

Erkennung und Vorhersage von Fehlern in EPK-Geschäftsprozessmodellen

Jan Mendling

BPM Group, Queensland University of Technology
j.mendling@qut.edu.au

Abstract: Dieser Beitrag fasst die wesentlichen Ergebnisse meiner Doktorarbeit mit dem Titel *Detection and Prediction of Errors in EPC Business Process Models* zusammen. Die Motivation der Arbeit gründet auf der Beobachtung, dass Geschäftsprozessmodelle bedeutende Designartefakte im Geschäftsprozessmanagement-Lebenszyklus sind, und daher unter Qualitätsgesichtspunkten betrachtet werden sollten. In diesem Zusammenhang ist es von entscheidender Bedeutung für die spätere informationstechnische Unterstützung, dass die Modelle formal fehlerfrei sind und insbesondere keine Deadlocks enthalten. Vor diesem Hintergrund liefert meine Doktorarbeit vier Beiträge. Als erstes wird eine neue Formalisierung der operationalen Semantik von Ereignis-gesteuerten Prozessketten (EPKs) vorgestellt. Auf Basis dieser Formalisierung wird dann ein Korrektheitskriterium für EPKs definiert, das sich dafür eignet Ausführungsprobleme zu erkennen. Als Drittes wird eine Reihe von Metriken zur Charakterisierung der Struktur und des Verhaltens von EPK-Modellen beschrieben, die einen vermuteten Zusammenhang mit der Fehlerwahrscheinlichkeit eines Prozessmodells haben. Als Viertes wird eine Auswahl von über 2000 EPK-Modellen aus der Praxis herangezogen, um über die Schätzung einer logistischen Regressionsfunktion den Zusammenhang zwischen Metriken und Fehlerwahrscheinlichkeit zu prüfen.

1 Überblick

Dieser Beitrag fasst die wesentlichen Ergebnisse meiner Doktorarbeit mit dem Titel *Detection and Prediction of Errors in EPC Business Process Models* [Men07] zusammen. Geschäftsprozessmanagement ist ein grundlegendes Gestaltungskonzept, das strategische und operative Anforderungen einer Unternehmung mit informationstechnischen Konzepten verbindet. Unter dem Begriff Geschäftsprozessmanagement werden sowohl Software-Pakete von Anwendungsherstellern als auch Beratungsdienstleistungen von Consulting-Firmen vermarktet. Das auf Informationstechnologie spezialisierte Marktforschungsunternehmen Gartner sieht die Geschäftsprozessmodellierung als wesentlichen Baustein des Geschäftsprozessmanagements unter den Top Ten strategischen Technologien 2008.

Die Motivation der Arbeit gründet auf der Beobachtung, dass Geschäftsprozessmodelle bedeutende Designartefakte im Geschäftsprozessmanagement-Lebenszyklus sind, und daher unter Qualitätsgesichtspunkten betrachtet werden sollten. In diesem Zusammenhang ist es von entscheidender Bedeutung für die spätere informationstechnische Unterstützung von Prozessen, dass die Modelle formal fehlerfrei sind und insbesondere keine Deadlocks

enthalten. Vor diesem Hintergrund liefert meine Doktorarbeit vier Beiträge. Als erstes wird die operationale Semantik von Ereignis-gesteuerten Prozessketten (EPKs) vollständig formalisiert. EPKs werden von vielen großen Unternehmen zur Dokumentation von Geschäftsprozessen eingesetzt und bieten mit Oder-Verknüpfungselementen eine grundlegende Erweiterung gegenüber Petri-Netzen. Zum Zweiten wird auf Basis dieser Formalisierung ein Korrektheitskriterium für EPKs definiert, das sich dafür eignet Ausführungsprobleme zu erkennen. Zur Verifikation von EPKs in der Praxis werden zwei Analysewerkzeuge vorgestellt. Als Drittes wird eine Reihe von Metriken zur Charakterisierung der Struktur und des Verhaltens von EPK-Modellen beschrieben, die einen vermuteten Zusammenhang mit der Fehlerwahrscheinlichkeit eines Prozessmodells haben. Als Viertes nutze ich eine Auswahl von über 2000 EPK-Modellen aus der Praxis, um über die statistische Schätzung einer logistischen Regressionsfunktion den Zusammenhang zwischen Metriken und Fehlerwahrscheinlichkeit zu prüfen. Diese Funktion wird mit Hilfe eines Holdout-Sample von 113 EPKs aus Lehrbüchern validiert. Es zeigt sich, dass 90% der Modelle korrekt als fehlerhaft oder fehlerfrei klassifiziert werden. Die hohe Anzahl der Fehler in den Praxis-Modellen unterstreicht einerseits die Wichtigkeit von Qualitätsbetrachtungen. Andererseits liefert die Arbeit einen wesentlichen Betrag zum Verständnis von Fehlern in Prozessmodellen und zur Entwicklung von Werkzeugunterstützung zur Vermeidung von Fehlern.

