Normierung und Standardisierung von Provenance-Modellen

PubFlow Workshop (23.03.2012, Kiel)

Andreas Schreiber <Andreas.Schreiber@dlr.de>
Deutsches Zentrum für Luft- und Raumfahrt e.V.

Überblick

- Einleitung
- Provenance-Modelle
- Methodik
- Speicherung von Provenance-Daten
- Nationale Standardisierung
- Ausblick

Historie im Bereich Provenance

- EU-Projekt Grid Provenance (2004-2006)
  http://www.gridprovenance.org/

- BMBF D-Grid-Projekt AeroGrid (2007-2010)
  http://www.aero-grid.de

- Anwendungen im DLR
  - Verteilte Simulationen
  - Elektronisches Laborbuch
  - Medizinische Studien
  - Entwurfsprozesse von Flugzeugen
  - Softwareentwicklungsprozesse

Deutsches Zentrum für Luft- und Raumfahrt e.V.
in der Helmholtz-Gemeinschaft

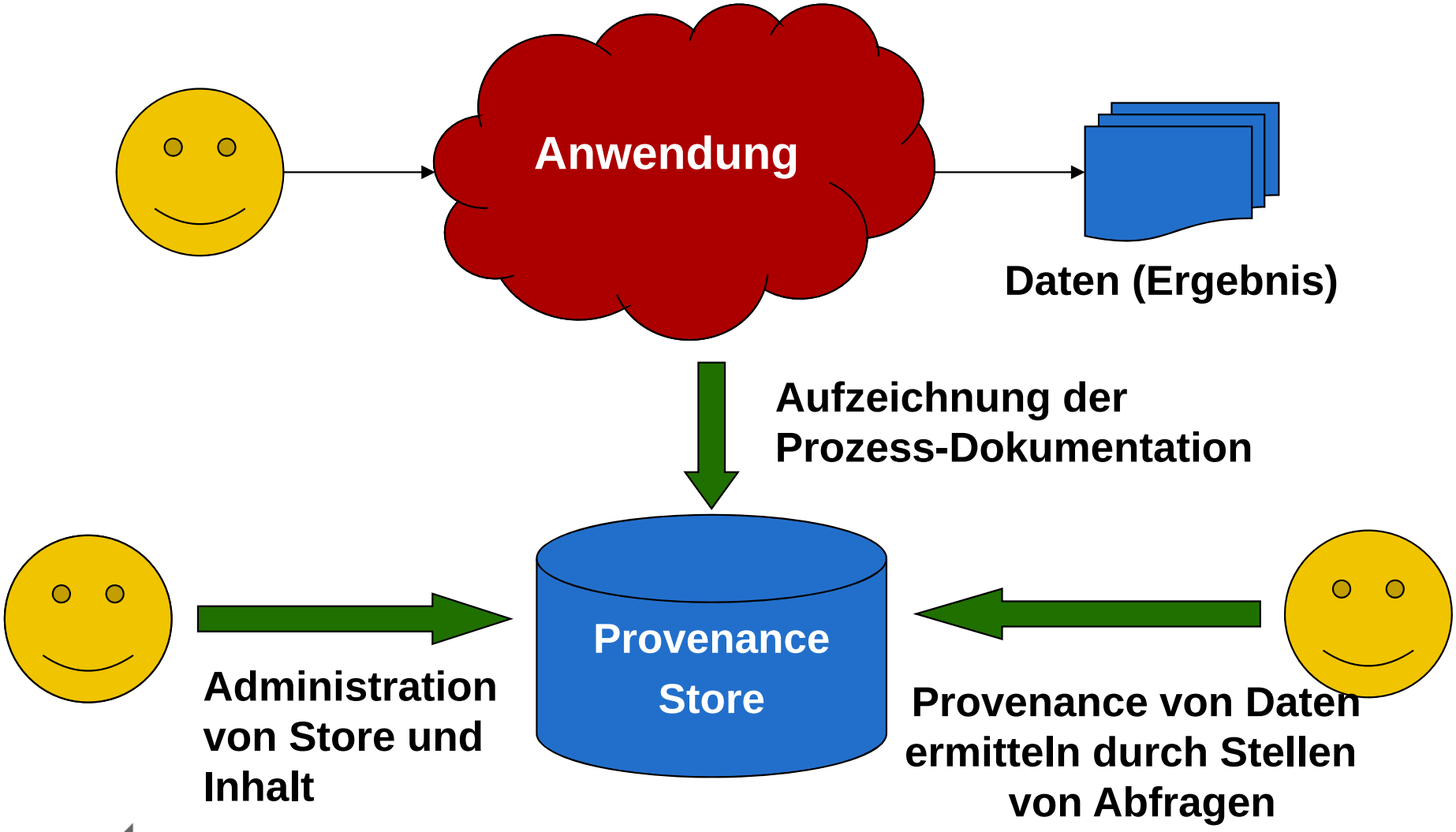
Einleitung

Provenance-Beispiel

Die Provenance einer Flasche Wein umfasst beispielsweise

- die Trauben, aus denen er hergestellt wurde,
- den Ort, wo die Trauben gewachsen sind,
- der Prozess der Weinherstellung,
- die Art, wie der Wein gelagert wurde,
- die Beteiligten, zwischen denen der Wein transportiert wurde (z.B. erst vom Hersteller zum Großhändler, dann zum Händler) und
- das Auktionshaus, das den Wein versteigert hat.

