# PODSL - Domänenspezifische Datenmodellierung auf Basis von Prozessen

Tobias Schneider, Stefan Jablonski

Lehrstuhl für Datenbanken und Informationssysteme
Universität Bayreuth
Universitätsstr. 30
95447 Bayreuth
Tobias.Schneider@uni-bayreuth.de
Stefan.Jablonski@uni-bayreuth.de

**Abstract:** In den letzten Jahrzehnten ist das gespeicherte Datenvolumen in Wissenschaft und Industrie exorbitant gestiegen. Dabei werden Daten zunehmend an einer zentralen Stelle für eine bestimmte Anwendungsdomäne gespeichert aber auch zwischen Teilnehmern innerhalb einer Domäne ausgetauscht. Dadurch entsteht ein erheblicher Bedarf an domänenspezifischen Datenstandards. Da innerhalb einer Domäne bestimmte Prozesse für die Datenerhebung maßgeblich sind, führen wir mit Hilfe von PODSL als Modellierungssprache domänenspezifische Datenmodelle auf Basis von Prozessen ein. Die Flexibilität der Datenmodelle wird durch die Metamodellierung von PODSL und dem Konzept der Vererbung ermöglicht. Die Anwendung von PODSL zur Erstellung von domänenspezifischen Datenstandards wird an Beispielen aus der Biodiversitätsinformatik und dem Gesundheitswesen demonstriert. Abschließend wird auf die Anwendung von mit PODSL formulierten Datenstandards beim Datenaustausch und in der Softwareentwicklung eingegangen.

## 1 Einleitung

In den letzten Jahrzehnten hat die Bedeutung der Datenspeicherung exorbitant zugenommen. So hat das Volumen der in einem Projekt gespeicherten Daten auf der einen Seite häufig eine Größenordnung erreicht, in denen herkömmliche Auswertungsmethoden nicht mehr zum Erfolg führen. Auf der anderen Seite wurde durch das Internet eine Möglichkeit zum globalen Datenaustauch geschaffen. Als Folge daraus wurden in verschiedenen wissenschaftlichen Bereichen Infrastrukturen für den globalen Datenaustausch entwickelt, wie z.B. die Global Biodiversity Information Facility (GBIF) und Encyclopedia of Life (EOL) in der Domäne der Biodiversitätsinformatik. Dies veranschaulicht, dass das Datenintegrationsproblem eine zentrale Bedeutung in diesem Anwendungsbereich besitzt. Dieses ist als das Problem definiert, Daten aus verschiedenen Quellen zu kombinieren und Nutzern eine einheitliche Sicht auf diese Daten zur Verfügung zu stellen [Le02]. In der Biodiversitätsinformatik übernehmen dabei diese Infrastrukturen als sogenannte

Megascience-Plattformen die wichtige Aufgabe der Langzeitarchivierung von Daten [THR12].

All diesen Systemen ist gemein, dass Sie von einer Vielzahl von Anwendern genutzt werden, die innerhalb derselben Anwendungsdomäne arbeiten. Die Anforderungen dieser Projekte an die Datenspeicherung sind dabei sehr unterschiedlich. Dies führt zu Datenverlusten bei zentralen Plattformen zur Datenspeicherung. Darüber hinaus sind die Anwendungsdomänen einem steten Wandel unterworfen, der sich in kontinuierlichen Änderungen der Anforderungen an ein zentrales Datenschema widerspiegelt. Eine weitere Herausforderung ist die große und sich ständig ändernde Anzahl an Teilnehmern einer Plattform. Die Motivation zur Teilnahme an einer Plattform zum Datenaustausch ist im Allgemeinen bei den jeweiligen Teilnehmern sehr unterschiedlich, da diese aus der Perspektive ihres jeweiligen Projekts Daten erheben, sammeln, speichern oder auswerten möchten. Diese Projekte haben meistens nur gemein, dass sie derselben Anwendungsdomäne angehören und eine gemeinsame Infrastruktur nutzen. Die Infrastruktur gibt ein Datenschema vor, das von allen Teilnehmern des Systems verwendet werden muss. Aufgrund der Teilnehmerstruktur werden solche Infrastrukturen im Folgenden als **offen** bezeichnet, wohingegen Infrastrukturen mit einem eingeschränkten und stabilen Teilnehmerkreis als **geschlossen** bezeichnet werden.

Offene Infrastrukturen stellen besondere Herausforderungen an die Datenintegration. So ist es möglich, dass Daten falsch interpretiert werden, wenn die Speicherung der Daten möglich ist, aber die Daten in einen anderen Kontext gesetzt werden und somit ein Bedeutungswandel stattfindet. Um dieses Problem zu lösen, wurde eine Reihe von domänenspezifischen Datenstandards entwickelt wie z.B. ABCD und DwC für die Biodiversitätsinformatik [TD09] oder aber die HL7-Standards für klinische Informationssysteme [HL13]. Die Entwicklung von domänenspezifischen Standards ist zeitaufwändig, da sie nicht zuletzt von subjektiven Meinungen von einzelnen Teilnehmern und Organisationen geprägt ist [Mo05]. In der Praxis ist aber die mangelhafte Umsetzung von Anforderungen im Bezug auf das Datenschema der Grund für das Scheitern vieler Projekte [Mo05]. Eine weitere Schwierigkeit ergib sich aus dem Problem der alternativen Datenmodelle [MS94] nach dem ein gegebenes Modellierungsproblem auf verschiedene Weise gelöst werden kann. Damit ist das Datenschema die Schlüsselstelle eines Datenstandards und maßgeblich für die Bewertung und Nutzbarkeit eines Standards zum Datenaustausch.

