which take the non-binary values as basis for the evaluation. However, the results of the matchers were interpreted as binary, because the range of the confidence values differs strongly among the matchers. In this chapter we introduce a fully non-binary evaluation procedure, where the matcher output is ranked according to its confidence values.

The chapter is organized as follows: Section 6.1 introduces the idea of the ranking based-evaluation and provides examples to illustrate the characteristics of the metric. While Section 6.2 presents the experimental results of the ranking-based evaluation and conducts a detailed performance analysis of the matchers of the PMMC 2015, Section 6.3 concludes our findings and provides recommendations for the application of the ranking-based evaluation.

Some of the work presented in this chapter has already been published in Kuss et al. (2017).

6.1 Introduction to the Ranking-based Evaluation

One important aspect of the above proposed evaluation techniques is their restriction to interpret the evaluated alignments of the matchers as binary alignments. Even though some of the matcher's alignments feature a rich distribution of different confidence values, these alignments have been interpreted for conceptual reasons as binary alignments. Therefore, the proposed approach above does not account for the fact that automatically generated alignments are also often annotated with confidence values. For that reason it is better to analyze how close the confidence value distribution in the gold standard is to the confidence value distribution in the generated alignment. The question whether or not a correspondence is correct then needs to be replaced by the question in how far the confidence estimated by a matching technique resembles the confidence in the gold standard. In this approach, we follow this idea and propose an evaluation procedure for comparing a non-binary gold standard against a non-binary alignment. The probabilistic gold standard comprises a ranking of the probability of the correspondences in the gold standard. Low support values stand for a low rank. Similarly, some matcher provide confidence values for each correspondences. Thus the confidence values yield a ranking of all computed correspondences. To calculate the correlation of the correspondences to the reference alignment, the values of the reference alignment as well as the confidence values are only used for generating a ranked gold standard and a ranked matcher output. The rank-correlation then measures the correlation between the ranked matcher output and the ranked reference alignment. Thus, the confidence and support values just affect the ranking of a correspondence, the values itself are not considered in the final calculations of the correlation coefficient.

Note that the fact whether an alignment is annotated with a confidence value is sometimes mixed up with the question whether or not the alignment is generated by a first-line or by a second-line matcher. According to Gal and Sagi (2010), a first-line matcher is defined as a matcher that uses the models themselves as input, while a second-line matcher uses the output of one or several first-line matchers as input, e.g., a set of matrices that store confidence values. For that reason one can assume that the result of a first-line matcher is always a binary alignment, while the result of a second-line matcher might be a non-binary alignment. However, many matching systems are a combination of first-line and second-line matchers

Matcher	Minimum	Maximum	Average	Std. Dev.
AML	0.60	1.00	0.80	0.15
AML-PM	0.30	1.00	0.73	0.23
Know-Match-SSS	0.76	1.00	0.90	0.12
LogMap	0.75	1.00	0.92	0.08
Match-SSS	0.95	1.00	0.99	0.01
pPalm-DS	0.77	1.00	0.85	0.07
TripleS	0.70	1.00	0.84	0.13

Table 6.1: Range of confidence values of process matchers participating in the PMMC 2015 and the OAEI 2016/2017

and, therefore, generate non-binary alignments. The ranking-based evaluation procedure is applicable to any non-binary alignment.

6.1.1 Evaluating with Confidence Values

Table 6.1 shows properties of the non-binary alignments which were generated by the participants of the Process Model Matching Contest 2015 (Antunes et al., 2015) – AML-PM, Match-SSS, pPalm-DS, TripleS – and by the participants of the Process Model Matching track at the Ontology Alignment Evaluation Initiative (OAEI) 2016 (Achichi et al., 2016) – AML, LogMap. The non-binary alignments of the missing matching techniques were not available to us. We therefore excluded them from the analysis.

Table 6.1 shows for each matching system the confidence values of the correspondences with the lowest and highest confidence as well as the average and the standard deviation. The minimum confidence values vary strongly among the matching techniques. We can observe that the alignments of AML-PM contain correspondences with rather low confidence values (i.e., as low as 0.3). Apparently, such a low confidence value was not sufficient for LogMap to include a correspondence in the final alignment. The lowest confidence value included by LogMap is 0.75. This illustrates that the meaning of a confidence value differs considerably among different matchers. An evaluation procedure that analyzes confidence values needs to take this into account in an appropriate way.

One strategy to do so is to normalize the confidence values. To this end, a range for the normalization has to be determined. The confidence values are then extended (i.e., projected) to this defined range. One intuitive choice for such a nor-

malization range is to normalize the confidence values of the matcher to the range of the support values of the alignments in the non-binary gold standard. In case of 10 annotators, these support values range between 0.1 and 1.0. However, in fact, choosing a range for the normalization is arbitrary. Furthermore, some matchers might be closer to the chosen range than others. This also means that the normalization affects some matchers more than others. In Table 6.1 it can be observed that there are matchers using a high threshold (i.e., they have a range between 0.95 and 1.0). For such matching techniques, this small range of confidence values has to be stretched to the full range of support values of the non-binary gold standard. A normalization which affects some matchers stronger than others, would not result in a reasonable assessment. Therefore, also a correlation-based assessment does not deliver meaningful results, due to different ranges of confidence values. The disadvantages associated with normalization can be avoided by applying a different strategy. Instead of comparing the absolute values, the confidence values of the matchers can be used to transform the matcher's output into a ranked list (set) of correspondences. In this way, the confidence values only define a rank of the considered correspondence.

6.1.2 Foundations of the Ranking-based Evaluation

In the following, we introduce and define the ranking-based evaluation procedure for process model matching techniques. The core idea is to compare two non-binary alignments (i.e., the matcher output and a non-binary gold standard) based on comparing the rankings of their correspondences.

Given two process models $\mathcal{M}_1$ and $\mathcal{M}_2$, let $\mathcal{G}$ be a non-binary alignment between $\mathcal{M}_1$ and $\mathcal{M}_2$ that represents the manually created gold standard and $\mathcal{A}$ be a non-binary alignment between $\mathcal{M}_1$ and $\mathcal{M}_2$ that was generated by a matching techniqe. In the following, we show how to compute and use the Spearman's rank-correlation coefficient (Spearman, 1904) to measure the quality of $\mathcal{A}$ given the manually created gold standard $\mathcal{G}$. Let n be the number of correspondences with a confidence value higher than zero in $\mathcal{G}$ or $\mathcal{A}$, i.e.,

$$n = |\{(a_1, a_2) \in act(\mathcal{M}_1) \times act(\mathcal{M}_1) \mid \mathcal{A}(a_1, a_2) > 0 \vee \mathcal{G}(a_1, a_2) > 0\}|.$$

To compute the rank-correlation, the following steps need to be performed independently for both $\mathcal{G}$ and $\mathcal{A}$.

Normalized Ranks The n correspondences in $\mathcal{G}$ and $\mathcal{A}$ have to be ranked according to their confidence values (in increasing order). This leads to a rank of 1 through n for each correspondence. If there are correspondences with the same confidence value, their ranks are normalized. In these cases, which we refer to as ties, the rank of each correspondence with the same confidence value is given by the arithmetic mean of the ranks occupied by these correspondences.

Correction Term for Ties The number of times each value is observed in the alignment is counted. This is denoted by $t_{\mathcal{A},k}$ with respect to $\mathcal{A}$ and $t_{\mathcal{G},k}$ with respect to $\mathcal{G}$. The index k is used here to refer to the different values (or ranks). As a result of this counting, we obtain $\sum_k t_{\mathcal{A},k} = \sum_k t_{\mathcal{G},k} = n$. In the final formula, we need to use the correction terms $T_{\mathcal{A}} = \sum_k \left((t_{\mathcal{A},k})^3 - t_{\mathcal{A},k} \right)$ and $T_{\mathcal{G}} = \sum_k \left((t_{\mathcal{G},k})^3 - t_{\mathcal{G},k} \right)$.

We can now use the following formula to compute Spearman's rank-correlation coefficient, where d_i denotes the difference between the normalized ranks of the i-correspondence from those correspondences that have a positive confidence value in $\mathcal{G}$ or $\mathcal{A}$:

$$\rho = \frac{n^3 - n - \frac{1}{2}T_{\mathcal{G}} - \frac{1}{2}T_{\mathcal{A}} - 6\sum_{i=1}^{n} d_i^2}{\sqrt{(n^3 - n - T_{\mathcal{G}})\,(n^3 - n - T_{\mathcal{A}})}}\,.$$

Table 6.2 illustrates how to compute the correlation coefficient for an illustrative example by showing the resulting values for all intermediate steps. The starting point is a set of 15 correspondences, of which 13 are part of the gold standard $\mathcal{G}$. The alignment $\mathcal{A}$ represents the output of a fictional matching technique and includes confidence values between 0.75 and 1.0. From the values it becomes clear, that $\mathcal{A}$ includes two correspondences that are not part of $\mathcal{G}$, i.e. the correspondences with a confidence value of 0.0. What is more, $\mathcal{A}$ includes several correspondences with a confidence value of 0.0 that have a confidence value of above 0.0 in the gold standard $\mathcal{G}$. To compare $\mathcal{G}$ and $\mathcal{A}$, both alignments are first ranked. Since several correspondences have the same confidence value, the ranks need to be normalized. This results in a total of 9 different ranks for $\mathcal{G}$ and 5 different ranks for $\mathcal{A}$. Based on the frequency of these different ranks (see $t_{\mathcal{G},k}$ and $t_{\mathcal{A},k}$), the rank differences can be computed. The final rank-correlation is -0.07102. Note that, while a high correlation coefficient is in general desirable, this number is hard to interpret in

isolation. A negative correlation might indicate a high number of false-positives, but also that many correspondences with a high rank from $\mathcal{G}$ have a low rank in $\mathcal{A}$. We will point this out in more detail with the experiments in Section 6.2.

To include all possible pairs of correspondences into the calculations, i.e., additionally all true-negative correspondences, would highly increase the total number of correspondences considered. In our considered data sets the average size of a process model is about 24 activities (cf. Table 3.2). For our data sets about 95% of the total number of all possible correspondences are true-negatives. Therefore, the true-negatives dominate the number of correspondences in the gold standard as well as matcher output. Similarly, like in the Accuracy measure, the differences of the rank-correlation between the matchers would decrease, because all matchers share the very high amount of the true-negative correspondences. Therefore, this would not lead to an increase of information.

Gold Standard $\mathcal{G}$					Alignment $\mathcal{A}$					Rank Difference	
Conf.	Rank	Norm.	$t_{\mathcal{G},k}$	$t^3_{\mathcal{G},k}$-$t_{\mathcal{G},k}$	Conf.	Rank	Norm.	$t_{\mathcal{A},k}$	$t^3_{\mathcal{A},k}$-$t_{\mathcal{A},k}$	d_i	d_i^2
0.000	1	1.5	2	6	1.00	15	14	3	24	-12.5	156.25
0.000	2	1.5			0.75	8	8.5	2	6	-7	49
0.125	3	4.5	4	60	0.75	9	8.5			-4	16
0.125	4	4.5			0.00	1	4	7	336	0.5	0.25
0.125	5	4.5			0.00	2	4			0.5	0.25
0.125	6	4.5			0.80	10	10	1	0	-5.5	30.25
0.250	7	7	1	0	0.81	11	11.5	2	6	-4.5	20.25
0.375	8	8	1	0	0.81	12	11.5			-3.5	12.25
0.500	9	9	1	0	0.00	3	4			5	25
0.625	10	10	1	0	0.00	4	4			6	36
0.750	11	11	1	0	0.00	5	4			7	49
0.875	12	12	1	0	0.00	6	4			8	64
1.000	13	14	3	24	0.00	7	4			10	100
1.000	14	14			1.00	13	14			0	0
1.000	15	14			1.00	14	14			0	0
n=15				$T^{\mathcal{G}} = 90$					$T^{\mathcal{A}} = 372$		$\sum_{i=1}^{n} d_i^2 = 558.5$

Table 6.2: Example of correlation coefficient calculation for an alignment $\mathcal{A}$ computed by a matching technique and the gold standard $\mathcal{G}$. The resulting correlation coefficient is ρ = -0.07102

Table 6.3 uses the output of eight exemplary matching techniques to further illustrate how different characteristics of the alignments affect the correlation coefficient. More specifically, it illustrates the effect of three particular characteristics:

- *Differing range of confidence values*: The matching techniques producing the alignments $\mathcal{A}_1$ and $\mathcal{A}_2$ compute different ranges of confidence values. In alignment $\mathcal{A}_1$, the lower bound is 0.3 and the upper bound is 0.8. In alignment $\mathcal{A}_2$, both the lower bound and the upper bound are higher (0.78 and 1). The

lower bound is even considerably higher, which makes the confidence values hardly comparable. However, in the context of the rank-based evaluation, both matching techniques yield the same result. This is the case because the calculation is based on the ranks and not on the absolute confidence values. This example illustrates the idea and the value of the ranking-based evaluation procedure.

- *Missing correspondences*: The alignments $\mathcal{A}_3$, $\mathcal{A}_4$, $\mathcal{A}_5$, and $\mathcal{A}_6$ illustrate the effect of missing correspondences in the produced alignments. We observe that $\mathcal{A}_3$ and $\mathcal{A}_4$ yield quite similar results although alignment $\mathcal{A}_3$ includes all correspondences from $\mathcal{G}$ and $\mathcal{A}_4$ misses the correspondence from line 3. The missing correspondence (which is interpreted as a correspondence with confidence 0.0), results in a slight decrease of the overall correlation coefficient calculated for $\mathcal{A}_4$. However, since the missing correspondence has a very low confidence value (i.e., 0.125), the decrease is marginal. Alignment $\mathcal{A}_5$ shows a case where a more important correspondence is missing (i.e., a correspondence with a confidence value of 0.5 in $\mathcal{G}$). Here, we see that this missing correspondence has a quite considerable effect on the final correlation coefficient because it drops to 0.512. Quite expectedly, this effect is even more severe for alignment $\mathcal{A}_6$, where a correspondence with a confidence value of 1.0 is not included. Note that we inserted a row of zeros to make sure that the lowest rank is always associated with the same value. Without this row, matcher $\mathcal{A}_3$ and matcher $\mathcal{A}_4$ would have the same rank-correlation. However, in our experiments this changed the results only marginally (the fourth decimal digit).

- *Additional correspondences*: Alignment $\mathcal{A}_7$ includes a correspondence that is not part of the gold standard $\mathcal{G}$. However, since it has the lowest rank, the correlation coefficient is only affected marginally. For alignment $\mathcal{A}_8$, we observe a case where the corresponding matching technique has computed an incorrect correspondence with the highest possible confidence value (i.e., 1.0). The final correlation coefficient is affected accordingly.

All in all, the examples from Table 6.3 show that if matching techniques miss or incorrectly identify correspondences with low confidence values in $\mathcal{G}$, the correlation coefficient is only marginally affected. However, if a matching technique computes or misses incorrect correspondences with high confidence values in $\mathcal{G}$,

the correlation coefficient is affected severely. Since this appropriately reflects the desired performance of process model matching techniques, this is a favorable outcome.

$\mathcal{G}$	$\mathcal{A}_1$	$\mathcal{A}_2$	$\mathcal{A}_3$	$\mathcal{A}_4$	$\mathcal{A}_5$	$\mathcal{A}_6$	$\mathcal{A}_7$	$\mathcal{A}_8$
0.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.000	–	–	–	–	–	–	0.70	1.00
0.125	0.30	0.78	0.76	0.00	0.78	0.78	0.78	0.78
0.125	0.30	0.78	0.78	0.78	0.78	0.78	0.78	0.78
0.125	0.30	0.78	0.78	0.78	0.78	0.78	0.78	0.78
0.125	0.30	0.78	0.78	0.78	0.78	0.78	0.78	0.78
0.125	0.30	0.78	0.78	0.78	0.78	0.78	0.78	0.78
0.250	0.40	0.80	0.80	0.80	0.80	0.80	0.80	0.80
0.500	0.41	0.88	0.88	0.88	0.00	0.88	0.88	0.88
1.000	0.80	1.00	1.00	1.00	1.00	0.00	1.00	1.00
ρ	1.000	1.000	0.953	0.929	0.512	0.210	0.996	0.503

Table 6.3: Behavior of rank-correlation illustrated by the output of eight exemplary matchers.

The rank-correlation is not a valid assessment metric for matchers which only compute binary alignments. This would result in correspondences with only two ranks. This is not sufficient to measure a ranking-based correlation. In the next section we apply the rank-correlation to all matchers which included confidence values in their matcher output. This is the case for seven matching systems of the University Admission data set and five of the Birth Registration data set. In this way, we aim to assess if the matchers generate the same distribution of correspondences like in the non-binary gold standard.

6.2 Experiments of the Ranking-based Evaluation

In the following, we compute the rank-correlation for all matchers which compute confidence values. In our case, all matchers which computed ≥ 3 confidence values, are considered in the ranking-based evaluation. (This is valid for seven matchers of the University Admission data set and five of the Birth Registration data set.)