Die folgenden Seiten bieten einen illustrativen Überblick über die einzelnen Schritte dieser Arbeit. Abschnitt 2 erläutert die vorgeschlagene Ausführungssemantik von EPKs anhand eines Beispiels. Danach skizziert Abschnitt 3 die Verifikation von EPKs mit Hilfe des Erreichbarkeitsgraphen. Anschließend verdeutlicht Abschnitt 4 die Berechnung von Metriken für EPKs. Abschnitt 5 veranschaulicht zum Abschluss die Vorhersage von Fehlerwahrscheinlichkeiten mit Hilfe der logistischen Regression.

2 Ausführungssemantik der EPK

Ereignis-gesteuerte Prozessketten (EPKs) sind eine Sprache zur Modellierung von Geschäftsprozessen. Eine EPK besteht aus Funktionen und Ereignissen, die einander abwechseln sollen. Funktionen (dargestellt als abgerundete Rechtecke) stellen Geschäftsaktivitäten dar während Ereignisse (dargestellt als Sechsecke) die Vor- und Nachbedingungen von Funktionen erfassen. Darüber hinaus gibt es drei Arten von Konnektoren, die komplexe Verzweigungsregeln beschreiben: Der XOR-Konnektor (Symbol $\times$) repräsentiert eine Entscheidungssituation als Verzweigung und deren entsprechende Zusammenführung. Der UND-Konnektor (Symbol $\wedge$) signalisiert eine Verzweigung in Parallelverarbeitung und eine anschließende Synchronisation. Diese beiden Elemente können mit Hilfe von Petri-Netzen beschrieben werden. Der ODER-Konnektor stellt eine inklusive Auswahl dar, das heißt mindestens eine und maximal alle von verschiedenen Alternativen können ausgeführt werden. Die Schwierigkeit liegt darin, das Synchronisationsverhalten der ODER-Zusammenführung zu formalisieren. Verschiedene Ansätze wurden in der Vergangenheit vorgestellt (siehe [AH05, Kin06, WEAH05]). Deren Problem ist es jedoch, dass verschiedene ODER-Konnektoren auf nicht intuitive Weise miteinander interagieren können.

Die in meiner Dissertation vorgeschlagene Formalisierung der EPK-Semantik basiert auf

