Querying the Semantic Web for Concept Identifiers to Annotate Research Datasets

André Langer, Christoph Göpfert and Martin Gaedke
Chemnitz University of Technology, Germany
andre.langer@informatik.tu-chemnitz.de

The Fourteenth International Conference on Advances in Semantic Processing
SEMAPRO 2020
25 – 29 October 2020 - Nice, France
Background
Professorship for Distributed and Self-organizing Systems
Chemnitz University of Technology (TUC), Germany

Short Résumé

- Born in 1983
- Scholarships from DAAD, SDW and e-fellows
- Research Stay at UCSB, USA in 2006
- Graduated in Media Computer Science at TUC in 2007
- Project Lead for TV Media Online Services until 2016
- Since 2016 PhD Student at TUC, Germany, PhD topic:
  “PIROL: Publishing Interdisciplinary Research Over Linked Data”

- Research Interest:
  Research Data Management, Linked Data, User Interface Experience

André Langer
PhD Student

https://vsr.informatik.tu-chemnitz.de/research
Professorship for Distributed and Self-organizing Systems
Chemnitz University of Technology (TUC), Germany

“At VSR, we contribute to enrich the way people live and work in the hyper-connected society by improving human-machine collaboration.”

Current Project Context:
DFG Collaborative Research Center „Hybrid Societies“ (2020-2024)

• Establish an Institutional Research Data Repository for Research on Humans Interacting with Embodied Technologies
• Enable Interdisciplinary Discovery and Reuse of Research Artifacts through a common language
• Allow semantic research data annotation based on sophisticated user input web interfaces
Problem Description
Data from: UNICITY: A depth maps database for people detection in security airlocks

Explore at zenodo.org  Explore at figshare.com  Explore at search.datacite.org

Unique identifier
https://doi.org/10.5281/zenodo.2556679

Dataset updated Feb 4, 2019

Dataset provided by
Haute école d'ingénierie et d'architecture de Fribourg
Fastcom Technology SA
Idiap Reserach Institute

Authors
Joël Durnoulin; Olivier Canévet; Michael Villamizar; Hugo Nunes; Omar Abou Khaled; Elena Mugellini; Fabrice Moscheni; Jean-Marc Odobez

License
Attribution 4.0 (CC BY 4.0)
License information was derived automatically

Description
UNICITY: A depth maps database for people detection in security airlocks.
„From Strings to Things“

Persistent Subject Identifier

Some Basic Properties

Unambiguous Author Identifier

Ambiguity

Missing Typification

No Inference Possibilities
Research Question
Which knowledge bases exist that provide relevant concepts to describe research datasets and how can we query them?
Approach
Identification of interdisciplinarily relevant concepts

1. Examined established vocabularies for attribute groups
2. Examined UI of established research dataset repositories
3. Examined meta descriptions of existing research datasets

Research area, topic, resource type, (file) format/media type, rights/license, discipline, measurement technique/device, material, audience, demographic characteristics, examined objects, research and evaluation methods, research objective, metrics, measurement characteristics, models
# Sources for research data concept identifiers

## Ontology Catalogs
- NCBO BioPortal
- LOV
- AberOWL
- ORR
- OLS
- Ontobee
- IBC AgroPortal
- SmartCity OC
- ...

## Authority Services
- EU NALs/Eurovoc
- LC
- DNB
- RAMEAU
- UNESCO
- AGROVOC
- GEMET
- SSW
- ...

## Instance Datasets
- LODCache
- LOD-a-lot
- Dbpedia
- Wikidata
- BTC
- YAGO

## Other Sources
- Static Files
- Dictionary services
- Semantic Search Engines
Results

• Appropriate data sources already exist that provide concepts of a certain focus with a concept identifier

• Scattered situation

• Varying data granularity and quality

• Still no or insufficient data providers available for some aspects of interdisciplinary relevance such as methods, devices, objectives

• Variety of representation formats and APIs
Querying Research Concepts

```
SELECT DISTINCT ?concepturi ?conceptlabel
WHERE {
  FILTER (CONTAINS(str(?typeLabel), str(?query))).
  # ... - other additional concept scope filter patterns -
}
```
Evaluation
Evaluation Strategy

• **Prototypical Implementation** of *Metadata Input Application to annotate a Research Dataset* with a QueryEngine that dynamically retrieves concepts from external data sources as [Proof-of-Concept](#).

• **Data and Service Quality Metrics** for each data provider group measured based on four example use cases.
## Evaluation Results

<table>
<thead>
<tr>
<th>Concept Group</th>
<th>LOV</th>
<th>BioPortal</th>
<th>EuroVoc</th>
<th>Wikidata</th>
<th>DBpedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>27</td>
<td>37*</td>
<td>4</td>
<td>34</td>
<td>28</td>
</tr>
<tr>
<td>License</td>
<td>11</td>
<td>42*</td>
<td>41</td>
<td>435</td>
<td>108</td>
</tr>
<tr>
<td>File Format</td>
<td>128</td>
<td>51*</td>
<td>172</td>
<td>4201</td>
<td>432</td>
</tr>
<tr>
<td>Research Method</td>
<td>16</td>
<td>149*</td>
<td>0</td>
<td>16</td>
<td>5</td>
</tr>
</tbody>
</table>

### Extent of retrieved instances per requested Class Label

<table>
<thead>
<tr>
<th>Concept Group</th>
<th>LOV</th>
<th>BioPortal</th>
<th>EuroVoc</th>
<th>Wikidata</th>
<th>DBpedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>1.5s</td>
<td>1.5s*</td>
<td>1.0s</td>
<td>2.7s</td>
<td>0.2s**</td>
</tr>
<tr>
<td>License</td>
<td>1.5s</td>
<td>1.5s</td>
<td>1.4s</td>
<td>5.3s</td>
<td>0.2s**</td>
</tr>
<tr>
<td>Media Type</td>
<td>1.8s</td>
<td>2.9s</td>
<td>1.0s</td>
<td>5.8s</td>
<td>0.5s**</td>
</tr>
<tr>
<td>Research Method</td>
<td>1.5s</td>
<td>3.9s</td>
<td>1.0s</td>
<td>13.2s</td>
<td>0.2s**</td>
</tr>
</tbody>
</table>

### Processing Time per requested Class Label in s
Conclusion
Contribution

• **Analysis of data sources** presented that provide concepts and corresponding Linked Data identifiers to describe research datasets

• Implementation of a web-based prototype presented to **Dynamically query external services**

• **Varying service and data quality** shown
Inspired and Interested?

Andre.Langer@Informatik.TU-Chemnitz.de

@myVSR  /myVSR