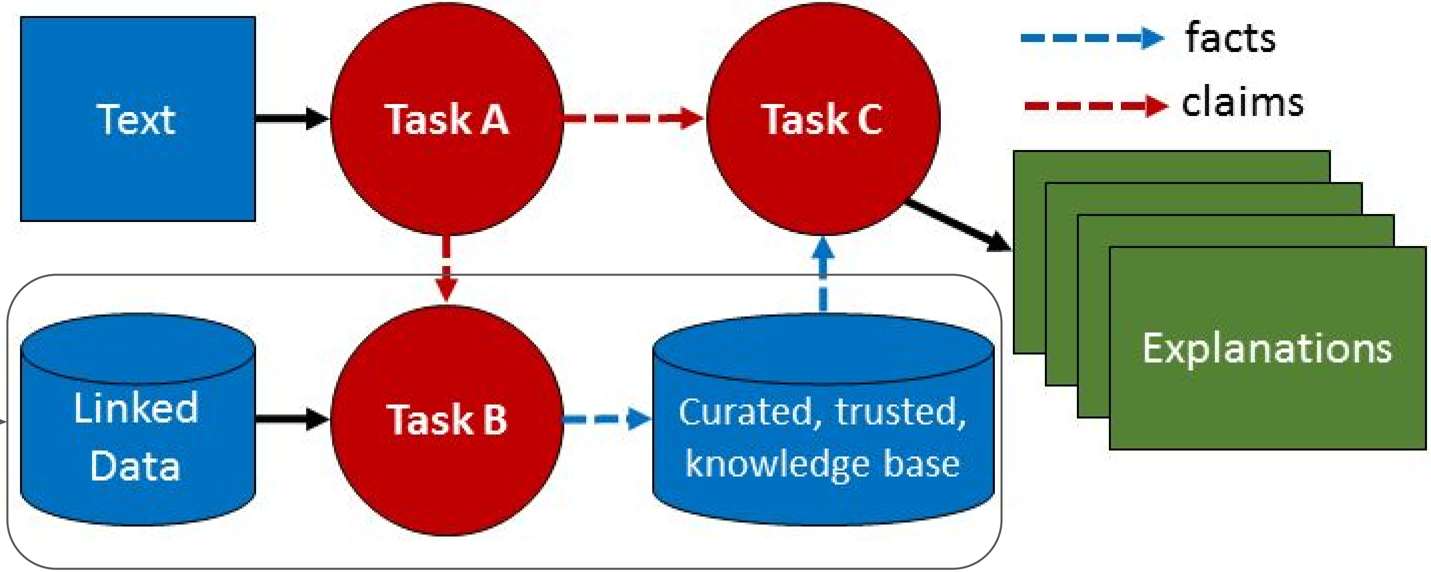
Supporting Fact Checking Applications using Structured Open Web Data

Steven Lynden, AI Cloud Team, AIST

Where does it fit in?



(Task A) Claims extraction from text.

(Task B) Knowledge-driven information gathering.

(Task C) Trust-based explanation finding.

Plan of this talk

- Introduce Linked Open Data and other structured data on the Web.
- Describe how I utilised various data sources to for a the “Movie Critiques” scenario for BackDrop.
- Discuss some issues with this process, and introduce some previous work which may be applicable going forward.
- Throw out some potential ideas for future work in the short term.

Linked Open Data

- Use URIs/IRIs to identify things

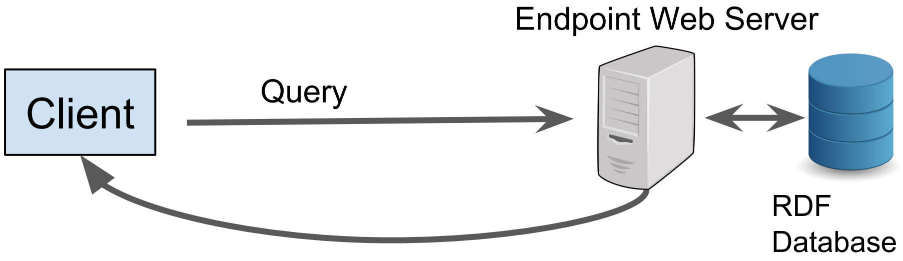
- Use HTTP IRIs
  - So that things can be looked up (dereferenced)

- Provide useful information about resource being identified
  - Using standards such as RDF.