Als Lösung für diese Herausforderung bietet sich die Analyse der Prozesse in der Anwendungsdomäne an. Die Entwicklung von Datenmodellen aus Prozessmodellen wurde in der Dissertation „Domänenspezifische Evaluation und Optimierung von Datenstandards und Infrastrukturen“ [Sc13] ausführlich vorgestellt und dient als Grundlage für die folgenden Ausführungen. Die Prozessmodellierung als Grundlage der Analyse der Anforderungen an ein Datenmodell bietet den Vorteil, dass die Darstellung in Prozessen zunächst einen einfachen Zugang zu dieser komplexen Problemstellung ermöglicht [Sc13]. Des Weiteren sind Prozesse gut verständlich und als Technik weit verbreitet und akzeptiert. Darüber hinaus werden innerhalb einer Anwendungsdomäne bestimmte Prozesse regelmäßig ausgeführt. Wenn bei diesen Prozessen Daten erhoben werden, ist es von entscheidendend, dass die Anforderungen an die Datenerhebung im

Schema eines Datenstandards repräsentiert sind. Durch die Strukturierung in Prozessen können die Anforderungen klar strukturiert werden und in ein Datenschema für die Anwendungsdomäne übertragen werden. Spezialinteressen von Domänenexperten werden durch die Prozessmodellierung erkennbar und können diskutiert werden [Sc13]. Die Entwicklung eines Datenstandards aus einem Prozessmodell folgt der in Abbildung 1 dargestellten Vorgehensweise. Dazu wird die Entwicklung eines Datenschemas aus einem Prozessmodell in Abschnitt 2 mit Hilfe der perspektivenorientierten Prozessmodellierung (POPM) [JB96] eingeführt.


Abbildung 1: Modellierungspfad bei der Entwicklung eines domänenspezifischen Datenstandards

In Abschnitt 3 werden weitere Anforderungen an einen domänenspezifischen Datenstandard und Konzepte zur Lösung dieser Herausforderungen mit der Process Oriented Data Schema Language (PODSL) eingeführt. In Abschnitt 4 wird das Metamodell von PODSL eingeführt, welches die Modellierungssprache für die Erzeugung von Datenschemata ist. Die Erstellung von Datenschemata mit PODSL wird in Abschnitt 5 für die Domäne der Biodiversität und den Krankenhausbereich demonstriert. Ein Ausblick auf die Nutzung von mit PODSL entwickelten Datenschemata in Infrastrukturen und in der Softwareentwicklung wird in Abschnitt 6 vorgenommen. In Abschnitt 7 werden die Ergebnisse aus den vorangegangenen Abschnitten zusammengefasst und es wird ein Ausblick auf zukünftige Entwicklungen gegeben.

## 2 Von der Prozessmodellierung zur Datenmodellierung

In diesem Abschnitt wird die Entwicklung von Datenschemata auf Basis von Prozessen eingeführt. Dazu soll zunächst mit POPM eine bewährte Prozessmodellierungssprache vorgestellt werden. POPM konnte sehr gute Ergebnisse in der praktischen Anwendung in der Domäne der Biodiversitätsinformatik [Sc13] und in Krankenhäusern aufweisen [FJS07]. Anschließend wird die Ermittlung der Anforderungen an die Datenspeicherung aus Prozessen demonstriert und mit der aspektorientierten Datenmodellierung eine Methode zur Entwicklung von Datenschemata auf Basis von Prozessen eingeführt.

### 2.1 Prozessmodellierung mit POPM

Im Rahmen der perspektivenorientierte Prozessmodellierung (POPM) [Ja95] wird ein Prozess durch verschiedene Perspektiven definiert. Wenn ein Prozess modelliert wird, werden Antworten auf die Fragen Was?, Wer?, Womit?, Wie? und Wann? gesucht. Die Antworten auf diese Fragen werden in Form von Perspektiven in das Modell eingebracht. Die Perspektiven stehen dabei orthogonal zueinander [JB96]. Somit überlappen sich die Informationen der Perspektiven nicht. In POPM werden folgende Basisperspektiven definiert [Ja95] [JB96]:

- Funktionale Perspektive (Was?): Beschreibt die funktionalen Einheiten eines Prozesses, die ausgeführt werden sollen.
- Datenorientierte Perspektive (Womit?): Beschreibt, wo innerhalb eines Prozesses Daten erzeugt oder aber konsumiert werden. Im Rahmen dieser Perspektive können neben Daten, die in Dokumenten erfasst werden, auch physische Erzeugnisse verstanden werden.
- Organisatorische Perspektive (Wer?): Beschreibt, wer für die Ausführung des Prozesses verantwortlich ist. Dies muss nicht zwangsläufig eine natürliche Person sein, sondern kann auch eine Organisation oder aber eine Maschine sein [Bu98].
- Operationale Perspektive (Wie?): Beschreibt die Werkzeuge, die bei der Ausführung eines Prozesses verwendet werden.
- Verhaltensorientierte Perspektive (Wann?): Diese Perspektive legt fest, in welcher Reihenfolge die Prozesse innerhalb eines Prozessmodells ausgeführt werden sollen.

Die Auflistung der Perspektiven ist nicht abschließend. Nach den Modellierungsanforderungen der Domäne können weitere Perspektiven wie z.B. die Kausalitätsperspektive (Warum?) eingeführt werden [JB96].

## 2.2 Ermittlung der Anforderungen an die Datenspeicherung aus Prozessen

Die Erfassung von Daten ist das Ergebnis eines Prozesses der Datenaufnahme. So entstehen im Rahmen der Biodiversitätsinformatik die Daten nicht einfach durch das Füllen von Datenstrukturen, sondern werden in Begehungen erhoben. In Rahmen einer Begehung analysiert ein Biologe beispielsweise die Vegetation einer Wiese und erstellt dadurch eine Artenliste, welche die taxonomischen Bezeichnungen der identifizierten Fundobjekte enthält. Zur Demonstration der Ableitung von Anforderungen an das Schema soll folgender Prozess (Abbildung 2) aus der Biodiversitätsinformatik betrachtet werden:

**Prozess der Geländekartierung**: *Ein Kartierer ist auf der Suche nach biologischen Objekten z.B. nach Pilzen in einem zuvor definierten Gebiet. Hat er einen Pilz gefunden, dokumentiert er die taxonomische Bezeichnung des Fundes sowie den Ort und Zeitpunkt der Kartierung.*

In Abbildung 2 wird in der datenorientierten Perspektive deutlich, dass in jedem Prozessschritt mit PED1 ein Dokument benötigt wird, welches das Ergebnis des Prozesses speichert. Elemente der datenorientierten Perspektive zur Speicherung des Prozessergebnisses werden als ProcessExecutionDocument (PED) bezeichnet.