Provenance Life Cycle



Bausteine der Provenance-Welt

Provenance-Modell

- Modellierung der Provenance-Daten

Methodik

- Einführen von Provenance für Prozesse

Software-Infrastruktur

- Speicherung und Abfrage der Provenance-Daten

Provenance-Modelle

Standards für Provenance-Modelle

- **Open Provenance Model (OPM)**
  - Offenes und interoperables Modell für Provenance-Daten
  - Entwickelt seit 2006
  - Version 1.0 der Spezifikation im Dezember 2007
  - Version 1.1 im Juni 2009
  - Informationen: http://openprovenance.org

- **W3C Provenance Data Model (PROV-DM)**
  - Entwickelt seit Juni 2011 durch W3C Provenance Working Group
  - Bisher nur als Draft
  - Informationen: http://www.w3.org/2011/prov

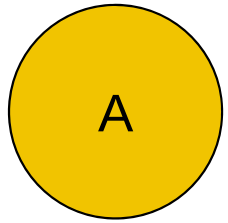
Open Provenance Model (OPM)

- Erlaubt, das Zustandekommen von Dingen zu beschreiben
- Ermöglicht eine **Prozessorientierte** und eine **Datenflußorientierte** Sicht
- Basiert auf der Vorstellung eines **annotierten Kausalitätsgraphen** (gerichteter azyklischer Graph, DAG)

Open Provenance Model

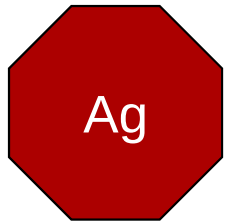
Open Provenance Model
Nodes

Artefakt *(Artifact)*
- Unveränderlicher Zustand
- Kann eine physikalische Verkörperung in Form eines physikalischen Objekts haben oder eine digitale Repräsentation in einem Computer sein



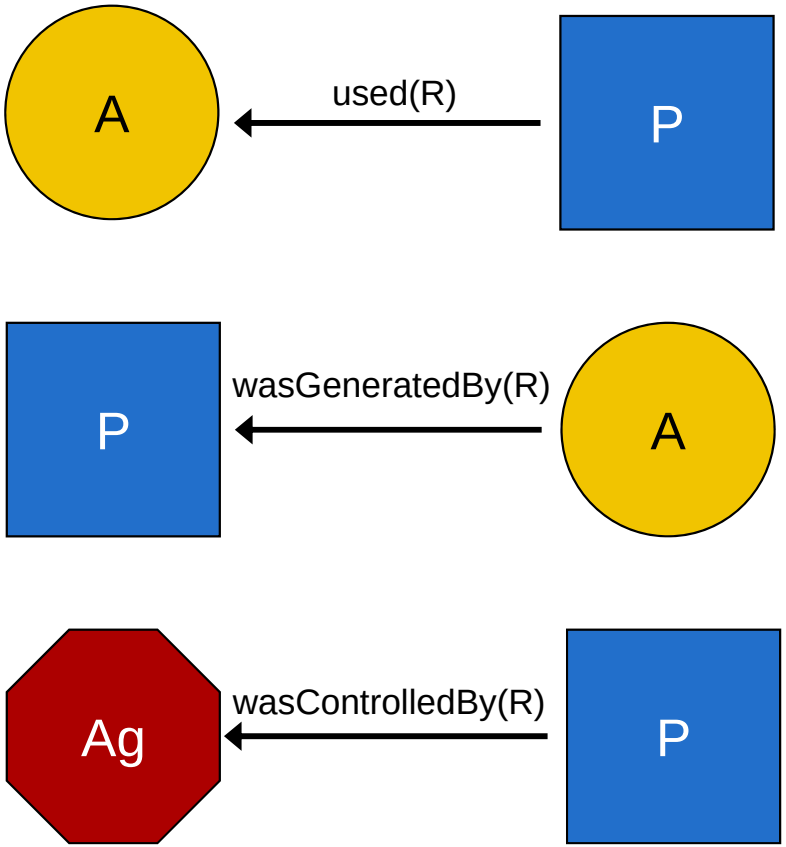
Prozess *(Process)*
- Eine Aktion oder Serie von Aktionen ausgeführt auf oder verursacht durch Artefakte
- Resultiert in neuen Artefakten

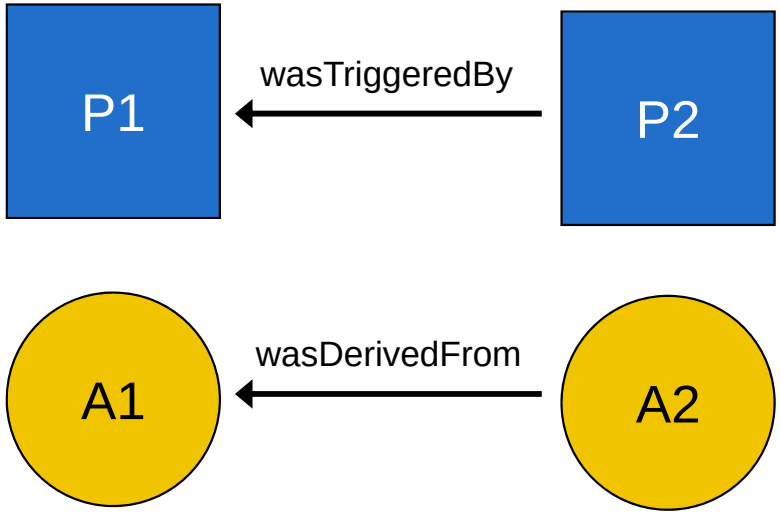
Agent
- Kontextabhängige Instanz, die als Katalysator für den Prozess wirkt
- Ermöglicht, erleichtert oder kontrolliert die Ausführung



Open Provenance Model
Edges

Edges können Rollen haben „(R)“ als textuelle Beschreibung.