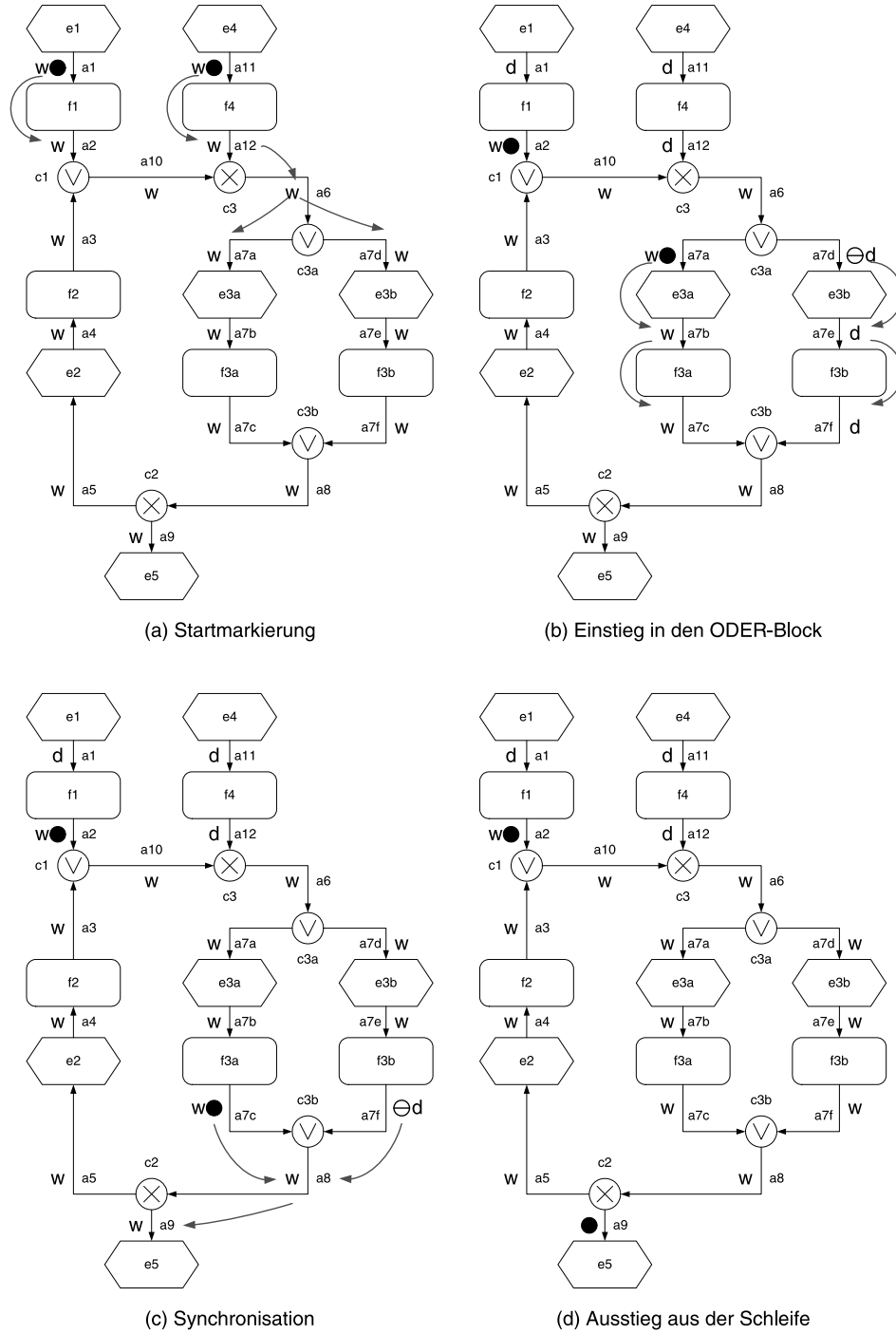


Abbildung 1: Schalten einer EPK.

der Annahme, dass wohlstrukturierte ODER-Blöcke nicht verklemmen sollten. Abbildung 1 zeigt eine entsprechende EPK mit einem solchen ODER-Block in einer Schleife. Andere Formalisierungen haben mit diesem Modell Probleme, da die ODER-Zusammenführung über die Schleife in Selbstreferenz steht. Die neue Formalisierung nutzt positive und negative Tokens ähnlich wie [LSW98], um den ODER-Block vor einer solchen Selbstreferenz abzuschirmen. Darüber hinaus wird für jede Kante festgehalten, ob als nächstes ein positives Token ankommen kann. In diesem Fall ist der Kontext-Status *wait*, im anderen Fall *dead*. Das Beispiel in Abbildung 1 stellt dar, wie eine EPK mit der neuen Semantik durchlaufen werden kann.

3 Verifikation von EPKs

Auf Basis der im vorherigen Abschnitt skizzierten Semantik lassen sich fehlerhafte EPKs mit Hilfe des Erreichbarkeitsgraphen identifizieren. Dabei werden für alle möglichen Startbelegungen alle erlaubten Schaltfolgen der EPK-Semantik durchgespielt. Das Ergebnis ist ein nicht-deterministisches Transitionssystem, das sämtliche Zustände und Zustandsänderungen der EPK beinhaltet. Dieses wird auch Erreichbarkeitsgraph genannt. Anhand des Erreichbarkeitsgraphen lassen sich Deadlocks identifizieren. Dies sind Zustände ohne weitere Übergangsmöglichkeiten, die nicht die Eigenschaften eines Endzustandes erfüllen. Im einfachsten Fall sind dies UND-Zusammenführungen, die nicht genügend Kontrolltoken erhalten, um schalten zu können.