- Refer (link) to other resources using HTTP IRI-based names when publishing data on the Web

Sources of linked data

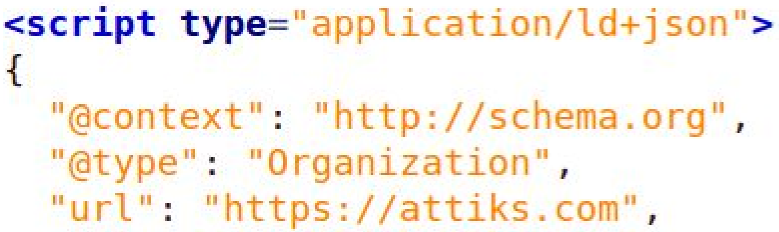
- Endpoints provide access to specific data sources



- Raw RDF data in various formats, e..g RDF/XML

```
rdf:resource="http://www.w3.org/2002/07/owl#FunctionalProperty"/></res:bin
ding>
   </res:solution>
   <res:solution rdf:nodeID="r1">
     <res:binding
rdf:nodeID="r1c0"><res:variable>Concept</res:variable><res:value
rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/></res:b
inding>
```

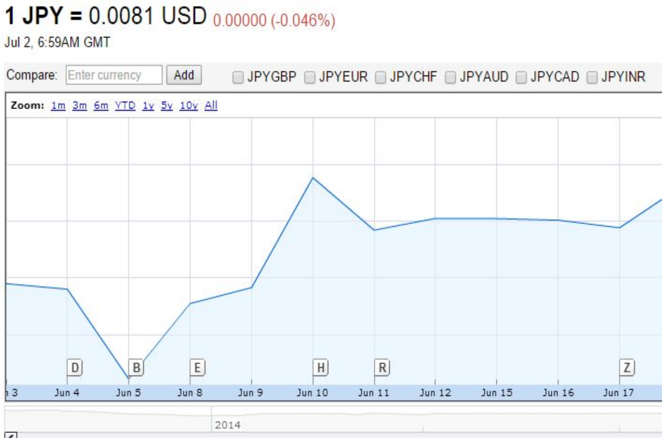
- Embedded serialization formats in HTML
  - JSON-LD
  - RDFa



```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Organization",
  "url": "https://attiks.com",
```

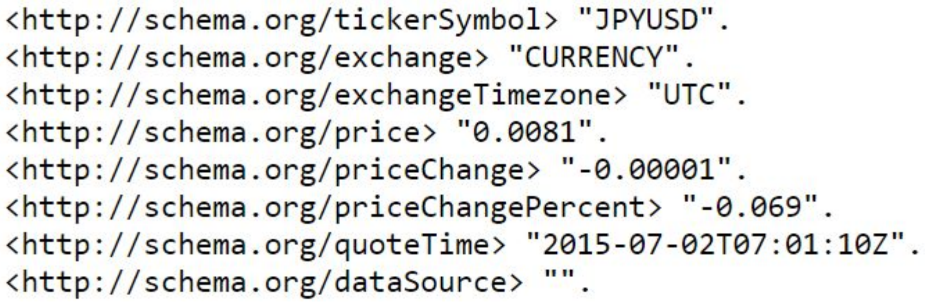
Microdata

- Not an RDF serialization but allows structured data in HTML5.
- Utilised by Google and other search engines to produce, for example, rich snippets in search results.



1 JPY = 0.0081 USD 0.00000 (-0.046%)

Jul 2, 6:59AM GMT

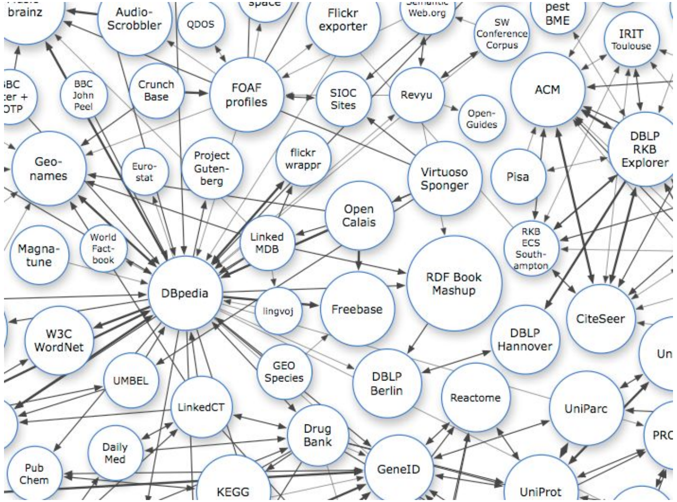


```
<http://schema.org/tickerSymbol> "JPYUSD".
<http://schema.org/exchange> "CURRENCY".
<http://schema.org/exchangeTimezone> "UTC".
<http://schema.org/price> "0.0081".
<http://schema.org/priceChange> "-0.00001".
<http://schema.org/priceChangePercent> "-0.069".
<http://schema.org/quoteTime> "2015-07-02T07:01:10Z".
<http://schema.org/dataSource> "".
```