Edges werden in der Vergangenheitsform bezeichnet, um zu verdeutlichen, dass es sich um einen vergangenen Prozessschritt handelt.

Open Provenance Model
Annotations

- Hinzufügen weiterer Informationen zum Graphen
- Annotiert werden können
    - der ganze Graph,
    - einzelne Nodes,
    - Edges und
    - Annotationen.

- Annotationen sind eine Liste von Key-Value-Paaren

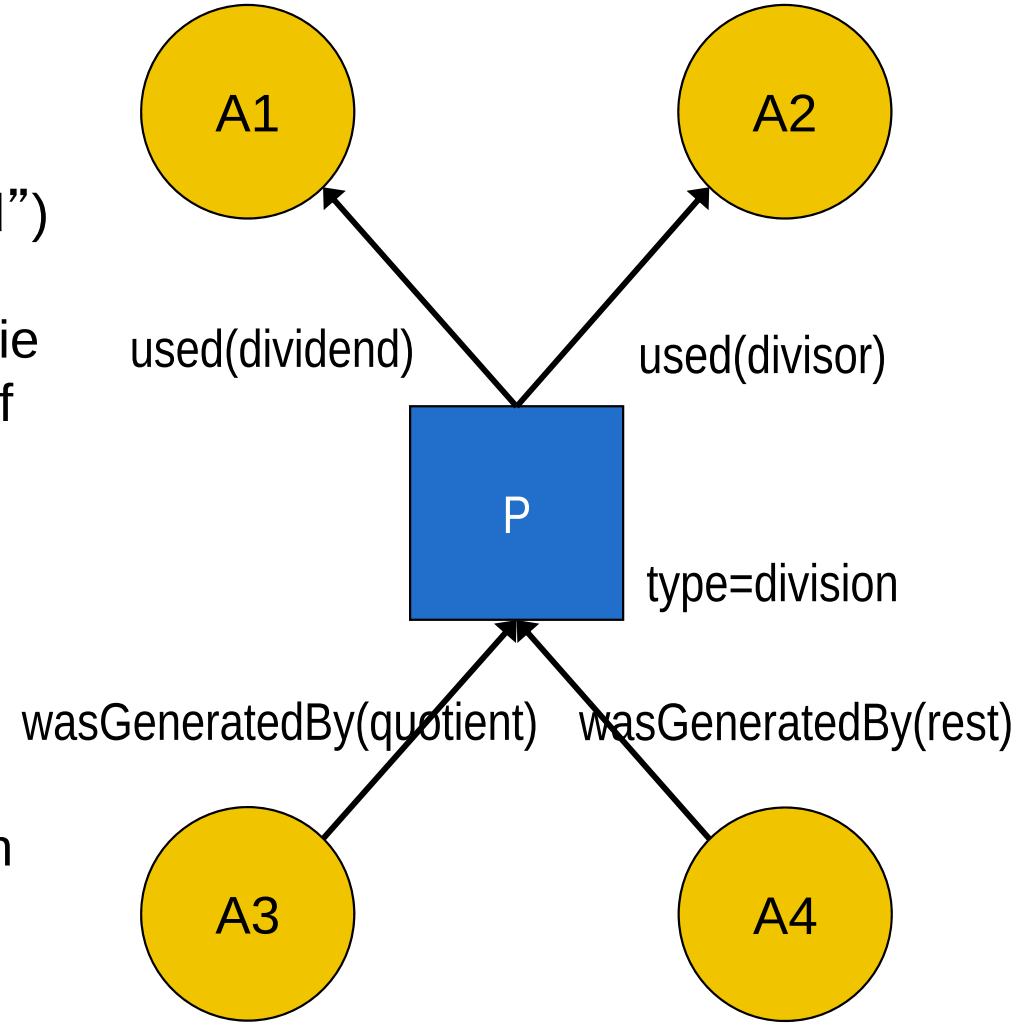
Open Provenance Model

Semantik

- Ein Prozess benutzt (“used”) Artefakte und generiert (“generated”) Artefakte
- Die Rollen der Edges bezeichnen die Funktion der Artefakte im Bezug auf den Prozess
- Edges und Nodes können typisiert sein