Die konkrete Ausführung des Prozesses führt zu einer Aussage wie:

*Josef Simmel hat am 27.3.2012 um 14.12 Uhr ein bestimmtes biologisches Objekt als Quercur robur (Eiche) identifiziert, welches sich an den GPS Koordinaten 49.4628332, 11.3526638 befindet.*

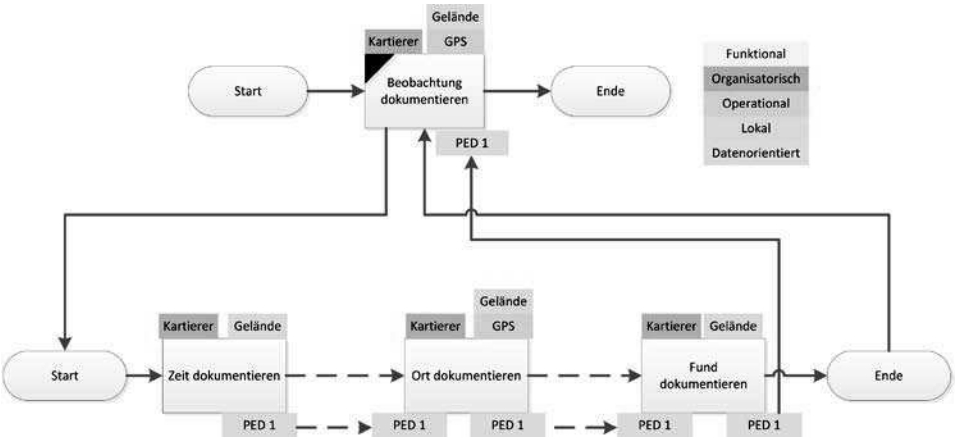


Abbildung 2: Prozess der Geländekartierung

Dementsprechend muss ein PED über ein Datenschema verfügen, welches die bei der Prozessausführung entstehenden Daten erfassen kann. Im konkreten Beispiel kann am Prozessmodell aus der organisatorischen Perspektive abgelesen werden, dass der Name des Kartierers ein solches Datum ist. Analog dazu kann au der funktionalen Perspektive in den Subprozessen abgeleitet werden, dass im Schema des PED Felder zur Erfassung von Zeit und Ort benötigt werden. Durch die Prozessperspektiven werden somit Anforderungen an das Schema eines PED spezifiziert. Für die Entwicklung von domänenspezifischen Datenstandards folgt daraus, dass sobald die Prozesse einer Domäne formuliert sind auch die Anforderungen an den Datenstandard bekannt sind.

## 2.3 Aspektorientierte Dokumentenmodellierung

Im vorangegangenen Abschnitt wurde gezeigt, dass sich die Anforderungen an ein Datenschema direkt aus einem Prozessmodell ableiten lassen. In diesem Abschnitt wird eine strukturierte Methode zur Erstellung von Schemata eingeführt, welche auf der Zerlegung eines Prozesses in Perspektiven beruht. Dazu werden den Prozessperspektiven im Schema des PED thematisch unabhängige Bereiche gegenübergestellt, die die Speicherung der Daten bei der Prozessausführung zur Aufgabe haben. Diese Bereiche werden als **Dokumentenaspekte** bezeichnet.

**Definition:** *Ein Dokumentenaspekt eines PED ist ein Bereich des Schemas zur eindeutigen Speicherung von Daten eines bestimmten Themenbereichs, der orthogonal zu allen anderen Dokumentenaspekten eines PEDs steht [Sc13].*

Orthogonal bedeutet in diesem Zusammenhang, dass sich die Aspekte nicht überlappen, also eine thematisch disjunkte Gliederung eines Dokumentes ermöglichen. Die Auswahl der Aspekte ist dabei von der betrachteten Domäne und der Art der Prozesse dieser Domäne abhängig.

Folgende Aspekte konnten in Projekterfahrungen für die die Domäne der Biodiversitätsinformatik identifiziert werden [Sc13]:

- **Funktionaler Aspekt**: Speicherung der Primärdaten. Das sind die Daten zu dessen Zweck der Prozess der Datenerhebung ausgeführt wurde (z.B. Daten von Messungen, taxonomische Bestimmungen).
- **Organisatorischer Aspekt:** Speicherung des Verantwortlichen der Prozessausführung
- **Operationaler Aspekt:** Speicherung der verwendeten Werkzeuge und Hilfsmittel
- **Datenorientierter Aspekt:** Speicherung von Referenzen auf andere Daten und Dokumente, die während eines Erhebungsprozesses erfasst wurden oder Speicherung von Referenzen auf physische oder virtuelle Objekte des Erhebungsprozesses (z.B. Mitnahme von Belegen, Multimediaobjekte, externe Daten)
- **Temporaler Aspekt:** Speicherung des Zeitpunkts der Prozessausführung
- **Lokaler Aspekt**: Speicherung des Ausführungsorts
- **Verhaltensorientierter Aspekt**: Im Verhaltensorientierten Aspekt wird die zeitliche Abfolge zwischen Prozessen erfasst. Somit werden in diesem Aspekt verschiedene Aussagen mit ihrer zeitlichen Reihenfolge verknüpft.

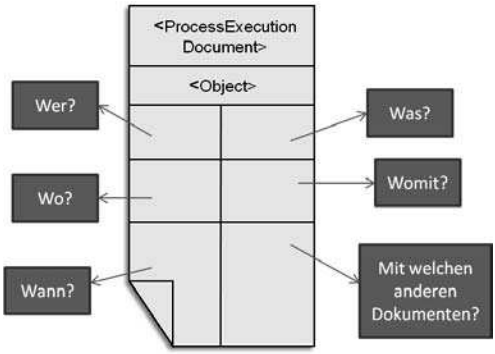


Abbildung 3: Orthogonale Zerlegung eines PED´s nach Aspekten

Den Dokumentenaspekten sind die elementaren Fragen nach Was?, Wer?, Wie?, Wann? und Wo? zugeordnet (Abbildung 3). Die Auflistung der Dokumentenaspekte ist analog zu den Prozessperspektiven nicht abschließend. Je nach Anwendungsdomäne kann es notwendig sein, weitere Dokumentenaspekte aufzunehmen. Um die Daten einer Prozessausführung in einem PED zu speichern, sind folgende Regeln einzuhalten [Sc13]:

- Für jede Prozessperspektive muss ein korrespondierender Dokumentenaspekt existieren.
- Der lokale und temporale Dokumentenaspekt muss existieren, um den Zeitpunkt und der Ort der Prozessausführung abzubilden.