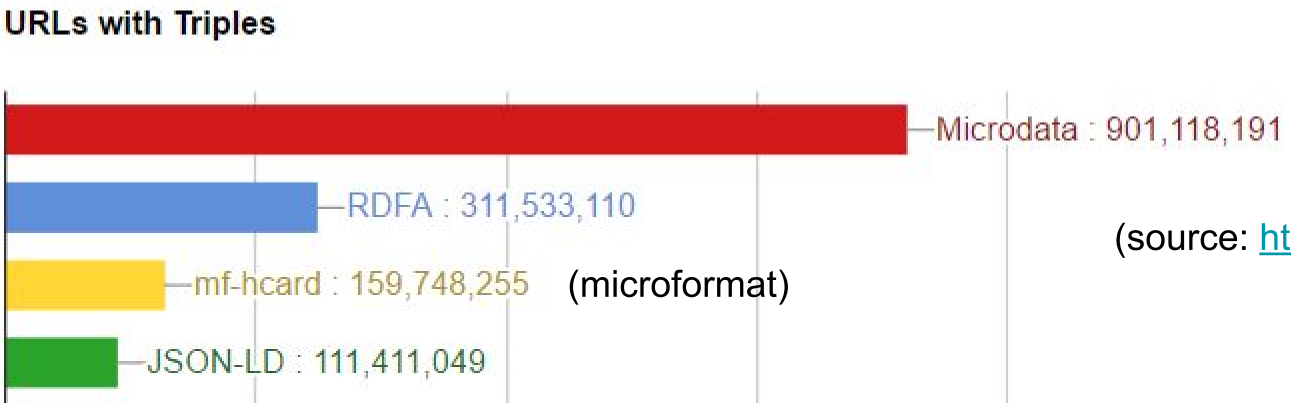
Availability

SPARQL endpoints

http://sparqles.ai.wu.ac.at/
SPARQL Endpoint Monitoring
(>550) endpoints



Embedded Structured Data (38% of pages)



(microformat)

(source: http://webdatacommons.org)

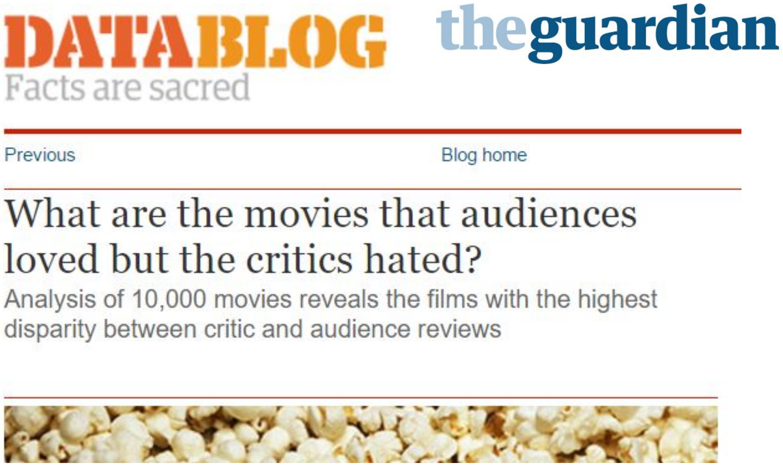
Utilisation - recent experience

- Fact checking application using open data about movies

The idea is that claims can be broken down and semi-automatic fact checking by answering questions such as:

- According to which sources are controversial films preferred by critics?
  - Does that change over time?
- According to which sources are Micheal Bay films a box office success?
- What makes a movie controversial?
- Are attitudes to LGBT films changing over time?

