HTW Chur
Hochschule für Technik und Wirtschaft
University of Applied Sciences

Graphdatenbank als Basis für ein Linked Data -AIS

Niklaus Stettler (Alexandra Weissgerber, Bruno Wenk)

1. März 2019

Inhalt

Der Auftrag

Aufbau eines Piloten für ein auf einer Graphdatenbank basierenden AIS

- Entwicklung von Lösungen für die Steuerung der Zugriffsberechtigungen
- Entwicklung von Lösungen für die Veränderung von grossen Mengen von Datensätzen (Änderung der Zugriffsberechtigung) im AIS.
- Entwicklung von Lösungen für eine sehr differenzierte Zugriffsrechtesteuerung (Unterschiedliche Schreibrechte pro Attribut eines Knoten)

Realisiert wurde der Pilot mit der Graphdatenbank ‘GraphDB von Ontotext’

Problematik Rechteverwaltung

Spezielle Herausforderungen bei der Rechteverwaltung:

Unproblematisch ist die Rechteverwaltung pro Records.

Herausfordernd jedoch, sobald:

- Eine Entität
  - Öffentlich sichtbar ist (weil sie in einer Beziehung zu sichtbaren Entitäten steht), als auch
  - Verborgen sein soll (weil eine mit ihr verbundene Entität nicht öffentlich ist).
- Ein Attribut einer Entität
  - Von allen Schreibberechtigten bearbeitet werden darf
  - Während ein anderes Attribut nur von einem eingeschränkten Kreis bearbeitet werden darf

In diesen Situationen kann ‘Property-Gaph’ von Graphdatenbanken (Kanten in einem Graph haben Attribute) hilfreich sein.

Problematik Vielfalt von Beziehungstypen

Das konzeptionelle Modell von RiC enthält sehr viele Beziehungstypen, was die Modellierung erschwert.

‘Property-Gaph’ von Graphdatenbanken (Kanten in einem Graph haben Attribute) könnte die Zahl der Beziehungen erheblich reduzieren.

Ziele unseres Piloten auf der Basis von GraphDB

Auf Basis von GraphDB soll ein Pilot entwickelt werden, der ermöglicht:

- Managen von Sperrfristen
- Steuerung der Zugriffsrechte auf Ebene der Attribute zu Entitäten
- Massenmigration von Daten (Z.B. Freigabe von Datenbeständen nach Ablauf der Sperrfrist)
- Nutzung von Normdateien zur Integration von Wissensbeständen von externen Quellen

Abgrenzung

Für den Pilot war ein Datenmodell zu entwickeln, das die zu testenden Funktionalitäten ermöglicht.

Dieses Datenmodell sollte sich am konzeptionellen Modell von RiC orientieren.
Für unsere Zwecke vereinfachen wir wesentlich. Insbesondere nicht umgesetzt sind:

- Entitäten zur Beschreibung der Funktionen (Prozesse)
- Differenzierungen betr. der Zugriffsberechtigungs-Einschränkungen

Inhalt

1. Der Auftrag
2. Lösungsansatz

Die Daten

- Vollständiger Medatadatendump aus dem AIS des BAR

  Die Daten des BAR haben einen Umfang, der mit den uns zur Verfügung stehenden Mitteln nicht ohne Probleme verarbeitet werden konnte. Da unser Pilot lediglich zur Klärung einiger konzeptioneller Fragen dienen sollte, haben wir aus dem gesamten Dump einen relativ kleinen Auszug genutzt.

Die Graphdatenbank: Freeversion der GraphDB von Ontotext

Keine systematische Evaluation einer Graphdatenbank, da wir nur Potential des Ansatzes erkunden wollten

Unsere Kriterien:

- Kostenlose Freeversion
- Webtauglich und lokal multiplattformfähig
- Die wichtigsten Funktionalitäten einer Graphdatenbank sollten abgedeckt sein
- Enthält ein Tool zur Bereinigung / Bearbeitung der Daten (OntoRefine)
- Enthält ausreichend Hilfsmittel (tutorials, webseminare, Möglichkeiten der Beratung)

Was GraphDB auszeichnet:

Property-Graph:

- Attribute werden nicht an Kanten vergeben
- Statt dessen operiert GraphDB mit ‘Beziehungsknoten’
- Dadurch kann SPARQL als Abfragesprache genutzt werden

Rechteverwaltung in GraphDB wird realisiert, indem:

- Daten werden in unterschiedlichen Repositorien abgelegt werden
- Verschiedene Repositorien verfügen über eigene Zugriffsregelungen
- Trippel können repositorienübergreifend abgelegt werden

Lösungsansatz für die Rechteverwaltung: Zusätzliche Entität im Datenmodell: Relation

Zur Attributierung der Beziehungen haben wir eigene Knoten, die die Beziehung beschreiben, umgesetzt.