6.2.1 Results of the Ranking-based Evaluation

As a result of applying our evaluation procedure to the output of the considered matching techniques, we obtained a respective rank-correlation coefficient for each matcher. We computed the results with the free software environment R, which is a statistical programming language (R Core Team, 2013). Tables 6.4 and 6.5 summarize these results. It shows the evaluation metrics and the rank (R) for three different evaluation procedures:

- *nB-nB* (non-binary – non-binary): The non-binary evaluation procedure introduced in this paper. The performance is captured using the rank-correlation coefficient (ρ).
- *B-nB* (binary – non-binary): The probabilistic evaluation procedure comparing the binary output of a matcher against a non-binary gold standard. The performance is captured using the probabilistic F-Measure (ProFM), probabilistic Precision (ProP), and probabilistic Recall (ProR).
- *B-B* (binary – binary): The classical evaluation procedure comparing the binary output of a matcher against a binary gold standard. The performance is captured using the F-Measure (FM), Precision (Prec), and Recall (Rec).

	nB-nB		**B-nB**				**B-B**			
Matcher	R	ρ	R	ProFM	ProP	ProR	R	FM	Prec	Rec
AML	1	**.245**	1	**.424**	.806	.288	1	**.702**	.719	**.685**
Match-SSS	2	.223	5	.314	**.828**	.194	2	.608	**.807**	.487
LogMap	3	.153	2	.418	.680	.302	5	.481	.449	.517
Know-Match-SSS	4	.120	3	.409	.676	.293	3	.544	.513	.578
TripleS	5	-.008	6	.300	.519	.211	4	.485	.487	.483
AML-PM	6	-.266	4	.407	.411	**.404**	6	.385	.269	.672
pPalm-DS	7	-.295	7	.276	.230	.346	7	.253	.162	.578

Table 6.4: Results for the seven considered matchers from the University Admission data set from the PMMC 2015 and the OAEI 2016/2017 for three evaluation procedures

The results from Tables 6.4 and 6.5 reveal that there is a weak correlation between the output of some matchers and the non-binary gold standards. Three matchers at the University Admission data set and one of the Birth Registration

	nB-nB		B-nB				B-B		
Matcher	R	ρ	R	ProFM	ProP	ProR	R	FM	Prec	Rec
Match-SSS	1	**.524**	5	.476	**.974**	.315	5	.332	**.922**	.202
Know-Match-SSS	2	.471	1	**.527**	.865	.379	3	.385	.800	.254
TripleS	3	.277	2	.515	.651	.426	4	.384	.613	.280
pPalm-DS	4	.127	4	.493	.469	.521	1	**.459**	.502	.422
AML-PM	5	-.068	3	.509	.513	**.505**	2	.392	.423	.365

Table 6.5: Results for the five considered matchers from the Birth Registration data set from the PMMC 2015 for three evaluation procedures.

data set even have a negative correlation coefficient. This outcome can be explained by the characteristics of the matchers as well as the characteristics of the gold standards. To understand how the characteristics of the matchers can explain this outcome, consider the metrics from the other two evaluation procedures (i.e., B-nB and B-B). All three matchers with a negative correlation coefficient have a particularly low Precision. Apparently, a negative correlation coefficient primarily relates to a high number of false-positives. A notable characteristic of the non-binary gold standard that contributed to the weak correlation is the high number of correspondences with a low support value. The non-binary gold standard of the University Admission data set for example contains a total of 831 correspondences, of which about 20% have the lowest rank, i.e., at most one of the eight annotators has voted for them. It is, thus, not surprising that many matchers miss these correspondences. While the penalty for missing them is rather low, the recall values reveal that this also explains the overall correlation coefficient.

Table 6.6 provides the number of correspondences in the non-binary gold standard of the University Admission data set as well as the matcher output and their union with n correspondences. It can be observed that many matchers compute a very low fraction of correspondences, which are part of the non-binary gold standard, e.g., 110 for M-SSS (140 - (861-831)). The matcher pPalm-DS, which computes a high number of correspondences, computes 181 correct correspondences, 647 are wrong correspondences. Thus, the matcher pPalm-DS misses a high fraction of correspondences and at the same time computes a high number of wrong correspondences.

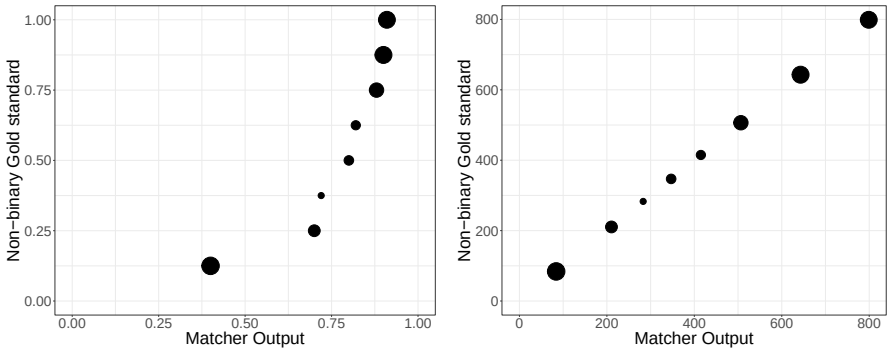
Let us consider two scenarios. The first scenario: a matcher does not compute a correspondence with the lowest confidence value. This correspondence is then

Matcher	**# Alignments in Matcher Output**	n
AML	221	912
AML-PM	579	1178
Know-Match-SSS	261	949
LogMap	267	950
Match-SSS	140	897
pPalm-DS	828	1477
TripleS	230	978

Table 6.6: Number of computed alignments with the corresponding Union with the non-binary gold standard of the matchers exemplary for the University Admission data set

evaluated as a zero for the confidence for the matcher, thus it is included in the group of the lowest rank. In the ranking of the non-binary gold standard, the correspondence is part of the second lowest ranking (the lowest rankings are those which are not in the non-binary gold standard but are in the matcher output). Consequently, the rank-correlation is only affected marginally because both ranks of the non-binary gold standard and the matcher output are close to each other. In contrast, consider the second scenario: the matcher computes a correspondence (with a relatively high confidence value) which is not part of the non-binary gold standard. Such a correspondence has the lowest rank in the non-binary gold standard but a relatively high rank in the matcher output ranking. One reason is that a relative high number of correspondences in the non-binary gold standard have low support values.

In sum, not computing a correspondence in the non-binary gold standard is less critical for the rank-correlation compared to computing a correspondence which is not part of the non-binary gold standard. This effects become larger with larger confidence value in the matcher output. Thus, a matcher which computes a wrong correspondence with high confidence, significantly decreases the rank-correlation; see Table 6.3. This is a desired feature, as the non-binary gold standard can be regarded to contain all reasonable correspondences, by construction.



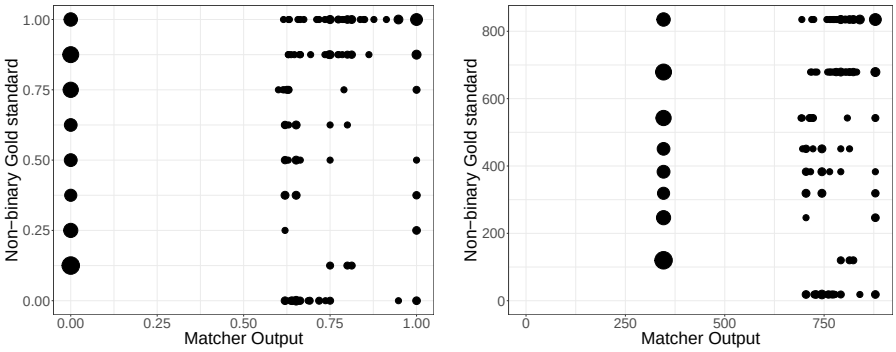
(a) Confidence value (b) Ranking

Figure 6.1: Plots for a matcher with a rank-correlation of 1

6.2.2 Visualization of the Results

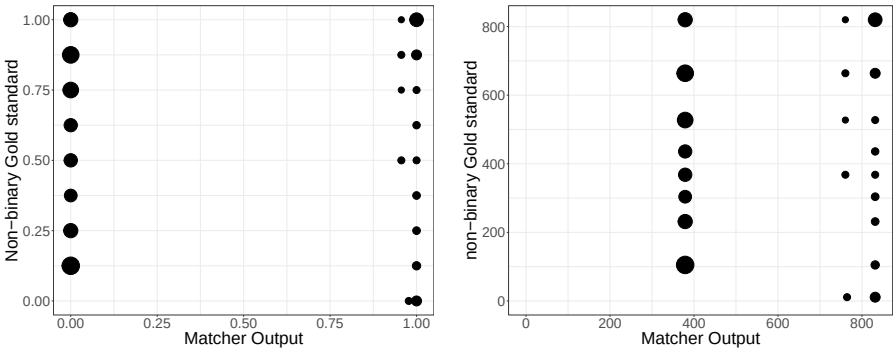
Abstracting from the absolute values, we see that the correlation coefficient allows us to rank the matchers according to their performance. We can observe that the ranking obtained through the evaluation presented here does not always deviate from the ranking we obtain when using the other evaluation procedures. In fact, the matcher AML is always considered to perform best and the matcher pPalm-DS is always considered to be worst for the University Admission data set. However, for the Birth Registration data set, we observe that the rankings of the matchers change considerably.

To understand the variations, consider Figure 6.2, which visualizes the output of the different matching techniques by plotting the confidence values of the generated correspondences against the confidence values of the gold standard. The horizontal axis indicates the confidence values of the gold standard, the vertical axis the confidence values of the matching technique. The size of the dots indicates the number of the correspondences with this particular combination. The bigger the dot, the more correspondences with this combination exist. As discussed earlier, the rank-correlation is a linear measure of dependency after the ranking has been applied. Figure 6.1 shows a "perfect matcher". Note that the standard correlation for this perfect matcher is only 0.911 while the rank-correlation is 1. This highlights the importance of considering the ranks instead of the absolute values. Therefore, the optimal result after ranking is a point cloud resembling the linear line shown in Figure 6.1(b). Looking into the details, we can make the following observations:



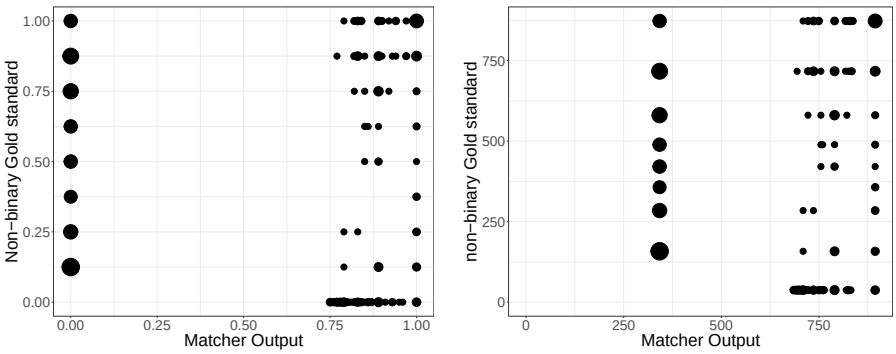
(a) AML: Confidence value

(b) AML: Ranking



(c) MSSS: Confidence value

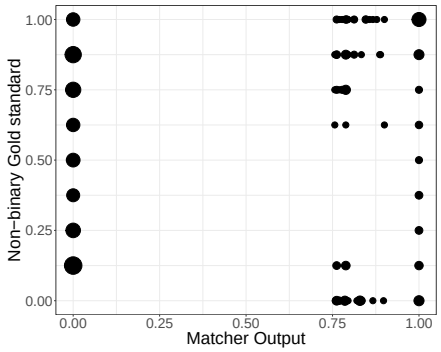
(d) MSSS: Ranking



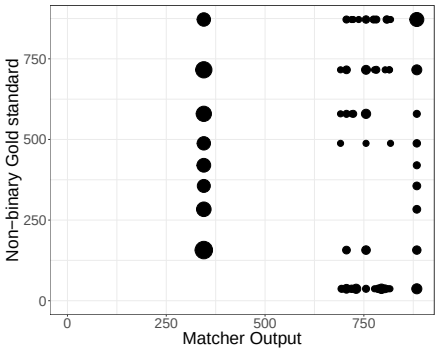
(e) LogMap: Confidence value

(f) LogMap: Ranking

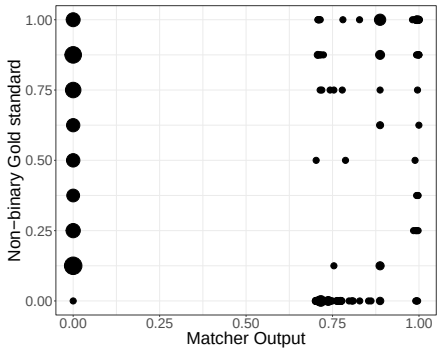
Figure 6.2: Visualization of rank-correlation results for University Admission data set.



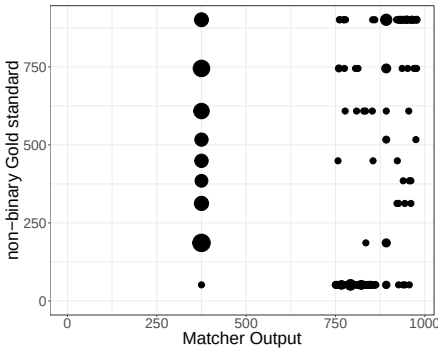
(g) KMSSS: Confidence value



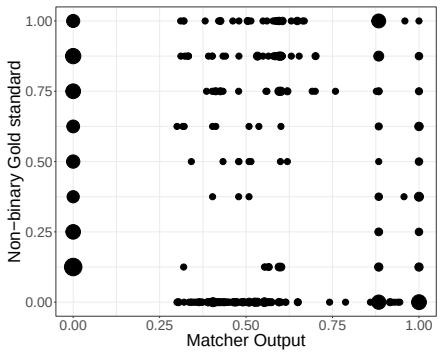
(h) KMSSS: Ranking



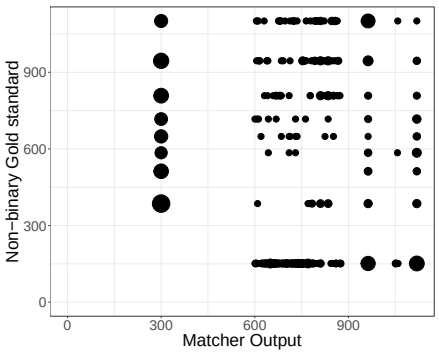
(i) TripleS: Confidence value



(j) TripleS: Ranking

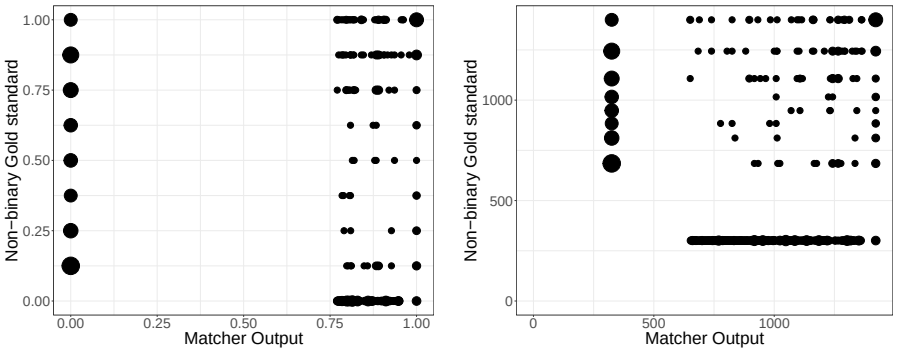


(k) AML-PM: Confidence value



(l) AML-PM: Ranking

Figure 6.2: Visualization of rank-correlation results for University Admission data set (continued).



(m) pPalm-DS: Confidence value

(n) pPalm-DS: Ranking

Figure 6.2: Visualization of rank-correlation results for University Admission data set (continued).

- *AML*: The matcher computes a solid number of correspondences that also have high confidence values according to the gold standard (see upper right corner). At the same time, AML also misses a considerable number of correspondences. The left side of the plot clearly shows that there are many correspondences where the output of the matching technique is 0.0 while the gold standard contains confidence values above zero. Note that gap between 0.0 and 0.6 results from the range of the confidence values generated by AML.
- *MSSS*: The plots for this matching technique illustrate that MSSS identifies a rather small number of correspondences with quite high accuracy. The upper right corner shows that particularly safe indisputable correspondence are identified. Nonetheless, the left side of the plot also highlights that the matcher misses several correspondences, of which quite a notable number have a confidence of above 0.75 in the non-binary gold standard.
- *LogMap*: At first glance, the plot for the matcher LogMap seems to resemble the results from AML. However, the clear difference in the correlation coefficient shows that there is, in fact, a notable difference between the two. The biggest difference can be found in the lower right corner of the plot. LogMap identifies a even higher number of correspondences that are not part of the gold standard with high certainty. This is respectively reflected in the lower rank coefficient.
- *KMSSS*: The matching technique KMSSS mainly produces correspondences with a confidence value of above 0.75 in the gold standard (see upper right

corner). This indicates that this matcher focuses, similarly to MSSS, on rather obvious correspondences. The main difference to MSSS is a lower cutoff value. As a result, the total number of generated correspondences increases. This, however, does not result in a better correlation coefficient.