"*Attitudes towards LGBT films are changing due as gay looses its edge due to wider societal acceptance.*" http://www.hollywoodreporter.com/news/critics-notebook-hollywoods-big-queer-842638



DATABLOG
Facts are sacred

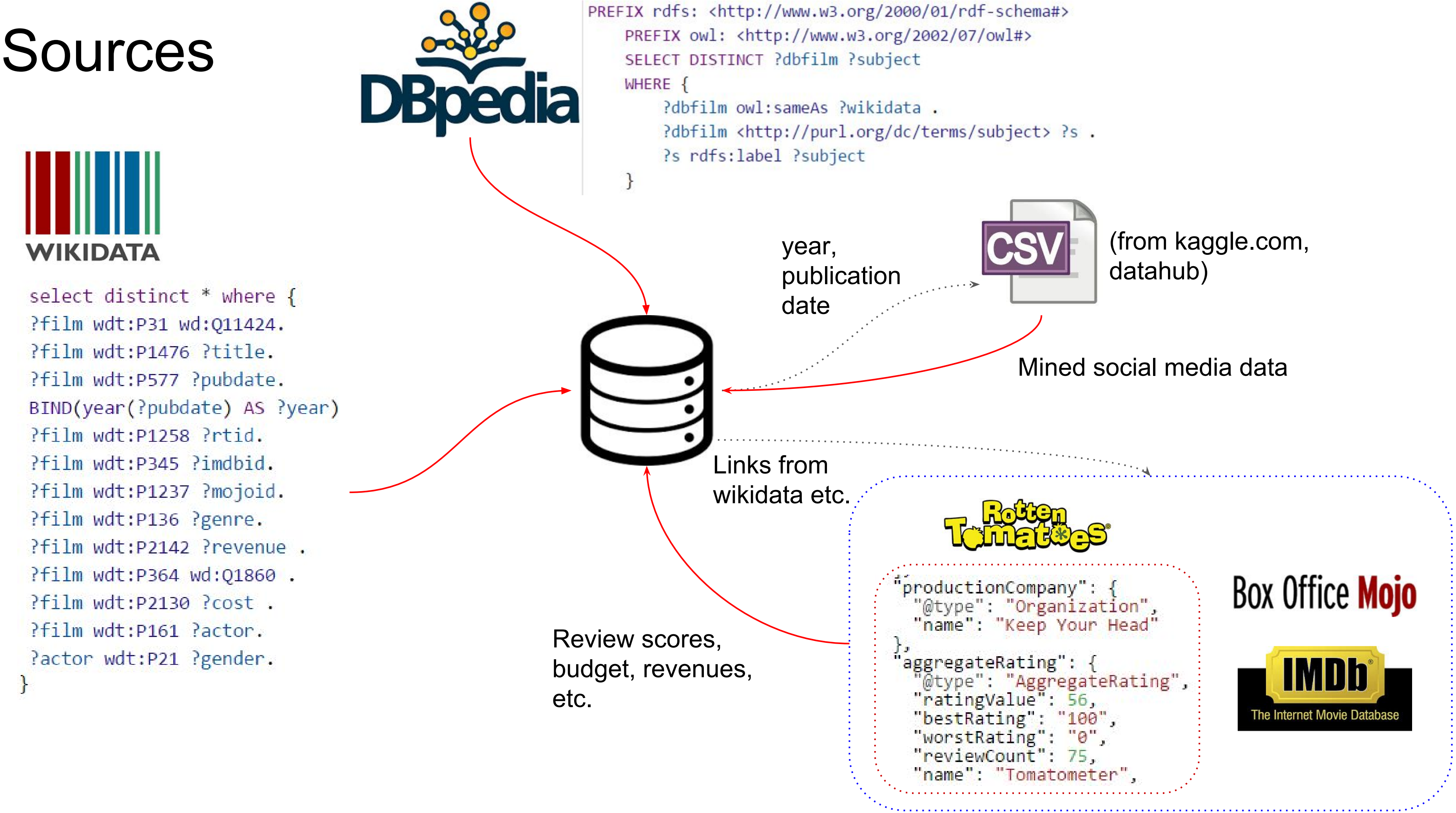
theguardian

Previous Blog home

What are the movies that audiences loved but the critics hated?

Analysis of 10,000 movies reveals the films with the highest disparity between critic and audience reviews

Sources



```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
    PREFIX owl: <http://www.w3.org/2002/07/owl#>
    SELECT DISTINCT ?dbfilm ?subject
    WHERE {
        ?dbfilm owl:sameAs ?wikidata .
        ?dbfilm <http://purl.org/dc/terms/subject> ?s .
        ?s rdfs:label ?subject
    }
```

```
select distinct * where {
?film wdt:P31 wd:Q11424.
?film wdt:P1476 ?title.
?film wdt:P577 ?pubdate.
BIND(year(?pubdate) AS ?year)
?film wdt:P1258 ?rtid.
?film wdt:P345 ?imdbid.
?film wdt:P1237 ?mojoid.
?film wdt:P136 ?genre.
?film wdt:P2142 ?revenue .
?film wdt:P364 wd:Q1860 .
?film wdt:P2130 ?cost .
?film wdt:P161 ?actor.
?actor wdt:P21 ?gender.
}
```

year, publication date

(from kaggle.com, datahub)

Mined social media data

Links from wikidata etc.