Diese Anforderungen an ein PED sind zur Erfassung eines Prozesses notwendig, da ein unvollständiges Schema zu Datenverlust führt und somit für die Dokumentation eines Prozesses ungeeignet ist. Damit ist das PED das zentrale Dokument zur Speicherung der Ausführung eines Prozesses und Grundlage für die Speicherung der Prozessausführung in der Metastruktur von PODSL.

### 2.4 Genauigkeit der Erfassung der Prozessausführung

Es genügt nicht nur, dass alle Prozessperspektiven in einem Dokumentenaspekt repräsentiert sind. Diese Repräsentation muss auch mit einer bestimmten Genauigkeit erfolgen, damit das Schema des PED den Prozess ausreichend erfassen kann. Ein PED zur Erfassung des Prozesses der Geländekartierung findet sich in Abbildung 4. Das Schema des PED´s enthält für alle Prozessperspektiven aus Abbildung 2 Dokumentenaspekte zur Aufnahme der Daten. Das Schema des PED´s kann diese Dokumentenaspekte nicht mit der erforderlichen Genauigkeit erfassen, da die Uhrzeit der Prozessausführung nicht erfasst wurde. Diese gehört aber zu den Anforderungen der Geländekartierung. Die Anforderungen des Prozesses an das Dokument zur Erfassung des Prozesses sind damit nicht vollständig erfüllt.

| | |
|---|---|
| **Kartierer:** | Josef Simmel |
| **Zeitpunkt:** | 14.07.2012 |
| **Sammelort(Klartext):** | Großer Waldweg westlich Bruckhäusl (MTB 6939/2) und angrenzende Waldbereiche. |
| **Latitude:** | 12.291050911 |
| **Longitude:** | 49.070976257 |
| **GPS-Chip:** | SiRF III |
| **Satteliten:** | 3 |
| **Fehlertoleranz** | 30 m |
| **Taxonomische Bezeichnung:** | Salix fragilis L. |

Abbildung 4: Beispieldatensatz mit einem Mangel im temporalen Dokumentenaspekt [Sc13]

Die klare Untergliederung eines PED´s in Dokumentenaspekte ermöglicht es, diese Fehler im Design zu identifizieren. Die erforderliche Genauigkeit muss dabei durch Interaktion mit Domänenexperten ermittelt werden, wobei das Prozessmodell als Diskussionsgrundlage dient.

## 3 Anforderungen an einen domänenspezifischen Datenstandard und Umsetzung in PODSL

Kernaufgabe eines domänenspezifischen Datenstandards ist der Datenaustausch zwischen den verschiedenen Teilnehmern einer offenen Infrastruktur innerhalb einer Domäne. Durch diesen Anwendungshintergrund wird eine Reihe von Anforderungen definiert, die ein domänenspezifischer Datenstandard in diesem Kontext erfüllen muss. Im folgenden Abschnitt werden die wichtigsten Anforderungen diesbezüglich aufgeführt und mit den Modellierungskonzepten von PODSL gelöst. Für weitere Anforderungen und ihre Lösung mit PODSL wird auf [Sc13] verwiesen.

### 3.1 Technologieunabhängigkeit

In einer offenen Infrastruktur werden verschiedene Technologien zur Datenspeicherung eingesetzt. Ein domänenspezifischer Datenstandard befindet sich dementsprechend in einer Konfliktsituation, wie in Abbildung 5 (links) dargestellt ist. Der Datenstandard muss mit Datenspeichern und Programmen auf Basis von verschiedenen Technologien kommunizieren. Diese sind nur bedingt zueinander kompatibel oder plattformspezifisch. Der Datenstandard muss den Austausch von Daten über diese Technologiegrenzen hinweg ermöglichen. So werden z.B. in der Biodiversitätsinformatik mit ABCD und DwC einerseits Datenstandards verwendet, die in XML spezifiziert sind. Andererseits müssen dieselben Daten mit OBOE [Ma07] in einer Ontologie oder in der objektorientierten Programmierung und in relationalen Datenbanken verwendet werden können. Für die Zukunft sind weitere Technologien wie die Verwendung von NoSQL-Datenbanken denkbar.

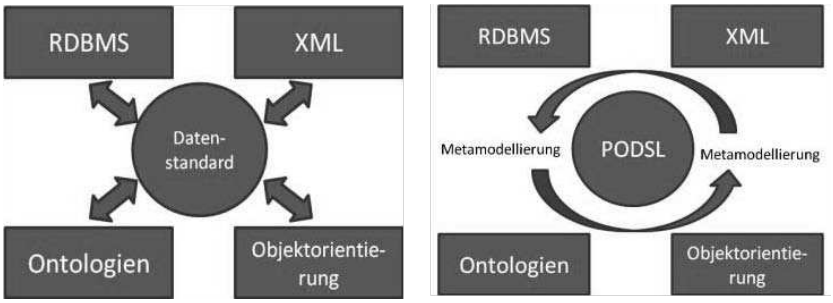


Abbildung 5: Konfliktsituation eines Datenstandards in einer offenen Infrastruktur und Lösung durch Metamodellierung [Sc13]

Die Herausforderung in einer offenen Infrastruktur ist, dass der Datenaustausch über Technologiegrenzen erfolgen muss. Eine Lösung hierfür bietet die Metamodellierung, wie in Abbildung 5 (rechts) dargestellt ist. Datenmodelle aus unterschiedlichen Technologiebereichen werden durch ein moderierendes Metamodell ineinander abgebildet. Ein Standard für die Metamodellierung wird mit der Meta Object Facility (MOF) durch die Object Management Group (OMG) spezifiziert. Nach der Spezifikation der MOF [Gr11] muss ein Metamodell über eine Mindestanzahl von zwei Ebenen verfügen, wobei theoretisch beliebig viele Ebenen unterstützt werden können. Für die Formulierung von PODSL wurde eine dreischichtige Metastruktur verwendet, welche mit Hilfe des Open MetaModeling Environment (OMME) entwickelt wurde [Vo11]. OMME unterstützt die Erweiterung eines Modells mit Vererbung und Powertypes [Vo11], welche bei der Formulierung von PODSL benötigt wurden [Sc13]. Die dreischichtige Metastruktur von PODSL besteht aus der Metaebene (M2), der Ebene der Modelle (M1) und der Populationsebene (M0). Die Metastruktur von PODSL ist in Abschnitt 4 genau beschrieben.

