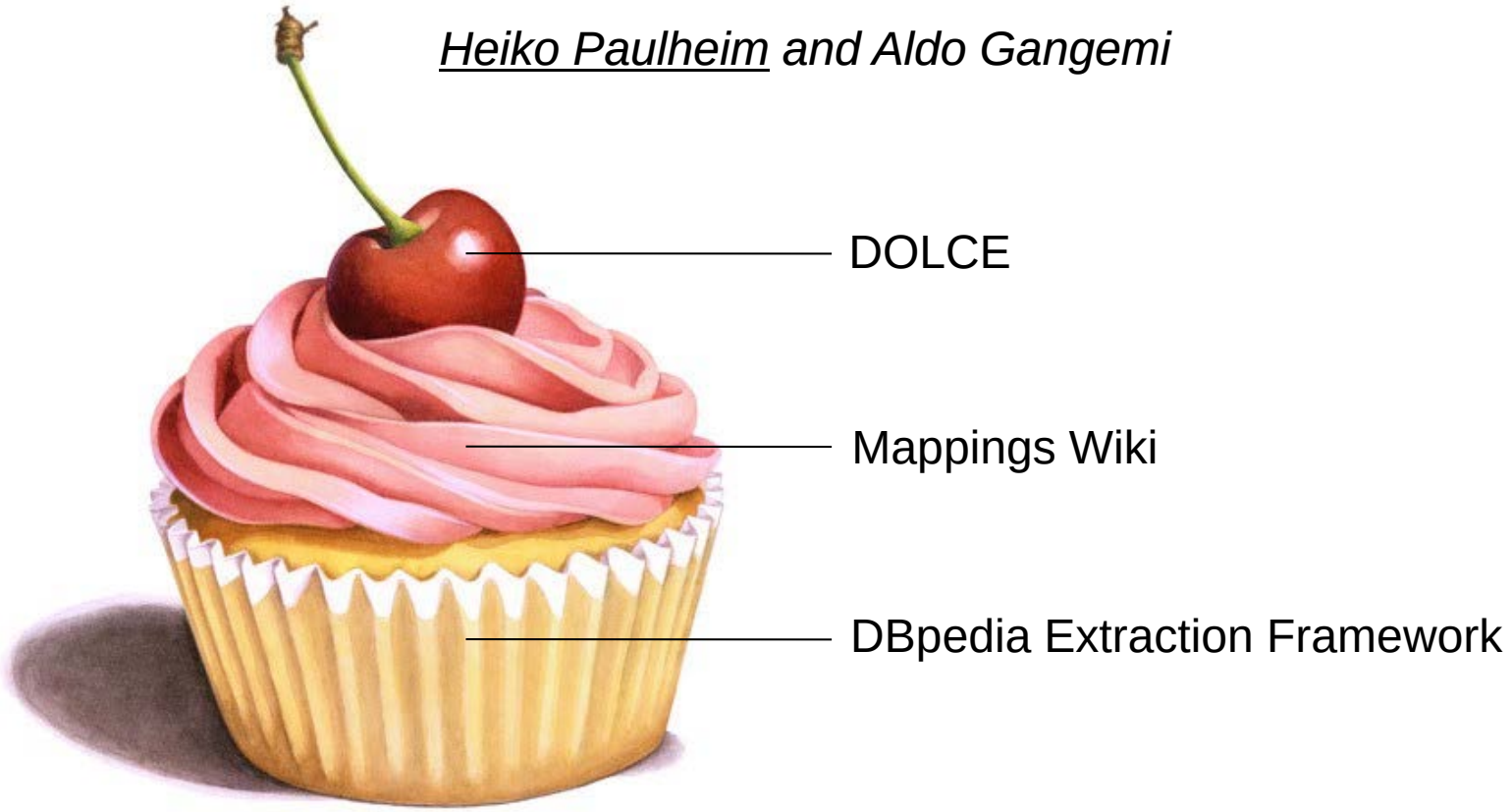
Serving DBpedia with DOLCE

More than Just Adding a Cherry on Top

Heiko Paulheim and Aldo Gangemi



DOLCE Mappings: Served Since DBpedia 2014

DOLCE Mappings: Served Since DBpedia 2014

rdf:type	owl:Thing foaf:Person dbo:Person dul:Agent dul:NaturalPerson wikidata:Q215627 wikidata:Q483501 wikidata:Q5 dbo:Agent

