3-5-2015

Zur Nutzung von Techniken der Natürlichen Sprachverarbeitung für die Bestimmung von Prozessmodellähnlichkeiten – Review und Konzeptentwicklung

Tim Niesen

Constantin Houy

Zur Nutzung von Techniken der Natürlichen Sprachverarbeitung für die Bestimmung von Prozessmodellähnlichkeiten – Review und Konzeptentwicklung

Tim Niesen, Constantin Houy

Institut für Wirtschaftsinformatik (IWi) im Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI GmbH) und Lehrstuhl für Betriebswirtschaftslehre, insbesondere Wirtschaftsinformatik (Prof. Dr. Peter Loos) an der Universität des Saarlandes

{tim.niesen,constantin.houy}@iwi.dfki.de

Abstract. Die Modellierung von Geschäftsprozessen stellt für viele Unternehmen einen unverzichtbaren methodischen Ansatz dar, um Wissen über geschäftliche Aktivitäten zu bündeln und zu strukturieren. Häufig existieren Geschäftsprozessmodelle in verschiedenen Varianten, die unterschiedlichen Modellierungskonventionen folgen. Zur Handhabung solcher Varianten werden Ähnlichkeitsmaße benötigt, die den Abstand zweier Modelle in Bezug auf Struktur, Semantik oder Verhalten quantifizieren. Methoden der Natürlichen Sprachverarbeitung (engl. *Natural Language Processing*, NLP) können ein inhaltliches Verständnis durch die Analyse von Modellbezeichnern (sog. *Labels*) und somit eine Suche nach thematisch ähnlichen Modellen ermöglichen.
Der vorliegende Work-in-Progress-Artikel möchte zwei Beiträge leisten: Erstens wird ein strukturiertes Literaturreview durchgeführt, um einen Überblick über den State-of-the-Art im Bereich *Business Process Similarity* zu geben. Dieser Überblick liefert Erkenntnisse über die Verbreitung vorhandener Ansätze in der Literatur. Zweitens wird ein Konzept präsentiert, das unter Verwendung von NLP-Techniken eine inhaltliche Erschließung und einen thematischen Ähnlichkeitsvergleich von Prozessmodellen erlaubt.

Keywords: Geschäftsprozessmodellierung, Business Process Similarity, Natural Language Processing, NLP, Natürliche Sprachverarbeitung

1 Ausgangssituation und Zielsetzung

Geschäftsprozessmodellierung hat sich für viele Unternehmen zu einer unverzichtbaren Methode entwickelt, um geschäftliche Aktivitäten zu organisieren, zu strukturieren und zu dokumentieren. In Informationssystemen abgebildete Geschäftsprozesse werden vielfach als intellektuelle Vermögenswerte betrachtet, die Wettbewerbsvorteile einer Unternehmung ausmachen und entscheidend zu ihrem Markterfolg beitragen [1]. Prozessmodelle werden üblicherweise in Sammlungen gespeichert (sog. *Modell-*

12th International Conference on Wirtschaftsinformatik,
March 4-6 2015, Osnabrück, Germany

Niesen, T.; Houy, C. (2015): Zur Nutzung von Techniken der Natürlichen Sprachverarbeitung für die Bestimmung von Prozessmodellähnlichkeiten – Review und Konzeptentwicklung, in: Thomas. O.; Teuteberg, F. (Hrsg.): Proceedings der 12. Internationalen Tagung Wirtschaftsinformatik (WI 2015), Osnabrück, S. 1829-1843

Repositories) und sind von dort aus einer weiteren Verwendung zugänglich [2]. Dabei nimmt die Größe dieser Sammlungen, u. a. bedingt durch forcierte Modellierungsvorhaben, Anpassungen von bestehenden Modellvarianten oder auch in Folge von Unternehmenszusammenschlüssen, beständig zu. Dieser Umstand erschwert deren Wartung und Nutzbarkeit, weshalb automatisierte Verfahren zum Management von Prozessmodellen nötig werden [3]. Die Anwendung von Methoden zur Messung von Ähnlichkeiten zwischen Prozessmodellen (*Business Process Similarity*) kann zur effizienten Analyse von Modell-Repositories beitragen. Sie bildet außerdem die Grundlage für eine Vielzahl weiterer Verfahren wie beispielsweise die Konformitätsprüfung zwischen Referenzprozessen und tatsächlichen Prozessvarianten (engl. *Conformance Checking*), die Suche nach „verwandten" Prozessmodellen zu einem gewählten Modell oder die Komplexitätsreduktion innerhalb einer Sammlung [4,5].

Bestehende Verfahren zur Bestimmung von Ähnlichkeiten zwischen Prozessmodellen oder deren Bestandteilen nutzen typischerweise strukturelle, semantische und verhaltensbasierte Modellmerkmale [4]. Diese Merkmale werden in den meisten Fällen isoliert voneinander betrachtet. In der vorliegenden Arbeit wird angenommen, dass gerade eine Verknüpfung von strukturellen Eigenschaften mit textuellen Informationen aus den Bezeichnern eines Modells jedoch ein tieferes Verständnis der modellierten Zusammenhänge ermöglicht und zu einer genaueren und schnelleren Bestimmung von ähnlichen Modellen führen kann.

Vor diesem Hintergrund ist es die *Zielsetzung* der vorliegenden Arbeit, die Potenziale der Anwendung von Methoden aus dem Bereich der Natürlichen Sprachverarbeitung (*Natural Language Processing*, NLP) für die Bestimmung von Ähnlichkeiten zwischen Prozessmodellen zu untersuchen. Zu diesem Zweck wird zunächst eine Erhebung des aktuellen Forschungsstandes mithilfe eines strukturierten Literaturreviews [6] im Bereich *Business Process Similarity* durchgeführt und eine entsprechende Aufbereitung der Ergebnisse hinsichtlich der verwendeten Methoden präsentiert. Aufbauend auf diesen Ergebnissen wird dann ein fachliches Konzept entwickelt [7], das bestehende Verfahren zur Ähnlichkeitsbestimmung durch NLP-Techniken erweitern soll. Vor diesem Hintergrund werden folgende Forschungsfragen adressiert:

(F1) Welche Verfahren zur Bestimmung von Modellähnlichkeiten stehen aktuell zur Verfügung und wie häufig finden sie Beachtung in der Literatur?