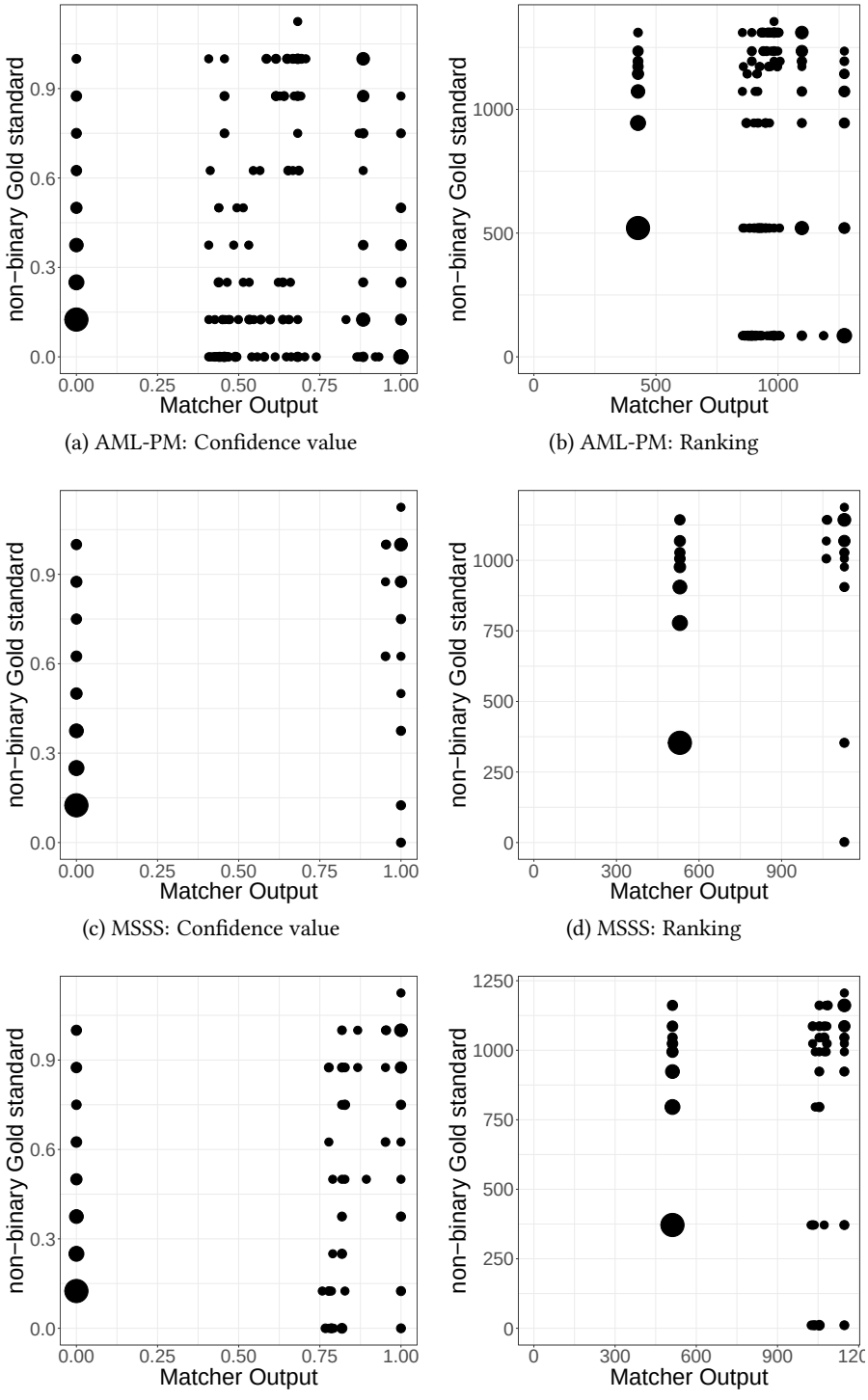
- *TripleS*: The plot for TripleS illustrates that, on the one hand, it misses a high number of correspondences (see left-hand side of the figure). On the other hand, it is not able to identify correspondences with high confidence values in the gold standard with sufficient certainty. As opposed to many other matchers, the big dot in the right upper corner is missing.

- *AML-PM*: The results for the matching technique AML-PM look quite similar to the results of AML, although the range of the generated confidence values is bigger (0.3 to 1.0). What is notable is the high number of false-positives with a high confidence value, indicated by the large dots in the lower right corner. This dramatically decreases the rank-correlation coefficient for AML-PM. Note that AML-PM by far identifies the highest number of correspondences with a confidence of 1.0 that are not included in the gold standard.

- *pPalm-DS*: The matcher pPalm-DS generates a high number of correspondences that are not part of the gold standard (see lower right corner). At the same time, however, it also misses a high number of correspondences (see left-hand side of the plot). Note that this technique generated by far the highest number of correspondences.

Similar behavior can be observed for the Birth Registration data set, as shown in Figure 6.3. However, the data set contains a very high number of correspondences which only one annotator classified as an alignment. The matchers miss a high number of such correspondences which is indicated by the big dot in the lower left corner. However, in the rank-correlation this does not result in a strong effect, because the matchers miss the correspondences with the lowest rank. This explains the better results of the rank-correlation for the Birth Registration data set compared to the University Admission data set. In the following, we explain the results for the Birth Registration data set in more detail.

- *MSSS*: The matcher MSSS identifies a rather small number of correspondences with quite high accuracy. At the same time misses many correspondences of the non-binary gold standard. However, the matcher only computes a



(a) AML-PM: Confidence value

(b) AML-PM: Ranking

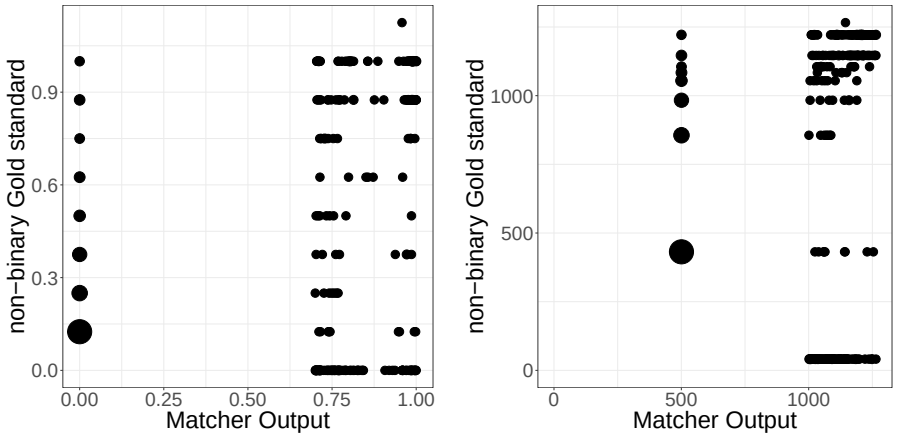
(c) MSSS: Confidence value

(d) MSSS: Ranking

(e) KMSSS: Confidence value

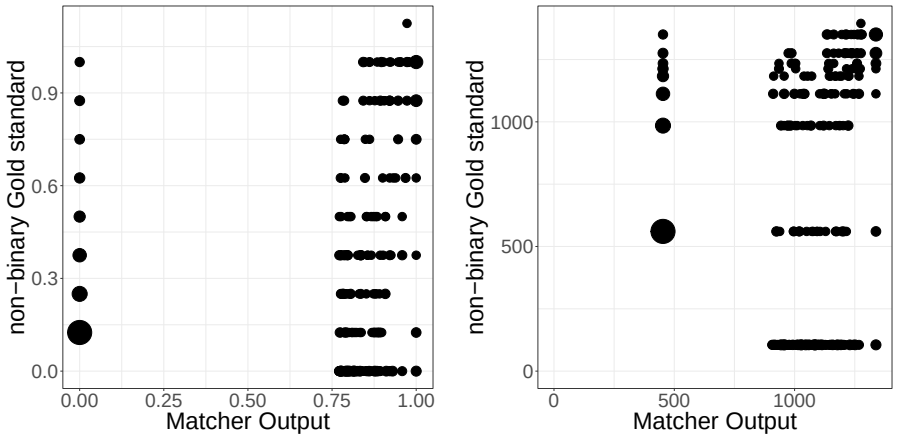
(f) KMSSS: Ranking

Figure 6.3: Visualization of rank-correlation results for the Birth Registration data set.



(g) TripleS: Confidence value

(h) TripleS: Ranking



(i) pPalm-DS: Confidence value

(j) pPalm-DS: Ranking

Figure 6.3: Visualization of rank-correlation results for the Birth Registration data set (continued).

very low number of correspondences which are not part of the gold standard (lower right corner in both plots). That explains the rather good results in the rank-correlation. However, it computes its false-positives with its highest rank, which decreases the rank-correlation considerably.

- *KMSSS*: The matcher KMSSS computes a low number of false-positive alignments but mainly with its lowest rank. Therefore, the matcher yields a rather high correlation coefficient.
- *TripleS*: Similarly like already observed for the the University Admission data set, the matcher misses a high number of correspondences and at the same time generates a high number of false-positive alignments.
- *AML-PM*: Similarly the matcher generates a high number of false-positives with a high confidence value. This indicates that the confidence values which the matcher computes are not accurate and do not resemble the distribution of the non-binary gold standard.
- *pPalm-DS*: Similarly like already observed at the University Admission data set the matcher pPalm-DS generates a high number of correspondences that are not part of the gold standard (see lower right corner). At the same time, however, it also misses a high number of correspondences (see left-hand side of the plots). This is a surprising observation since this matcher had a rather high performance for this data set at the *binary* gold standard used at the PMMC 2015.

The described metric provides an accurate correlation measure for matchers which provide confidence values. The rank-correlation implies a normalization of the confidence values, computed by a matcher, by normalizing the matcher output into a ranked collection of the computed correspondences. Compared to the ranked reference alignment, it provides a correlation which is not based on the absolute values, but on the rank of the correspondences. However, the rank-correlation is not a suitable measure for matcher which do not compute confidence scores since they only provide binary results, thus computes only correspondences with two ranks. This is not enough to compute stable results for a ranking-based evaluation. (Binary values do not impose a natural ranking of the computed correspondences.)

All in all, this analysis highlights a major difference of the presented evaluation procedure to existing ones: The confidence of the matcher is taken into account.

If a matcher identifies a correspondence that is not part of the gold standard with high certainty, the penalty is much higher than if the certainty is low. This is an important difference to both the B-nB and B-B evaluation procedures where the output of the matcher is considered as zero or one.

This particular feature of our evaluation procedure also explains the different rankings. Matching techniques that identify false-positives with high certainty receive a bigger penalty than matching systems that identify false-positives with low certainty. This complements the metrics, introduced in Chapter 5. In this way it can be observed, whether the confidence values of the matchers reflect the extent of confidence of the computed correspondences, or whether the confidence values of the matchers do not correlate with the confidence values in the non-binary gold standard.

Both, false-positives and false-negatives have a high negative impact on the measured results. Therefore, the metric strictly measures any differences to the reference alignment. The metric takes into account the importance of the correspondences (generated and in the reference alignment) it does not measure a correlation for the computed values, but for the ranking.

If the results of one single matcher has to be validated, then the results of the ProFM is easier to interpret than the ranking-based evaluation. The ranking based evaluation depends on many variables which makes an interpretation for improvement more difficult. However, the ranking-based evaluation can indicate if matchers compute FP alignments with high confidence or if matchers miss alignments with high confidence; e.g., the evaluation can validate the confidence values, computed by a matcher. The evaluation procedure requires a variety of confidence values of the matchers as well as in the gold standard. Otherwise too many correspondences share the same rank.

6.3 Conclusions

In this chapter, we introduced a fully non-binary evaluation method, which takes the confidence values of the matchers as well as the support values of the non-binary gold standard as basis for the evaluation. However, the values itself are not considered in the calculations, instead they are considered to transform the matcher output as well as non-binary gold standard into a ranked collection of correspondences.

The rank-correlation directly translates the properties of the non-binary gold standard in that it assumes that the non-binary gold standard is almost complete. With other words, the non-binary gold standard is designed to avoid the randomness/arbitrariness of the gold standard. Consequently, the non-binary gold standard is robust with respect to different annotators, i.e., changing the annotators, or increasing the number of annotators, won't alter the non-binary gold standard much, while it might have a significant effect on the binary gold standard. Therefore, the rank-correlation strictly measures the existence of false-positive or missing correspondences with a high rank in the reference alignment. This also helps to understand if the confidence values of the matchers reflect a realistic value.

With the different evaluation measures, presented in this thesis, we assess the quality of the generated correspondences of the matchers from different angles and allow for an application-dependent evaluation. In the next chapter, we will introduce a conceptually new evaluation measure, which classifies the matching task as well as matcher output into matching patterns. These matching patterns provide additional information about specific strength and weaknesses of a matcher, by dividing the matching task into categories with differing complexity level. We introduce the categories in the next chapter.

7

Evaluation by Automatic Classification to Matching Patterns

The experiments at the Process Model Matching Contests showed that currently no matching technique has a high performance on all tested data sets (Achichi et al., 2016, 2017; Antunes et al., 2015). Consequently, the evaluation needs to offer insights into strengths and weaknesses of each matching technique. On the one hand, this enables to use matching techniques for specific applications which fit the patterns of a matching technique. On the other hand, the matchers can be tuned to specific application scenarios and therefore weaknesses can be eliminated. One important instrument to perceive strengths and weaknesses of a matcher is to analyze the correspondences which the matchers compute. In this way, it can be analyzed which types of correspondences the matchers can identify reliably and

which are especially challenging. It further indicates what kind of correspondences lead to false-positive or false-negative alignments in the matcher output. Currently, this is assessed manually, by manually analyzing the matcher output. However, this manual assessment comes with high efforts.

To automatize this expensive task, we propose a conceptually different evaluation approach. The basis of our idea is to group the alignments of the matcher output as well as the gold standard into different categories. In our example, we utilize five different such categories. For each of the category, the widely-used measures Precision, Recall and F-Measure are computed. These metrics for each category enable us then to analyze a matcher's performance in greater detail. In this way the matching task is divided into groups with specific attributes. This way, we gain insights which matcher performs well on "trivial" correspondences or can also identify "difficult" correspondences. Among the more complex correspondences, we learn for each individual matcher which correspondences can be found and which are challenging. In particular, it is possible to predict the performance of matching techniques for specific applications. Therefore, these insights allow to differentiate for each specific application, which matching technique is suitable for the specific matching task. Sometimes information about the data set are available in advance and therefore desired features of a matching technique can be known in advance. In our experiments we will show that, for some data sets, it is already sufficient to compute stable results if matchers focus to classify syntactically identical labels ("trivial" correspondences) in process models.

The remainder of the chapter is organized as follows. Section 7.1 introduces the matching patterns which are automatically assigned and provides examples to illustrate these patterns in Section 7.2. In Section 7.3, the evaluation metric is introduced. In Section 7.4, evaluation experiments are obtained and the results of the evaluation by matching patterns are discussed in detail. Section 7.5 provides a conclusion of the introduced evaluation method.

Some of the work presented in this chapter has already been published in Kuss and Stuckenschmidt (2017).

7.1 Introduction to the Automatic Classification into Matching Patterns

Currently, the evaluation focuses on grading the evaluated matchers. Thus, the evaluation is designed to provide a matcher's rank within a group of matchers. Most evaluation methods are not designed to provide a detailed analysis of an individual matcher output. To obtain such detailed information, currently it is required to manually process and interpret the matcher output to identify possibilities for improvement. In contrast, we propose a new evaluation technique which provides such information for an individual matcher output, without the need for manual processing. In the first step of the proposed evaluation method, the correspondences in the gold standard are assigned to the different categories. The same is done for the matcher output in the second step. Both steps are done automatically – no user input is required. Then, each category is treated as its own matching problem where standard metrics can be applied. By automatic annotation to matching patterns, the matching problem itself is classified into categories. These categories divide the matching problem into different levels of complexity. On the one hand, this helps gaining insights about the complexity of the matching task itself, on the other hand, this helps understanding strengths and weaknesses of a matcher. Equally important, the categorization provides possibilities for the improvement of a matcher's performance.

Furthermore, the categorization helps to understand the performance of a matcher for specific matching tasks. Sometimes a matcher needs to satisfy different tasks. For example, when finding similar process models in a database it may be required to be able to identify a high fraction of correspondences which are only semantically identical, thus have only limited syntactical overlap. However, for some applications it is simply required to compute a high fraction of identical labels. For an efficient evaluation it is necessary that the evaluation is able to take the specific application scenarios into account.

Currently all correspondences are evaluated together, although the kind of correspondences in a data set differs notably. We know, for instance, that there is a high fraction of "trivial" correspondences in some data sets. Moreover, the complexity differs significantly with regard to different syntactical overlap which the activities share. To make this more clear, consider the following example:

- **C1**: *Send application to selection committee – Forward documents*

The correspondence C1 is a complex correspondence, because both activities have no syntactical identical word in common.

- **C2**: *Invite applicant for appointment – Invite applicant for interview*

In contrast, C2 has a much lower complexity level, since both activities have a high syntactic overlap. In this case, after stemming is applied, $2/3$ of the words are syntactically identical. Therefore, it is not reasonable to evaluate all correspondences as a whole. In fact, the evaluation needs to differentiate between different complexity levels, to allow for a detailed perfomance analysis. On the one hand to better assess the performance of a matcher and on the other hand to find strengths and weaknesses of matchers.

The question which arises is which categories actually are useful and which can be assigned automatically. We want to assess this by the fraction of syntactical overlap of the activities in the process models. Therefore, we propose a stepped complexity depending on the syntactical overlap, which we explain in the following:

- **Trivial correspondences:**

  A natural choice is to test if matchers are able to detect trivial correspondences, e.g., correspondences which are syntactically identical after basic stemming has been applied. (This classification can be done automatically.)