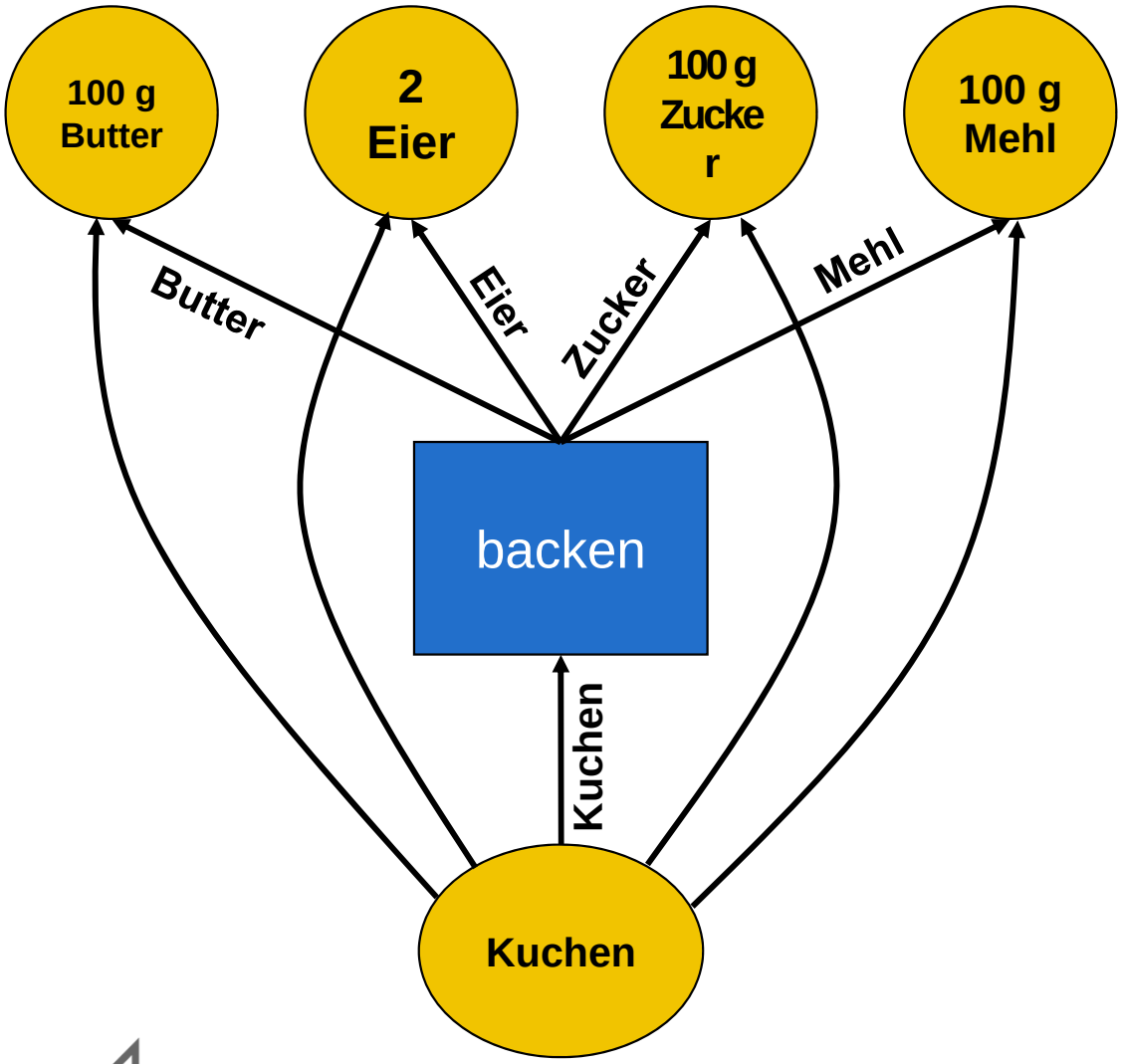
Kausalkette

- P wurde verursacht durch A1 und A2
- A3 und A4 wurden verursacht durch P



Open Provenance Model

Kuchen backen



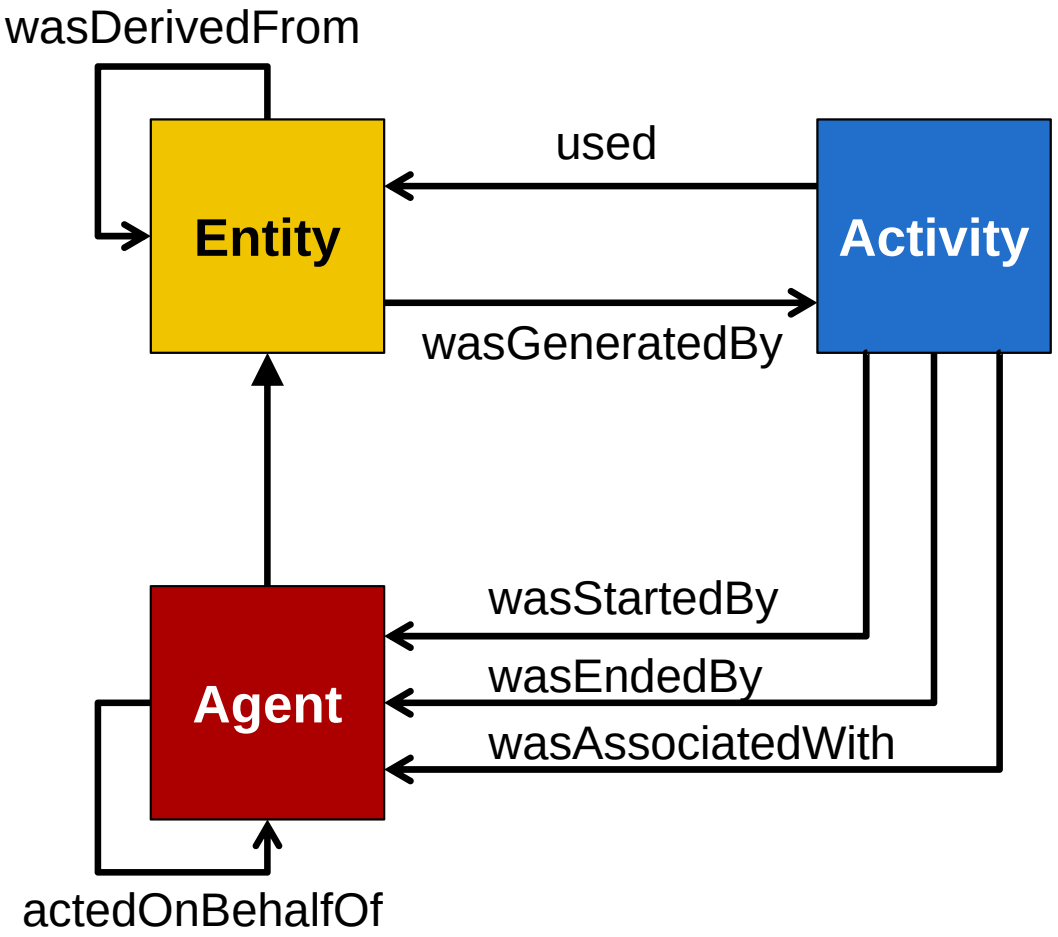
W3C Provenance Data Model (PROV-DM)