(F2) Wie lassen sich Techniken aus dem Bereich NLP gewinnbringend zur Bestimmung von Modellähnlichkeiten einsetzen?

Der vorliegende Beitrag ist wie folgt *strukturiert*: nach dieser Einleitung werden in Kapitel zwei die begrifflichen Grundlagen in den Bereichen *Business Process Similarity* und *Natürliche Sprachverarbeitung* dargelegt. Kapitel drei expliziert den verwendeten Forschungsansatz, bevor in Kapitel vier die gewonnenen Erkenntnisse des strukturierten Literaturreviews präsentiert werden. In Kapitel fünf wird das unter Berücksichtigung dieser Erkenntnisse erarbeitete Konzept zur inhaltlichen Erschließung von Prozessmodellen vorgestellt. Im Anschluss werden in Kapitel sechs die Potenziale und Grenzen des Konzeptes diskutiert. Kapitel sieben schließt den Beitrag mit einem Resümee sowie einem Ausblick auf zukünftige Arbeiten.

2 Begriffliche Grundlagen

2.1 Business Process Similarity

Methoden des Forschungsbereichs *Business Process Similarity* weisen einige Parallelen zu verschiedenen anderen Themenkomplexen auf, insbesondere zu *Process Equivalence* und *Process Matching*. Im Folgenden wird daher zunächst eine thematische Abgrenzung relevanter Begriffe vorgenommen. Als *Process Matching* wird die Identifikation von zwei „gleichwertigen" Aktivitäten bzw. Prozessbestandteilen innerhalb eines Modells bezeichnet, d. h. zu einem Bestandteil in Modell A wird ein entsprechender Bestandteil in einem zu vergleichenden Modell B gesucht [8]. Im Ergebnis entsteht ein sogenanntes *Mapping* zwischen den jeweiligen Elementen. Dieses kann beispielsweise als Grundlage für die Bestimmung von Ähnlichkeiten zwischen zwei Modellen dienen. Die Bestimmung eines solchen Mappings ist vor der Anwendung von Methoden zur Ähnlichkeitsbestimmung zwingend notwendig, da diese auf Basis erkannter Korrespondenzen zwischen Prozessbestandteilen arbeiten [4]. Der Themenbereich *Process Equivalence* zielt auf eine Beurteilung von Prozessen hinsichtlich des beobachtbaren Verhaltens, d. h. hinsichtlich ihres Outputs und nicht hinsichtlich ihrer zugrundeliegenden Struktur [9]. Zwei vollkommen unterschiedlich modellierte Prozesse, die den gleichen Output erzeugen, können somit in diesem Sinne als äquivalent bezeichnet werden. Folglich kann Process Equivalence unabhängig von den konkreten syntaktischen Elementen einer Modellierungsmethodik beurteilt werden [10].

Methoden aus dem Forschungsbereich *Business Process Similarity* stellen Metriken bereit, die den Abstand zwischen zwei zu vergleichenden Prozessmodellen quantifizieren. Ein *Ähnlichkeitsmaß* auf einer Menge von Prozessmodellen M ist definiert als eine Funktion $sim: M \times M \rightarrow [0,1]$ [11]. Diese Methoden werden in einer Vielzahl von Anwendungsfällen benötigt, z. B. bei der Zusammenlegung von Prozessmodellsammlungen im Rahmen von Unternehmensfusionen [12], zur Komplexitätsreduktion [5] oder zur effizienten Verwaltung von Prozessmodell-Repositories [13]. In diesem Zusammenhang spielt insbesondere die Möglichkeit einer gezielten Suche nach zu einem Eingabemodell (sog. *Query*-Modell im Information-Retrieval-Kontext) ähnlichen Modellen eine wichtige Rolle.

In der bestehenden Literatur werden Ähnlichkeitsmaße zumeist hinsichtlich ihrer Ausrichtung auf bestimmte Aspekte eines Prozessmodells klassifiziert [14]:

1. **Strukturelle Ähnlichkeit** berücksichtigt die Struktur eines Prozessmodells in Bezug auf dessen Kontrollfluss, die Anzahl seiner Elemente oder deren Reihenfolge im Modell.
2. **Semantische Ähnlichkeit** bezieht sich auf die Semantik eines Modells und versucht, Rückschlüsse auf seinen Inhalt und seine Bedeutung zu ziehen.
3. **Verhaltensbasierte Ähnlichkeit** untersucht die Ausführungssemantik eines Modells, d. h. mögliche Ausführungssequenzen (sog. *Traces*) von Aktivitäten [5].

2.2 Natural Language Processing

Methoden der *Natürlichen Sprachverarbeitung* befassen sich mit der Analyse von natürlich sprachlichen Texten, die nicht gesondert für eine technische Verarbeitung aufbereitet oder strukturiert wurden [15]. Zu diesem Zweck wird eine Vielzahl von Methoden eingesetzt, von denen einige im Folgenden kurz erläutert werden sollen. Die folgende Darstellung erhebt keinen Anspruch auf Vollständigkeit und dient nur der thematischen Einordnung zentraler Begriffe. Um die Bedeutung eines Wortes abschätzen zu können, ist es häufig nützlich, Informationen über dessen Wortklasse zu gewinnen. Sollen etwa nur Nomen betrachtet werden, so ist eine Klassifikation nach Wortarten (*parts of speech*, PoS) erforderlich. Den Prozess einer solchen Klassifizierung bezeichnet man als *Part-of-Speech-Tagging* [16]. Durch die Analyse von Mehrwortgruppen, sogenannten *N-Grammen*, können zusammengehörige Sinneinheiten identifiziert werden [17]. Auf diese Weise kann z. B. aus den Einzelbegriffen „data“ und „transfer“ das Bi-Gramm „data transfer“ erkannt werden, was ein tieferes Textverständnis ermöglicht. Um der Tatsache gerecht zu werden, dass den verschiedenen sinntragenden Wörtern (engl. *terms*) eines Dokuments eine unterschiedlich große Bedeutung zukommt, werden *Gewichtungsmaße* verwendet [18]. Diese setzen die absolute Häufigkeit eines Terms in einem Dokument in Beziehung zu seiner relativen Häufigkeit in der gesamten Dokumentensammlung. Eines der verbreitetsten Gewichtungsmaße ist das Produkt aus Termhäufigkeit (*term frequency*, TF) und inverser Dokumentenhäufigkeit (*inverse document frequency*, IDF): TF-IDF [19]. Methoden wie *Tokenization* – die Trennung von Sätzen in einzelne Sinneinheiten auf Wortebene – und *Stemming* – die Abstraktion von Begriffen auf ihren Wortstamm zur Berücksichtigung verschiedener Flexionsformen – sind häufig als Verarbeitungsschritte vor der Anwendung weiterführender NLP-Techniken notwendig.