- **Correspondences which share one word or the verb, e.g., one word is syntactical identical:**

  To achieve a detailed analysis, it is useful to have categories with different complexity level. Complexity can be measured by the fraction of syntactical overlap of correspondences. If correspondences share only one word, then such a correspondence may be complex to detect, because the other words in the activities do not syntactically overlap and are only semantically identical

or similar. However, some matchers may already compute a correspondence if both activities have only one word in common. This results in a low Precision for this category. With this proposed category we can detect such shortcomings. We are further interested to learn if it makes a difference for the matching results if the common word is a verb or any other word. Most matchers combine the bag-of-word with the Lin-similarity. That means each words of a label are compared with the corresponding words of the corresponding label and the highest score for each combination is computed. Hence, sometimes the similarity score is already high, if two labels share one word. Therefore, we want to learn if matchers have a high false-positive rate in this category. (This classification can be done automatically.)

- **Correspondences which share two or more words:**

  To achieve a "stepped" complexity level, it is interesting to learn how the results change with decreasing complexity. Therefore, we propose a category which includes correspondences with two or more identical words. This category is then more complex than the "trivial" category. It is interesting to observe if matchers already fail to have a high F-Measure in this category. Therefore, we gain stepped complexity levels of the proposed categories. Moreover, this naturally leads to the advantage that all correspondences of the data set are assigned to a category (exclusively). This is an important aspect for the evaluation, to fully obtain insight which kind of correspondences matchers compute. If matchers compute a high fraction of correspondences which are not part of any category, those correspondences would not be classified and therefore the information is lost. In this way, we circumvent this problem. (This classification can be done automatically.)

- **Complex correspondences (like synonyms):**

  Most difficult to match are correspondences which have no syntactical overlap (as in our example C1 above). This is the case for example for synonyms. However, to test matchers with the automatic generation of synonyms leads to the problem that a database is required to access synonyms. Mostly WordNet (Miller, 1995) is used in such cases. To generate test cases with synonyms automatically, the synonyms have to be obtained from such a database. If matchers use the same database as for the synonym generation, they are rewarded. This is contrary to the idea of objective evaluation experiments.

Therefore, we propose to classify a category, which contains alignments which have no syntactical overlap but can be assigned automatically. Such a category then contains complex alignments which have only a semantic similarity. In this way, we avoid the problem of artificially generated categories. In fact, we obtain a complex category with real-world data, without any manipulation, which allows to learn if matchers can deal with synonyms. To complement this complex category we propose categories with stepped complexity levels, like the categories explained above. (This classification can be done automatically.)

In the next section we introduce the categories, which we automatically assign. Furthermore, we illustrate the categories with examples from the data sets of the Process Model Matching Contest 2015 (Antunes et al., 2015).

7.2 The Categories

The core of the evaluation via matching patterns is to automatically assign correspondences of the reference alignment as well as the computed alignments to groups with specific attributes. After the automatically generated classification to one of the categories, the well-known metrics Precision, Recall and F-Measure (Manning et al., 2008) are calculated for each of theses categories separately. As a consequence, the matching task is divided into groups with specific attributes. In the following, we define the categories and illustrate these with examples. The categories are chosen to provide a deeper knowledge about specific attributes which are important features of a matching technique. Note that the specific numbering of the categories is not related to the complexity level of the corresponding category.

Definition 11 (Normalization)**.** *For the classification, an activity is normalized, if (1) all stop words are removed, (2) stemming has been applied and (3) case sensitivity is ignored.*

Example of stop-words are “of”, “for” and “the”. Examples for stemming are “checking” and “checks” transformed into “check”. (This is a basic step for matchers.)

Definition 12 (Categorization, category)**.** *Let A_1, A_2 be the activity sets for two process models P_1 and P_2. A categorization is a partition into disjoint sets $C(i)$, i.e., $C(i) \subseteq A_1 \times A_2$ for all i with $\cup_i C(i) = A_1 \times A_2$ and $C(i) \cap C(j) = \emptyset$ for all $i \neq j$. Any $C(i)$ is a category.*

It is important to note that each correspondence is assigned to one category exclusively. All correspondences are assigned to a category.

In general, the categories should be chosen carefully. The following characteristics have to hold when choosing the categories:

- the classification has to be assigned automatically,
- the categories have to resemble difficult matching problems,
- the categories have to resemble trivial matching problems,
- there needs to be a reasonable number of correspondences in each category. Too many categories lead to too few alignments in each category; too few categories lack information.

Therefore, we propose the following categories (the examples are extracted from the gold standard of the data sets of the PMMC 2015 (Antunes et al., 2015)):

Category "trivial": This category contains alignments which are identical after normalization.

All remaining correspondences, which are not in Category "trivial", are assigned to one of the following categories:

Category I "no word identical": Alignments which have no word in common after normalization are assigned to this category. Examples:

Example 1: *Evaluate – Assessment of application*

Example 2: *Hand application over to examining board – Send documents to selection committee*

[The stop word "to" is ignored and not counted as an identical word.]

Example 3: *Talk to applicant – Do the interview*

Example 4: *Shipping – Delivery and Transportation Preparation*

Example 5: *Shipment – Transportation Planning and Processing*

Category II "one verb identical": Alignments which are assigned to this category have exactly one identical verb after normalization. No other words are identical. Examples:

Example 6: ***Send*** *documents by post – **Send** all the requirements to the secretarial office for students*

Example 7: ***Wait*** *for results – **Waiting** for response*

[This example illustrates two specific characteristics: the verb is normalized (stemming), the stop word (in this case "for") is ignored.]

Example 8: ***Send*** *acceptance – **Send** commitment*

Example 9: ***Check*** *data – **Check** documents*

Category III "one word identical": This category consists of alignments which have exactly one word (but not a verb) in common after normalization. Examples:

Example 10: *Talk to **applicant** – Appoint **applicant***

Example 11: *Hand **application** over to examining board – Send **application** to selection committee*

[In this example the stop word "to" is ignored.]

Example 12: *Apply **online** – Fill in **online** form of application*

Example 13: ***Invoice*** *approval – **Invoice** Verification*

Category IV "$\geq$ two words identical": This category consists of correspondences which share ≥ 2 words. Examples are:

Example 14: ***Send application*** *– **Send application** form and documents*

Example 15: ***Send*** *documents to **selection committee** – **Send** application to **selection committee***

Example 16: ***Receiving*** *the written **applications** – **Receive application***

Example 17: ***Time Sheet*** *Approval – **Time Sheet** Permit*

One important aspect is the different complexity level of the described categories. In the following, we discuss each category with increasing complexity level of the categories.

The *Cat. trivial* contains identical labels after normalization. Only basic syntactical matching techniques are required to identify such correspondences. This is

important to assess since matching techniques are required to achieve very precise results in this category.

The *Cat. IV* contains only alignments which share two or more identical words. Therefore this category is a category with less complex alignments compared to Cat. I through Cat. III.

Cat. II and Cat. III have a rather high complexity level, since these categories have just one word / one verb in common. Both categories can further indicate if a matcher produces already a high fraction of alignments if one word or the verb between two labels are identical.

Cat. I, however, is the most complex category among the introduced categories, since these alignments have no word in common. They have no syntactical overlap. Consequently these alignments just have a semantic connection, like this is the case for synonyms. To identify alignments from this category correctly, a matcher requires advanced semantic knowledge.

Note that each alignment is assigned to exactly one of these categories exclusively, i.e., the alignments cannot be assigned to several categories. The above described categories are a partitioning of all possible correspondences in a data set, i.e., each possible correspondence is assigned to exactly one of the five categories above, because any non-trivial correspondence has either no identical words, one identical verb, one identical word which is not a verb or two or more identical words.

Figure 7.1 illustrates a simplified example of a reference alignment which is assigned to the above described categories. The figure shows two example process models, which illustrate the application process of Master students at two universities. A matcher's task is to identify correspondences of one process model in the other process model. The correspondences of this matching task are marked with different gray scales for each of the introduced categories above. For each introduced category there is one example in the figure. However, the Cat. IV is illustrated with two examples of the reference alignment. Note that the illustrated figure is an example of a reference alignment. For the evaluation procedure also the alignments computed by a matcher are classified to the matching patterns.

Figure 7.2 illustrates the conceptual structure of the automatically assigned categories. The matching problem is divided into "trivial" and "non-trivial" alignments. "Trivial" alignments are any alignments which are identical or identical after nor-

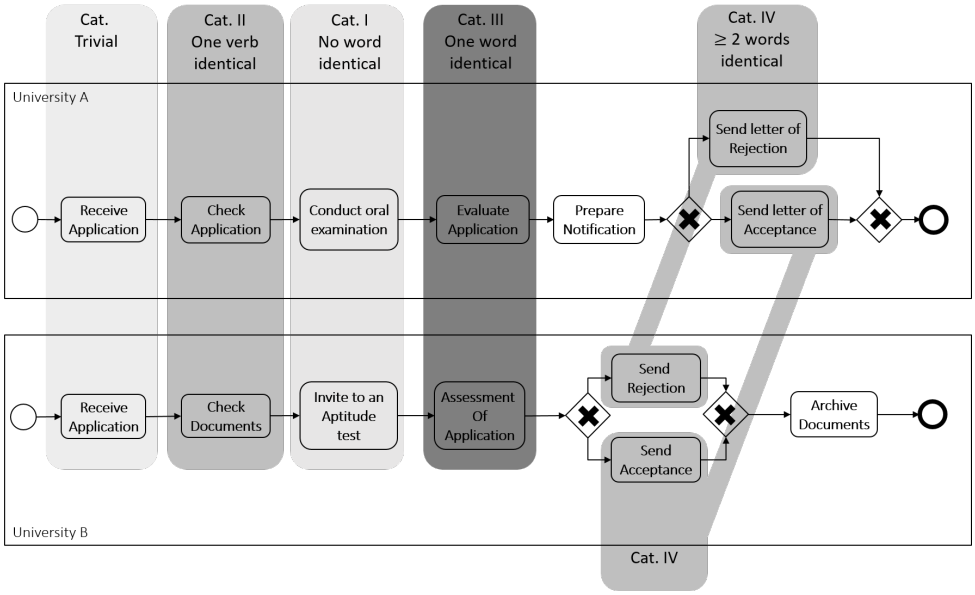


Figure 7.1: Example of a categorized reference alignment

malization. "Non-trivial" alignments are all other alignments. The "non-trivial" category consists of four sub-categories.

7.3 Metrics for the Categories

We define the category-dependent metrics as follows:

Definition 13 (Category-dependent Precision, category-dependent Recall, category-dependent F-Measure)**.** *Let set G be the gold standard, O be the matcher output and $C(i)$ be the categories. The set $G(i)$ is the collection of reference alignments assigned to category $C(i)$, i.e., $G(i) = G \cap C(i)$ for all i. Similarly, $O(i)$ is the collection of correspondences computed by a matcher and assigned to category $C(i)$; i.e., $O(i) = O \cap C(i)$ for all i.*

The category-dependent Precision, $cP(i)$, is defined as

$$cP(i) = \frac{|G(i) \cap O(i)|}{|O(i)|}$$

and the category-dependent Recall, $cR(i)$, is given by

$$cR(i) = \frac{|G(i) \cap O(i)|}{|G(i)|} .$$

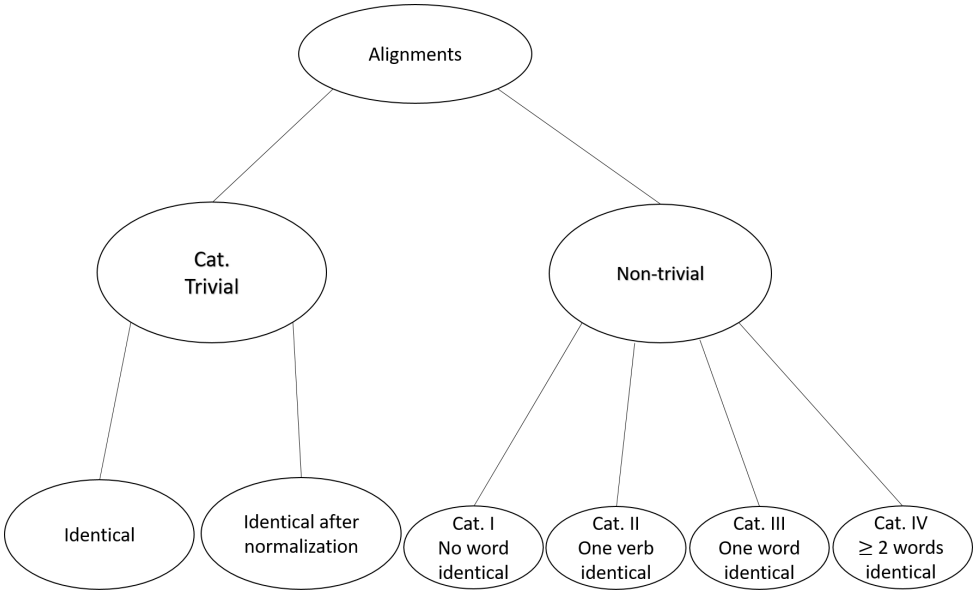


Figure 7.2: Structural dependencies of the categories

The category-dependent F-Measure, $cFM(i)$, is then

$$cFM(i) = 2 \cdot \frac{cP(i) \cdot cR(i)}{cP(i) + cR(i)} .$$

cP is the fraction of correctly computed alignments to all computed alignments in the category. cR is the fraction of correctly computed alignments to all correct correspondences (with respect to the gold standard) in category i. Both, category-dependent Precision and Recall are values between 0.0 and 1.0. A category-dependent Precision of 1.0 means that all computed correspondences in the corresponding category are contained in the gold standard, i.e., $O(i) \subseteq G(i)$. In contrast, a Recall of 1.0 means that all correspondences of the gold standard are computed, i.e. $G(i) \subseteq O(i)$. The cFM(i) is the harmonic mean of the category-dependent Precision (cP) and Recall (cR). All alignments in the gold standard as well as the matcher output are assigned to exactly one category exclusively, i.e. there is no overlap between these categories. After this, category-dependent Precision, Recall and F-Measure of the alignments are calculated. That means, the categories are evaluated separately and independently.

7.4 Experiments

As demonstrated in the previous experiments, we apply the proposed evaluation procedure to the data sets and participating matchers of the Business Process Model Matching Contest 2015 and the Process Model Matching Track at the OAEI 2016 and 2017. We will show, which insights the evaluation via matching patterns offers. We will learn more about the characteristics of the applied data sets and highlight strengths and weaknesses of matchers. We want to find out if matchers are able to detect "complex" correspondences and if "trivial" correspondences can be found reliably. Moreover, we aim to acquire knowledge if the observations are consistent for all data sets. We will further apply the matching patterns to the probabilistic evaluation, introduced in Section 5.1.

Additionally to the already introduced data sets of Section 5.4, we apply the evaluation procedure to the Asset Management data set. This data set consists of 36 model pairs of a SAP Reference Model collection which describe processes in the area of finance and accounting. This data set was first introduced and applied at the PMMC 2015.

In our experiments, the matching patterns are assigned automatically to the gold standard, as well as to the matcher output of the matchers which participated in the PMMC 2015 and the PMMT at the OAEI 2016 and 2017. Then category-dependent Precision, Recall and F-Measure are computed for each category separately. After application of the matching patterns to the gold standard as well as to the alignments computed by the matchers, the following results are computed.[1]

7.4.1 Computational Results

In the following, we provide the experimental results of the categorization of the matching task for all data sets and gold standards of the PMMC 2015 and the Process Model Matching Track of the OAEI 2016 and 2017. We further compare the binary to the non-binary results.