About: birth place

An Entity of Type : ObjectProperty, from Named Graph : http://dbpedia.org/resource/classes#, within Data Space : dbpedia.org

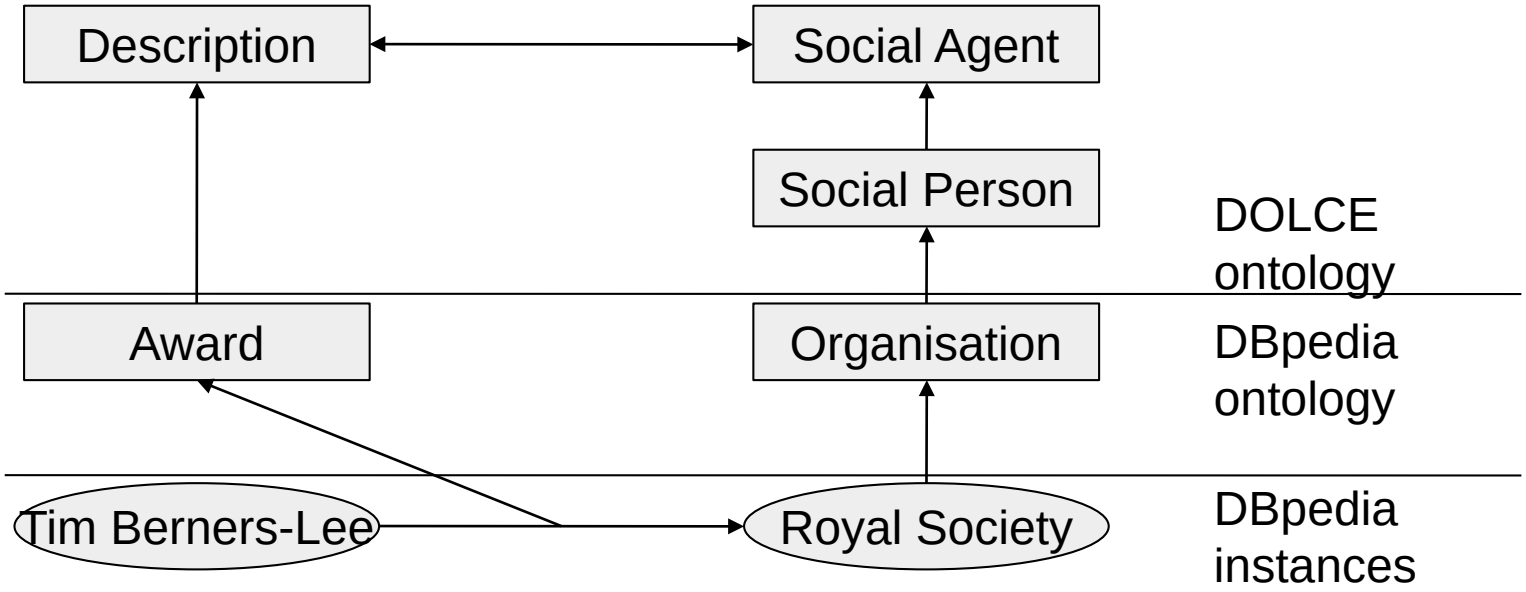
where the person was born

Property	Value
rdf:type	rdf:Property owl:ObjectProperty
rdfs:comment	where the person was born
rdfs:domain	dbo:Person
rdfs:isDefinedBy	http://dbpedia.org/ontology/
rdfs:label	birth place
rdfs:range	dbo:Place
rdfs:subPropertyOf	dul:hasLocation
owl:equivalentProperty	wikidata:P19
wdrs:describedby	dbo:data/definitions.ttl
http://www.w3.org/ns/prov#wasDerivedFrom	http://mappings.dbpedia.org/index.php/OntologyProperty:birthPlace
is http://open.vocab.org/terms/defines of	http://dbpedia.org/ontology/
is http://open.vocab.org/terms/describes of	dbo:data/definitions.ttl