3 Vorgehensweise und Forschungsansatz

Zur Erhebung des aktuellen Forschungsstandes im Bereich *Business Process Similarity* wurde in dieser Arbeit die *Forschungsmethode* des strukturierten Review angewandt [6]. Dazu wurde im Juli 2014 eine Recherche in den internationalen Wissenschaftsdatenbanken *ScienceDirect*, *Scopus* und *SpringerLink* durchgeführt und diese um Treffer in den Datenbanken von *WiSo-Net* und *Google Scholar* ergänzt. Die auf diese Weise identifizierten Arbeiten wurden hinsichtlich zuvor formulierter Auswahlkriterien bewertet. So wurden z. B. solche Beiträge von der weiteren Untersuchung ausgeschlossen, die sich nicht oder nur am Rande mit der konkreten Identifikation von Modellähnlichkeiten befassen oder aus „entfernten“ Fachbereichen stammen. Tabelle 1 auf der nächsten Seite zeigt die zur Recherche verwendeten Suchbegriffe sowie die jeweils erzielten Trefferzahlen.[1]

[1] Einige Suchbegriffskombinationen lieferten bei *Google Scholar* eine große Menge von Resultaten (> 6.300). In diesen Fällen wurden lediglich die ersten zehn Trefferseiten untersucht, da meistens bereits nach drei Seiten ein starker Präzisionsverlust festgestellt werden konnte. Dies

Tabelle 1. Suchbegriffe zur Literaturrecherche mit Trefferanzahl

	Google Scholar	*Scopus*	*Science ence-Direct*	*Springer Link*	*WiSo-Net*
Suchbegriff(e)	*Trefferanzahl*				
„business process similarity“	117	14	10	39	1
„business process model similarity“	158	8	13	35	2
„process model similarity“	302	22	21	84	2
prozessmodelle ähnlichkeit	6.340	0	0	151	28
geschäftsprozessmodelle ähnlichkeit	1.500	0	0	16	3
geschäftsprozessmodelle matching	277	0	0	8	1
„business process model matching“	4	0	2	4	0
„business process model“ matching	5.160	48	234	2.415	2
business „process equivalence“	404	11	36	104	1
business „process model equivalence“	2	1	1	2	0
„business process variants“	292	27	33	90	5
„business process model variants“	15	1	1	3	0
„workflow similarity“	153	13	17	33	0

Nach Ausschluss nicht relevanter Arbeiten verblieb eine Menge von 86 einschlägigen Beiträgen, die systematisch in Bezug auf methodische Ansätze untersucht und hinsichtlich ihrer Kernaussagen zusammengefasst wurden. Zwei Review-Beiträge wurden von der weiteren Betrachtung ausgeschlossen, da sie selbst keine neuen Erkenntnisse präsentieren, sondern bestehende Arbeiten systematisieren und kommentieren (die Gesamtanzahl untersuchter Beiträge ist somit $n = 84$). Aus Platzgründen wird in diesem Beitrag eine reduzierte Darstellung präsentiert und auf die Zusammenfassungen der einzelnen Arbeiten verzichtet.[2] Kapitel 4 diskutiert in knapper Form die Ergebnisse des Literaturreviews. Eine Übersicht über die zeitliche Verteilung der Beiträge ist in Tabelle 2 dargestellt. Sie zeigt die über die Jahre zunehmende Beachtung des Themenbereichs in der Literatur.

Tabelle 2. Übersicht zur Verteilung der identifizierten Beiträge nach Erscheinungsjahr

Jahr	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	Σ
Beiträge	1	0	5	5	4	8	12	14	18	9	8	84

Aufbauend auf den im Rahmen des Reviews gewonnen Erkenntnissen erfolgt im Anschluss die Entwicklung eines Konzeptes zur inhaltlichen Erschließung von Prozessmodellen (siehe Kapitel 5). Dieses Konzept repräsentiert ein Artefakt im Sinne gestal-

lässt sich damit erklären, dass *Google Scholar* auf eine hohe Trefferquote (engl. *recall*) optimiert ist, dabei aber einen Verlust in der Präzision (engl. *precision*) akzeptiert.

[2] Die Zusammenfassungen können beim Erstautor gerne angefragt werden.

tungsorientierter Wirtschaftsinformatikforschung [7]. Es wird ein bisher neues fachliches Modell entwickelt, welches als Vorstufe zur Implementierung eines Softwaresystems anzusehen ist. Vor diesem Hintergrund wird das Konzept somit zum einen der Anforderung nach wissenschaftlicher Fundierung durch die Berücksichtigung des aktuellen Forschungsstandes gerecht. Zum anderen zeigt es aufgrund der Möglichkeiten für die Handhabung von Prozessmodellvarianten in Organisationen durch die angestrebte Implementierung den konkreten Praxisbezug der Idee auf [20].

4 Literaturreview

4.1 Vorstellung der Review-Ergebnisse

Die 84 im Rahmen des Literaturreviews identifizierten Beiträge wurden hinsichtlich der verwendeten Methoden zur Bestimmung von Modellähnlichkeiten untersucht und systematisiert. Die Systematik wurde aus der grundsätzlichen Einteilung in **strukturelle**, **semantische** und **verhaltensbasierte Ähnlichkeit** abgeleitet [14] und während der Untersuchungen kontinuierlich um Kriterien und Unterkriterien erweitert, die innerhalb der identifizierten Beiträge diskutiert werden. Nach abschließender Aufstellung aller Kriterien wurden die Beiträge zur Erstellung der finalen Übersicht einer zweiten Analyse unterzogen. Im Folgenden werden die verwendeten Kriterien beschrieben und die in den Tabellen 3 und 4 überblicksartig dargestellten Ergebnisse diskutiert.[3]

Tabelle 3. Methodenübersicht zur Ähnlichkeitsbestimmung

Ansatz	*Anzahl*
Strukturähnlichkeit	**67**
Graph-Ähnlichkeit	44
Graph-Editierdistanz	19
strukturelle Eigenschaften	32
Semantische Ähnlichkeit	**50**
Labels	46
Ontologien	11
semantische Annotation	14
String-Editierdistanz	22
WordNet	12
Verhaltensbasierte Ähnlichkeit	**23**
Causal Footprints	6
Process Logs	5
Traces	18

[3] Eine Zuordnung jedes Ansatzes zu den 84 einzelnen Quellen wird hier aus Platzgründen nicht vorgenommen. Eine tabellarische Übersicht kann beim Erstautor angefragt werden. In den Tabellen sind nur solche Kriterien enthalten, die in mindestens zwei Beiträgen erwähnt werden. Insbesondere die sog. *Latente Semantische Indexierung* wird zwar an späterer Stelle in diesem Artikel diskutiert, aber nur in *einem* der untersuchten Beiträge tatsächlich verwendet.