Tables 7.1-7.3 illustrate the results for each data set. The first column provides a list of all participating matchers. They are listed in alphabetic order. In the second column, the F-Measure (FM) over all matching patterns is reported as the

[1] The implementation of the matching patterns, containing the automatic annotation can be accessed here:`https://github.com/kristiankolthoff/PMMC-Evaluator/tree/master/src/main/java/de/unima/ki/pmmc/evaluator/annotator`

Approach	FM	Cat. trivial [44.3%][103]			Cat. I no word iden. [29.3%][68]			Cat. II one verb iden. [11.6%][27]			Cat. III one word iden. [7.3%][17]			Cat. IV ≥ two words iden. [7.3%][17]		
		cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM
AML	**.698**	.844	.959	.862	**.952**	**.595**	**.623**	**.833**	.300	**.311**	**.667**	.372	**.344**	.157	.576	**.183**
AML-PM	.385	.844	.963	.864	.458	.397	.334	.187	**.633**	.217	.045	**.500**	.069	.112	**.970**	.151
BPLangM	.397	**.939**	.816	.864	–	–	–	.462	.344	.262	.152	**.526**	.175	.084	.348	.094
DKP	.538	.844	.968	.867	.267	.048	.070	–	–	–	–	–	–	.136	.227	.099
DKP-lite	.534	.844	.968	.867	–	–	–	–	–	–	–	–	–	.136	.227	.099
I-Match	.472	.907	.942	**.924**	–	–	–	.400	.074	.125	–	–	–	**.500**	.059	.105
KnoMa-Proc	.394	.833	.931	.845	–	–	–	.078	.133	.067	.068	.346	.092	.052	.409	.066
KMSSS	.544	.846	**1.0**	**.883**	.450	.172	.151	.500	.289	.251	.357	.205	.164	.142	.636	.152
LogMap	.481	.844	.978	.872	–	–	–	.467	.167	.127	.082	.372	.094	.092	.530	.110
MSSS	.608	.844	.968	.867	.500	.069	.057	**.833**	**.500**	**.489**	–	–	–	.143	.091	.083
OPBOT	.601	**.978**	.706	.774	.713	**.468**	**.433**	**.562**	.322	.290	.432	**.500**	**.333**	.128	.530	**.164**
pPalm-DS	.253	.843	.986	.874	–	–	–	.053	.344	.072	.029	.410	.046	.062	**.939**	.086
RMM-NHCM	**.668**	**.954**	.930	**.928**	**.821**	.374	.397	.452	**.456**	**.292**	**.550**	.372	**.302**	**.178**	.439	**.166**
RMM-NLM	**.636**	.843	**1.0**	.881	.486	.324	.303	–	–	–	–	–	–	**1.0**	.091	.091
RMM-SMSL	.543	.844	.912	.839	**.778**	**.423**	**.439**	.152	.311	.121	–	–	–	.087	.121	.058
RMM-VM2	.293	.825	.767	.759	–	–	–	.044	.367	.065	.040	.372	.058	.081	**.742**	.110
TripleS	.485	.843	**1.0**	.881	–	–	–	.077	.156	.072	**.625**	.179	.185	.025	.121	.029

Table 7.1: Results of University Admission data set

Approach	FM	Cat. trivial [45.9%][102]			Cat. I no word iden. [34.2%][76]			Cat. II one verb iden. [0.9%][2]			Cat. III one word iden. [8.1%][18]			Cat. IV ≥ two words iden. [10.8%][24]		
		cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM
AML-PM	**.677**	**.996**	**1.0**	**.998**	**1.0**	**.059**	**.059**	**.667**	**1.0**	**.667**	.231	**.396**	**.219**	.571	**.742**	**.565**
BPLangM	.646	**.996**	.951	.970	**.300**	**.176**	**.149**	**1.0**	.500	.500	.200	.354	**.161**	**.646**	.706	**.528**
KnoMa-Proc	.355	.251	.968	.367	–	–	–	.083	**1.0**	.119	.100	.062	.042	.220	.570	.237
KMSSS	.579	**.996**	**1.0**	**.998**	–	–	–	–	–	–	**.333**	.062	.067	.342	.552	.297
MSSS	.619	**.996**	**1.0**	**.998**	–	–	–	–	–	–	–	–	–	.417	.127	.119
OPBOT	.639	**.996**	**1.0**	**.998**	.250	.026	.033	.500	**1.0**	.500	.286	**.469**	**.252**	.640	**.891**	**.653**
pPalm-DS	.474	**.996**	**1.0**	**.998**	–	–	–	**1.0**	**1.0**	**1.0**	.243	.312	**.161**	.301	**.909**	.333
RMM-NHCM	**.661**	**.996**	**1.0**	**.998**	–	–	–	–	–	–	**.500**	.062	.083	**.667**	.303	.290
RMM-NLM	**.653**	**.996**	**1.0**	**.998**	–	–	–	–	–	–	–	–	–	**1.0**	.194	.217
RMM-SMSL	.354	.990	.582	.659	–	–	–	–	–	–	**.333**	.177	.144	.333	.109	.089
RMM-VM2	.603	**.996**	.962	.976	**1.0**	**.059**	**.059**	**.667**	**1.0**	**.667**	.131	**.417**	.126	.450	.612	.418
TripleS	.578	**.996**	**1.0**	**.998**	–	–	–	–	–	–	.111	.062	.048	.372	.633	.324

Table 7.2: Results of Asset Management data set

micro value, i.e. it is computed over all test cases. The remaining columns provide the category-dependent Precision (cP), Recall (cR) and F-Measure (cFM) for each matcher in each category. cP, cR and cFM are macro values, independently computed for each of the matching patterns. For each category, the tables further show in the heading the fraction of correspondences from the whole data set as well as the total number of correspondences of a category in the gold standard. The best three matchers are highlighted in each category. One central observation is the distribution of the correspondences in the reference alignments. This aids in understanding the complexity level of the applied data sets.

Approach	FM	Cat. trivial [4.5%][26]			Cat. I no word iden. [75.0%][437]			Cat. II one verb iden. [1.5%][9]			Cat. III one word iden. [9.9%][58]			Cat. IV ≥ two words iden. [9.1%][53]		
		cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM
AML	.420	.759	.846	.800	.427	**.364**	**.393**	.133	.222	.167	.438	.362	**.396**	.632	**.453**	**.527**
AML-PM	.392	.190	.792	.239	.386	.329	.329	.071	.048	.036	.496	.336	.308	.772	.366	.382
BPLangM	.418	.891	.594	.562	.517	.254	.314	**.500**	**.333**	**.250**	**.554**	.346	**.340**	.742	.253	.295
I-Match	.358	**.950**	.731	**.826**	.746	.236	.358	**.667**	.222	**.333**	.400	.103	.164	.667	.151	.246
KnoMa-Proc	.262	.563	.698	.509	.215	.279	.229	.200	.190	.100	.130	.130	.082	.519	.342	.300
KMSSS	.385	.908	.688	.701	**.791**	.239	.308	–	–	–	.450	.148	.143	.773	.352	.355
LogMap	.358	.339	.731	.463	.726	.261	.384	–	–	–	.357	.086	.139	**.818**	.170	.281
MSSS	.332	**1.0**	.667	.696	**.973**	.174	.243	–	–	–	**.583**	.130	.119	**1.0**	.144	.158
OPBOT	**.565**	.882	**.854**	**.831**	.676	**.422**	**.483**	.250	**.286**	.152	**.714**	**.485**	**.470**	.688	.451	.444
pPalm-DS	**.459**	.894	**.875**	**.871**	.454	**.356**	.354	.100	**.286**	.111	.452	**.426**	.335	.706	**.587**	**.504**
RMM-NHCM	**.456**	.923	.698	.717	.717	.319	**.389**	**.400**	.190	**.167**	.552	.262	.261	.623	.292	.283
RMM-NLM	.309	**1.0**	.443	.487	**.952**	.163	.223	–	–	–	.389	.130	.119	**1.0**	.154	.174
RMM-SMSL	.384	.667	.292	.307	.506	.329	.346	.095	**.286**	.111	.304	.194	.147	.633	.390	.353
RMM-VM2	.433	.894	**.854**	.805	.391	.339	.339	.050	.190	.056	.413	**.432**	.335	.652	**.470**	**.469**
TripleS	.384	.445	.719	.450	.651	.268	.309	–	–	–	.433	.142	.131	.679	.367	.361

Table 7.3: Results of Birth Registration data set

Table 7.1 illustrates the results for the University Admission data set (including the participants of the PMMT at the OAEI 2016 and 2017). As can be observed, the University Admission data set consists of a very high fraction of trivial correspondences. Almost half of the correspondences (44,3%) in the gold standard are trivial correspondences. It can further be observed that most matchers focus on identifying trivial correspondences. Just few matchers can identify a reasonable number of complex correspondences. Similar behavior can be observed for the Asset Management data set in Table 7.2 with 45,9% trivial correspondences. Again, most matchers focus on identifying these trivial correspondences. No matcher can achieve good results for Cat. I. For Cat. II and Cat. III there is a similar picture. However, for the Asset Management data set, the number of correspondences in Cat. II is too low to draw meaningful conclusions. Moreover, the matchers compute a high fraction of false-positives in Cat. IV, which we can observe by the very low Precision in this category. Thus, the matchers compute a high fraction of false-positives when ≥ 2 words are identical.

For the Birth Registration data set (Table 7.3), we can make different observations. 75% of all correspondences are correspondences of Cat. I. This shows that this data set is by far the most complex of these three data sets. Similar to the Asset Management data set, only a low fraction of correspondences of the gold standard have only the verb in common (Cat. II). Furthermore, it can be observed that many matchers fail in identifying the trivial correspondences of this data set. One explanation may be the gold standard for this data set because we find that

the binary gold standard of the Birth Registration data set does not fully cover all trivial correspondences or contains wrong trivial alignments. Another observation is that matcher need to take structural dependencies into account, to differentiate between wrong and correct trivial alignments.

The Cat. trivial of the Asset Management and the Birth Registration data sets only contains correspondences which are exactly identical without any normalization. This is not the case for the University Admission data set. Therefore, we further distinguish between the kind of trivial correspondences, i.e., if these correspondences are "identical" or "identical after normalization"; see Figure 7.2. We find that in the University Admission data set, about 7% of Cat. trivial consists of correspondences which are trivial after normalization, like stemming. This is a very small fraction and illustrates that for the detection of most correspondences of this category not even a normalization is required. However, the sub-division of Cat. trivial, helps to understand if matchers are able to detect trivial correspondences, which require normalization. We found that only the matchers RMM-NHCM and KMSSS achieved a F-Measure of 1.0 for "trivial" alignments *after normalization*. This indicates that most of the tested matchers cannot detect "trivial" correspondences which require a normalization.

7.4.2 Exemplary Observations and Findings

With the evaluation through matching patterns it is possible to identify characteristics, strengths and weaknesses of a matcher. The results clearly show that most matchers focus on finding correspondences with low complexity, i.e., Cat. trivial and Cat. IV. The matchers clearly lack identifying complex correspondences. This is especially evident for the Asset Management data set which contains special technical terms. For detecting non-trivial correspondences, a matching technique requires knowledge about these terms. It can be observed that the matcher BPLangMatch, in contrast to the other matchers, is able to identify difficult correspondences of this specific data set (Cat. I). At the Asset Management data set, Cat. II consists only of two correspondences and therefore it is impossible to draw conclusions for this category. The matcher AML achieves very good results for Cat. I (cFM of 0.623) at the University Admission data set. In general, the matcher OPBOT achieves considerably good results over all categories and test cases. Moreover, the matcher OPBOT achieves considerably good results for Cat. I in the Admission data set.

Therefore, it is not surprising that this matcher reaches the best F-Measure on the Birth Registration data set. (For both, the binary as well as non-binary evaluation.)

Approach	University Admission										Asset Management										Birth Registration									
	trivial		I		II		III		IV		trivial		I		II		III		IV		trivial		I		II		III		IV	
	[103]		[68]		[27]		[17]		[17]		[102]		[76]		[2]		[18]		[24]		[26]		[437]		[9]		[58]		[53]	
	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN	FP	FN
AML	12	6	2	27	1	22	3	11	45	8	–	–	–	–	–	–	–	–	–	–	7	4	213	278	13	7	27	37	14	29
AML-PM	12	5	32	48	60	12	113	10	117	1	1	0	0	75	2	0	19	11	13	4	48	5	203	287	9	8	24	38	6	32
BPLangM	6	24	37	68	8	20	55	9	70	10	1	5	15	71	0	1	16	13	8	7	2	8	72	311	3	6	16	37	6	41
DKP	12	4	19	60	0	27	0	17	36	14	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
DKP-lite	12	4	0	68	0	27	0	17	36	14	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
I-Match	9	6	1	68	3	25	0	17	1	16	–	–	–	–	–	–	–	–	–	–	1	7	35	334	1	7	9	52	4	45
KnoMa-Proc	12	10	1	68	24	23	41	12	134	9	209	7	0	76	13	0	15	17	69	8	12	7	463	306	10	7	35	53	15	37
KM-SSS	12	0	16	61	7	18	9	13	83	6	1	0	0	76	0	2	3	17	61	10	2	7	23	326	1	9	7	52	4	41
LogMap	12	2	0	68	4	23	39	11	92	8	–	–	–	–	–	–	–	–	–	–	37	7	43	323	1	9	9	53	2	44
Match-SSS	12	4	4	64	2	17	0	17	9	17	1	0	3	76	0	2	0	18	8	21	0	8	6	347	1	9	3	53	0	48
OPBOT	3	21	11	32	8	21	12	10	60	8	1	0	18	74	3	0	27	9	21	2	2	4	74	248	5	5	17	29	12	24
pPalm-DS	13	3	5	68	143	15	317	10	216	2	1	0	4	76	0	0	35	13	163	1	4	3	181	274	10	7	32	34	17	19
RMM-NHCM	8	3	7	42	13	16	5	11	36	9	1	0	0	76	0	2	1	17	3	15	1	7	49	295	3	7	12	43	8	37
RMM-NLM	13	0	25	45	0	27	0	17	0	17	1	0	0	76	0	2	0	18	0	18	0	13	10	352	1	9	7	53	0	46
RMM-SMSL	11	17	6	34	59	15	8	17	43	15	1	56	0	76	0	2	14	14	5	22	4	16	130	291	13	7	19	50	8	39
RMM-VM2	8	20	2	68	114	20	169	11	104	5	1	7	0	75	1	0	39	10	16	9	3	4	190	280	21	7	32	31	13	28
TripleS	13	0	0	68	52	22	2	14	51	16	1	0	0	76	0	2	19	16	56	7	25	6	60	313	1	9	10	52	7	40

Table 7.4: False-positive (FP) and false-negative (FN) alignments for the three data sets and all matchers, assigned to the categories

Observing the performance of the matchers at the three different data sets, it seems that the matchers are optimized to the specific data sets. This is a disadvantage of making the gold standards publicly available, as it was for example the case in the PMMC 2015. For example, while the matchers focus on finding correspondences from Cat. trivial in the University Admission data set and Asset Management data set, in contrast, at the Birth Registration data set matchers aim at identifying correspondences from Cat. I. This can also be observed by the number of false-positive and false-negative alignments for each category (Table 7.4). The matchers compute a high number of false-positive alignments in Cat. I for the Birth Registration data set, i.e., the matchers aim at identifying correspondences from this category. For the Asset Management data set, however, most matchers do not compute alignments from Cat. I at all. This can be explained by the fact that Cat. I is on the one hand the most difficult category, but on the other hand, to succeed at the Birth Registration data set it is necessary to compute correspondences from this category. The reason is the very high fraction of correspondences on the whole Birth Registration data set for Cat. I (about 75%). Furthermore, the Asset Management data set contains a high number of technical terms. Therefore, Cat. I is over proportionally complex at this data set.

The classification of the false-positives and false-negatives into the categories allows a more fine-grained understanding about a matcher's performance. It enables to directly identify where sources for errors of the matchers are. Moreover, it allows for an application-dependent evaluation, thus "tuning" of the matchers. In the Asset Management data set, for example, the matcher KnoMa-Proc computes a very high number of false-positive alignments in Cat. trivial. The matcher RMM-SMSL misses many trivial correspondences (56) from the Asset management data set. Moreover, we can observe that in the Birth Registration data set the binary gold standard seems to contain some errors. This can be observed by the high number of false-positives and false-negatives which all matchers compute in this category. In the next section, we will describe the results for the matchers with the non-binary evaluation and see that we can verify this observation.

7.4.3 Results of the Matching Patterns using Probabilistic Evaluation

When we apply the evaluation via matching patterns to the probabilistic evaluation, introduced in Section 5.1, we can make some interesting observations and observe characteristics about the non-binary gold standard. Firstly, we can observe that the distribution of the gold standard changes considerably compared to the binary gold standard. Moreover, in the Birth Registration data set we learn that the binary gold standard misses a high number of, e.g., trivial correspondences. It is interesting to observe that the absolute values of many matchers increase with the probabilistic evaluation in the Birth Registration data set. This is especially surprising because we expect a decrease of Recall, since the non-binary gold standard covers a much broader range of correspondences. We can moreover show that the poor results in the "trivial" category result from mistakes in the annotation of the *binary* gold standard. The performance for the Cat. trivial in the non-binary evaluation is much more reasonable. Hence, the *binary* gold standard of the Birth Registration data set has many shortcomings. Moreover, we can observe an increase of the performance for Cat. I at the Birth Registration data set (e.g., DKP). This also indicates the shortcomings of the binary gold standard of the Birth Registration data set.

Moreover, we can observe that the fraction of correspondences changes considerably for the University Admission data set. Especially the fraction of "trivial" correspondences decreases (relatively). However, the absolute numbers (of corre-

spondences) cannot be compared directly since they are weighted differently in the non-binary gold standard.