## 3.2 Vollständigkeit

Für einen domänenspezifischen Datenstandards ist die Vollständigkeit der Erfassung der Anwendungsdomäne von entscheidender Bedeutung, da eine unzureichende Erfassung zu Datenverlusten und damit zur Ablehnung durch die Nutzer des Modells führt. In einem domänenspezifischen Datenstandard müssen dementsprechend alle wichtigen Elemente zur Beschreibung dieser Domäne vorhanden sein. Um die Qualität eines Datenschemas zu messen, wird die Fehlerklassifikation nach Moody [Mo98] verwendet. Diese Klassifikation wurde in [Sc13] mit der Entwicklung der Process Oriented Schema Evaluation (POSE) zu einem System zur Messung der Vollständigkeit von Datenschemata weiterentwickelt.

Bei der Evaluation von Datenschemata nach Moody wird zwischen Fehlern der 1.-3. Art unterschieden (vgl. Abbildung 6) unterschieden [Mo98]. Elemente, die im Datenmodell existieren, denen aber keine konkrete Anforderung zugeordnet werden kann, werden als Fehler erster Art bezeichnet. Nutzeranforderungen die bei der Modellierung des Datenmodells nicht berücksichtigt wurden, werden als Fehler zweiter Art bezeichnet. Als Fehler dritter Art werden Elemente bezeichnet, die nicht vollständig einer Nutzeranforderung entsprechen. Demgegenüber steht der Bereich der korrekt modellierten Anforderungen.

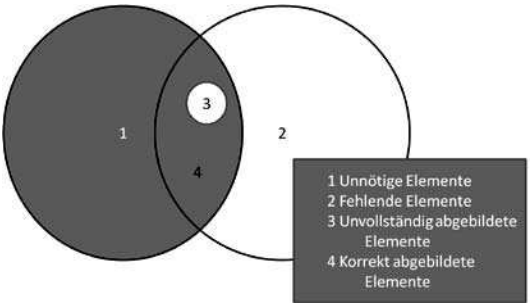


Abbildung 6: Vollständigkeitskriterium nach Moody [Mo98]

Ziel bei der Modellierung einer Anwendungsdomäne ist es, Fehler erster, zweiter und dritter Art zu vermeiden. In einem domänenspezifischen Datenstandard können in einer offenen Infrastruktur die Fehler erster nicht Art vermieden werden, da durch die verschiedenen Teilnehmer eine Vielzahl von Anforderungen an den Datenstandards gestellt werden, die aber nicht für alle Teilnehmer relevant sind. Für die Entwicklung eines domänenspezifischen Datenstandards mit PODSL bedeutet dies, dass für alle wichtigen Prozesse der Anwendungsdomäne die Möglichkeit der Datenspeicherung in einem PED besteht. Dementsprechend ist es für die Erstellung eines Datenstandards mit PODSL von entscheidender Bedeutung, die zentralen Prozesse der Anwendungsdomäne zu kennen und im Datenmodell abzubilden.

## 3.3 Flexibilität

Flexibilität hat die Anpassung von Schemata an neue Anforderungen zum Ziel und ist ein wesentliches Element einer Modellierungssprache [CSW08]. In einer offenen Infrastruktur treten im Laufe der Zeit immer wieder neue Anforderungen an die Vollständigkeit eines Datenschemas auf, die in dieses integriert werden müssen. Zusätzlich können auch bei einer umfangreichen Analyse der Anwendungsdomäne nicht immer alle Prozesse direkt in einem Datenstandard unterstützt werden. Dementsprechend ist die Flexibilität und Erweiterbarkeit eines mit PODSL erstellten domänenspezifischen Datenstandards von entscheidender Bedeutung. Dabei kann die Möglichkeit zur Erweiterung des Datenstandards sowohl zentral zur Abbildung der Änderung der Anforderungen über die Zeit als auch zur Spezifikation von lokalen Erweiterungen verwendet werden. Insbesondere der letzte Punkt ist in einer offenen Infrastruktur ein zentrales Kriterium. Darüber hinaus müssen diese Erweiterungen zu vorhergehenden Versionen eines Datenstandards kompatibel sein. Diese Anforderung kann durch die Einbettung in eine Metastruktur, die Vererbung unterstützt, mit der Möglichkeit der Spezialisierung erreicht werden. So ist definiert, wie diese neuen Elemente im Kontext des Metamodells zu interpretieren sind. Ausgangspunkt für eine Modellerweiterung sind ausschließlich bestehende Modellelemente, die als Basiskonzept für neue Konzepte dienen.

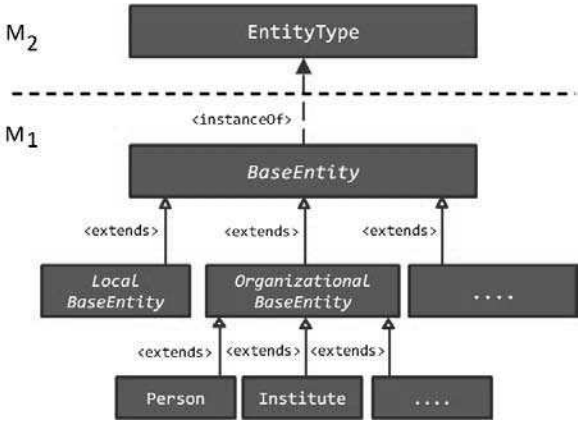


Abbildung 7: Ableitung neuer Entitäten aus BaseEntity

In PODSL wird die Flexibilität durch die Erweiterung der Konzepte auf der M1-Ebene (siehe Abschnitt 4.2) realisiert. Dazu existiert für alle Modellelemente bei der Modellierung mit PODSL auf der M1-Ebene ein Basiselement, von dem alle anderen Elemente abgeleitet werden. Die Erstellung eines domänenspezifischen Datenstandards beruht damit auf der Spezialisierung bereits bekannter Elemente auf der M1-Ebene. In Abbildung 7 ist dies am Beispiel für Entitäten dargestellt. Alle Entitäten werden von der Basisklasse ‚BaseEntity' abgeleitet und verfügen somit über alle wesentlichen Eigenschaften. Zusätzlich werden bereits an dieser Stelle die Dokumentenaspekte dahingehend berücksichtigt, dass Entitäten den Dokumentenaspekten eindeutig zugeordnet werden. Dies löst das Kompatibilitätsproblem bei Modellerweiterungen.