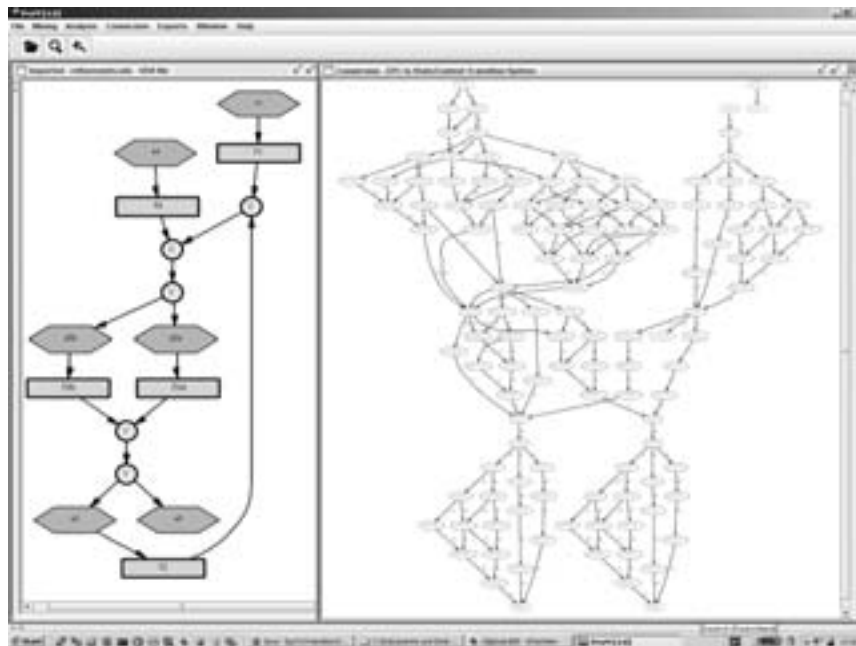


Abbildung 2: Erreichbarkeitsgraph der EPK aus Abbildung 1.

Zur Verifikation von EPKs bedarf es eines geeigneten Korrektheitskriteriums. Soundness [Aal97], das klassische Korrektheitskriterien für Workflow-Netze, kann hier nicht direkt genutzt werden, da es ein eindeutiges Startelement und ein eindeutiges Endelement braucht. Dies ist bei EPKs nicht gegeben, sie können mehrere Startereignisse und mehrere Endereignisse haben. Aus diesem Grund wird das neue Kriterium EPK Soundness [MA07] vorgeschlagen. Dieses verlangt, dass es eine Menge von Startbelegungen gibt, für die eine fehlerfreie Ausführung garantiert ist. Zudem muss jedes Startereignis in mindestens einer dieser Startbelegungen benutzt werden.

Abbildung 2 zeigt einen Bildschirmausschnitt eines Analysewerkzeuges, das zur Prüfung von EPK Soundness entwickelt wurde. Dieses Werkzeug berechnet für eine EPK den Erreichbarkeitsgraphen und stellt ihn als Transitionssystem dar. Zudem wird EPK Soundness geprüft und die Information über die Korrektheit der EPK in einem weiteren Fenster bereitgestellt. Das Werkzeug ist als Erweiterung in den Analyse-Baukasten ProM integriert (siehe [DMV⁺05]), der als Open-Source-Software frei im Internet bereitgestellt wird.

4 Berechnung von Metriken für EPKs

Im Software-Engineering und in verschiedenen anderen Bereichen werden Metriken genutzt, um wichtige Aspekte eines Programms oder eines Artefaktes zu quantifizieren. Es gibt bereits einige Arbeiten, die diese Idee auf Prozessmodelle übertragen, bspw. [LY92, Nis98, Mor99, RV04, Car05, BG05, CGP⁺05, CMNR06, LG06, MVD⁺08]. Ein Problem dieser Arbeiten ist allerdings, dass prozess-spezifische Konzepte wie Strukturiertheit und Nebenläufigkeit kaum erfasst werden. Der Fokus liegt auf einfachen Metriken, die verschiedene Modellelemente abzählen.