Approach	FM	Cat. trivial [7.3%][86]			Cat. I no word iden. [75.6%][896]			Cat. II one verb iden. [3.1%][37]			Cat. III one word iden. [7.0%][83]			Cat. IV ≥ two words iden. [6.2%][73]		
		cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM
AML	.490	.764	.701	.731	.433	**.488**	.459	.284	**.352**	.314	.539	**.580**	**.559**	.553	**.619**	.584
I-Match	.504	**1.0**	.670	**.802**	.816	.363	.502	**1.0**	.222	**.364**	.524	.195	.284	.688	.301	.419
LogMap	**.551**	.882	**.814**	**.847**	.847	.411	**.554**	**1.0**	.019	.036	.535	.204	.295	**.873**	.312	.46
AML-PM	.509	.829	**.878**	**.853**	.488	.484	.486	.111	.093	.101	.385	.42	.402	.754	.557	**.641**
BPLangMatch	.511	.944	.611	.742	.645	.399	.493	.273	.167	.207	.59	.509	**.546**	.733	.375	.496
KnoMa-Proc	.296	.603	.661	.631	.217	.45	.293	.011	.019	.014	.156	.204	.177	.414	.449	.431
Know-Match-SSS	.527	.947	.647	.769	**.886**	.376	.528	**1.0**	.019	.036	**.676**	.221	.333	.695	.415	.52
Match-SSS	.476	**1.0**	.629	.772	**.966**	.316	.476	**1.0**	.019	.036	**1.0**	.173	.294	**1.0**	.199	.332
OPBOT	**.576**	.905	.692	.785	.64	**.489**	**.555**	.200	.148	.170	.665	**.633**	**.649**	.583	**.636**	.609
pPalm-DS	.493	.867	**.710**	.781	.442	**.488**	.464	.515	.315	**.391**	.379	.540	.445	.545	**.761**	**.635**
RMM-NHCM	**.565**	.948	.661	.779	.781	.436	**.559**	.077	.037	.05	**.722**	.367	.487	.786	.5	**.611**
RMM-NLM	.443	**1.0**	.452	.623	**.931**	.3	.454	**1.0**	.019	.036	.549	.173	.263	**1.0**	.256	.407
RMM-SMSL	.464	.75	.326	.454	.526	.449	.485	.373	**.352**	**.362**	.322	.252	.283	.552	.449	.495
RMM-VM2	.466	.906	.701	.791	.426	.444	.435	.2	**.370**	.26	.441	**.558**	.492	.522	.597	.557
TripleS	.515	.439	.679	.533	.721	.432	.54	**1.0**	.019	.036	.51	.221	.309	.619	.443	.517

Table 7.5: Results of Birth Registration data set using probabilistic evaluation

Approach	FM	Cat. trivial [12.3%][108]			Cat. I no word iden. [50.1%][439]			Cat. II one verb iden. [15.6%][137]			Cat. III one word iden. [9.1%][80]			Cat. IV ≥ two words iden. [12.8%][112]		
		cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM	cP	cR	cFM
AML	**.424**	.92	.991	.945	**.701**	.078	**.12**	**1.0**	.043	.068	**1.0**	.164	.2	.696	.429	.439
I-Match	.271	**.97**	.99	**.979**	–	–	–	.5	.029	.029	–	–	–	.5	.004	.007
LogMap	**.418**	.92	**1.0**	.949	–	–	–	.661	.058	.067	.256	.17	.125	.696	.683	**.616**
AML-PM	.407	.92	.994	.946	.441	.072	.099	.588	**.327**	**.343**	.283	**.441**	**.273**	.649	**.818**	**.665**
BPLangMatch	.376	**.996**	.789	.846	.05	.018	.024	.808	.083	.127	.356	**.341**	**.278**	.635	.575	.526
DKP	.343	.92	**1.0**	.949	**1.0**	**.400**	**.467**	–	–	–	–	–	–	**.809**	.358	.345
DKP-lite	.342	.92	**1.0**	.949	–	–	–	–	–	–	–	–	–	**.809**	.358	.345
KnoMa-Proc	.406	.949	.983	**.956**	.319	.121	.1	.144	**.152**	.085	.515	**.776**	**.552**			
Know-Match-SSS	.409	.92	**1.0**	.949	.378	.041	.051	.719	.098	.119	**.857**	.149	.161	.593	.529	.483
Match-SSS	.314	.92	**1.0**	.949	.259	.012	.015	**.941**	.109	**.137**	–	–	–	.019	.001	.001
OPBOT	.376	.939	.62	.681	.631	.06	.097	.695	.053	.074	.483	.19	.164	.739	.374	.352
pPalm-DS	.276	.92	.986	.942	.143	.002	.004	.092	.124	.073	.069	.203	.077	.477	**.855**	.534
RMM-NHCM	**.448**	**.966**	.868	.874	.634	.046	.077	**1.0**	**.172**	**.258**	**.800**	.177	.2	**.903**	.441	.477
RMM-NLM	.311	.92	**1.0**	.949	.263	.043	.061	–	–	–	–	–	–	–	–	–
RMM-SMSL	.356	.914	.915	.902	**.777**	.058	**.098**	.15	.121	.063	–	–	–	**.870**	.485	.517
RMM-VM2	.320	.934	.843	.88	–	–	–	.086	.138	.1	.143	.321	.153	.591	**.711**	**.57**
TripleS	.300	.920	**1.00**	.949	–	–	–	.073	.04	.033	.625	.073	.075	.247	.131	.093

Table 7.6: Results of University Admission data set using probabilistic evaluation

7.5 Conclusions

We propose a conceptually new evaluation procedure by automatically dividing the matching task as well as the matcher output into patterns with specific attributes. The proposed evaluation via matching patterns provides an in-depth evaluation about a matcher's performance, including specific strengths and weaknesses. It

replaces the need for manual processing the matcher output, which is very time-consuming and supports a fast improvement of matching techniques.

Our proposed category-dependent evaluation has the following properties:

- informs about the data set, e.g., the complexity of the matching task,
- assesses the gold standard indirectly, e.g., quality and quantity of manual annotations,
- identifies characteristics as well as strengths and weaknesses of a matcher,
- enables to optimize a matcher to specific application scenarios.

By identifying the strengths and weaknesses of a matcher, the proposed evaluation technique may aid the progress of matching techniques.

Moreover, it allows for an application dependent evaluation, as the evaluation procedure can aid in improving matching techniques to obtain desired attributes. The evaluation procedure further is an efficient way to automatically process the matcher output. It delivers insights in what kind of false-positive and false-negative alignments matchers generate and therefore enables for an quantitative as well as qualitative analysis. Moreover, it offers information about the complexity of the data set. In current literature, complexity is associated with different level of granularity and the fraction of $1 : m$ or $n : m$ correspondences. In our evaluation, we analyze the syntactical overlap of the activities. We found, for example, that for the University Admission data set, a matcher can achieve a rather high performance, if it only computes the "trivial" correspondences.

The detailed performance measure, by the categories, allows to *predict* the results of matchers for future applications. This *prediction* helps to choose the best matchers for each *specific application*. The approach can be extended by different matching patterns. Furthermore, standard metrics can be applied. The only limitation is that they can be assigned automatically.

To overcome this limitation, we propose synthetic test cases, which complement the matching patterns. To keep the manual effort as low as possible, we propose a framework were the test scenarios are generated semi-automatically. The test scenarios allow to test and tune matchers to specific applications and challenges. As a result we get synthetic test scenarios, with real-world data. They can be complemented by different synthetic test cases which can be generated automatically

as well as semi-automatically. However, even the models are generated from real-world data, the manipulation of the models may lead to artificial circumstances. However, the aggregation of the correspondences of the gold standard as well as matcher output provides test scenarios like synthesized data sets, with real-world data. Therefore, matchers can be tuned and tested for a specific application scenario.

We will introduce the synthetic scenarios in the outline in Section 8.2.1 and show some future possibilities regarding a prediction of the performance of matchers. To offer the evaluation of all metrics as well as all evaluation frameworks and synthetic data sets, introduced in this thesis, we implement an "*Evaluation Portal*" which can be assessed by researchers in the field of process model matching. All described metrics and evaluation methods can be assessed via this portal. We will describe the evaluation platform in greater detail in Section 8.2.2.

8

Summary, Conclusions and Outlook

In this chapter, we summarize the main contributions of the thesis in Section 8.1. Moreover, we give an outlook about how to use, access and complement the introduced evaluation procedures and metrics in Section 8.2. As an outlook for further evaluation procedures, we moreover introduce synthetic test cases, which may aid in a deeper understanding of the functionality of matching techniques and allows to tune matchers to fulfill specific attributes in Section 8.2.1. In Section 8.2.2 we introduce the evaluation platform, which allows to access all metrics, introduced in this thesis. Moreover, we provide an approach to predict the performance of matching techniques for specific data sets with the results of the experiments in Section 8.2.3. Section 8.2.4 states additional possible future research directions in the field of process model matching evaluation.

8.1 Thesis Summary

In this thesis, we introduced a probabilistic evaluation procedure for process model matching techniques. In this context we introduced a non-binary gold standard, where we calculated and included the support values of the annotators of the non-binary gold standard. We adapted the standard notions of Precision and Recall to comprise non-binary values and introduced Bounded variants of the measures, which can be computed with the determination of a specific threshold. Moreover, we introduced a distance-based performance measure. This metric takes the arguability of correspondences with low support values explicitly into account. If matchers compute a correspondence with low support value it is marginally sanctioned. With this metrics we conducted experiments with the data sets and matchers from the Process Model Matching Contest 2015 and the Process Model Matching Track at the OAEI 2016 and 2017.

The metrics proposed in this thesis allow for a more fine-grained evaluation, since they offer insights, for instance whether matchers focus on identifying correspondences with low support or with high support values. Therefore, we can acquire knowledge if matchers are suitable for a specific application scenario. In our evaluations with the non-binary gold standard, we found that the ranking of the performance of the matchers for the tested data sets change considerably for some matchers. Moreover, we learned that the binary gold standard did not contain many correspondences which are reasonable and even have a high support value in the non-binary gold standard. In the probabilistic evaluation with the Birth Registration data set, we observed that some matchers improve their performance absolutely. This is especially surprising, since we expect a decrease of Recall, due to the high coverage of correspondences in the non-binary gold standard. However, this indicates that the binary gold standard does not contain many correspondences which are actually correct and highlights the risk of a binary evaluation.

Furthermore, we introduced a ranking-based evaluation method for process model matching where the confidence values of the matchers as well as the support values of the non-binary gold standard are considered. However, in this completely non-binary evaluation procedure the confidence values itself are not considered. The confidence values are only considered to compile a ranked collection of correspondences. Besides the performance measurement, this aids in understanding to which extent the results and confidence values of matchers correlate with the

support values of the non-binary gold standard. It helps to understand whether the confidence values of a matcher reflect a similar distribution of the non-binary gold standard. In the experiments, we learned that the matchers from the Process Model Matching Contest only have a small rank-correlation with the non-binary gold standard. Some matchers even have a negative rank-correlation. This results from the high number of false-positives which some matchers compute with a high confidence score. This means that the confidence scores which some matchers compute, do not reflect the actual distribution properly. Thus, the confidence scores which some matchers compute, do not properly give the extent in how far one should trust in the correspondence.

Furthermore, we introduced a category-dependent evaluation procedure. The matching task as well as the matching results are split into categories with different attributes and complexity levels. In this way, we provide an in-depth evaluation which offers insights about specific strengths and weaknesses of matchers. Moreover, the category-dependent evaluation provides diverse insights about the data set itself. It further enables to tune matchers to specific application scenarios. We found that the University Admission data set and the Asset Management data set consist of a high fraction of "trivial" correspondences. We learned that in both data sets the matchers focus on computing "trivial" correspondences. The matchers fail in identifying complex correspondences, which have a weak syntactical overlap. We further classified all false-positive and false-negative correspondences, computed by the matchers into the introduced categories and therefore pointed out specific areas of improvement of matchers.

The introduced evaluation methods and metrics were applied in the context of the Process Model Matching Track at the Ontology Alignment Evaluation Initiative 2016 and 2017 (Achichi et al., 2016, 2017). There the results of the participating ontology matchers are evaluated with the introduced evaluation procedures, described in this thesis.

Table 8.1 briefly summarizes the approaches introduced in this thesis. It further indicates which measures have been applied to the Process Model Matching Track at the OAEI 2016 and 2017. The approaches which have not been applied (ReD and Ranking-based evaluation), have been published after the OAEI 2017 has been conducted.

In addition to the above described approaches, we propose synthetic generated test scenarios which also take structural dependencies of the process models into

Approach	Attributes	Applied at OAEI 2016	Applied at OAEI 2017
Non-binary Gold standard	Definition of a non-binary GS, which takes the uncertainty of a GS into account. Uncertain correspondences are included and assigned with a support value.	x	x
ProP, ProR, ProFM	Adaption of Precision, Recall, F-Measure to consider non-binary values, in this way the uncertainty of the GS is considered in the metrics.	x	x
Bounded ProP,ProR,ProFM	Adaption of the probabilistic measure, which allows to exclude correspondences under a chosen threshold.	x	–
Relative Distance (ReD)	Distance-based measure, which takes the non-binary GS as basis; in this measure, correspondences with low support-values are marginally sanctioned.	–	–
Ranking-based Evaluation	Correlation-based evaluation, which transforms the matcher output and GS into a ranked collection of correspondences, then the correlation is applied. (completely NB)	–	–
Matching Patterns	Category-dependent evaluation, which divides the matching task and results into different levels of complexity.	–	x

Table 8.1: Summary of the introduced Evaluation Approaches

account, to complement the category-dependent evaluation. This is introduced in Section 8.2.1.

Furthermore, we provide an evaluation framework, which is an "Evaluation Portal" where researches can access the evaluation methods and metrics which we introduced in this thesis. We describe the "Evaluation Portal" in Section 8.2.2.

In addition to the above stated findings, the category-dependent evaluation enables to predict the performance of matchers for specific data sets. We provide an overview of such a prediction in Section 8.2.3. There we will try to learn how complex the matching task is, to give a recommendation about which matchers may be most successful on a given matching task.

8.2 Future Research

In the following, we discuss future work in the evaluation of process model matching techniques and give an outlook of approaches for future evaluation experiments. We introduce semi-automatically generated synthetic scenarios, which we propose for future evaluation experiments. Furthermore, we want to provide an outlook in how far the evaluation results of the matching patterns may allow for a prediction of the performance of matchers for specific data sets.

8.2.1 Semi-automatically Generated Synthetic Test Scenarios

The categorization to matching patterns introduced in Section 7.1 is annotated automatically. This automatic annotation naturally leads to a limitation of the problem classes, since not each possible category can be assigned automatically so far. The aim of the synthetic scenarios is that they resemble typical difficulties of process model matching tasks. Moreover, we want to learn if matchers consider structural information or background information of the process models, such as textual descriptions. Currently most matchers compare the activity labels, however there are also matching approaches which use the control flow information of business process models, like by Klinkmüller and Weber (2017). To complement the matching patterns by test cases which include such scenarios and test cases which cannot be generated automatically, we propose synthetic scenarios.

The manual effort to generate those scenarios has to be as low as possible. Therefore, we propose semi-automatically generated synthetic test cases. To complement the above introduced matching patterns, we chose synthetic scenarios which reflect typical difficulties in data sets. However, some of the synthetic test cases cannot be assigned completely automatically. One example is the generation of synonyms, as we explained in Section 7.1. Therefore, we generate transformation rules depending on the matching scenario and depending on application requirements of the matching techniques. On the one hand, we want to learn if matchers also take the structure of process models into account, or if they use different background information or if they are solely label-based. The original models can be manipulated manually and automatically. Then as a next step the manipulated models can be compared to the original models. *In this way, the gold standard is obtained automatically.*

8.2.1.1 Transformation Rules of the Synthetic Test Scenarios

In (Ferrara et al., 2010) the organizers introduce synthetic scenarios. Those test cases are artificially generated test cases. They provide additional information about the ability of matchers to solve specific matching tasks. One example is to test if ontology matcher are able to detect synonyms. The synthetic datasets itself are artificially generated and therefore cannot resemble real-world data sets. Moreover, for the automatic generation of such test cases libraries such as WordNet (Miller, 1995) are utilized to find (for example) synonyms. However, most matchers

rely on libraries like WordNet, therefore, we propose the manual generation of the synonyms. Moreover we propose to manipulate real-world data sets in a way that they fulfill typical difficulties in process model matching tasks. To simplify this we propose semi-automatically generated test scenarios, where part of the generation is done manually. In the following, we introduce the synthesized data sets and present which insights can be obtained.

In the following, we describe those transformation rules and possible scenarios and indicate which of the test scenarios are manipulated *semi-automatically*, thus require *manual input*.

Semi-automatic generated test scenarios:

1. Generate $1 : m$ correspondences:

   a) Insert activities to the process model:

      The original models are manipulated in a way that $1 : m$ correspondences emerge.

      Example: *evaluate* $\rightarrow$ *check application, score qualification, sum scores*

   b) Summarize activities to one activity: the opposite step (to summarize activities) has the same effect. In this way, $1 : m$ correspondences are generated.

      Example: *print out and sign application form, collect additional required documents* $\rightarrow$ *prepare application*

2. Manually replace words by synonyms

3. Manipulate activities: Adding stop words, use abbreviations, adding typos

   In real-world data typos exist and abbreviations are used. To test if matchers offer such a functionality we add typos or manipulate models by using abbreviations.

4. Process models with (one) identical label: Matchers which do not take the structure of process models into account, generate $n : m$ mappings. The matchers would match each identical label with each other. The evaluation of such matchers, would also indicate if matchers take structural dependencies into account or if they can deal with the structural information only. The distribution of the results can indicate this.

Automatic generated test scenarios: [1]

1.* Background information are removed, the models itself are not changed. The results are compared to the original models. If the results change it indicates if matchers take background information into account.

2.* Delete words from activity
Example: *send letter of rejection* → *send rejection*

3.* a) Mapping of part of the process model with the original process model
Example: Mapping of half model of process model 1 with half model of process model 2. It helps to understand if matchers take structural properties of the process models into account.

4.* Flip the process model vertically. The activities are exchanged vertically. This tests if matchers take structural dependencies of the process models into account.

8.2.1.2 Conclusions

We introduced synthetic scenarios, which complement the matching patterns from Chapter 7.1 in different ways. On the one hand the synthetic scenarios complement the matching patterns by test cases which cannot be extracted automatically from the data sets. On the other hand the synthetic scenarios assess if matchers take structural information or background information into account or if they work solely label-based.

We will include the synthetic data sets to the "Evaluation Portal", which we introduce in the next section. There the synthetic data sets can be downloaded and the matchers can be applied on the new data sets. Then the results can be evaluated via this platform.

8.2.2 Evaluation Portal

To give researchers open access to all introduced evaluation procedures and metrics we establish an evaluation framework, where researchers can upload their matcher output and evaluate them against the introduced metrics. As a basis for the experiments the data sets from the Process Model Matching Contests are available.

[1] With a * noted steps are done completely automatically.

Moreover, the synthetic test cases can be downloaded. The Portal can be accessed online[2].