More than Just a Cherry on Top

- DOLCE adds a layer of formalization
  - high level axioms
  - additional domain and range restrictions
  - fundamental disjointness (e.g., physical object vs. social object)
- Enriches DBpedia ontology
- Can be used for consistency checking

More than Just a Cherry on Top

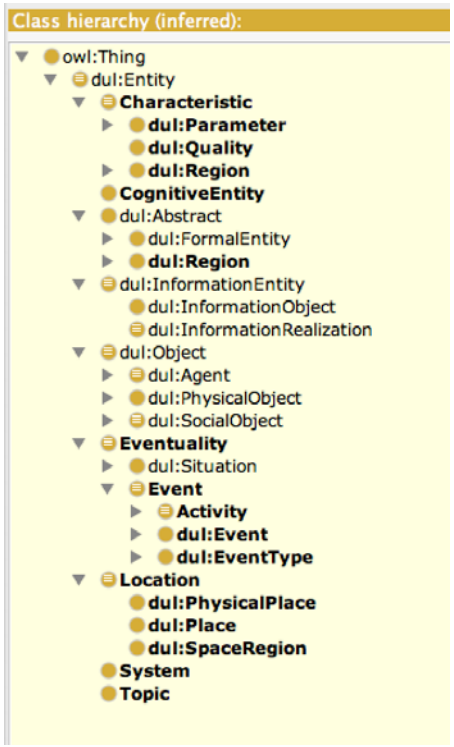


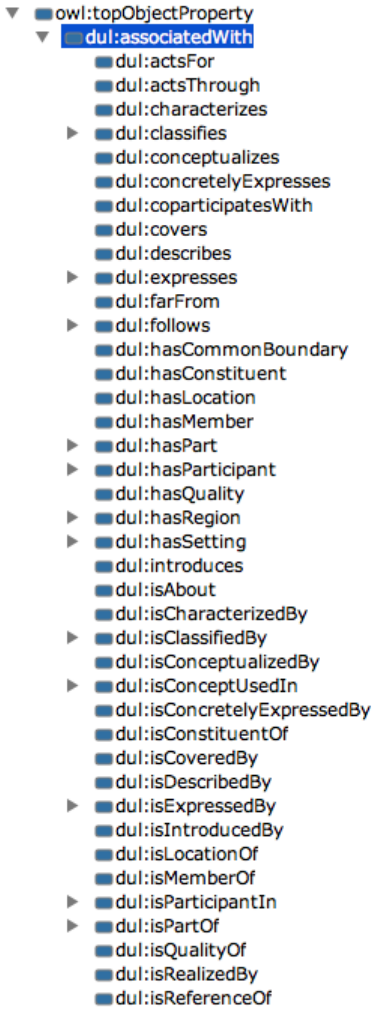
DBpedia in a Nutshell

- Raw extraction from Wikipedia infoboxes
- Infobox types and keys mapped to ontology
  - Crowdsourcing process (aka *Mappings Wiki*)
  - 735 classes
  - ~2,800 properties
- Almost no disjointness
  - only 24 disjointness axioms
  - many of those are corner cases
    - MovingWalkway vs. Person

DOLCE in a Nutshell

- A top level ontology
- Defines top level classes and relations
- Including rich axiomatization





DOLCE in a Nutshell

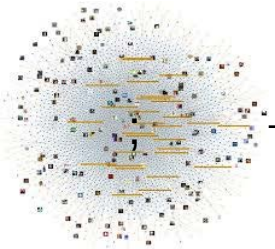
- Original DOLCE ontologies were too heavy weight
  - thus: hardly used on the semantic web
  - remember: a *little* semantics goes a long way!
- DOLCE-Zero
  - Simplified version
  - Contains both DOLCE and D&S (Descriptions and Situations)
  - D&S introduces some high level design patterns

Systematic vs. Individual Errors in DBpedia

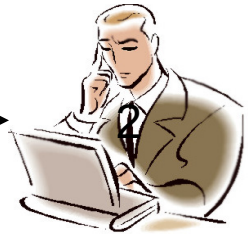
- Systematic errors
  - occur frequently following a pattern
  - e.g., organizations are frequently used as objects of the relation *award*
- are likely to have a common root cause
  - wrong mapping from infobox to ontology
  - error in the extraction code
  - ...