Strukturähnlichkeit

- *Graph-Ähnlichkeit* subsumiert Verfahren, die auf Grundlage von Graph-Darstellungen Ähnlichkeiten von Modellen berechnen. Hierunter fällt z. B. der Vergleich von Vektoren, welche die jeweiligen Graphen repräsentieren.
- *Graph-Editierdistanz* bezeichnet Verfahren, welche die Ähnlichkeit zwischen Graph-Darstellungen von Prozessmodellen als Anzahl der nötigen Änderungsoperationen (Einfügen/Löschen von Elementen) berechnen, um Graph A in Graph B zu transformieren.
- *Strukturelle Eigenschaften* fassen Verfahren zusammen, welche beispielsweise die Reihenfolge oder die Anzahl der enthaltenen Elemente in Prozessmodellen zur Bestimmung der Ähnlichkeit betrachten.

Semantische Ähnlichkeit

- Unter dem Punkt *Labels* werden Verfahren beschrieben, die mit den Wörtern eines Modell-Labels arbeiten und z. B. durch Berechnung der String-Editierdistanz oder durch Nutzung von *WordNet* Ähnlichkeiten ableiten.
- Häufig werden in diesem Zusammenhang auch unternehmensspezifische *Ontologien* verwendet, die zum Abgleich von Begriffen eingesetzt werden können.
- Der Punkt *semantische Annotation* fasst Verfahren zusammen, die eine (manuelle) Anreicherung des Modells mit zusätzlichen Informationen benötigen. Beispiele hierfür sind Annotationen zu Inputs/Outputs von Aktivitäten oder zu Ressourcen.

Verhaltensbasierte Ähnlichkeit

- *Causal Footprints* bezeichnen approximierte Darstellungen des Modellverhaltens, die z. B. durch Abstraktion von konkreten Aktivitäten und Konnektoren entstehen. Sie eignen sich auch zum Vergleich von Modellen, die durch unterschiedliche Modellierungsmethoden beschrieben sind.
- Verfahren zur Analyse von *Process Logs* untersuchen konkrete Prozessinstanzen und beziehen sich daher auf das tatsächliche Verhalten eines Prozesses.
- *Traces* bezeichnen mögliche Ausführungspfade eines Prozessmodells und können direkt aus der Semantik eines Modells bestimmt werden. Sie geben die Reihenfolge an, in der Aktivitäten ausgeführt werden können.

Tabelle 3 liefert auch einen Überblick über die Verteilung der drei verschiedenen Ansätze zur Ähnlichkeitsbestimmung: strukturelle Ansätze treten in 67 von 84 untersuchten Beiträgen auf und stellen damit den größten Anteil (79, 8%), semantische Ansätze werden in 50 Beiträgen betrachtet (59,5 %) und verhaltensbasierte Ansätze stellen mit 23 Beiträgen den kleinsten Anteil (27,4 %). In vielen Beiträgen werden zwei oder mehr Ansätze auch in unterschiedlichem Umfang kombiniert, um bessere Ergebnisse zu erzielen.

Neben den genannten Hauptkriterien wurden weitere Aspekte in den identifizierten Beiträgen untersucht (Tabelle 4). So bilden Ansätze aus dem Bereich der Natürlichen Sprachverarbeitung („**NLP-Ansätze**“) und Spezifika verschiedener „**Modellierungsmethoden**“ weitere Blöcke mit Unteraspekten. Schließlich sind in dem Block „**Darstellungen**“ drei häufig verwendete Modelldarstellungsformen zusammengefasst.

Tabelle 4. Weitere Systematisierungsaspekte

Ansatz	*Anzahl*
Verwendete NLP-Ansätze	**14**
N-Gramme	2
Part-of-Speech-Tags	6
Stemming	8
TF-IDF	2
Tokenization	5
Verwendete Modellierungsmethoden	**30**
methodenunabhängig	16
BPMN	5
EPK	6
Petri Netz	12
Verwendete Darstellungen	**58**
Baumdarstellung	15
Graph-Darstellung	46
VSM	11

Verwendete NLP-Ansätze

Die Aspekte *N-Gramme*, *Part-of-Speech-Tags*, *Stemming*, *TF-IDF* und *Tokenization* beziehen sich auf die in Kapitel 2.2 beschriebenen NLP-Verfahren.

Im Hinblick auf die Entwicklung eines Konzeptes zur Nutzung von NLP-Techniken zur Ähnlichkeitsbestimmung ist insbesondere der Anteil derjenigen Arbeiten interessant, in denen diese Techniken verwendet werden. Hierzu kann festgestellt werden, dass lediglich 14 von 84 Beiträgen (16,7 %) NLP-Techniken einsetzen. Die geringe Verbreitung von NLP-Methoden ist insbesondere bei Verfahren auffällig, die auf dem Vergleich von Text-Labels basieren. Zwar untersuchen 46 Beiträge (54,8 %) den Inhalt dieser Labels, aber lediglich 13 davon nutzen dazu NLP-Techniken (28,3 %). Vielfach wird an dieser Stelle lediglich ein Vergleich der String-Editierdistanz zwischen Labels durchgeführt, was vor allem bei der Verwendung von unterschiedlichen *Label-Styles* in den Modellen zu schlechten Resultaten und vermeintlich geringer Übereinstimmung führt. Abschnitt 5.1 diskutiert dieses Problem detaillierter und beschreibt einige Ansatzpunkte, an denen der Einsatz von NLP-Techniken einen Mehrwert gegenüber bestehenden Methoden bringen kann.