In the following, we give a short introduction to the functionality and features of the evaluation platform.

At first, the User has to register to the portal and to log into the system. After this short registration process, the User can access the evaluation portal. Figure 8.1, shows the start page of the evaluation portal, after the log-in has been performed. At this stage the User can choose a data set, on which he or she wants to perform the evaluation experiments. In this case the University Admission data set and the Birth Registration data set are available for evaluation.

After the choice of the data set has been performed, the corresponding matcher output can be uploaded for the evaluation (cf. Figure 8.2 and Figure 8.3).

After the matcher output has been successfully uploaded, the User is asked for the choice of the metrics for the evaluation experiments (Figure 8.4). Then the evaluation can be conducted. The uploaded results are compared to the choice of gold standards. After the information have been computed, the results page appears, which summarizes all selected metrics and data. This is illustrated in the screen-shot in Figure 8.5.

[2] http://alkmaar.informatik.uni-mannheim.de/pmmc

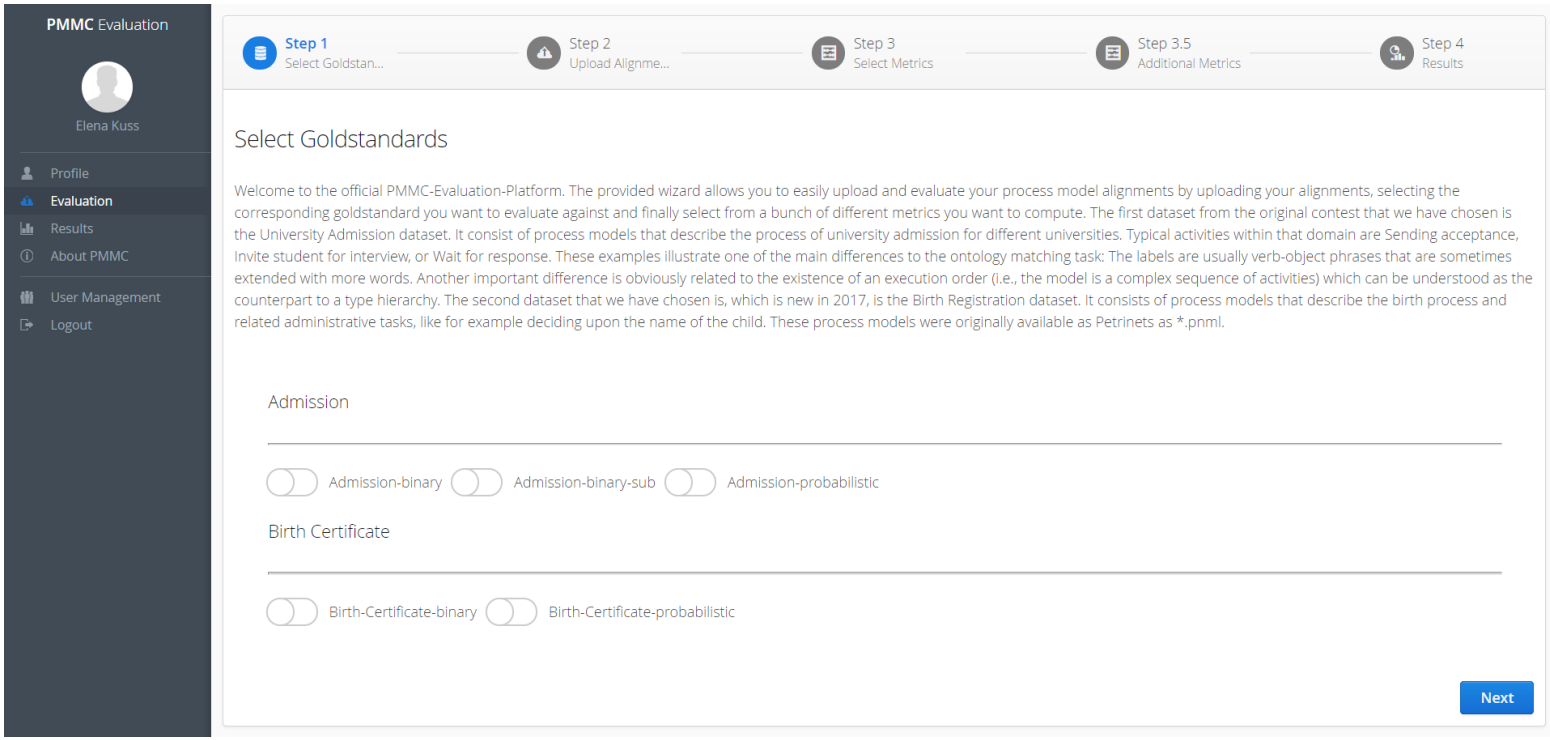


Figure 8.1: Start of the Evaluation Portal

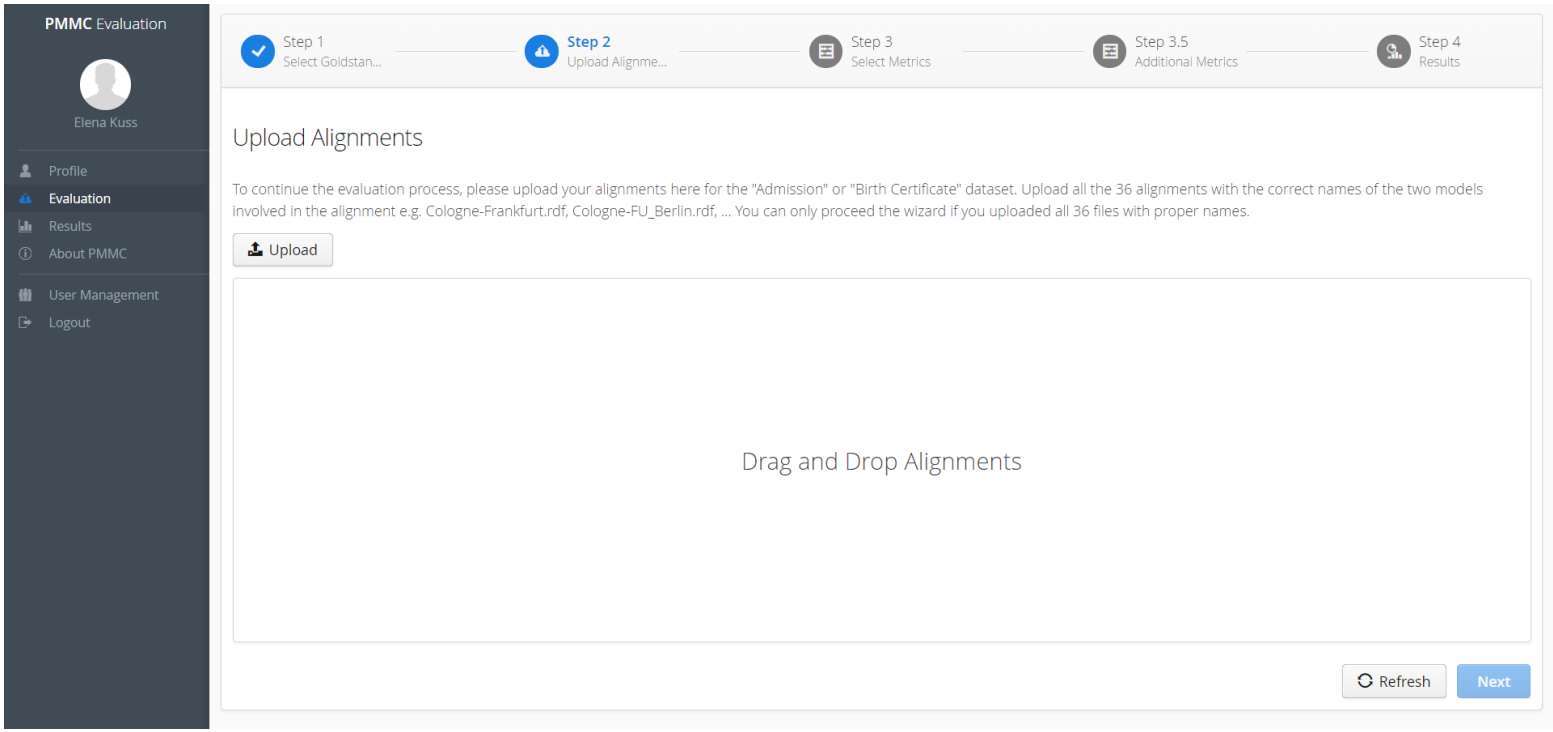


Figure 8.2: Process of uploading the matcher output

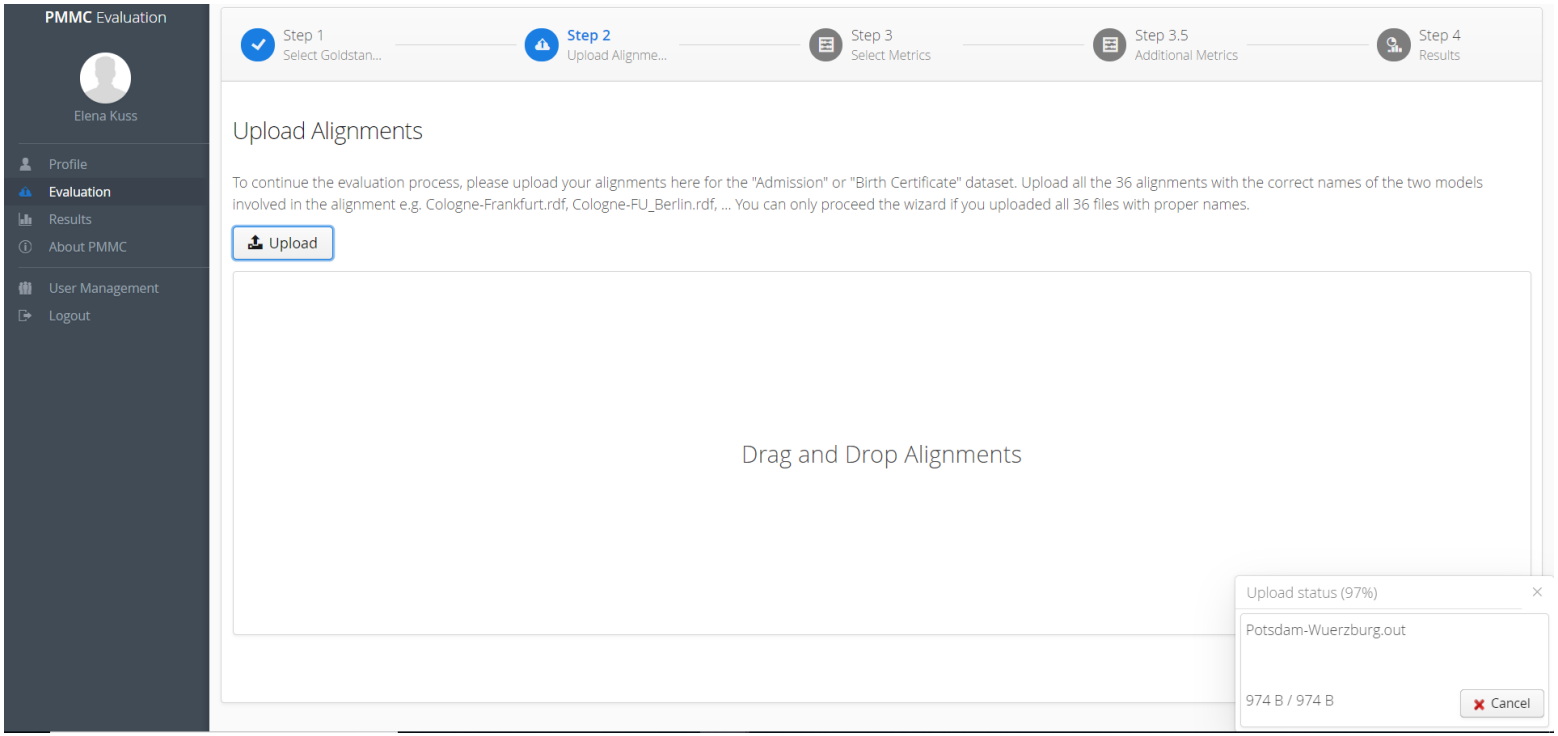


Figure 8.3: Upload of the matcher output, which should be evaluated

Figure 8.4: Choice of evaluation metrics

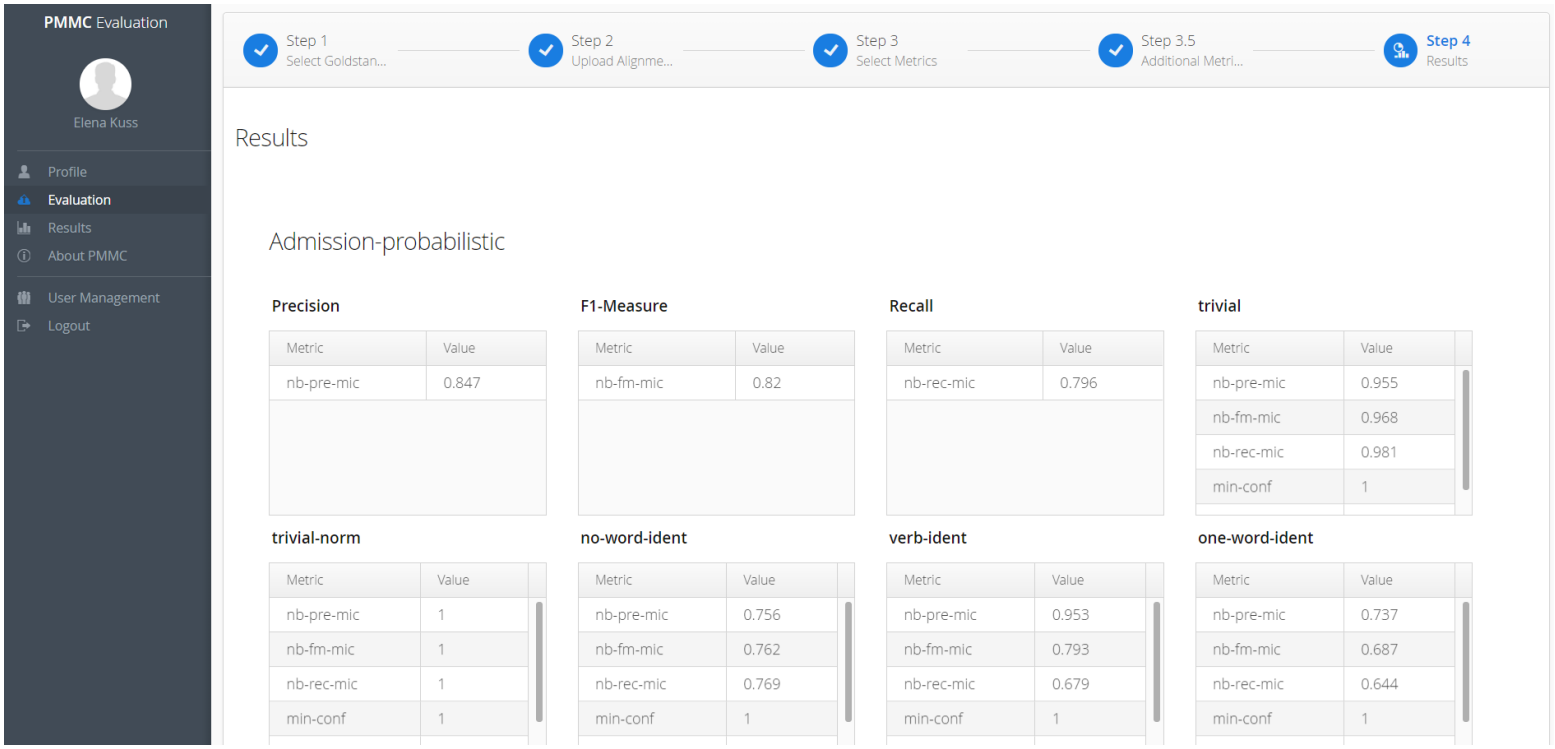


Figure 8.5: Results page with choice of metrics

8.2.3 Predicting the Performance of Matchers

The introduced matching patterns from Chapter 7.1 allow for a detailed analysis of the performance of matching techniques, without manually processing the matcher output. It helps understanding strengths and weaknesses of matchers and to tune matchers to specific application scenarios. However, the insights gained by this evaluation procedure may also be a basis for a prediction of the performance of matching techniques. Moreover, it can help to classify the complexity of the matching task itself.

The insights gained by the classification to matching patterns can be directly applied for such a prediction, if the structure (fraction of matching patterns) of the matching task is known in advance. Then the following can be applied:

- to apply the best matcher of the dominant group of the matching task
- to apply an ensemble of matchers which have the best performance in the most dominant group(s)
- to apply an ensemble of matchers which have the best performance of the most dominant groups, thus complement each other

In schema- and ontology matching as well as process matching, an ensemble of matchers was proposed to achieve better matching results (Eckert et al., 2009; Gal and Sagi, 2010; Meilicke et al., 2017). In our case, the ensemble of matchers is then selected depending on the performance of the matchers in the most dominant categories of the applied data set.

In many applications the "structure" of the data set, in terms of the distribution of complexity of the correspondences of the matching task may be known in advance. However, in some matching context the structure of the data set may *not* be known in advance. One approach to learn more about the data set, especially about the complexity of the matching task, may be to use the results of the matchers which solved a given matching task. Then the results can be assigned into the categories, introduced in Section 7.2. The matchers results, classified into the categories, can serve as a hint to get information about the complexity of the matching task.