Overall Workflow

- Overall workflow
- For each statement
  - add the statement plus all subject/object types to the ontology
  - check consistency
  - present inconsistent statements and explanations for inspection



Reasoning



RDF Graph

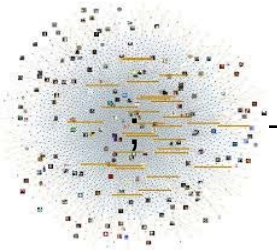
User

Overall Workflow

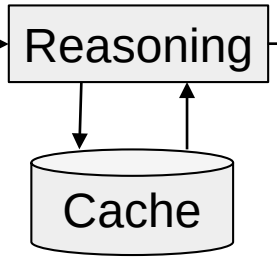
- Inspecting a single statement takes 2.6 seconds
  - on a standard laptop
- DBpedia 2014 has 15,001,543 statements
  - in the dbpedia-owl namespace

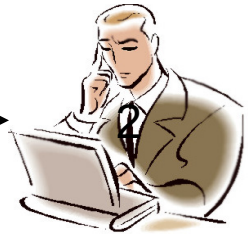
→ consistency checking would take 451 days

- Solution
  - cache results for signatures (predicate + subject types + object types)
  - there are only 34,554 different signatures!



RDF Graph

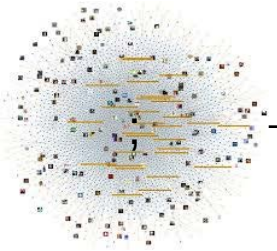




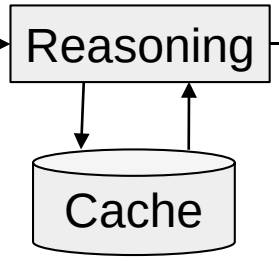
User

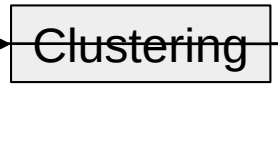
Overall Workflow

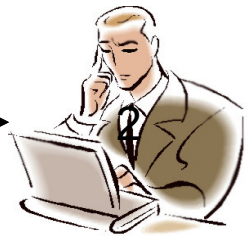
- Overall, we find 3,654,255 inconsistent statements (24.4%)
  - cf.: only 97,749 (0.7%) without DOLCE
- Too much to inspect!
  - We are looking for *systematic* errors
  - Cluster explanations w/ DBSCAN
  - Each cluster represents a systematic error