Realisiert haben wir das insbesondere für Relation AgentRelation

Mit der Einführung des Beziehungsknotens gelingt es, den Beziehungen Attribute zuzuweisen – was die Zahl der notwendigen Beziehungstypen gegenüber RiC wesentlich reduziert.

Lösungsansatz für Datenaufbereitung

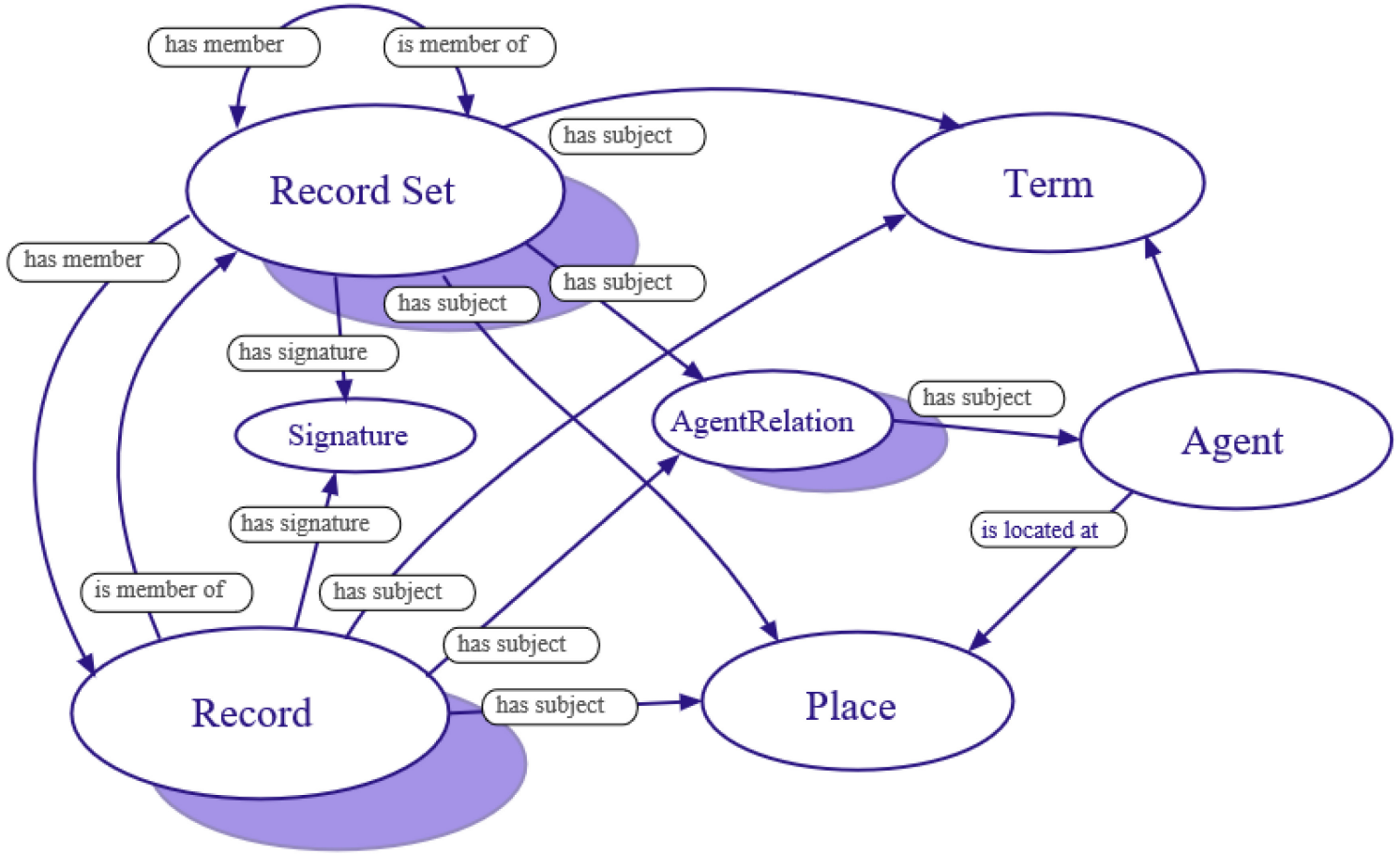
Die Daten des BAR basieren v.a. auf ISAD(G).