Konzepte

Nodes
- Entity
- Activity
- Agent

Für Entities und Agents zusätzlich
- plan
- collection

Edges
- association
- responsibility



Methodik

Vorgehensweise

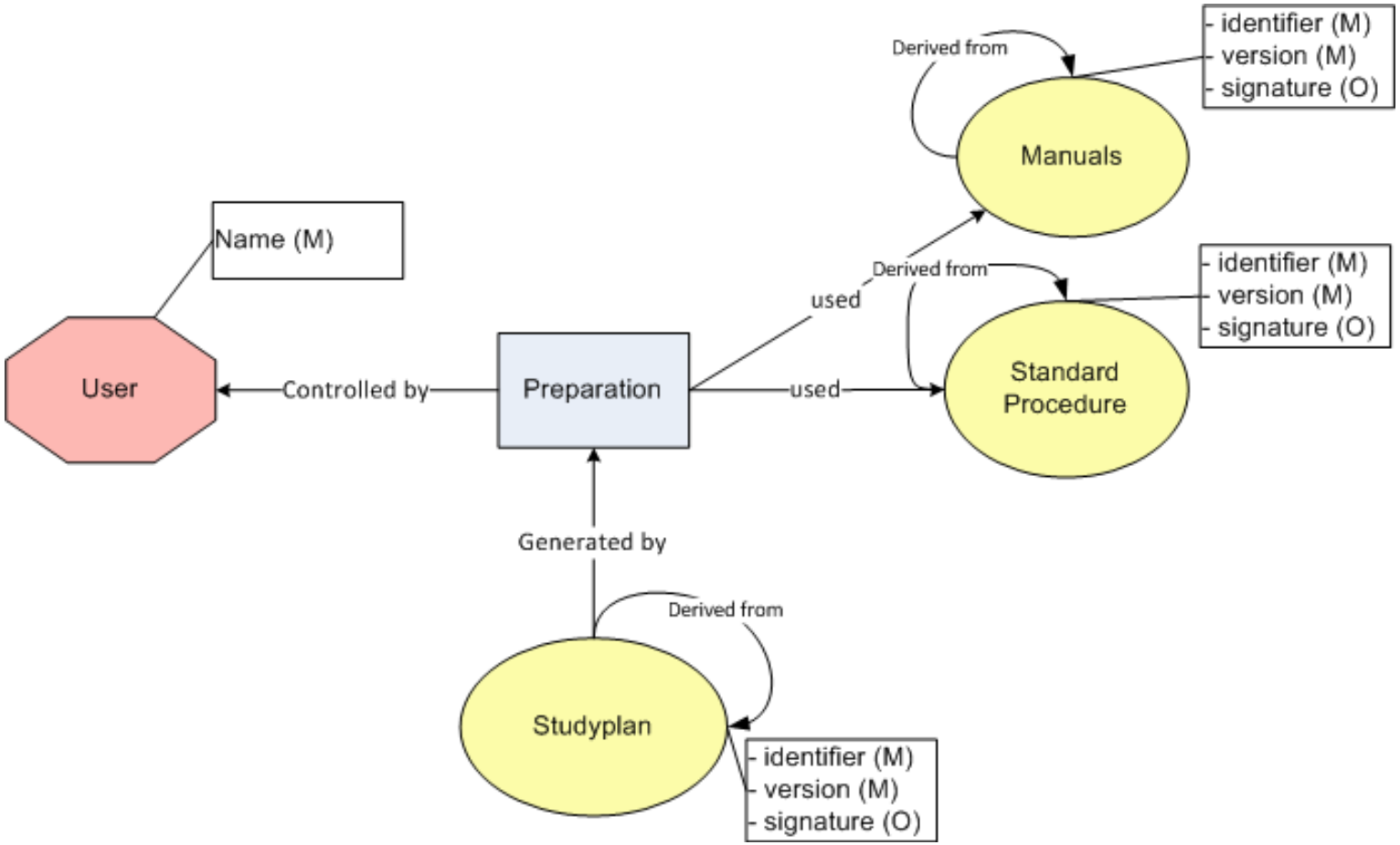
1. Sammeln von Fragen, welche beantwortet werden sollen
   - *Wer ist verantwortlich für die Implementierung von X?*
   - *Welches Element ist der logische Vorgänger von Element X?*