To follow this approach, we compute the results of an ensemble of matchers to learn if we can in this way get information about the distribution of the data set. To learn if the approach can provide such information, we test it at the data sets of the PMMC 2015. Similarly like proposed in Meilicke et al. (2017), we chose an ensemble

of computed correspondences. In our case we compute all correspondences where two matchers agreed on. As a next step, we apply the categorization to the matching patterns, which we introduced in Section 7.1.

For our experiments, we take the ensemble of the results of all matchers which participated in the PMMC 2015 and the Process Model Matching Track of the OAEI 2016 and 2017. The results are given in Table 8.2 - 8.4.

Category	# Alignments	Distribution in %	Actual Distribution in %
TRIVIAL	108	17.0	44.4
Cat. I	56	8.1	29.3
Cat. II	142	20.5	11.6
Cat III	143	20.7	7.3
Cat. IV	229	33.1	7.3

Table 8.2: Characteristics of the University Admission data set (with n=2)

Category	# Alignments	Distribution in %	Actual Distribution in %
TRIVIAL	36	5.8	4.5
Cat. I	422	68.5	75.0
Cat. II	25	4.1	1.5
Cat. III	77	12.5	9.9
Cat. IV	56	9.1	9.1

Table 8.3: Characteristics of the Birth Registration data set (with n=2)

Category	# Alignments	Distribution in %	Actual Distribution in %
TRIVIAL	103	38.7	45.9
Cat. I	1	0.4	34.2
Cat. II	12	4.5	0.9
Cat. III	39	14.7	8.1
Cat. IV	110	41.5	10.8

Table 8.4: Characteristics of the Asset Management data set (with n=2)

In Tables 8.2 – 8.4, we give the results for $n = 2$, e.g., computed correspondences of at least two matchers. In the Tables 8.5 – 8.7 the results of $n = 3$ are given. The first column states the five categories, the second column gives the number of

Category	# Alignments	Distribution in %	Actual Distribution in %
TRIVIAL	108	25.3	44.4
Cat. I	51	10.9	29.3
Cat. II	55	11.85	11.6
Cat. III	58	12.5	7.3
Cat. IV	182	39.2	7.3

Table 8.5: Characteristics of the University Admission data set (with n=3)

Category	# Alignments	Distribution in %	Actual Distribution in %
TRIVIAL	27	6.8	4.5
Cat. I	269	67.8	75.0
Cat. II	12	3.0	1.5
Cat. III	46	11.6	9.9
Cat. IV	43	10.8	9.1

Table 8.6: Characteristics of the Birth Registration data set (with n=3)

Category	# Alignments	Distribution in %	Actual Distribution in %
TRIVIAL	103	51.5	45.9
Cat. I	1	0.5	34.2
Cat. II	4	2.0	0.9
Cat. III	16	8.0	8.1
Cat. IV	76	38.0	10.8

Table 8.7: Characteristics of the Asset Management data set (with n=3)

correspondences which are computed for $n = 2$ and $n = 3$, respectively. The third column provides the fraction on the whole data set for the corresponding category. The last column states the fraction of the categories from the gold standard.

For both ($n = 2$ and $n = 3$), we can observe that for two out of three data sets the results do not totally reflect the realistic distribution of the kind and fraction of correspondences of the data sets. However, for the Birth Registration data set (Table 8.3) the results provide a very accurate impression about the distribution of the correspondences of the data set. In the results in Tables 8.2 – 8.4, we consider each correspondence, which is computed by at least two matchers. We also computed the results for $n = 3$ to reduce the number of false-positives. This improves the

results, e.g., the fraction of the categories, and therefore better reflects the real distribution for the University Admission data set and Asset Management data set (cf. Table 8.5 and Table 8.7).

However, the results are still not that convincing compared to the Birth Registration data set. Especially Cat. I and Cat. IV do not reflect the realistic distribution. The reason are the many false-positive correspondences which many of the matchers compute. This is especially evident for Cat. IV in the University Admission data set and partially for the Asset Management data set. We have observed this property already in Section 7.4.1. This category consists of correspondences with ≥ 2 syntactic identical words. Therefore, many matchers compute a high fraction of false-positives in this category.

To make this more clear, consider the following example of two activities which should *not* be matched. The example is extracted from a matcher output:

"*Fill out application form*" – "*Receive application form*"

Those activities describe a different action, therefore the example is no correct correspondence. Some matchers compute a high fraction of such false-positive correspondences. The reason is that the similarity measures of the matchers calculate a high similarity score because of the two identical words "application" and "form", even "Fill out" and "Receive" have a low similarity score. However, the average similarity score of the three words is above the threshold, which many matchers use. Thus, some matchers compute a high fraction of such false-positive correspondences.

We can observe this also in the results above on the very high fraction of correspondences of Cat. IV (cf. Table 8.2 – Table 8.4). This indicates that the prediction of the performance of matchers for unknown data sets might only consider the ensemble results of the matchers with the best performance, not of all matchers. In practice, this is also more feasible, since not all matchers are available for execution.

8.2.4 Additional Future Research Directions

In the previous section, we gave an introduction to future research directions in the field of process model matching evaluation and highlighted the possibilities to use evaluation results to predict the performance of matching techniques.

In this section, we outline additional possible research directions in process model matching evaluation.

In future experiments the evaluation of process model matching techniques should offer a broader range of data sets, and may contain synthetic scenarios as we introduced in this chapter. Those synthetic scenarios are generated from real-world data and therefore provide a realistic setting. This can complement the existing data sets also in terms of structural properties of the process models.

In future work, it would be interesting to apply such approaches which translate the process models into ontologies. This would help to understand if correspondences, computed by a matcher are close or related, or totally unrelated and thus simply wrong. Such an approach for Ontology Matching was introduced in Ehrig and Euzenat (2005).

To translate process models into ontologies has been conducted already for the Process Model Matching Track at the Ontology Alignment Evaluation Initiative (Achichi et al., 2016). There, ontology matcher were applied to process model matching. However, the translation into ontologies did not aim to provide all information of the process model into an ontological structure. The translation was performed to apply ontology matchers to the matching of the process models on a label-based comparison.

However, there have been further approaches to translate process models into a hierarchical representation, as in Vanhatalo et al. (2009). Such a translation would allow for an analysis of the matcher output, in particular the false-positives, computed by a matcher. In this way, the false-positives can be analyzed to learn whether both activities are related or simply wrong. Nevertheless, in real-world data, the process models are not always consistent. The process models are, for instance, not always modeled totally consistent with regard to the structure and hierarchy. Sometimes, for instance, “Swimlanes” are not considered correctly. Moreover, there are many different formats for the process models, like in the data sets which we used for this thesis, in EPC, BPMN or Petri-Nets.

For each format a different translation into ontologies is required. This additionally makes such a translation sophisticated.

However, for correctly modeled process models it would be an interesting approach to better differentiate the incorrect computed alignments.

This may also help to determine a threshold for a specific matcher. Moreover, it could aid in an improvement of matching results. However, this would not just result in a more detailed evaluation. To install such constraints for a matcher, to not match activities from a different hierarchy, may improve the matching results of future matching techniques.

Bibliography

M. Achichi, M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, I. Harrow, V. Ivanova, E. Jiménez-Ruiz, E. Kuss, P. Lambrix, H. Leopold, H. Li, C. Meilicke, S. Montanelli, C. Pesquita, T. Saveta, P. Shvaiko, A. Splendiani, H. Stuckenschmidt, K. Todorov, C. Trojahn, and O. Zamazal. Results of the ontology alignment evaluation initiative 2016. In *CEUR workshop proceedings*, volume 1766, pages 73–129. RWTH, 2016.

M. Achichi, M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, I. Harrow, V. Ivanova, et al. Results of the ontology alignment evaluation initiative 2017. In *OM 2017-12th ISWC workshop on ontology matching*, pages 61–113. No commercial editor., 2017.

G. Antunes, M. Bakhshandeh, J. Borbinha, J. Cardoso, S. Dadashnia, C. D. Francescomarino, M. Dragoni, P. Fettke, A. Gal, C. Ghidini, P. Hake, A. Khiat, C. Klinkmüller, E. Kuss, H. Leopold, P. Loos, C. Meilicke, T. Niesen, C. Pesquita, T. Péus, A. Schoknecht, E. Sheetrit, A. Sonntag, H. Stuckenschmidt, T. Thaler, I. Weber, and M. Weidlich. The process model matching contest 2015. In *6th International Workshop on Enterprise Modelling and Information Systems Architectures*, 2015.

R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

Z. Bellahsene, A. Bonifati, F. Duchateau, and Y. Velegrakis. On evaluating schema matching and mapping. In *Schema matching and mapping*, pages 253–291. Springer, 2011.

G. A. Bowen. Naturalistic inquiry and the saturation concept: a research note. *Qualitative research*, 8(1):137–152, 2008.

U. Cayoglu, R. Dijkman, M. Dumas, P. Fettke, L. García-Bañuelos, P. Hake, C. Klinkmüller, H. Leopold, A. Ludwig, P. Loos, et al. Report: The process model matching contest 2013. In *International Conference on Business Process Management*, pages 442–463. Springer, 2013a.

U. Cayoglu, A. Oberweis, A. Schoknecht, and M. Ullrich. Triple-s: A matching approach for Petri nets on syntactic, semantic and structural level. Technical report, Karlsruhe Institute of Technology (KIT), 2013b.

M. Cheatham and P. Hitzler. Conference v2. 0: An uncertain version of the oaei conference benchmark. In *International Semantic Web Conference*, pages 33–48. Springer, 2014.

R. Dijkman, M. Dumas, and L. García-Bañuelos. Graph matching algorithms for business process model similarity search. In *International Conference on Business Process Management*, pages 48–63. Springer, 2009.

H.-H. Do, S. Melnik, and E. Rahm. Comparison of schema matching evaluations. In *Net. ObjectDays: International Conference on Object-Oriented and Internet-Based Technologies, Concepts, and Applications for a Networked World*, pages 221–237. Springer, 2002.

K. Eckert, C. Meilicke, and H. Stuckenschmidt. Improving ontology matching using meta-level learning. In *European Semantic Web Conference*, pages 158–172. Springer, 2009.

M. Ehrig and J. Euzenat. Relaxed precision and recall for ontology matching. In *Proc. K-Cap 2005 workshop on Integrating ontology*, pages 25–32. No commercial editor., 2005.

C. C. Ekanayake, M. La Rosa, A. H. Ter Hofstede, and M.-C. Fauvet. Fragment-based version management for repositories of business process models. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 20–37. Springer, 2011.

H.-E. Eriksson and M. Penker. Business modeling with uml. *New York*, pages 1–12, 2000.

J. Euzenat. Semantic precision and recall for ontology alignment evaluation. In *IJCAI*, pages 348–353, 2007.

J. Euzenat, C. Meilicke, H. Stuckenschmidt, P. Shvaiko, and C. Trojahn. Ontology alignment evaluation initiative: six years of experience. In *Journal on data semantics XV*, pages 158–192. Springer, 2011.

T. Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8): 861–874, 2006.

G. F. Ferrara, I. Fundulaki, I. Harrow, V. Ivanova, E. Jiménez-Ruiz, P. Lambrix, H. Leopold, H. Li, C. Meilicke, et al. Results of the ontology alignment evaluation initiative 2010. In *ISWC workshop on Ontology Matching*, 2010.

A. Gal and T. Sagi. Tuning the ensemble selection process of schema matchers. *Information Systems*, 35(8):845–859, 2010.

F. Jabeen, H. Leopold, and H. A. Reijers. How to make process model matching work better? an analysis of current similarity measures. In *International Conference on Business Information Systems*, pages 181–193. Springer, 2017.

A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456. ACM, 2008.

C. Klinkmüller and I. Weber. Analyzing control flow information to improve the effectiveness of process model matching techniques. *Decision Support Systems*, 100:6–14, 2017.

C. Klinkmüller, I. Weber, J. Mendling, H. Leopold, and A. Ludwig. Increasing recall of process model matching by improved activity label matching. In *Business Process Management*, pages 211–218. Springer, 2013.

C. Klinkmüller, H. Leopold, I. Weber, J. Mendling, and A. Ludwig. Listen to me: Improving process model matching through user feedback. In *Business Process Management*, pages 84–100. Springer, 2014.

J. Kolb, H. Leopold, and J. Mendling, editors. *Enterprise Modelling and Information Systems Architectures, Proceedings of the 6th Int. Workshop on Enterprise Modelling and Information Systems Architectures, EMISA 2015, Innsbruck, Austria, September 3-4, 2015*, volume 248 of *LNI*, 2015. GI.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

E. Kuss and H. Stuckenschmidt. Automatic classification to matching patterns for process model matching evaluation. In *CEUR workshop proceedings*, volume 1979, pages 306–319. RWTH, 2017.

E. Kuss, H. Leopold, H. Van der Aa, H. Stuckenschmidt, and H. A. Reijers. Probabilistic evaluation of process model matching techniques. In *Conceptual Modeling: 35th International Conference, ER 2016, Gifu, Japan, November 14-17, 2016, Proceedings 35*, pages 279–292. Springer, 2016.

E. Kuss, H. Leopold, C. Meilicke, and H. Stuckenschmidt. Ranking-based evaluation of process model matching. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 298–305. Springer, 2017.

E. Kuss, H. Leopold, H. van der Aa, H. Stuckenschmidt, and H. A. Reijers. A probabilistic evaluation procedure for process model matching techniques. *Data & Knowledge Engineering*, 2018.

Y. Lee, M. Sayyadian, A. Doan, and A. S. Rosenthal. etuner: tuning schema matching software using synthetic scenarios. *The VLDB Journal—The International Journal on Very Large Data Bases*, 16(1):97–122, 2007.

H. Leopold, M. Niepert, M. Weidlich, J. Mendling, R. Dijkman, and H. Stuckenschmidt. Probabilistic optimization of semantic process model matching. In *International Conference on Business Process Management*, pages 319–334. Springer, 2012.

J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

L. Makni, N. Z. Haddar, and H. Ben-Abdallah. Business process model matching: An approach based on semantics and structure. In *e-Business and Telecommunications (ICETE), 2015 12th International Joint Conference on*, volume 2, pages 64–71. IEEE, 2015.

C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

C. Meilicke, H. Leopold, E. Kuss, H. Stuckenschmidt, and H. A. Reijers. Overcoming individual process model matcher weaknesses using ensemble matching. *Decision Support Systems*, 100:15–26, 2017.

G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

T. Murata. Petri nets: Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580, 1989.

M. Owen and J. Raj. Bpmn and business process management: Introduction to the new business process modeling standard. 2003.

G. Paolacci, J. Chandler, and P. G. Ipeirotis. Running experiments on amazon mechanical turk. 2010.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *the VLDB Journal*, 10(4):334–350, 2001.

T. Sagi and A. Gal. Non-binary evaluation for schema matching. In *Conceptual Modeling*, pages 477–486. Springer, 2012.

T. Sagi and A. Gal. Non-binary evaluation measures for big data integration. *The VLDB Journal—The International Journal on Very Large Data Bases*, 27(1):105–126, 2018.

H. Sfar, A. H. Chaibi, A. Bouzeghoub, and H. B. Ghezala. Gold standard based evaluation of ontology learning techniques. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing*, pages 339–346. ACM, 2016.

P. Shvaiko and J. Euzenat. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering*, 25(1):158–176, 2013.

A. Sonntag, P. Hake, P. Fettke, and P. Loos. An approach for semantic business process model matching using supervised machine learning. ecis, research–in–progress. 2016.

C. Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

T. Thaler, P. Hake, P. Fettke, and P. Loos. Evaluating the evaluation of process matching techniques. In *Multikonferenz Wirtschaftsinformatik*, pages 1600–1612, 2014.

A. M. Turk. Amazon mechanical turk. *Retrieved August*, 17:2012, 2012.

W. M. Van der Aalst. The application of petri nets to workflow management. *Journal of circuits, systems, and computers*, 8(01):21–66, 1998.

W. M. Van der Aalst. Formalization and verification of event-driven process chains. *Information and Software technology*, 41(10):639–650, 1999.

C. Van Rijsbergen. Information retrieval. dept. of computer science, university of glasgow. *The Computer Journal*, 14, 1979.

J. Vanhatalo, H. Völzer, and J. Koehler. The refined process structure tree. *Data & Knowledge Engineering*, 68(9):793–818, 2009.

M. Weidlich, R. Dijkman, and J. Mendling. The icop framework: Identification of correspondences between process models. In *International Conference on Advanced Information Systems Engineering*, pages 483–498. Springer, 2010a.

M. Weidlich, R. Dijkman, and M. Weske. Deciding behaviour compatibility of complex correspondences between process models. In *International Conference on Business Process Management*, pages 78–94. Springer, 2010b.

M. Weidlich, T. Sagi, H. Leopold, A. Gal, and J. Mendling. Predicting the quality of process model matching. In *Business Process Management*, pages 203–210. Springer, 2013a.

M. Weidlich, E. Sheetrit, M. C. Branco, and A. Gal. Matching business process models using positional passage-based language models. In *International Conference on Conceptual Modeling*, pages 130–137. Springer, 2013b.

M. Weske. Business process management architectures. In *Business Process Management*, pages 333–371. Springer, 2012.

E. Zavitsanos, G. Paliouras, and G. A. Vouros. Gold standard evaluation of ontology learning methods through ontology transformation and alignment. *IEEE Transactions on Knowledge and Data Engineering*, 23(11):1635–1648, 2011.