Damit können wir sehr viele Entitäten aus RiC nicht abdecken.
Um unseren Test trotzdem durchführen zu können, haben wir aus den ISAD(G)-Datensätzen Daten für folgende Entitäten extrahiert:

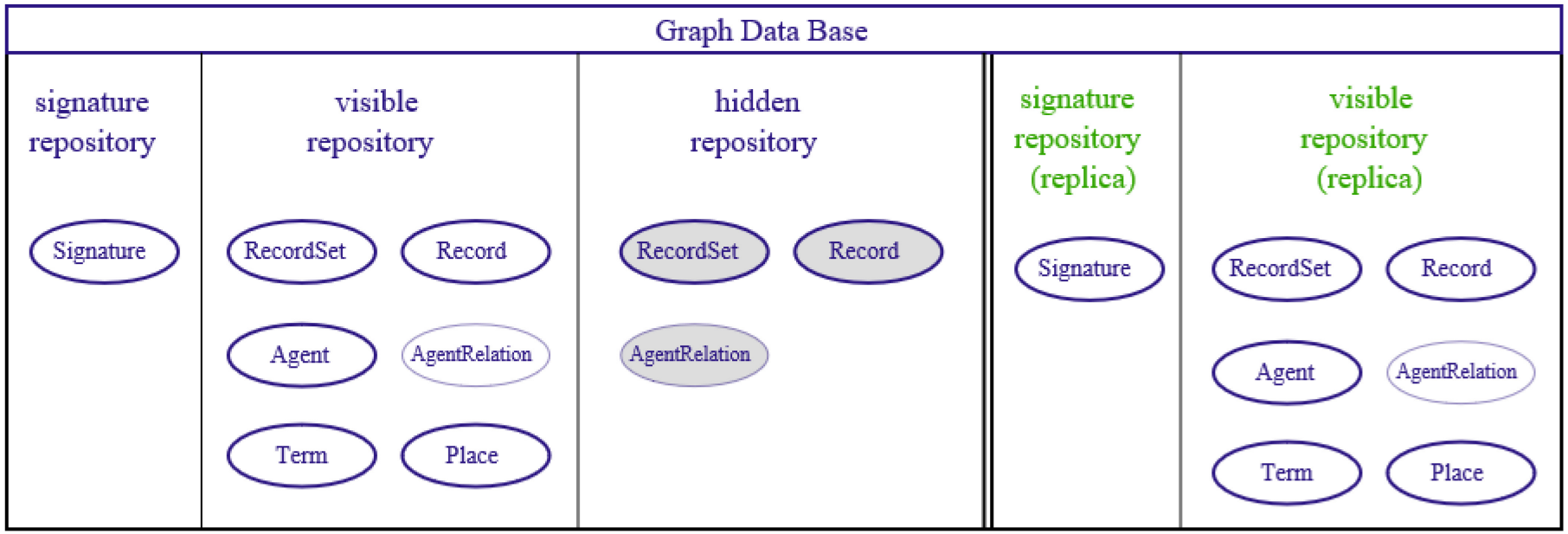
- Akteure
- Location

In unseren Tests haben wir das v.a. manuell vollzogen – maschinelle Extraktion wäre möglich

