

# Projekt-Bericht und Dokumentation im Kurs Wissensrepräsentation SoSe15

Lukas Hodel

Richard Remus

21. Juli 2015

# Contents

<b>1</b>	<b>Aufgabenstellung</b>	<b>3</b>
<b>2</b>	<b>Analyse Lobbyradar</b>	<b>4</b>
<b>3</b>	<b>Aufbau einer Ontologie</b>	<b>6</b>
3.1	Konvertierung der Daten in RDF . . . . .	6
3.2	T-Box . . . . .	6
3.3	A-Box . . . . .	6
<b>4</b>	<b>Anbindung zu DBpedia</b>	<b>6</b>
<b>5</b>	<b>Nachbetrachtung</b>	<b>6</b>

# 1 Aufgabenstellung

Die gegebene Aufgabenstellung lautete:

Führen Sie ein Semantic Web / Linked Data Projekt durch. Dies sollte mehrere der folgenden Aspekte umfassen.

- Verknüpfen von verschiedenen Datenquellen
- Konvertierung von Daten ins RDF-Format (mit entsprechender T-Box)
- Entwicklung von Ontologien
- Linked Data Anwendung
- Informationsextraktion/Web Scaping und Konvertierung in RDF
- Entity Resolution
- ..

Wir haben im Kurs bereits Erfahrungen mit dem [Lobbyradar des ZDF](#) und [DBpedia](#) gemacht.

Zuerst werden wir die im Lobbyradar enthaltenen Informationen in einen RDF-Graphen überführen. Danach werden wir eine Anbindung zu DBpedia entwickeln, in der Sparql-Queries zu den Entitäten des Graphen erstellt formuliert und abgeschickt werden können. Auch soll die Anwendung eine Suchwörterweiterung bieten. Die Ergebnisse der Queries sollen dann wiederum in den bestehenden Graphen eingepflegt werden, anschliessend wird ein Teilgraph angezeigt, in dessen Zentrum die gesuchte Entität steht und der die Relationen zu anderen Entitäten aufzeigt.

## 2 Analyse Lobbyradar

Lobbyradar enthält Knoten welche in JSON persistiert sind. Damit werden alle Entities und Relations abgebildet. Die Knoten enthalten schon für sich eine große Zahl an Properties:

```
{'_id': '54bd3c768b934da06340f4c6',
  'aliases': [],
  'created': 'datetime.datetime(2015, 1, 19, 17, 18, 46, 529000)',
  'data': [{'auto': 'True',
    'created': 'datetime.datetime(2015, 1, 19, 17, 19, 9, 420000)',
    'desc': 'Quelle',
    'format': 'link',
    'id': '3dc1416e38deac59076cd3f0d4e1235de79cb530207d06c28053134cc8aa7732',
    'key': 'source',
    'updated': 'datetime.datetime(2015, 1, 19, 17, 19, 9, 420000)',
    'value': {'remark': 'created by lobbyliste importer',
      'url': 'http://bundestag.de/blob/189476/8989cc5f5f65426215d7e0233704b20a/lobbylisteaktuell-data.pdf'}}],
  {'auto': 'True',
    'created': 'datetime.datetime(2015, 1, 19, 17, 19, 9, 420000)',
    'desc': 'Titel',
    'format': 'string',
    'id': '65873d29bd0ab6af91ef341689d28f4f0658cc851494a0ef34cfd07dd0cf5d42',
    'key': 'titles',
    'updated': 'datetime.datetime(2015, 1, 19, 17, 19, 9, 420000)',
    'value': 'Gesch\xe4fts\xfchrer'},
  {'auto': 'True',
    'created': 'datetime.datetime(2015, 1, 19, 17, 18, 46, 529000)',
    'desc': 'Titel',
    'format': 'string',
    'id': '1c98f43f0787b6dfcad25c38849ef3895bd26624a79c8fa9e6203c360d120fad',
    'key': 'titles',
    'updated': 'datetime.datetime(2015, 1, 19, 17, 18, 46, 529000)',
    'value': '2. Vorsitzender'}],
  'importer': 'lobbyliste',
  'name': 'Markus Hoymann',
  'search': ['markus hoymann'],
  'slug': 'markus hoymann',
  'tags': ['lobbyist', 'lobbyismus', 'executive'],
  'type': 'person',
  'updated': 'datetime.datetime(2015, 1, 19, 17, 19, 9, 426000)'}
```

Deshalb haben wir entschieden, von all diesen, nur die wichtigsten Eigenschaften abzubilden. Für Personen und Organisationen:

- id
- name
- created
- updated
- type (*Person, Organisation*)
- alias (*nur bei Organisationen*)
- tags (*als Interessengebiete*)

Für die Relationen zwischen diesen Entitäten benötigen wir auch eine Klassifikation. Die verschiedenen Bezeichnungen für die Ämter und Anstellungsverhältnisse der Personen, also ihre Positionen in der Lobbyradar-Datenbank, sind jedoch stark arbiträr, gehorchen keiner Syntax und liegen nur als Literal vorhanden sind. Es ist leider jedoch nicht ohne weiteres möglich einer Relation in RDF eine beliebige Bezeichnung zu geben. Dazu würden Hilfsknoten benötigen, welche den Graphen stark verkomplizieren würden. Da in der Datenbank auch viele Positionen mit semantisch gleichem Inhalt unterschiedliche Bezeichnungen haben, entweder weil keine einheitliche Benennung vorgegeben ist (z.B. *Bundesminister der Finanzen*, *Finanzminister*, *Minister der Finanzen*) oder weil die Importeure Tippfehler gemacht haben (*executive*, *ececutive*). Daher haben wir uns entschieden, die wichtigsten Relationen semantisch zu gruppieren und für diese zusammenfassende Properties zu erstellen. Diese Zuordnung mussten wir händisch erledigen, aus Zeitgründen konnten wir also die restlichen Daten nicht Berücksichtigen.

## 3 Aufbau einer Ontologie

### 3.1 Konvertierung der Daten in RDF

### 3.2 T-Box

### 3.3 A-Box

## 4 Anbindung zu DBpedia

## 5 Nachbetrachtung

- rdflib sollte beim hinzufügen von Knoten prüfen, ob die gegebenen Tripel sinnhaft sind und auf Knoten verweisen, die bereits da sind. `g.add()` ist da nicht *verbose* genug und sollte ruhig mal meckern und nicht still alles hinnehmen.