Review scores, budget, revenues, etc.

```
"productionCompany": {
  "@type": "Organization",
  "name": "Keep Your Head"
},
"aggregateRating": {
  "@type": "AggregateRating",
  "ratingValue": 56,
  "bestRating": "100",
  "worstRating": "0",
  "reviewCount": 75,
  "name": "Tomatometer",
```

Method

- Query Wikidata
  - Use the SPARQL query interface
  - Formed the bulk of the data and well linked to other sources
- Use links to RottenTomatoes, IMDB, BoxOfficeMojo
  - Tried structured data extraction tools
    - Any23 (not robust to errors, Google structured data tool not available as API)
    - BeautifulSoup (scraping tool)
      - Needed website-specific scripts
- Extracted movie categories/subcategories from DBpedia
- Further data from CSV files, e.g. from kaggle.com

Example data about a movie

```
+numberOfLikes("Christian Bale",23000)`TIME_PROV("2012","2017","http://facebook.com")

+appearsIn("The Dark Knight Rises","Christian Bale")`TIME_PROV("2012","2017","http://imdb.com")

+appearsIn("Terminator Salvation","Christian Bale")`TIME_PROV("2009","2017","http://imdb.com")

+budget("Terminator Salvation",200000000)`TIME_PROV("2009","2017","http://imdb.com")

+criticRating("Terminator Salvation",33)`TIME_PROV("2009","2017","http://rottentomatoes.com")

+criticRating("Terminator Salvation",54)`TIME_PROV("2009","2017","http://rottentomatoes.com")
```

Problems faced

- Writing the queries is difficult
  - Trial and error process
  - Usage restrictions of endpoints
- Messy data
  - Ended up using web scraping tools
  - Making sure all the data is relevant
- A lot of the data you want might not be readily available
  - Some of the data was obtained from downloaded CSV files, manually extracted the data
- In practice, the process required a lot of scripts and fiddling etc.
- How to automate such a process as much as possible

Relevant past works

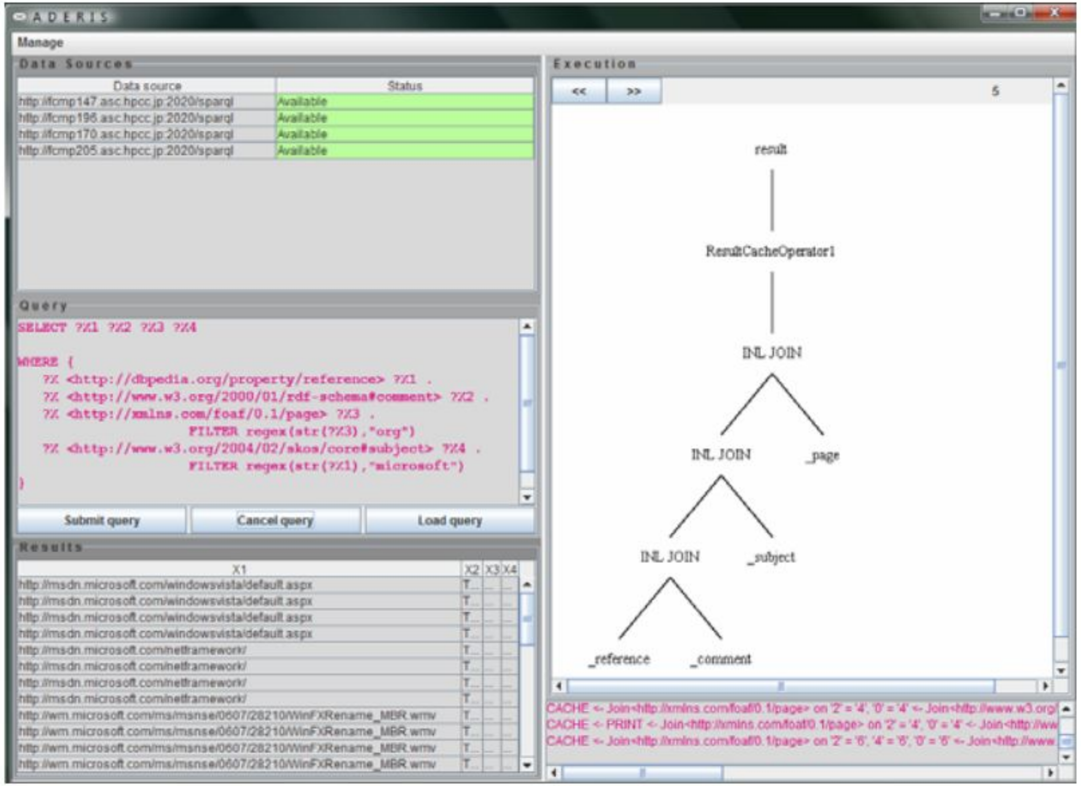
- Distributed query processing over SPARQL endpoints