Wenn in einer offenen Infrastruktur ein Teilnehmer eine lokale Erweiterung vornimmt, kann beim Datenaustausch stets auf eine Basisklasse zurückgegriffen werden. Darüber hinaus ist durch die Ableitung aus existierenden Strukturen für einen Teilnehmer der Infrastruktur die lokale Erweiterung eines anderen Teilnehmers leicht integrierbar.

## 3.4 Data Provenance

Unter Data Provenance sind alle Informationen zu verstehen, welche die Historie eines Datensatzes (beliebiger Technologie) beginnend bei der Originalquelle erfassen [SPG05]. Demnach wird mit Data Provenance nicht nur die Herkunft eines Datensatzes erfasst, sondern auch alle Transformationen, die ein Datensatz durchläuft. Data Provenance ist ein unverzichtbares Mittel für die Identifikation von Datensätzen und bei der Identifikation von Fehlern durch Datentransformationen oder bei der Datenintegration. Die Unterstützung von Data Provenance in einem domänenspezifischen Datenstandard hat das Ziel, Strukturen zur Verwaltung von Herkunft, Versionen und Veränderungen an Datensätzen zu erfassen.

Dazu muss ein Datensatz eindeutig identifiziert werden können. Da domänenspezifische Datenstandards mit PODSL zum Datenaustausch in einer offenen Infrastruktur verwendet werden sollen, ist es erforderlich, Datensätze auf globaler Ebene zu identifizieren. Dazu werden Elemente der M2 und M1-Ebene über Identifier in der für OMME üblichen Form [Vo11]

```
model:/repository/Modell/Konzeptname
```

referenziert. Datensätze auf M0-Ebene müssen global eindeutig referenzierbar sein und werden durch Identifier der Form

```
repository/ObjectID
```

identifiziert, wobei für die ObjectID ein Universally Unique Identifier (UUID) verwendet wird. Die Referenzierbarkeit von Elementen reicht in einer offenen Infrastruktur allerdings nicht aus, um Data Provenance zu realisieren. Zusätzlich müssen die Urheberschaft der Daten und alle Transformationen von Datensätzen dokumentiert werden.

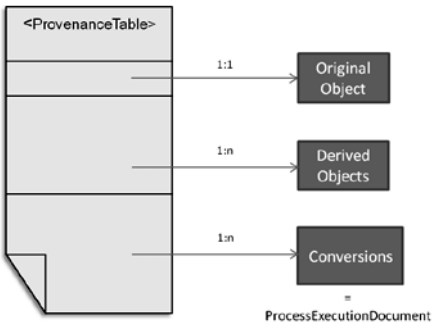

Abbildung 8: Struktur der ProvenanceTable

Dazu wird in PODSL mit der ProvenanceTable ein Konzept zur Speicherung von Provenance-Informationen in Form einer speziellen Entität geschaffen [Sc13]. Diese ist auf der M1-Ebene des Metamodells angesiedelt. Die Struktur der ProvenanceTable ist in Abbildung 8 dargestellt. Die ProvenanceTable enthält eine Referenz auf den Originaldatensatz und auf alle Datensätze, die aus diesem erzeugt wurden. Transformationen aus diesem Datensatz werden als Prozesse aufgefasst und über PED´s dokumentiert. Somit ist in der ProvenanceTable der Prozess der Datentransformation an sich dokumentiert. Diese PED´s werden im Bereich Conversions in der ProvenanceTable referenziert. Somit kann über die ProvenanceTable in PODSL die Herkunft eines Datensatzes und alle Veränderungen lückenlos dokumentiert werden.

# 4 Metastruktur von PODSL

Die Einbettung von PODSL in eine Metastruktur erfolgt über drei Ebenen (Abbildung 9). Für Elemente der Metastruktur wird gemäß dem Sprachgebrauch in OMME der Begriff Konzept verwendet [Vo11]. Dabei enthält die M2-Ebene die grundlegenden Konzepte der Modellierungssprache (Metaebene). M1 ist die Ebene der Modelle. Auf der M1-Ebene wird zwischen einem generischen Teil und einem domänenspezifischen Teil unterschieden. Der generische Teil wird als M1-Core bezeichnet. Domänenspezifische Erweiterungen werden aus M1-Core abgeleitet und mit PODSL-[Domäne] benannt. Auf der M0-Ebene wird die M1-Ebene durch konkrete Instanzen besiedelt. Die Besiedelung der M0-Ebene erfolgt auf Basis einer domänenspezifischen Erweiterung wie z.B. PODSL-Biodiv für die Biodiversitätsinformatik.

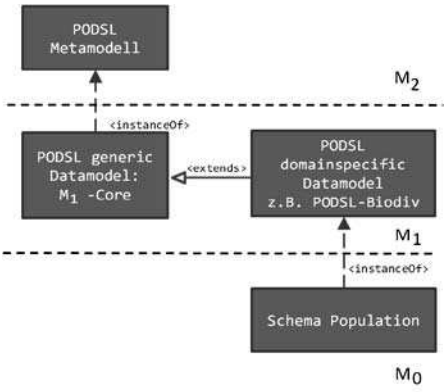


Abbildung 9: Metastruktur von PODSL [Sc13]

## 4.1 M2-Ebene

Auf M2 wird eine allgemeine Modellierungssprache zur Erstellung von Modellen mit PODSL spezifiziert. In dieser wird festgelegt, auf welche Weise in der M1-Ebene

modelliert werden kann. Die zentralen Konzepte der M2-Ebene von PODSL sind in Abbildung 10 dargestellt. Die Modellierung der Dokumentenaspekte ist dabei auf der Ebene der Relation angesiedelt, in welchen Entitäten mit anderen Entitäten in einem bestimmten Dokumentenaspekt in Beziehung gesetzt werden.