Vor diesem Hintergrund wurden eine Reihe neuer Metriken definiert und bestehende Metriken systematisch erfasst. Für jede dieser Metriken ergibt sich ein erwarteter Zusammenhang mit der Fehlerwahrscheinlichkeit eines Modells. Diese Erwartungen fußen insbesondere auf Überlegungen über die Kognition derer, die Modelle erstellen. Eine einfache Vermutung mit Blick auf die Anzahl der Modellelemente wäre, dass große Modelle eher Fehler aufweisen dürften als kleine, da der Modellierer in großen Modellen leichter den Überblick über das Zusammenspiel aller Elemente verliert. Demnach dürften sich eher Fehler einschleichen.

Für die EPK aus Abbildung 1 werden nun einige Metriken als Beispiel vorgestellt. Die Abzählmetriken für Knoten (16 Stück), d.h. Funktionen (5), Ereignisse (6) und Konnektoren (5), lassen sich leicht bestimmen. Die Dichte des Modells berechnet sich mit Anzahl der Kanten geteilt durch maximale Kantenzahl bei gegebener Knotenanzahl: das sind $17/(16 * 15) = 0,071$. Der maximale Grad eines Konnektors ist 3. Die Separierbarkeit setzt die Artikulationen des Prozessgraphen in Relation zur Knotenanzahl, das sind in diesem Fall $6/16 = 0,375$. Die Strukturiertheit misst wieviele Knoten durch einfache Reduktionsregeln entfernt werden können, hier $8/16 = 0,5$.

5 Fehlerwahrscheinlichkeit von EPKs

Die Vorhersage von Fehlern in Prozessmodellen anhand von Metriken wird in [MVD⁺08, MNA07] besprochen. Das statistische Werkzeug um Ereignisse vorherzusagen, in diesem Fall das Eintreten von Fehlern, ist die logistische Regression [HL00]. Die Abbildung 3 zeigt die logistische Funktion und ihren Verlauf zwischen 0 und 1. Die logistische Regressionsanalyse schätzt die Koeffizienten einer Linearkombination von Eingabevariablen (hier die Metriken), so dass für eine Auswahl von Prozessmodellen (hier 2003 EPK-Modelle aus der Praxis) der Erklärungswert einer abhängigen Variable (hier die Fehlerwahrscheinlichkeit) maximal ist. Wenn sich für ein Prozessmodell und der Linearkombination dessen Metriken ein Wert größer 0,5 ergibt, wird das Eintreffen des Ereignisses vorhergesagt. Die Fehlerhaftigkeit einer EPK wird mit Hilfe des EPK-Soundness-Kriteriums gemessen.

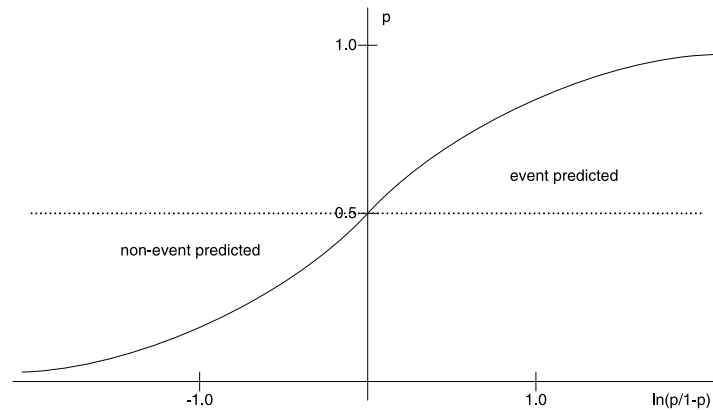


Abbildung 3: Logistische Funktion und Ereignisvorhersage.

Für die Erklärungskraft einer logistischen Regression gibt es eine Reihe von Gütekriterien. Eine davon ist das Pseudo- R^2 nach Nagelkerke, das Werte zwischen 0 (keine Erklärung) und 1 (vollständige Erklärung) annehmen kann. Es ergibt sich hier ein ausgesprochen hoher Pseudo- R^2 -Wert von 0,9. Daraus lässt sich schließen, dass die Fehlerhaftigkeit von Prozessmodellen im Wesentlichen von Strukturparametern beeinflusst werden. Dabei ist es interessant zu vermerken, dass nicht die einfachen Zählmetriken, sondern die komplexeren Messzahlen wie Strukturiertheit einen besonders hohen Einfluss haben.

