Projektdokumentation im Modul Semantic Web Anzahl der Publikationen pro Mitarbiter der HTWK-Leipzig

Marcel Kirbst

1. Juli 2014

Inhaltsverzeichnis

Recherchefragestellung

Eine Liste aller Angestellten der HTWK-Leipzig, geordnet nach der Anzahl der bisherigen Publikationen im deutschsprachigen Raum.

Inhaltliche Interpretation

- erfassen alle Mitarbeiter der HTWK in semantischer Datenbank
- erfassen aller Publikationen im deutschsprachigen Raum
- semantische Verknpfung dieser Daten

Umsetzung

- Virtuelle Maschine Oracle VirtualBox
- Betriebssystem Linux-Distribution Canonical Kubuntu 14.04
- Datenbank Virtuoso + OntoWiki

HTWK-Telefonverzeichnis

- aktuellste, ffentlich zugngliche Ressource die gefunden wurde
- abrufbar unter `http://www.htwk-leipzig.de/de/hochschule/telefonverzeichnis/`
- als JSON aufbereitete Daten von Herrn Henri Knochenhauer, B.Sc und Herrn Roy Meissner, B.Sc. verfgbar
- abrufbar unter `http://141.57.21.45:8080/info/staff`

Normdaten der Deutschen Nationalbibliothek

- sind ffentlich zugnglich, halbjhrlich aktualisiert, liegen direkt im RDF-Format vor
- jedoch sehr gro, (GND.rdf ca.9GB, DNBTitel.rdf ca.12GB), teilweise inkonsistent
- abrufbar unter `http://datendienst.dnb.de/cgi-bin/mabit.pl?userID=opendata&pass=opendata&cmd=login`

Importierbare Aufbereitung der Daten erforderlich

- HTWK-Mitarbeiterdaten: JSON $\Rightarrow$ RDF
- DNB-Daten passen nicht in die VM, Extraktion der Daten die HTWK-Mitarbeiter betreffen
- Implementierung eines JAVA-Programms

Verlinkung der Ressourcen direkt ueber Nachname, Vorname

- Vorteil: Mitarbeiterdaten lassen sich direkt mit den DNB-Titeldaten verknpfen
- Nachteil: sehr viele falsche Ergebnisse aufgrund doppelter Namen

Verlinkung der Ressourcen ueber eindeutige AutorenID

- (theoretischer) Vorteil: AutorenID ist theoretisch eindeutig, es existiert theoretisch nur eine AutorenID pro HTWK-Mitarbeiter
- Praxis: Daten (im Moment noch) inkosistent, teilweise mehrere AutorenIDs, jeweilis unterschiedliche Publikationen zugeordnet

SPARQL-Anfrage

siehe code

Ergebnis SPARQL-Anfrage

siehe code

Interpretation und Zusammenfassung

- Aufbereitung der Daten viel zeitintensiver als erwartet
- aufgrund teilweise inkonsistenter Daten noch keine tragfhigen Ergebnisse
- ggf manuelles Aufbereiten der HTWK-Daten (Geburtsjahr, DNB-ID hinzufgen) sollte auch zu besseren Ergebnissen fhren