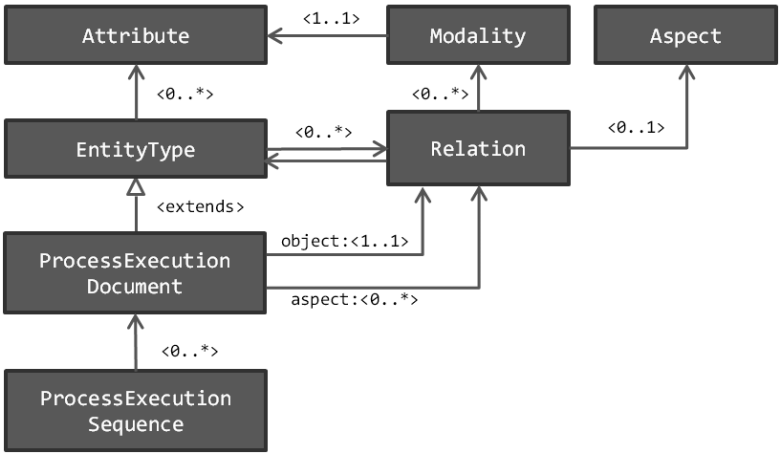


Abbildung 10: Schema der M2-Ebene von PODSL [Sc13]

## 4.2 M1-Ebene

Bei der M1-Ebene von PODSL wird zwischen der generischen Komponente M1-Core und den domänenspezifischen Erweiterungen PODSL-[Domäne] unterschieden. Im M1-Core Modell von PODSL werden Konzepte spezifiziert, welche in verschiedenen Domänen verwendet werden können. Dies sind z.B. Konzepte wie BaseEntity und BaseProcessExecutionDocument und davon abgeleitete generische Konzepte. So ist z.B. die ProvenanceTable in M1-Core über Zwischenschritte von BaseEntity abgeleitet. Darüber hinaus sind grundlegende Konzepte wie Person in M1-Core spezifiziert. Auf Basis der Konzepte von M1-Core werden die domänenspezifischen Erweiterungen durch Spezialisierung der Konzepte aus M1-Core gebildet. Die weitere Spezialisierung erfolgt auf Basis dieser domänenspezifischen Konzepte.

## 4.3 M0-Ebene

Auf der M0-Ebene werden konkrete Datensätze als Instanzen der Entitäten der domänenspezifischen Datenstandards auf M1-Ebene gebildet. Die M1-Ebene dient als eine logische Struktur für die Anwendungsdomäne. Daten sollen möglichst leicht in andere Formate konvertiert und integriert werden können. Die Persistenz der Daten findet in der Praxis in Datenbanken, XML, Strukturen der objektorientierten Programmierung und Ontologien statt.

# 5 Domänenspezifische Erweiterungen von M1-Core

Für die Entwicklung einer domänenspezifischen Erweiterung müssen die Prozesse der Anwendungsdomäne analysiert werden. Dies konnte im Rahmen des IBF-Projektes für die Domäne der Biodiversitätsinformatik erfolgen. Die Modellierung der wichtigsten Prozesse der Biodiversitätsinformatik wurde in [Sc13] vorgenommen. Auf Grundlage dieser Prozesse wurden Entitäten und PED´s aus M1-Core abgeleitet. Kennzeichnend für die Domäne der Biodiversitätsinformatik sind Spezialisierungen von Entitäten zur Beschreibung von Orten, Personen, Taxa und Messungen an biologischen Objekten, wie auch in Abbildung 2 am „Prozess der Geländekartierung“ zu erkennen ist. Das korrespondierende PED in PODSL-Biodiv ist nachfolgend aufgelistet:

```
concept BaseMonitoring extends BaseProcessExecutionDocument{
        identifier=BaseMonitoringIdentifier;
        object=ObservationObjectRelation;
        relation= ScientistRelation,ExecutionTimeRelation,ExecutionLocalityRelation}
```

Durch die Ableitung von „BaseProcessExecutionDocumnent“ erbt das PED „BaseMonitoring“ grundlegende Eigenschaften von PED`s, die z.B. DataProvenance ermöglichen. Als „object“ wird eine Relation vorgeschrieben, welche eine Beziehung zum Kartierungsobjekt herstellt. Die Subprozesse sind über weitere Relationen abgebildet, die Aspekten zugeordnet sind, die den Perspektiven des Kartierungsprozesses entsprechen. Für den Prozessverantwortlichen sieht die referenzierte Relation folgendermaßen aus:

```
concept ScientistRelation extends ResponsibleRelation{
        role="Scientist as a Gatherer in a process";
        aspect=OrganizationalAspect;
        identifier=ResponsibleGathererRelationIdentifier;
        target=Scientist;}
```

Dabei kann im PED spezifiziert werden, welche Anforderungen an die Genauigkeit eine Entität erfüllen muss. So ist für die Geländekartierung nicht jede Person automatisch qualifiziert, sondern nur Wissenschaftler. In PODSL-Biodiv gibt es dementsprechend ein Konzept „Scientist“, das von „Person“ abgeleitet wird.

Auf Basis dieser Erkenntnisse wurde PODSL-Biodiv als domänenspezifischer Datenstandard für die Biodiversitätsinformatik entwickelt und im IBF-Projekt umfangreich getestet. Es konnten alle im IBF-Projekt identifizierten Prozesse in PODSL-Biodiv erfasst werden. Zusätzlich umfasst PODSL-Biodiv den in der Biodiversität etablierten Datenstandard DarwinCore (DwC). Damit erfasst PODSL-Biodiv die Anforderungen des IBF-Projektes vollständig und zusätzliche weitere typische Prozesse der Biodiversitätsinformatik, so dass PODSL-Biodiv zur Anwendung in Projekten im Biodiversitätsbereich gut geeignet ist. Sollten zusätzliche Anforderungen auftreten werden durch Spezialisierung aus bekannten Konzepten die vollständige Erfassung wieder hergestellt. Über PODSL-Biodiv wird damit für die Domäne der Biodiversität über eine prozessorientierte Sichtweise ein umfangreiches Datenmodell zur Verfügung gestellt.