2. Identifikation der Aktoren, des Input und des Output für die Fragen.

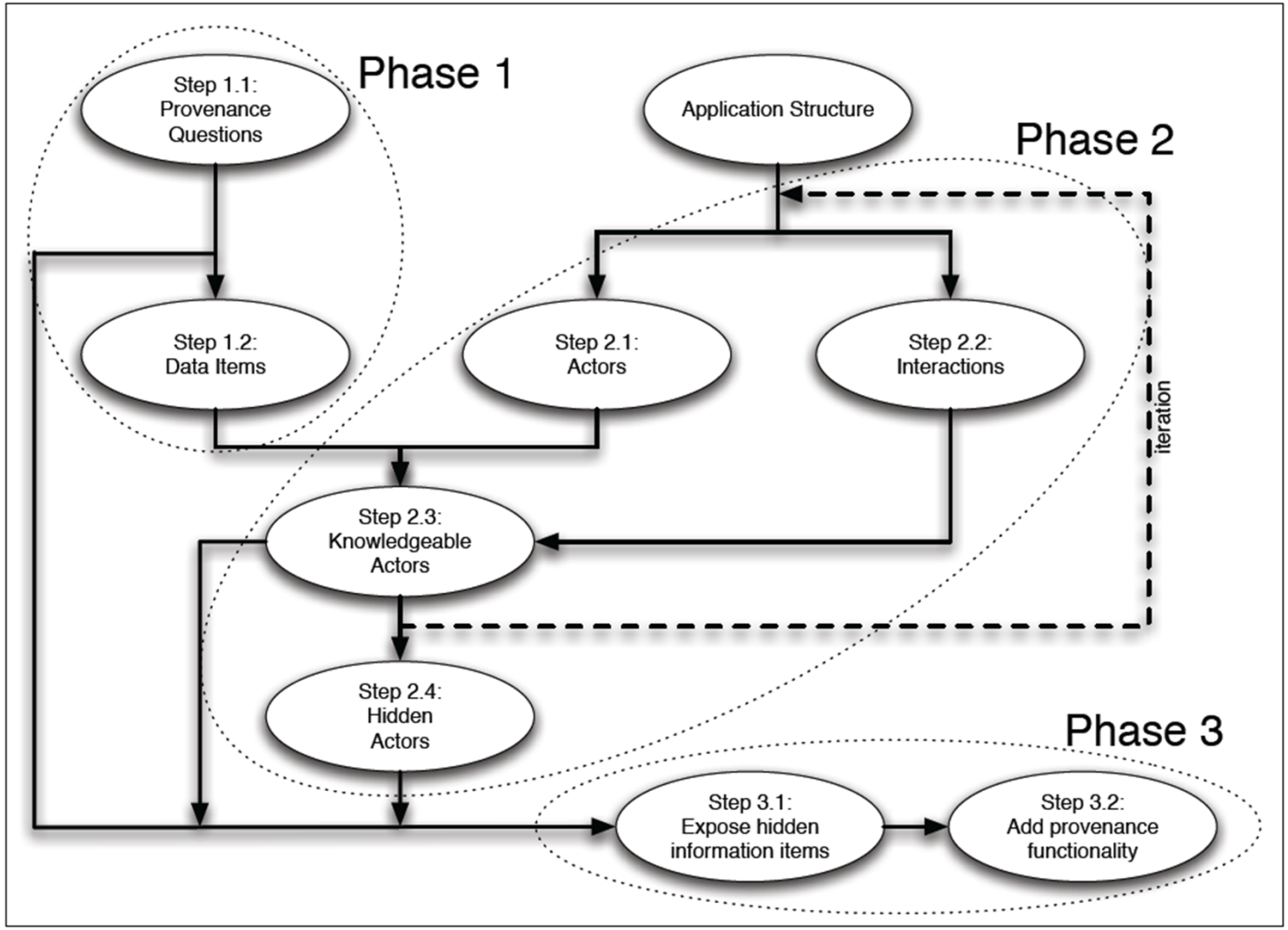
3. Ermittlung der beteiligten Prozesse

4. Entwicklung eines Provenance-Modells für die verschiedenen Prozesse

Beispielmodell
Laborbuch für Studien



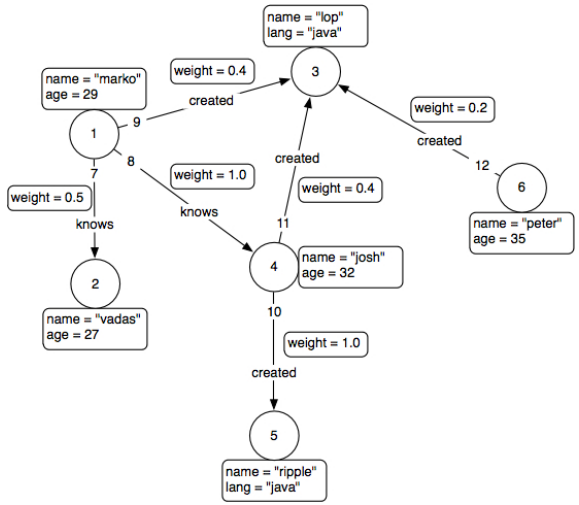
Methodik „PrIMe“
Anwendungen „Provenance-Aware“ machen