- Hybrid distributed RDF query processing

- Optimising user criteria during active discovery of RDF data

Adaptive distributed query processing over SPARQL endpoints

- Execution of queries over multiple endpoints
- Adaptive query processing
  - Change the query plan during execution based on properties of the data
  - Adapt to characteristics of the services being accessed, e.g. usage restrictions, speed etc.
- How many endpoints really useful?
- Query writing still challenging



Data source	Status
http://fcmp147.asc.hpcc.jp:2020/sparql	Available
http://fcmp198.asc.hpcc.jp:2020/sparql	Available
http://fcmp170.asc.hpcc.jp:2020/sparql	Available
http://fcmp205.asc.hpcc.jp:2020/sparql	Available

```
SELECT ?X1 ?X2 ?X3 ?X4

WHERE {
   ?X <http://dbpedia.org/property/reference> ?X1 .
   ?X <http://www.w3.org/2000/01/rdf-schema#comment> ?X2 .
   ?X <http://xmlns.com/foaf/0.1/page> ?X3 .
                FILTER regex(str(?X3),"org")
   ?X <http://www.w3.org/2004/02/skos/core#subject> ?X4 .
                FILTER regex(str(?X1),"microsoft")
}
```

X1	X2	X3	X4
http://msdn.microsoft.com/windowsvista/default.aspx	T...	...	...
http://msdn.microsoft.com/windowsvista/default.aspx	T...	...	...
http://msdn.microsoft.com/windowsvista/default.aspx	T...	...	...
http://msdn.microsoft.com/windowsvista/default.aspx	T...	...	...
http://msdn.microsoft.com/netframework/	T...	...	...
http://msdn.microsoft.com/netframework/	T...	...	...
http://msdn.microsoft.com/netframework/	T...	...	...
http://msdn.microsoft.com/netframework/	T...	...	...
http://wm.microsoft.com/ms/msnse/0607/28210/WinFXRename_MBR.wmv	T...	...	...
http://wm.microsoft.com/ms/msnse/0607/28210/WinFXRename_MBR.wmv	T...	...	...
http://wm.microsoft.com/ms/msnse/0607/28210/WinFXRename_MBR.wmv	T...	...	...
http://wm.microsoft.com/ms/msnse/0607/28210/WinFXRename_MBR.wmv	T...	...	...

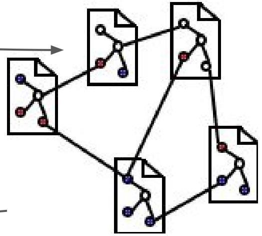
```
CACHE <- Join<http://xmlns.com/foaf/0.1/page> on '2' = '4', '0' = '4' <- Join<http://www.w3.org/
CACHE <- PRINT <- Join<http://xmlns.com/foaf/0.1/page> on '2' = '4', '0' = '4' <- Join<http://ww
CACHE <- Join<http://xmlns.com/foaf/0.1/page> on '2' = '6', '4' = '6', '0' = '6' <- Join<http://www.
```

Active discovery

```
SELECT DISTINCT *
WHERE {
        ?paper <http://data.semanticweb.org/ns/swc/ontology#isPartOf>
                <http://data.semanticweb.org/conference/iswc/2008/proceedings> .
        ?paper <http://swrc.ontology.org/ontology#author> ?p .
        ?p rdfs:label ?n .
}
```

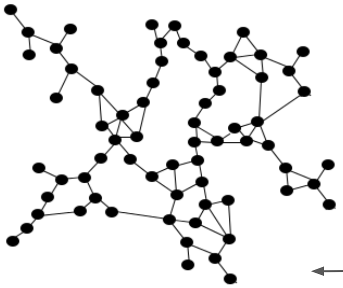
1 Initial dereferencing

e.g. http://data.semanticweb.org/conference/iswc/2008/proceedings is dereferenced and RDF data obtained.