Ein weitere Anwendungsdomäne für PODSL sind sie Prozesse in Krankenhäuser, die in [FJS07] erhoben werden konnten. Auf Basis dieser Prozesse können im organisatorischen Dokumentenaspekt wichtige Rollen in dieser Domäne wie Arzt, Pfleger, Verwaltungsangestellter und Patient identifiziert und weiter spezialisiert werden. Domänenspezifische Prozesse sind in dieser Domäne z.B. die Aufnahme und Entlassung eines Patienten, die Untersuchung eines Patienten mit Subprozessen wie Anamnese. Die Datenmodellierung mit PODSL wird in einem aktuellen Projekt des Lehrstuhls mit dem Klinikum Bayreuth intensiv getestet.

# 6 Ausblick

Im folgenden Abschnitt wird beschrieben, wie mit PODSL erstellt Datenstandards in der Praxis genutzt werden können. Dies ist zum einen die Erstellung von Mappings beim Datenaustausch – zum anderen die Verwendung in Softwareprodukten zur individuellen Anpassung.

Hintergrund der Entwicklung von domänenspezifischen Datenstandards mit PODSL ist ihre Anwendung in Infrastrukturen zum Datenaustausch. Dabei weisen mit PODSL entwickelte domänenspezifische Datenschemata aufgrund ihrer Flexibilität erhebliche Vorteile gegenüber anderen Datenmodellen auf. Zusätzlich wird in PODSL Data Provenance bereits direkt im Datenschema berücksichtigt. Dies ist in der Domäne der Biodiversitätsinformatik besonders wichtig, da hier heterogene Daten über globale Infrastrukturen wie z.B. dem GBIF oder dem LTER-Netzwerk ausgetauscht werden.

Ein weiterer Vorteil der Anwendung von PODSL ist die Möglichkeit einer besseren Nutzeranpassung in Softwareprodukten. Durch die Ableitung aus Basisklassen mit PODSL können lokale Erweiterungen besser in Oberflächen eingebunden werden. Zusätzlich können durch (semi-)automatische Softwareentwicklung grafische Oberflächen erzeugt werden, die direkt aus dem Datenmodell abgeleitet werden. Dazu wird für eine Anwendung eine Reihe von protypischen Oberflächen zu Verfügung gestellt und diese durch die Datenstruktur von PODSL angepasst. Dies ermöglicht die einfache Erzeugung von nutzerspezifischen Oberflächen und Anwendungen.

# 7 Fazit

Die Prozesse einer Anwendungsdomäne enthalten bereits alle Anforderungen an einen domänenspezifischen Datenstandard. Diese bilden die Grundlage für die Datenmodellierung. Mit der aspektorientierten Datenmodellierung wurde eine Methode eingeführt, mit der die Erstellung von Schemata für Datenstandards auf Basis von Prozessen erfolgt. Die Nutzung eines domänenspezifischen Datenstandards ist mit hohen Anforderungen an die Technologieunabhängigkeit, Vollständigkeit, Flexibilität und Data Provenance verbunden. Mit PODSL wurde eine Methode zur Entwicklung domänenspezifischer Datenstandards eingeführt, die diesen Anforderungen gerecht wird. In der praktischen Anwendung werden mit PODSL entwickelte Datenstandards beim

Datenaustausch in einer offenen Infrastruktur und als Grundlage für die Entwicklung von Anwendungen genutzt. Dabei ist PODSL für die individuelle Anpassung von Softwareprodukten ein hervorragendes Werkzeug, da sich Oberflächen direkt aus der Datenstruktur generieren lassen. Durch mit PODSL erstellte Datenstandards können dementsprechend die Herausforderungen einer offenen Infrastruktur gelöst werden.

# 8 Literaturverzeichnis


[Bu98] Bussler, C., Organisationsverwaltung in Workflow-Management-Systemen. 1998: Dt. Univ.-Verlag.

[CSW08] Clark, T., P. Sammut, and J. Willans, Applied metamodelling: a foundation for language driven development. 2008.

[FJS07] Faerber, M., S. Jablonski, and T. Schneider. A Comprehensive Modeling Language for Clinical Processes. In ECEH. 2007.

[Gr11] Group, O.M. Meta Object Facility (MOF) Core Specification Version 2.4.1. OMG Available Specification 2011 [cited 2013 10-October-2013]; Available from: http://www.omg.org/spec/MOF/2.4.1/.

[HL13] International, H.L.S. Introduction to HL7 Standards. 2013 12-October-2013]; Available from: http://www.hl7.org/implement/standards/index.cfm?ref=nav.

[Ja95] Jablonski, S. Functional and behavioral aspects of process modeling in Workflow Management Systems. in Proceedings of the ninth Austrian-informatics conference on Workflow management: challenges, paradigms and products: challenges, paradigms and products. 1995: R. Oldenbourg Verlag GmbH.

[JB96] Jablonski, S. and C. Bussler, Workflow management: modeling concepts, architecture and implementation. 1996.

[Le02] Lenzerini, M. Data integration: A theoretical perspective. in Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2002: ACM.

[Ma07] Madin, J., et al., An ontology for describing and synthesizing ecological observation data. Ecological informatics, 2007. 2(3): p. 279-296.

[Mo98] Moody, D.L., Metrics for evaluating the quality of entity relationship models, in Conceptual Modeling–ER'98. 1998, Springer. p. 211-225.

[Mo05] Moody, D.L., Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. Data & Knowledge Engineering, 2005. 55(3): p. 243-276.

[MS94] Moody, D.L. and G.G. Shanks, What Makes a Good Data Model? Evaluating the Quality of Entity Relationship Models, in Proceedings of the13th International Conference on the Entity-Relationship Approach. 1994, Springer-Verlag. p. 94-111.

[Sc13] Schneider, T., Domänenspezifische Evaluation und Optimierung von Datenstandards und Infrastrukturen. 2013, Dissertation, University of Bayreuth.

[SPG05] Simmhan, Y.L., B. Plale, and D. Gannon, A survey of data provenance techniques. Computer Science Department, Indiana University, Bloomington IN, 2005. 47405.

[TD09] TDWG. TDWG Standards. 2009 12-October-2013]; Available from: http://www.tdwg.org/standards/.

[THR12] Triebel, D., G. Hagedorn, and G. Rambold, An appraisal of megascience platforms for biodiversity information. MycoKeys, 2012. 5: p. 45-63.

[Vo11] Volz, B.W., Werkzeugunterstützung für methodenneutrale Metamodellierung. 2011, Dissertation, University of Bayreuth.