Verwendete Modellierungsmethoden

Einige der identifizierten Ansätze arbeiten auf einer konkreten Modellierungsmethode und nutzen beispielsweise Eigenheiten in deren Semantik aus, um Ähnlichkeiten zu berechnen. Die drei am häufigsten verwendeten Modellierungsmethoden sind in Tabelle 4 aufgelistet. Die Klasse „methodenunabhängig“ signalisiert, dass ein Ansatz zwar am Beispiel einer konkreten Modellierungsmethode demonstriert wird, sich aber auch auf andere Methoden übertragen lässt.

Verwendete Darstellungen

- Häufig wird die ursprüngliche Graph-Darstellung eines Modells in eine *Baum-Darstellung* überführt, z. B. um potenziell unendliche Traces abzubilden.

- *Graph-Darstellungen* können genutzt werden, um Abhängigkeiten von Aktivitäten untereinander abzubilden oder um von einer konkreten Modellierungsmethode zu abstrahieren.
- *VSM* (*Vector Space Model*) bezeichnet die Darstellung von Modellinformationen in einem Vektorraummodell. Hierbei werden alle Modelle einer Sammlung auf Vektoren abgebildet und Ähnlichkeiten durch Vektordistanzen berechnet.

4.2 Diskussion zentraler Beiträge

Die Entwicklung des in Kapitel 5 vorgestellten Konzepts wurde insbesondere durch die im Folgenden präsentierten Arbeiten motiviert, welche eine inhaltliche Verarbeitung von Prozessmodellen anstreben. Diese werden deshalb an dieser Stelle noch einmal etwas ausführlicher besprochen. LEOPOLD ET AL. präsentieren ein Verfahren zur automatischen Identifizierung von durchgeführten Aktionen („actions“) und der betroffenen Prozessressource („business object“) in Textlabels [8]. Die Autoren definieren zunächst verschiedene Arten von *Label-Styles* und berücksichtigen unter Verwendung von Strukturinformationen unterschiedliche Granularitätsebenen eines Modells. Die so extrahierten Informationen ermöglichen ein genaueres Verständnis der im Modell formulierten Zusammenhänge als ein Wortvergleich zwischen Labels ohne dieses Hintergrundwissen. Eine Methode zur thematischen Eingrenzung des Suchraums ähnlicher Modelle zu einem Query-Modell findet sich bei QIAO ET AL [21]. Die Autoren schlagen ein zweistufiges Retrieval-Verfahren vor, das in einem ersten Schritt Metadaten eines Prozessmodells auswertet, um dessen inhaltliche Ausrichtung zu bestimmen (z. B. „Quality Management“ oder „Customer Service“). In einem zweiten Schritt wird ein struktureller Vergleich des Kontrollflusses zwischen Modellen der gleichen Kategorie durchgeführt. Die Autoren verwenden *Latente Semantische Indexierung* zur Erzeugung dieser Kategorien. NIEMANN ET AL. verwenden NLP-Techniken wie Lemmatisierung (die Rückführung von Worten auf ihre Grundform, z. B. Infinitiv bei Verben, Singular bei Nomen), PoS-Tagging und Stemming zur Vorverarbeitung von Label-Inhalten [22]. Durch die Betrachtung semantischer Informationen aus *WordNet* sowie struktureller Zusammenhänge zwischen Prozessbestandteilen wird ein umfassendes Ähnlichkeitsmaß abgeleitet, das zur Berechnung von Cluster-Paaren verwendet wird.

5 Konzept zur inhaltlichen Erschließung von Prozessmodellen

5.1 Limitationen aktueller und Potenziale neuer Ansätze

Die Ergebnisse des Literaturreviews zeigen, dass Methoden des NLP in bestehenden Ansätzen zur Ähnlichkeitsbestimmung bisher nur selten Anwendung finden. In den meisten Fällen beschränkt sich ein Vergleich von Text-Labels auf die Berechnung von Editierdistanzen, z. B. durch die Levenshtein-Distanz. Werden diese Distanzen auf dem kompletten Text eines Labels anstatt pro Wort berechnet, so ergibt sich bereits durch Änderung der Wortreihenfolge ein geringer Ähnlichkeitswert: z. B. führt

die naive Berechnung der Distanz zwischen Label A „*Berechnung Porto*“ und Label B „*Porto berechnen*“ zu einem Wert von 15, was eine große Differenz suggeriert, obwohl die Labels inhaltlich den gleichen Sachverhalt darstellen. Dieses Beispiel motiviert den Einsatz einer einfachen Stemming-Methode, die Wörter auf eine einheitliche Grundform zurückführt und so von konkreten Wortausprägungen abstrahiert: sowohl Label A als auch Label B lassen sich zur Wortmenge *{Berech, Porto}* vereinfachen und somit als identisch identifizieren.

Eine Besonderheit bei der Anwendung von NLP-Techniken auf die Labels eines Prozessmodells liegt in der üblicherweise knappen Formulierung. Insbesondere Verfahren wie PoS-Tagging verwenden zur Bestimmung einer Wortklasse den Kontext des zu analysierenden Begriffes und arbeiten daher traditionell auf Satzebene. Labels sind häufig jedoch lediglich als Halbsätze oder Reihung von Substantiven formuliert und stellen daher zum Teil keine geeignete Grundlage für die Anwendung solcher Verfahren dar. Auch die Qualität der Ergebnisse von anderen Verarbeitungstechniken steht und fällt mit der konkreten Formulierung der Label-Inhalte; je konsistenter, detaillierter und eindeutiger diese beschrieben sind, desto aussagekräftiger sind die durch die Analyse gewonnenen Informationen. Diese Aspekte spielen auch im Sinne einer allgemeinen Modellverständlichkeit eine Rolle und sollten daher bereits bei der initialen Modellierung berücksichtigt werden [23].

