Linking Semantic Desktop Data to the Web of Data

Laura Drăgan, Renaud Delbru, Tudor Groza, Siegfried Handschuh, Stefan Decker
Linked Data Everywhere
Linked Data Everywhere

Semantic Desktop

Web

Enabling networked knowledge.
Linked Data Everywhere

Semantic Desktop

Web

Enabling networked knowledge.
Linked Data Everywhere

Semantic Desktop

Web of Data
... BUTDisconnected

Semantic Desktop

Web of Data
Connecting the SD to the WoD
Web alias

= web resource representing the same real-world entity as the desktop resource
It's Hard Because

Different identifiers

nepomuk:/res/Angela

http://angelaonthe.net/foaf/me
It's Hard Because

Different vocabularies
It's Hard Because

The sheer size of the Web of Data
Approach

1. Candidate Selection
   - Query various Web of Data sources
   - Identify candidate URIs
   - Retrieve data for each of the candidate

2. Candidate Filtering
   - Compute similarity score.
   - Filter the candidates.
Candidate Selection

Determined set of sources
- Specific requirements
- Restricted domain

Semantic search engine
- Generic domain
- Unknown data sources
Query Modules

SWSE
- Sindice

SPARQL
- Dbpedia
- SW Dogfood Server

Custom
Creating the Query

Desktop URI: nepomuk:/res/95b7360c-61f1-4039-9cda-c77a4f6fb5c6

**rdf:type** [http://www.semanticdesktop.org/ontologies/2009/02/19/nmm#MusicAlbum](http://www.semanticdesktop.org/ontologies/2009/02/19/nmm#MusicAlbum)

**nao:prefLabel** One Night Only

**nie:contentCreated** 1998

**nmm:performer** Bee Gees

**nie:title** One Night Only

- “One Night Only”
- OR “1998”
- OR “Bee Gees”
- OR “One” OR “Night” OR “Only”
- OR “Bee” OR “Gees”
Creating the Query

Desktop URI: nepomuk:/res/95b7360c-61f1-4039-9cda-c77a4ff6fb5c6

rdf:type http://www.semanticdesktop.org/ontologies/2009/02/19/nmm#MusicAlbum

nao:prefLabel One Night Only
nie:contentCreated 1998
nmm:performer Bee Gees
nie:title One Night Only

“One Night Only” OR “1998” OR “Bee Gees” OR “One” OR “Night” OR “Only” OR “Bee” OR “Gees”
Creating the Query

**Desktop URI:** nepomuk:/res/95b7360c-61f1-4039-9cda-c77a4f6fb5c6

- nao:prefLabel: One Night Only
- nie:contentCreated: 1998
- nmm:performer: Bee Gees
- nie:title: One Night Only

“One Night Only”
OR “1998”
OR “Bee Gees”
OR “One” OR “Night” OR “Only”
OR “Bee” OR “Gees”
Candidate Filtering

1. Filter by type
2. Compute similarity score
3. Filter by score

(local, web)
Matching Module

(local, web)

Type matching

Yes

Compute score

score ≥ threshold

No

return 0

Yes

return score
"pimo:Person":
{
   "mapping" : [ 
   "foaf:Person",
   "foaf:Agent",
   "dbpedia:Person"
   ]
}
"nmm:performer##nco:fullname" : 
{
    "mapping" : [
        "dbpedia:artist",
        "foaf:maker##foaf:name",
        "dbpedia:artist##foaf:name"
    ],
    "weight":"0.7",
    "thresholds" : [
        "approx":"true",
        "MongeElkan:0.7",
        "Chapman:0.8"
    ]
}

Compute *score*
"nmm:performer##nco:fullname" : 
{
  "mapping" : [
    "dbpedia:artist",
    "foaf:maker##foaf:name",
    "dbpedia:artist##foaf:name"
  ],
  "weight":"0.7",
  "thresholds" : [
    "approx":"true",
    "MongeElkan:0.7",
    "Chapman:0.8"
  ]
}

Compute score

artist

“Bob Dylan”
Matching Parameters

String matching (SM)
- Exact matching versus approximate string matching
- Köln vs. Köln

Weighted properties (WP)
- Weighted participation of properties in the final score
- Email address more exact than name

Multi-valued properties (MVP)
- All matching values for a property contribute to the score
- e.g. Authors' names for a paper
Score Calculation

Driven by the local data

\[ \text{score} = \frac{\text{weighted sum of matching props}}{\text{total sum of all weighted props}} \]
Evaluation

Manually constructed gold standard
- Data collection
- Relevance judgements

IR measures
- Study effect of parameter & threshold settings

Performance
Data collection

**Desktop data**
- 50 people – nco:PersonContact
- 50 music albums – nmo:MusicAlbum
- 50 publications – nfo:PaginatedTextDocument
- 11,917 triples

**Web data**
- 20 candidates for each desktop resource → 3000 URIs
- 1,530,686 triples
Relevance Judgements

Do these two URIs represent instances of the same real-world object/person?

Desktop URI: nepomuk:/res/95b7360c-61f1-4039-9cda-c77a4f6fb5c6
  rdf:type http://www.semanticdesktop.org/ontologies/2009/02/19/nmm#MusicAlbum
  nao:prefLabel One Night Only
  nie:contentCreated 1998
  nmm:performer Bee Gees
  nie:title One Night Only

Web URI: http://dbpedia.org/resource/One_Night_Only_%28The_Bee_Gees_album%29
  owl:sameAs http://dbpedia.org/resource/One_Night_Only_%28The_Bee_Gees_album%29
  foaf:isPrimaryTopicOf http://en.wikipedia.org/wiki/One_Night_Only_%28The_Bee_Gees_album%29
  rdfs:label One Night Only (The Bee Gees album)
  dbp:redirect http://dbpedia.org/resource/One_Night_Only_%28The_Bee_Gees_album%29
  owl:sameAs http://rdf.freebase.com/ns/guid.9202a8c04000641f800000002fb5278

You matched 275 pairs!
Relevance Judgements

3000 pairs x 3 different experts

Fleiss' $K = 0.638 \pm 0.214$
Average pairwise agreement 92.252%

highest for publications
lowest for albums
IR Measures

- MAP
- NDCG
- P@k (k=1,2,3,4,5)

Baseline:
- exact match
- all properties count equally
- single value considered for each property
Results

Approximate string matching
- improves results for albums and people
- does not help for publications

Weights and multiple values
- when combined improve results for publications,
- but not for the other types
Performance

Average total check time is <1s
- Publications take longest 978ms
- People and albums – 52–53ms

Dependent of
- candidates retrieved
- number of properties
- complexity of properties
- number of values per property
Discussion

Precision vs. recall

Evaluation

- Personal [semantic] data for evaluation
- Object identifier vs. representation
- Quality of data makes a difference (doh!)
Conclusion

Identify *web aliases* for Semantic Desktop data

- Automatically (precision over recall)
- Promising results in evaluation
- Just the 1st step

What next?

- Data sync / update / enhance
- Personalized desktop services and apps