6 Schlussfolgerungen

Aus der Arbeit ergeben sich eine Reihe von allgemeineren Schlussfolgerungen für die Prozessmodellierung. Aus dem Einfluss einzelner Metriken lassen sich direkt Modellierungsempfehlungen ableiten um Fehler zu vermeiden. Beispielsweise sollte strukturiert modelliert werden und Modelle ab einer Größe von 50 Elementen in mehrere Teilmodelle zerlegt

werden. Die hohe Anzahl von Fehlern in der benutzten EPK-Sammlung aus der Praxis von über 10% unterstreicht die Wichtigkeit der Verifikation für das Qualitätsmanagement von Prozessmodellen. Werkzeughersteller sollten in Zukunft entsprechende Funktionalität in ihren Tools bereitstellen.

Literatur

- [Aal97] W.M.P. van der Aalst. Verification of Workflow Nets. In Pierre Azéma und Gianfranco Balbo, Hrsg., *Application and Theory of Petri Nets 1997*, Vol. 1248 of *Lecture Notes in Computer Science*, Seiten 407–426. Springer Verlag, 1997.
- [AH05] W.M.P. van der Aalst und A.H.M. ter Hofstede. YAWL: Yet Another Workflow Language. *Information Systems*, 30(4):245–275, 2005.
- [BG05] S. Balasubramanian und M. Gupta. Structural metrics for goal based business process design and evaluation. *Business Process Management Journal*, 11(6):680–694, 2005.
- [Car05] J. Cardoso. *Workflow Handbook 2005*, Kapitel Evaluating Workflows and Web Process Complexity, Seiten 284–290. Future Strategies, Inc., Lighthouse Point, FL, USA, 2005.
- [CGP⁺05] G. Canfora, F. García, M. Piattini, F. Ruiz und C.A. Visaggio. A family of experiments to validate metrics for software process models. *Journal of Systems and Software*, 77(2):113–129, 2005.
- [CMNR06] J. Cardoso, J. Mendling, G. Neumann und H. Reijers. A Discourse on Complexity of Process Models. In Johann Eder und Schahram Dustdar, Hrsg., *BPM Workshops 2006*, Vol. 4103 of *Lecture Notes in Computer Science*, Seiten 115–126, 2006.
- [DMV⁺05] B.F. van Dongen, A.K. Alves de Medeiros, H.M.W. Verbeek, A.J.M.M. Weijters und W.M.P. van der Aalst. The ProM framework: A New Era in Process Mining Tool Support. In G. Ciardo und P. Darondeau, Hrsg., *Application and Theory of Petri Nets 2005*, Vol. 3536 of *LNCS*, Seiten 444–454. Springer-Verlag, Berlin, 2005.
- [HL00] D.W. Hosmer und S. Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, 2nd. Auflage, 2000.
- [Kin06] E. Kindler. On the semantics of EPCs: Resolving the vicious circle. *Data & Knowledge Engineering*, 56(1):23–40, 2006.
- [LG06] R. Laue und V. Gruhn. Complexity Metrics for Business Process Models. In Witold Abramowicz und Heinrich C. Mayr, Hrsg., *9th International Conference on Business Information Systems 2006*, Vol. 85 of *Lecture Notes in Informatics*, Seiten 1–12, 2006.
- [LSW98] P. Langner, C. Schneider und J. Wehler. Petri Net Based Certification of Event driven Process Chains. In J. Desel und M. Silva, Hrsg., *Application and Theory of Petri Nets*, Vol. 1420 of *Lecture Notes in Computer Science*, Seiten 286–305, 1998.
- [LY92] G.S. Lee und J.-M. Yoon. An empirical study on the complexity metrics of Petri nets. *Microelectronics and Reliability*, 32(3):323–329, 1992.
- [MA07] J. Mendling und W.M.P. van der Aalst. Formalization and Verification of EPCs with OR-Joins Based on State and Context. In J. Krogstie, A.L. Opdahl und G. Sindre, Hrsg., *Proceedings CAiSE 2007*, Vol. 4495 of *Lecture Notes in Computer Science*, Seiten 439–453, 2007.