5.2 Konzeptvorstellung

Das erarbeitete Konzept ist inspiriert durch die Arbeit von QIAO ET AL. [21] und stellt eine inhaltliche Erschließung von Prozessmodellen in den Mittelpunkt. Ziel ist es, den Suchraum aller Modelle in Bezug auf ein Query-Modell zunächst thematisch einzugrenzen und anschließend beliebige Ähnlichkeitsvergleiche in der reduzierten Menge von Suchergebnissen durchführen zu können (vgl. Abbildung 1). Dieser Vorgehensweise liegen zwei Gedanken zugrunde: erstens ist eine thematische Suche nach Prozessmodellen häufig von Interesse, etwa zur Beantwortung der Frage, ob ein bestimmter Sachverhalt bereits modelliert wurde oder nicht. Zweitens ist ein struktureller Ähnlichkeitsvergleich von Modellen in vielen Fällen aufgrund der mit der Berechnung einhergehenden Komplexität nicht praktikabel (NP-Vollständigkeit des Graph-Matching-Problems [24]). Vor diesem Hintergrund kann eine Reduktion der zu vergleichenden Modellmenge die Anwendung nachfolgender Ähnlichkeitsmaße voraussichtlich stark beschleunigen. Im Gegensatz zur Arbeit von QIAO ET AL. soll ein Modell hierbei inhaltlich deutlich detaillierter betrachtet und nicht nur auf Grundlage einer groben Kategorisierung klassifiziert werden. Somit kann das Ergebnis der inhaltlichen Erschließung zum Retrieval von ähnlichen Modellen verwendet werden.

Das Konzept gliedert sich in zwei Ebenen mit insgesamt drei Phasen (Abbildung 1); die erste Ebene „Aufbereitung“ umfasst die Vorverarbeitung aller Modelle der Prozesssammlung während die zweite Ebene „Retrieval“ die Phasen zur inhaltlichen Eingrenzung des Suchraums in Bezug auf eine Suchanfrage und die anschließende Detailanalyse der Elemente des eingegrenzten Suchraums beinhaltet.

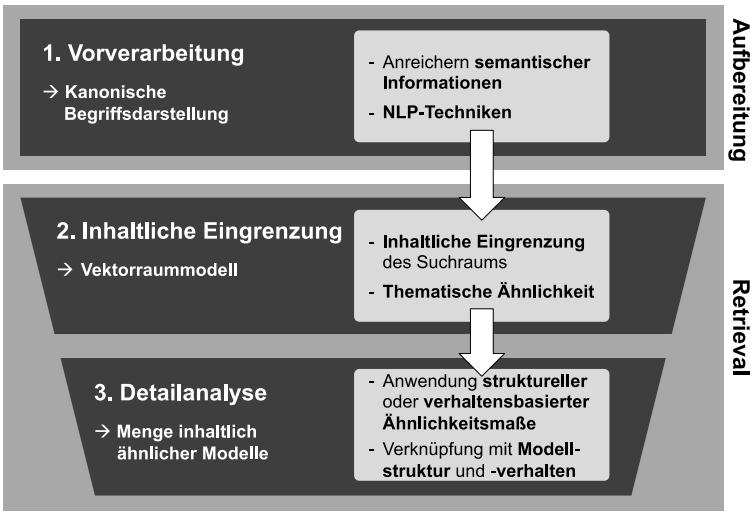

Abbildung 1. Konzept zur inhaltlichen Erschließung von Prozessmodellen

Im Folgenden werden die drei Phasen näher erläutert.

1. Vorverarbeitung. In der ersten Phase werden alle Modelle der zu betrachtenden Prozesssammlung durch die Anwendung von NLP-Techniken vorverarbeitet, um sie einem anschließenden Retrieval zugänglich zu machen und sie mit automatisch generierten semantischen Informationen anzureichern. Zu diesem Zweck wird jedes Modell in eine Vektordarstellung (VSM) transformiert: die Elemente eines Vektors sind hierbei sinntragende Wörter, die den Inhalt des im Modell formulierten Sachverhalts beschreiben (sog. *Indexterme*). Diese Indexterme müssen folgende Eigenschaften aufweisen: Sie müssen den Inhalt eines Modells möglichst *detailliert* beschreiben, über die gesamte Prozesssammlung *charakteristisch* sein und *eindeutig* innerhalb selbiger verwendet werden.

Um aus einem Prozessmodell mit textuellen Label-Bezeichnern eine Menge von Indextermen mit den genannten Eigenschaften abzuleiten, wird nachfolgend eine Vorgehensmethode mit drei Einzelschritten präsentiert:

1. Im ersten Schritt werden alle Labels eines Modells in Bezug auf die verwendeten Label-Styles hinsichtlich der Klassifikation von LEOPOLD ET AL. analysiert [8]. Als Ergebnis liegen im Anschluss für jedes Label Informationen über durchgeführte Aktionen (*actions*) und verwendete Prozessressourcen (*business objects*) vor. Eine PoS-Analyse liefert zudem Informationen über entsprechende Wortklassen.
2. Anschließend werden Wörter entfernt, die für die weitere Analyse nicht von Bedeutung sind. Dies sind zum einen Wörter, die über klassische Verfahren wie das Entfernen von Stoppwörtern (*stop word removal*) erkannt werden können und zum anderen Wörter, die nicht im Rahmen der *Label-Style*-Analyse klassifiziert werden konnten oder durch die PoS-Analyse z. B. als Präposition o. ä. erkannt wurden.
3. Im letzten Schritt werden die verbliebenen Wörter durch das Ersetzen von Synonymen mit einheitlichen Formulierungen harmonisiert. Dies kann auf verschiedene Arten geschehen, z. B. durch den Abgleich mit einem manuell erstellten Glossar

von unternehmensspezifischen Begriffen oder durch die Nutzung von Ressourcen wie *WordNet*. Abschließend werden die vereinheitlichten Wörter durch die Anwendung von Stemming-Verfahren auf ihren Wortstamm reduziert.

Im Ergebnis steht nach Anwendung der genannten Schritte eine konsistente Menge von Indextermen, die in Abbildung 1 als kanonische, d. h. standardisierte, Begriffsdarstellung bezeichnet ist.