Lösungsansatz: Datenmodell



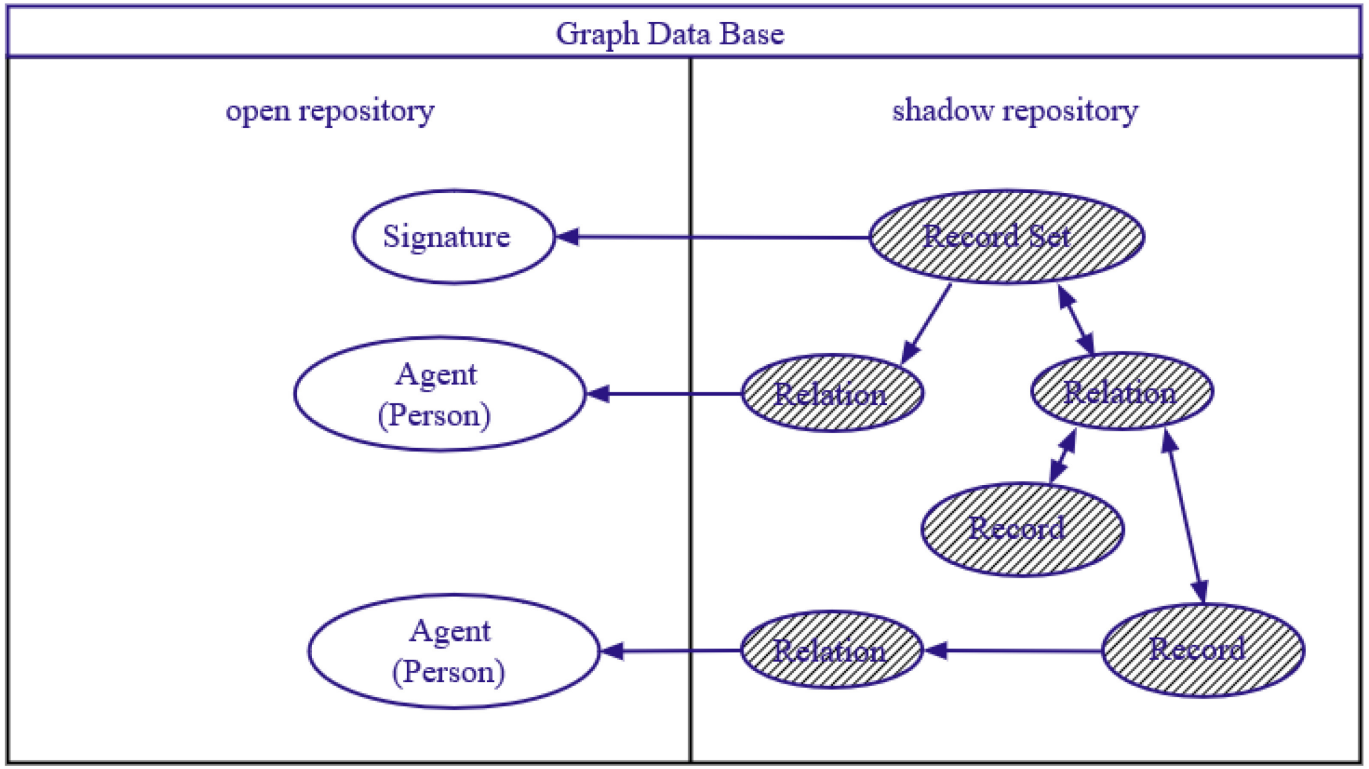
Lösungsansatz: die Repositorien



Ablage von Entitäten in unterschiedlichen Repositories:

z.B. : Record Set ist unter Verschluss

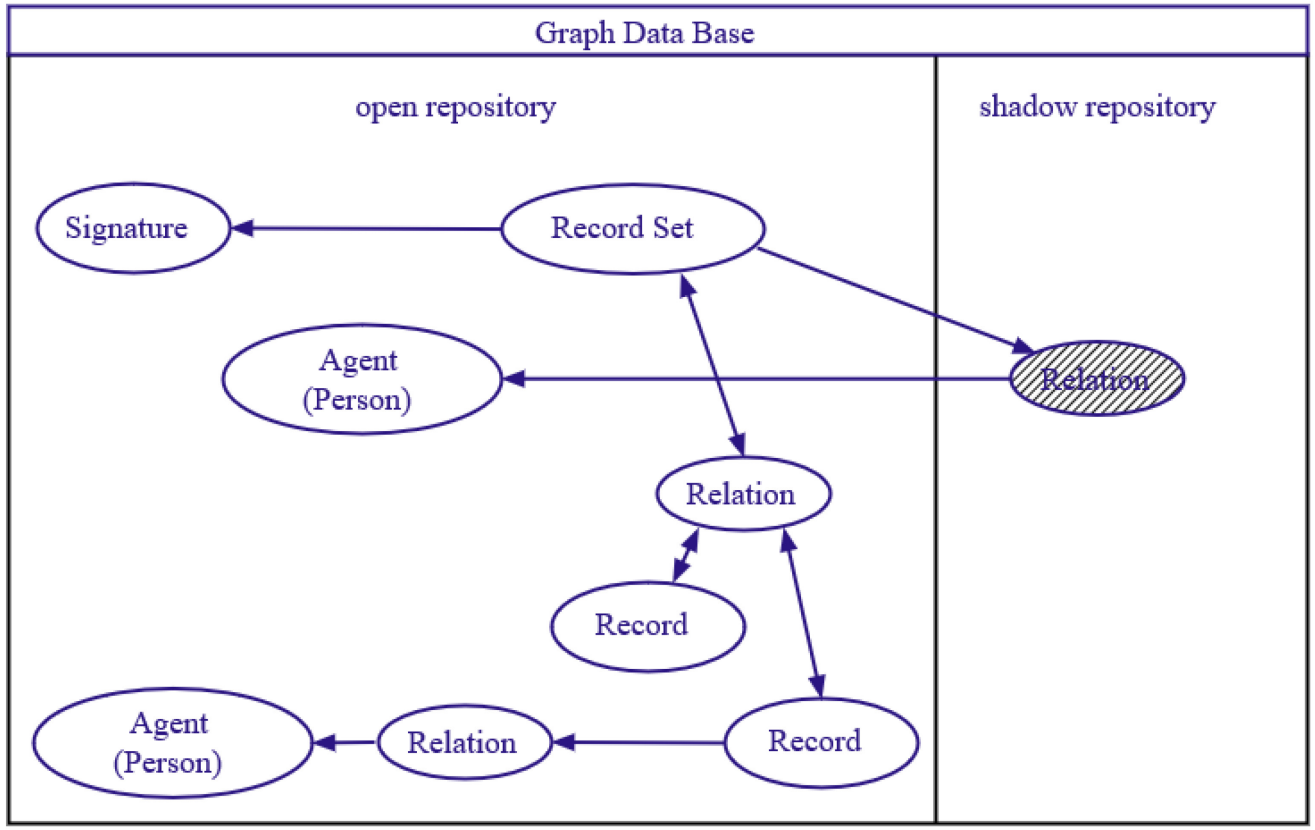
- Record Set ist unter Verschluss
- Standortsignatur ist im offenen Repositorium und damit bearbeitbar
- Agent ist im offenen Repositorium und kann damit im Kontext von öffentlich sichtbaren Records sichtbar sein.
- Hier ist der Agent ist nicht sichtbar, da kein vollständiges Trippel sichtbar ist.



Ablage von Entitäten in unterschiedlichen Repositories:

z.B. : Record Set und Records sind öffentlich, Agent unter Verschluss

- Record Set und seine Records sind öffentlich.
- Da die Personendaten jedoch noch unter Verschluss sind, darf der Agent nicht sichtbar sein.
- Die Sperrfrist auf der Beziehung unterdrückt das Trippels zum Agenten im offenen Repositorium.



Lösungsansatz: Massenmigration

Mit SPARQL-Abfragen kann gruppe von Entitäten ermittelt werden.

Bei diesen Entitäten kann ein Attribut (z.B. Sperrfrist) verändert werden

Die veränderten Entitäten können in visible Repository verschoben werden

Fazit

GraphDB leistet:

- Relationsknoten erlauben Steuerung auf Ebene der Trippel (statt Entitäten)
- Relationsknoten erlauben differenzierte Beschreibung der Relationen
- Zugriffsrechte für die verschiedenen Repositorien einfach zu steuern
- Massenmigration einfach auszugestalten

Lösungsansatz: Aufbrechen der Silos

Aufbrechen der Silos wird im archivischen Bereich in erster Linie über die Entität ‘Agent’ realisiert.

Daher haben wir bei der Datenaufbereitung Körperschaften, Personen und Familien soweit als möglich extrahiert.