- [Men07] J. Mendling. *Detection and Prediction of Errors in EPC Business Process Models*. Dissertation, Vienna University of Economics and Business Administration, 2007.
- [MNA07] J. Mendling, G. Neumann und W.M.P. van der Aalst. Understanding the Occurrence of Errors in Process Models based on Metrics. In R. Meersman und Z. Tari, Hrsg., *OTM Conference 2007, Proceedings, Part I*, Vol. 4803 of *Lecture Notes in Computer Science*, Seiten 113–130. Springer, 2007.
- [Mor99] S. Morasca. Measuring Attributes of Concurrent Software Specifications in Petri Nets. In *METRICS '99: Proceedings of the 6th International Symposium on Software Metrics*, Seiten 100–110, Washington, DC, USA, 1999. IEEE Computer Society.
- [MVD⁺08] J. Mendling, H.M.W. Verbeek, B.F. van Dongen, W.M.P. van der Aalst und G. Neumann. Detection and Prediction of Errors in EPCs of the SAP Reference Model. *Data & Knowledge Engineering*, 64(1):312–329, January 2008.
- [Nis98] M.E. Nissen. Redesigning reengineering through measurement-driven inference. *MIS Quarterly*, 22(4):509–534, 1998.
- [RV04] H.A. Reijers und I.T.P. Vanderfeesten. Cohesion and Coupling Metrics for Workflow Process Design. In J. Desel, B. Pernici und M. Weske, Hrsg., *Business Process Management: Second International Conference, BPM 2004, Potsdam, Germany, June 17-18, 2004. Proceedings*, Vol. 3080 of *Lecture Notes in Computer Science*, Seiten 290–305. Springer, 2004.
- [WEAH05] M.T. Wynn, D. Edmond, W.M.P. van der Aalst und A.H.M. ter Hofstede. Achieving a General, Formal and Decidable Approach to the OR-join in Workflow using Reset nets. In G. Ciardo und P. Darondeau, Hrsg., *Applications and Theory of Petri Nets 2005*, Vol. 3536 of *Lecture Notes in Computer Science*, Seiten 423–443, 2005.



Jan Mendling ist Postdoctoral Research Fellow an der Faculty of Information Technology der Queensland University of Technology in Brisbane, Australien. Zuvor arbeitete er als wissenschaftlicher Mitarbeiter am Institut für Wirtschaftsinformatik und Neue Medien der WU Wien, Österreich. Er erhielt seinen Dokortitel ebenfalls von der WU Wien für seine Arbeit unter der Betreuung von Prof. Dr. Gustaf Neumann (WU Wien) und Prof. Dr. Wil van der Aalst (TU Eindhoven). Sein Studium absolvierte er an der Universität Antwerpen, Belgien, und an der Universität Trier, von der er einen Abschluss als Diplom-Wirtschaftsinformatiker und Diplom-Kaufmann hat. Dr. Mendling hat

mehr als 80 Forschungsbeiträge veröffentlicht, unter anderem auf Top-Konferenzen wie BPM 2005, 2006 und 2007, CAiSE 2006, 2007 und 2008, EDOC 2006 und 2007, FASE 2008 und OTM 2007. Die Beiträge zur BPM 2007 und EDOC 2007 wurden für Special Issues der Konferenzen eingeladen. Des Weiteren sind seine Artikel in DKE, IJBPM, JEIM und ISeB erschienen. Darüber hinaus gehörte er über 40 internationalen Programmkomitees an, hat verschiedene Workshops geleitet, sowie an der Organisation der Konferenzen BPM 2006 in Wien und BPM 2007 in Brisbane mitgewirkt. Für die Zeitschriften ISF, ISR, IJTM, TST, IS, DSS, SCCE, DKE, IJEMISA, IJIKM, und SPIP war er als Gutachter tätig. Seine Forschungsinteressen liegen im Bereich Geschäftsprozessmanagement, Konzeptionelle Modellierung, Enterprise Systems und Qualitätsaspekte von Prozessmodellen.