2. Inhaltliche Eingrenzung. Nach der Identifikation relevanter Indexterme eines Modells und der anschließenden Aufbereitung können diese in der zweiten Konzeptphase zum Aufbau eines Vektorraums genutzt werden. Dieser besteht aus der Menge aller Vektoren über der Prozesssammlung, wobei ein Vektor pro Modell existiert. Die Anzahl der Dimensionen eines Vektors entspricht der Anzahl aller Indexterme der Sammlung und die Vektorenelemente repräsentieren das Auftreten der jeweiligen Indexterme. An dieser Stelle sind verschiedene Alternativen denkbar: neben einer binären Unterscheidung, ob ein Indexterm im Modell enthalten ist oder nicht, kann auch eine Gewichtung nach der Häufigkeit seines Auftretens vorgenommen werden. Textbasierte Information-Retrieval-Systeme verwenden diesbezüglich häufig eine Gewichtung nach dem TF-IDF-Verfahren. Aufgrund der – im Vergleich zu Vektorraummodellen über umfangreiche Textsammlungen – geringen Gesamtzahl von Indextermen (bedingt durch kurze Formulierungen in Labels) muss untersucht werden, ob das TF-IDF-Verfahren in diesem Zusammenhang zusätzlichen Nutzen bringen kann.

Zum inhaltlichen Ähnlichkeitsvergleich zwischen zwei Prozessmodellen wird der Abstand zwischen ihren Vektorendarstellungen berechnet. Ein weit verbreitetes Distanzmaß ist die *Kosinusähnlichkeit*, welche den Abstand zweier Vektoren über die Größe des zwischen ihnen gelegenen Winkels berechnet. Modelle werden bis zu einem gewissen Schwellenwert des Distanzmaßes als ähnlich betrachtet und ihrem Abstand entsprechend geordnet.

3. Detailanalyse. Die Ausführung der ersten beiden Phasen des Konzeptes resultiert in einer Menge von inhaltlich ähnlichen Modellen. Diese Menge kann in zweierlei Hinsicht genutzt werden. Erstens kann sie selbst bereits als geordnete Resultatmenge betrachtet werden, in der ein Benutzer zu einem Query-Modell inhaltlich ähnliche Modelle identifizieren kann. In diesem Fall ist keine weitere Analyse nötig und die inhaltliche Eingrenzung des Suchraums fungiert als eigenes Ähnlichkeitsmaß. Zweitens kann sie auch als Vorstufe für weitergehende Ähnlichkeitsvergleiche basierend auf der Modellstruktur oder Modellverhalten betrachtet werden. Die Einschränkung des Suchraums aus inhaltlicher Sicht reduziert die Anzahl der zu vergleichenden Modelle erheblich und kann somit den Einsatz struktureller und verhaltensbasierter Vergleiche in annehmbarer Laufzeit ermöglichen.

Das hier beschriebene Verfahren rückt – in Abgrenzung zu bestehenden Ansätzen zur thematischen Einschränkung des Suchraums – eine detaillierte inhaltliche Betrachtung von Prozessmodellen in den Vordergrund. Im Gegensatz zu [21] werden hier Ähnlichkeitsvergleiche nicht durch die Analyse von Metadaten eines Modells, sondern auf der Grundlage von vorhandenen Modellbezeichnern durchgeführt. Dadurch wird ein wesentlich detaillierteres Verständnis des modellierten Sachverhalts möglich, das über eine thematische Klassifikation hinausgehend neue Potentiale zur Erschließung von Modellsammlungen bietet.

6 Diskussion

Techniken der Natürlichen Sprachverarbeitung ermöglichen ein tieferes Verständnis der innerhalb von Prozessmodellen formulierten Sachverhalte. Das präsentierte Konzept unterstützt einen inhaltlichen Vergleich von Prozessmodellen. Durch das Ersetzen von Synonymen mit einheitlichen Indextermen und durch Abstraktion auf ihren Wortstamm können trotz unterschiedlicher Formulierungen Beziehungen zwischen Modellen erkannt werden. Als Vorstufe zu einer detaillierteren Analyse mit beliebigen Ähnlichkeitsvergleichsmaßen kann der vorgestellte Ansatz zu einer starken Laufzeit- und Komplexitätsreduktion beitragen, aber auch die alleinige Anwendung einer inhaltlichen Suche kann bereits interessante Zusammenhänge offenbaren.

Es ist anzumerken, dass im präsentierten Konzept stets das Vorhandensein einer inhaltlichen Ähnlichkeit zwischen zu vergleichenden Modellen unterstellt wird. Dies erscheint in den meisten Fällen sinnvoll, da davon auszugehen ist, dass beispielsweise ein Modellierer während der Umgestaltung eines bestehenden Modells durch die Suche nach ähnlichen Modellen in einer Sammlung gezielt nach „verwandten" Prozessen recherchieren möchte. Wird als Query-Modell etwa der Prozess einer Flugreservierung gewählt, so erscheint es intuitiv, dass Modelle, die sich „im weiteren Sinne" mit Flügen oder Reservierungen befassen, als ähnlich zur Eingabe betrachtet werden – der Toleranzbereich bei der Suche hängt von den Parametern der Ähnlichkeitsberechnung zwischen den Vektoren ab. Möglicherweise existieren aber auch Anwendungsfälle, in denen zu einem Query-Modell all diejenigen Modelle einer Sammlung identifiziert werden sollen, die Ähnlichkeit bezüglich ihrer Struktur aufweisen und deren Inhalt vernachlässigt werden kann.

Die weitere Entwicklung und Verfeinerung des vorgestellten Konzeptes ist Gegenstand aktueller Forschungsarbeiten des Instituts für Wirtschaftsinformatik (IWi) im DFKI. Neben einer detaillierteren Ausarbeitung einzelner Konzeptphasen ist deren konkrete Umsetzung in Form eines prototypischen Softwaresystems geplant. Vor diesem Hintergrund ist insbesondere eine Integration in die Software des Projektes *RefMod-Miner* denkbar, welche bereits eine umfangreiche Methodensammlung u. a. zur Analyse und zum Vergleich von Prozessmodellen zum Zweck einer induktiven Referenzprozessmodellentwicklung [25] implementiert.[4] Genauere Abschätzungen zur Konfiguration der Parameter und zu deren Performanz können erst nach Implementierung, Test und Evaluation mit geeigneten Modellkorpora abgegeben werden.