Partial answer

Contains RDF matched against triple patterns, used to answer the query.



triple pattern matching

2 Iterative dereferencing

IRIs are repeatedly selected, dereferenced and matching triples added to the local graph. The focus of this paper is how to select which IRIs to dereference from a potentially huge number

e.g. http://conference:iswc/2008/paper/37 (subject)

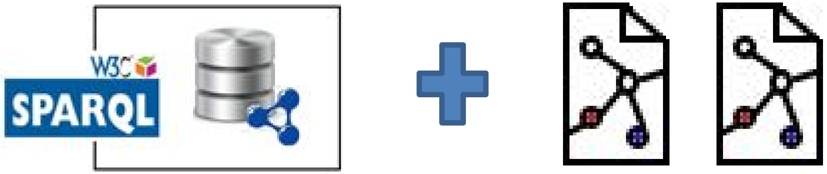
|

isPartOf (predicate)

|

http://data.semanticweb.org/conference/iswc/2008/proceedings (object)

Hybrid



SPARQL Endpoints

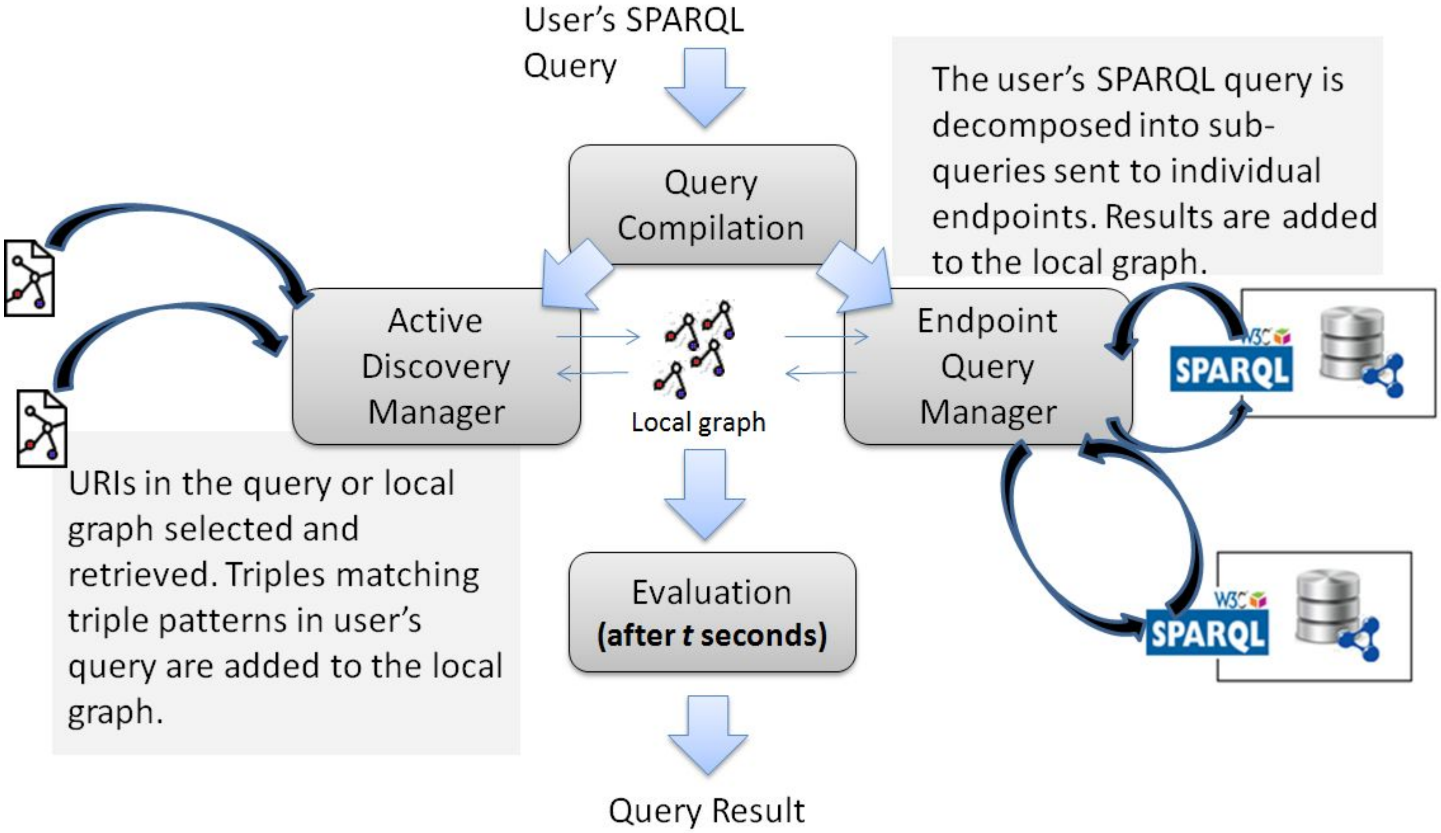
RDFa

RDF/XML,

- Increased coverage
- Freshness
- Mitigating usage restrictions

- Using SPARQL endpoints and Web documents (RDF/XML etc.) during query processing
- Web documents found by active discovery
    - Dereferencing URIs on-the-fly
- Potentially useful in a fact checking context to increase coverage

Hybrid query processing

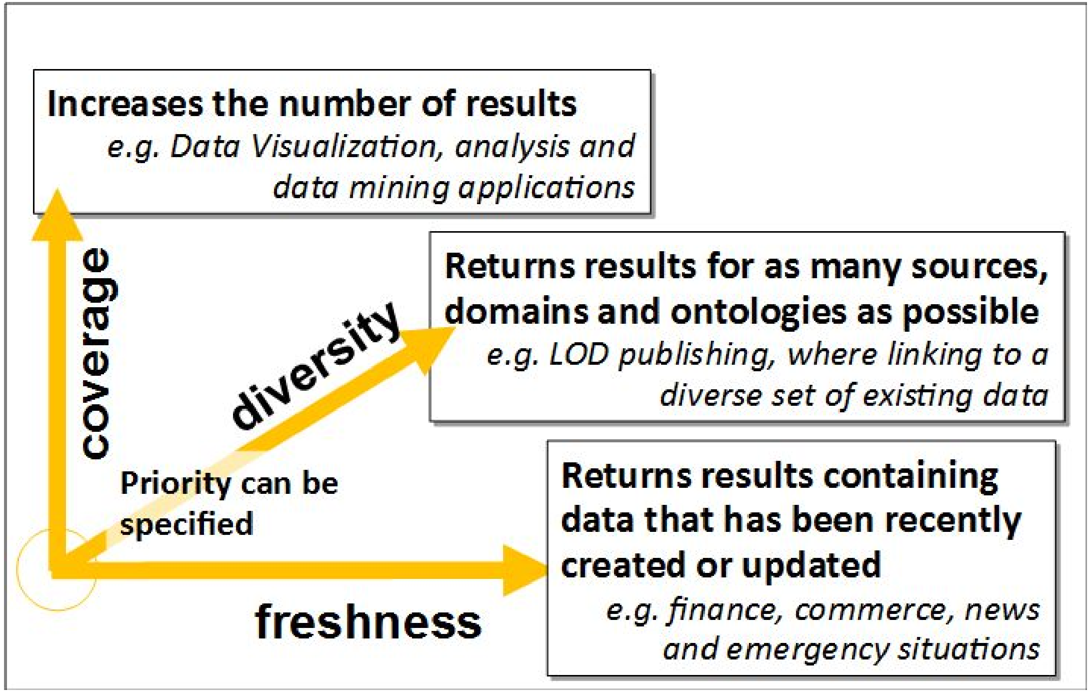


The user's SPARQL query is decomposed into sub-queries sent to individual endpoints. Results are added to the local graph.

URIs in the query or local graph selected and retrieved. Triples matching triple patterns in user's query are added to the local graph.

Optimisation of user criteria during active discovery

- Develop optimization techniques for common application/user requirements
  - **Time constraints: best-effort query processing** – optimization techniques for returning results within a time limit
  - **User criteria: coverage, freshness, diversity** – concepts from Information Retrieval (IR) optimized based on user requirements; simplify query construction



Increases the number of results
e.g. Data Visualization, analysis and data mining applications

Returns results for as many sources, domains and ontologies as possible
e.g. LOD publishing, where linking to a diverse set of existing data

Returns results containing data that has been recently created or updated
e.g. finance, commerce, news and emergency situations

Priority can be specified

Conclusions

- Aim to reuse previous work to solve some of the issues in finding relevant data for fact checking applications
- Some issues
    - Structured data e.g. Web Data Commons (Common Crawl Corpus)
        - How much of it is fit for purpose from a fact checking perspective?
    - Wikidata is probably an excellent starting point for many applications
        - Well linked to many different sources
    - How to find other relevant endpoints and data sources is an important problem
    - Once found, knowing how to query them