RDF Graph



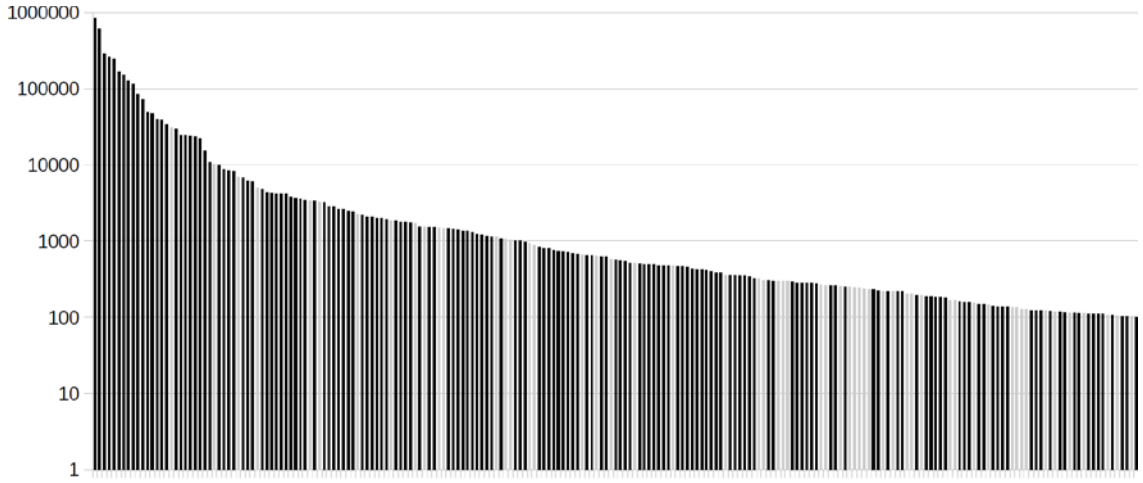




User

Clustering Inconsistent Statements

- Each explanation as a binary vector
  - with 0/1 for all axioms involved in the explanations
  - 1,467 axioms (dimensions) in total
- DBSCAN
  - Manhattan distance (i.e., number of axioms added/removed)
  - MinPts=100 (minimum frequency for a *systematic* error)
  - ε=4 (explanations in a cluster differ by two axioms at most)



Major Systematic Errors

- Inspection of the top 40 clusters
  - they contain 96% of all inconsistent statements
- *Overcommitment* (19) – using properties in different contexts
  - e.g., dbo:team is defined as a relation between persons and sports teams
  - but also used for relating participating teams to events
  - fix: relax domain/range constraints, or introduce new properties
- *Metonymy* (11) – ambiguous language terms
  - e.g., instances of dbo:Species contain both species as well as single animals
  - hard to refactor

Major Systematic Errors

- *Misalignment* (5) – classes/properties mapped to the wrong concept in DOLCE
  - occasionally occurs if intended and actual use differ
  - e.g., *dbo:commander* is more frequently used with events (e.g., battles) than military units – d0:hasParticipant rather than d0:coparticipatesWith
  - fix: change alignment
- *Version branching* (3) – semantics of dbo concepts have changed
  - e.g., dbo:team in DBpedia 3.9: career stations and teams, in DBpedia 2014: athletes and teams
  - fix: change alignment

A Look at the Long Tail

- DBSCAN identifies clusters and “noise”
  - i.e., statements that are not contained in clusters
- Manual inspection of a sample of 100 instances
  - 64 are erroneous
  - 30 are false negatives (i.e., correct statements)
  - 6 are questionable
- Typical error sources in the long tail
  - are expected to be *cross-cutting*, i.e., occurring with various classes and properties

A Look at the Long Tail

- Typical error sources in the long tail
- Link in longer text (23)
  - e.g., dbr:Cosmo_Kramer dbo:occupation dbr:Bagel .
- Wrong link in Wikipedia (9)
  - e.g., dbr#Stone_(band) dbo:associatedMusicArtist dbr#Dementia .
  - *Dementia* should link to the band, not the disease
- Redirects (7)
  - e.g., dbpedia#Ben_Casey dbo:company dbpedia#Bing_Crosby .
  - The target *Bing_Crosby_Productions* redirects to *Bing_Crosby*
- Links with Anchors (6)
  - e.g., dbr:Tim_Berners-Lee dbo:award dbr:Royal_Society .
  - the original link target is *Royal_Society#Fellows*
  - anchors are ignored by DBpedia

Occupation	Bagel Shop Worker Raincoat Salesman Entrepreneur (Kramerica Industries) Non-fiction Author Mall Santa Tennis Ball Boy

Conclusions

- We have shown that
  - DOLCE helps identifying inconsistent statements
  - Cluster analysis allows for identifying systematic errors
  - User interaction is minimized
    - we analyzed one statement each from 40 clusters
    - corresponding to 3,497,068 affected statements
- Outcomes for future DBpedia versions
  - DBpedia ontology changes
  - Mapping changes
  - DOLCE alignment changes
  - Bug reports

Serving DBpedia with DOLCE

More than Just Adding a Cherry on Top

Heiko Paulheim and Aldo Gangemi