7 Zusammenfassung und Ausblick

Der vorliegende Work-in-Progress-Artikel verfolgte die Zielsetzung, den aktuellen Stand der Forschung im Bereich *Business Process Similarity* zu untersuchen, um einen Überblick über die dort verwendeten Methoden zur Ähnlichkeitsbestimmung zu erhalten. Die Autoren möchten damit einen Beitrag zur Systematisierung des Forschungsstandes leisten und eine Grundlage für die Entwicklung neuer Ansätze zum

[4] http://refmod-miner.dfki.de

inhaltlichen Vergleich von Prozessmodellen schaffen. Zu diesem Zweck wurde ein strukturiertes Literaturreview in fünf wissenschaftlichen Datenbanken durchgeführt, das insgesamt 84 relevante Arbeiten analysierte. Die Untersuchung zeigte, dass Methoden der Natürlichen Sprachverarbeitung vergleichsweise selten genutzt werden, um Prozessmodelle auch inhaltlich zu erschließen. Um die Potenziale in diesem Bereich zu nutzen, wurde ein Konzept zur inhaltlichen Erschließung von Modellen entwickelt. Dieses beschreibt grundlegende Ideen zur Aufbereitung und Verarbeitung von Prozessmodellen durch ein Phasenmodell. Eine detaillierte Ausgestaltung der einzelnen Phasen in Form eines Feinkonzeptes wird momentan erarbeitet.

Um die konkrete Anwendbarkeit des Konzeptes zu demonstrieren und es in Bezug auf eine erreichbare Effizienzsteigerung beurteilen zu können, wird in Zukunft die prototypische Implementierung eines Softwaresystems angestrebt. Im Rahmen einer sich daran anschließenden Evaluation soll ein Vergleich mit bestehenden Methoden zur semantischen Ähnlichkeitsbestimmung durchgeführt werden.

Literatur

1. Zha, H., Wang, J., Wen, L., Wang, C., Sun, J.: A workflow net similarity measure based on transition adjacency relations. Comput. Ind. 61, 463–471 (2010).
2. Uba, R., Dumas, M., Garc, L.: Clone Detection in Repositories of Business Process Models. In: Rinderle-Ma, S. , Toumani, F., Wolf, K. (Hrsg.) Business Process Management (BPM 2011), LNCS 6896. S. 248–264. Springer, Berlin, Heidelberg (2011)
3. Woo, H.-G., Song, M.: A structural matching approach to manage large business process models. Proceedings of the 41st International Conference on Computers & Industrial Engineering. S. 1075–1080. Los Angeles (2011)
4. Becker, M., Laue, R.: A comparative survey of business process similarity measures. Comput. Ind. 63, 148–167 (2012)
5. Walter, J., Fettke, P., Loos, P.: Zur Identifikation von Strukturanalogien in Prozessmodellen. Tagungsband der Multikonferenz Wirtschaftsinformatik (MKWI 2012). S. 1703–1715 (2012)
6. Fettke, P.: State-of-the-Art des State-of-the-Art: Eine Untersuchung der Forschungsmethode „Review “ innerhalb der Wirtschaftsinformatik. Wirtschaftsinformatik 48, 257–266 (2006)
7. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS Q. 28, 75–105 (2004)
8. Leopold, H., Niepert, M., Weidlich, M., Mendling, J., Dijkman, R., Stuckenschmidt, H.: Probabilistic Optimization of Semantic Process Model Matching. In: Barros, A., Gal, A., and Kindler, E. (Hrsg.) Business Process Management (BPM 2012), LNCS 7481. S. 319–334. Springer, Berlin, Heidelberg (2012)
9. van der Aalst, W.M.P., Medeiros, A.K.A. De, Weijters, A.J.M.M.: Process Equivalence: Comparing Two Process Models Based on Observed Behavior. Business Process Management (BPM 2006), LNCS 4102. S. 129–144. Springer, Berlin, Heidelberg (2006)
10. Gerth, C.: Business Process Models. Change Management, Springer, Berlin, Heidelberg (2013)
11. Becker, M., Laue, R.: Analysing Differences between Business Process. In: Daniel, F., Barkaoui, K., Dustdar, S. (Hrsg.) Business Process Management Workshops, LNBIP 100. S. 39–49. Springer, Berlin, Heidelberg (2012)

12. Dumas, M., García-Bañuelos, L., Dijkman, R.M.: Similarity Search of Business Process Models. IEEE Data Eng. Bull. 32, 23–28 (2010)
13. Dijkman, R., Dumas, M., van Dongen, B., Käärik, R., Mendling, J.: Similarity of business process models: Metrics and evaluation. Inf. Syst. 36, 498–516 (2011)
14. Humm, B.G., Fengel, J.: Semantics-Based Business Process Model Similarity. In: Abramowicz, W., Kriksciuniene, D., Sakalauskas, V. (Hrsg.) Business Information Systems (BIS 2012), LNBIP 117. S. 36–47. Springer, Berlin, Heidelberg (2012)
15. Chowdhury, G.G.: Natural language processing. Annu. Rev. Inf. Sci. Technol. 37, 51–89 (2003)
16. Charniak, E.: Statistical Techniques for Natural Language Parsing. AI Mag. 18, 33–43 (1997)
17. Feldman, R., Sanger, J.: The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cambridge (2007)
18. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. J. Doc. 28, 11–21 (1972)
19. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24, 513–523 (1988)
20. Österle, H., Otto, B.: Konsortialforschung - Eine Methode für die Zusammenarbeit von Forschung und Praxis in der gestaltungsorientierten Wirtschaftsinformatikforschung. Wirtschaftsinformatik. 52, 273–285 (2010)
21. Qiao, M., Akkiraju, R., Rembert, A.J.: Towards Efficient Business Process Clustering and Retrieval: Combining Language Modeling and Structure Matching. In: Rinderle-Ma, S., Toumani, F., Wolf, K. (Hrsg.) Business Process Management (BPM 2011), LNCS 6896. S. 199–214. Springer, Berlin, Heidelberg (2011)
22. Niemann, M., Siebenhaar, M., Schulte, S., Steinmetz, R.: Comparison and retrieval of process models using related cluster pairs. Comput. Ind. 63, 168–180 (2012)
23. Reijers, H. a., Mendling, J.: A Study Into the Factors that Influence the Understandability of Business Process Models. IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans. 41, 449–462 (2011)
24. Abbas, S., Seba, H.: A module-based approach for structural matching of process models. Proceedings of the 2012 Fifth IEEE International Conference on Service-Oriented Computing and Applications (SOCA). S. 1–8. IEEE Computer Society Press, Washington, DC, USA (2012)
25. Ardalani, P., Houy, C., Fettke, P., Loos, P.: Towards a Minimal Cost of Change Approach for Inductive Reference Model Development. Proceedings of the 21st European Conference on Information Systems (ECIS). AIS, Utrecht, Netherlands (2013)