Software-Infrastruktur

Speicherung und Abfrage von Provenance-Daten

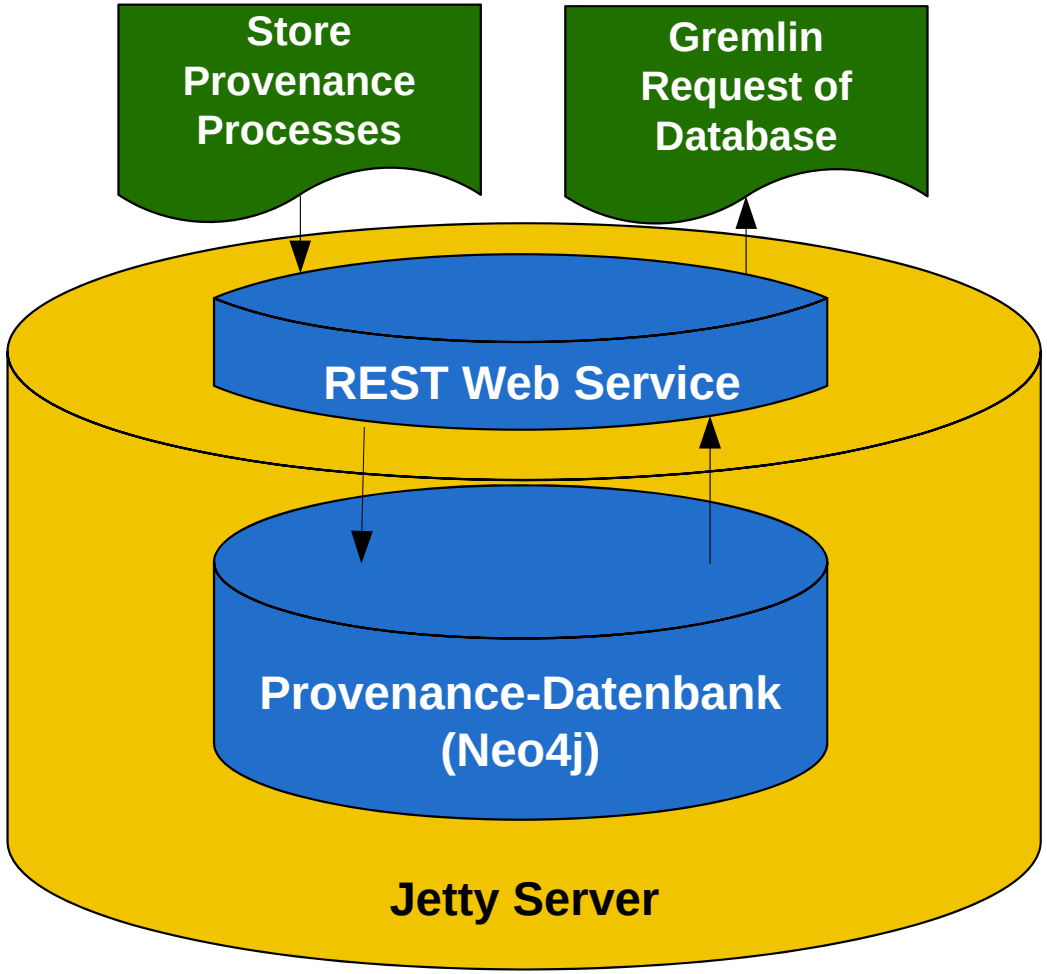
- Verschiedene Speichertechnologien möglich
  - Relationale DB, XML, RDF, SPARQL, …
- Zum Speichern der Provenance Graphen bieten sich Graph-Datenbanken an



- Implementierung für das Open Provenance Model: prOOst
  - Open Source (Apache 2 License)
  - Informationen: http://software.DLR.de/p/proost
  - Graph-basierte Datenbank Neo4j
  - Graph-Abfragesprache Gremlin
  - REST API

Provenance-Service prOOst

REST-API zur einfachen Nutzung in Anwendungen



Nationale Standardisierung

Normierung und Standardisierung
Deutsches Institut für Normung e.V. (DIN)

- **Normierung**
  - Erarbeitung **konsensbasierter** Normen und Standards
  - Normungsarbeit in definierten Prozessen
  - Beteiligt sind die interessierten Kreise (Hersteller, Verbraucher, Hochschulen, Behörden, ...)
  - Bezeichnung: ***DIN <Zählnummer>***

- **Standardisierung**
  - Erarbeitung von Spezifikationen (auch „Vornorm“)
  - Keine Norm aufgrund von Vorbehalten (kein Konsens)
  - Kann in Norm überführt oder zurückgezogen werden
  - Koordinierung über Standardisierungsportal des DIN http://www.spec.din.de/
  - Bezeichnung: ***DIN SPEC <Zählnummer>***

Projekt „Standardisierung eines erweiterbaren Modells für Provenance-Daten“

Projektdaten

- Förderprogramm „Transfer von Forschungs- und Entwicklungsergebnissen (FuE) durch Normung und Standardisierung“ des Bundesministeriums für Wirtschaft und Technologie
- Beginn: Juli 2012
- Laufzeit: 2 Jahre
- Durchgeführt durch DLR

Projekt „Standardisierung eines erweiterbaren Modells für Provenance-Daten“

Projektziele

- Evaluierung und ggf. Anpassung eines internationalen Provenance-Modells (geplant: W3C PROV-DM)

- Bereitstellung eines Provenance-Modells, dass praxisrelevant, praktikabel und einsatzbereit ist

- Erstellung einer DIN SPEC auf Grundlage des W3C PROV-DM

- Mitarbeit in internationalen Gremien (hier: W3C Working Group)

Bereitstellung eines Provenance-Modells

Praxisrelevanz

- Nachweis durch Evaluation mit praxis- und industrienahen Anwendungen verschiedener Branchen
- Durchführung von Workshops mit Anwendergruppen

Praktikabilität

- Möglichst einfach und anwendungsnah gestaltete Methodik und Schnittstellen zu Provenance-Datenbanken

Einsatzbereitschaft

- Bereitstellung einer Provenance-Datenbank als frei verfügbare Open-Source-Software
- Bereitstellung eines übersichtlichen Handbuchs

Ausblick

Ausblick

- Workshops mit Anwendergruppen
  - Ab Herbst 2012
  - Zusätzlich Gespräche mit Provenance-Interessierten und relevanten Projekten

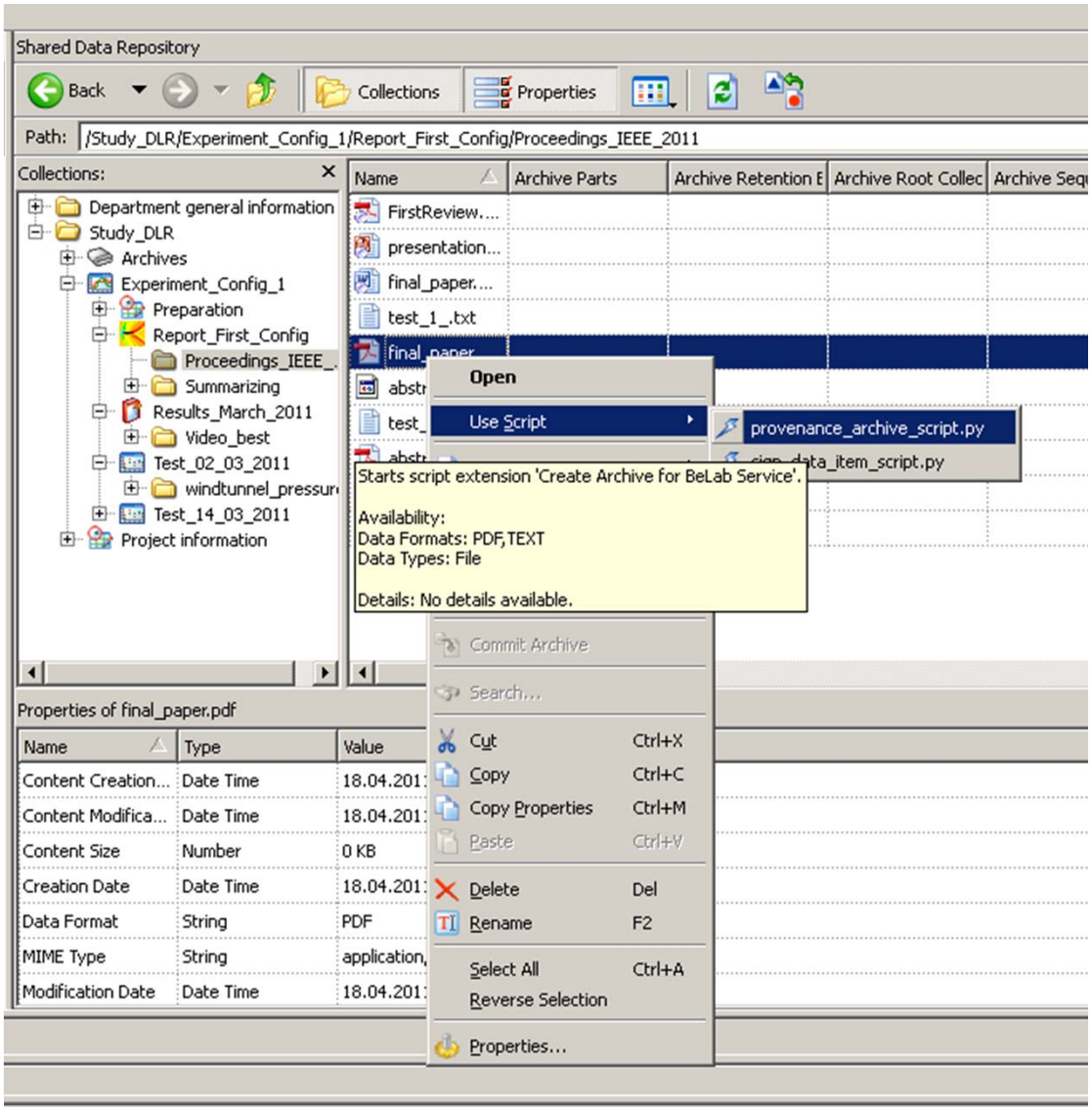
- Bereitstellung einer Provenance-Datenbank
  - Anpassung des Provenance-Datenbank prOOst an das W3C Provenance Data Model
  - Bereitstellung als Referenz-Implementierung
  - Test und Evaluation in Anwendungen

- Erstellung eines DIN SPEC Entwurfs
  - Mitte 2013 bis Mitte 2014

Elektronisches Laborbuch

Referenzanwendung für Test und Evaluation
- Open Source

Eigenschaften
- Prozessdokumentation
- Beweissicheres Archivieren
- Signieren von Daten



Fragen?

Zusammenfassung

Provenance wird W3C-Standard

DIN-Spezifikation wird erstellt

Evaluationsanwendungen gesucht

Andreas Schreiber

Andreas.Schreiber@dlr.de

http://www.dlr.de/sc