ODCleanStore

Jan Michelfeit
michelfeit.jan@gmail.com

Tomáš Knap, Dušan Rychnovský, Jakub Daniel, Petr Jerman, Tomáš Soukup

Faculty of Mathematics and Physics
Charles University in Prague

Outline

- Motivation – how to get clean data?
- What is ODCleanStore
- Examples of data processing
- Example of querying the data

Motivation

CLEAN DATA

Integration

Trust

Provenance

ODCleanStore

Motivation

- OpenData.cz
    - Initiative for transparent data infrastructure
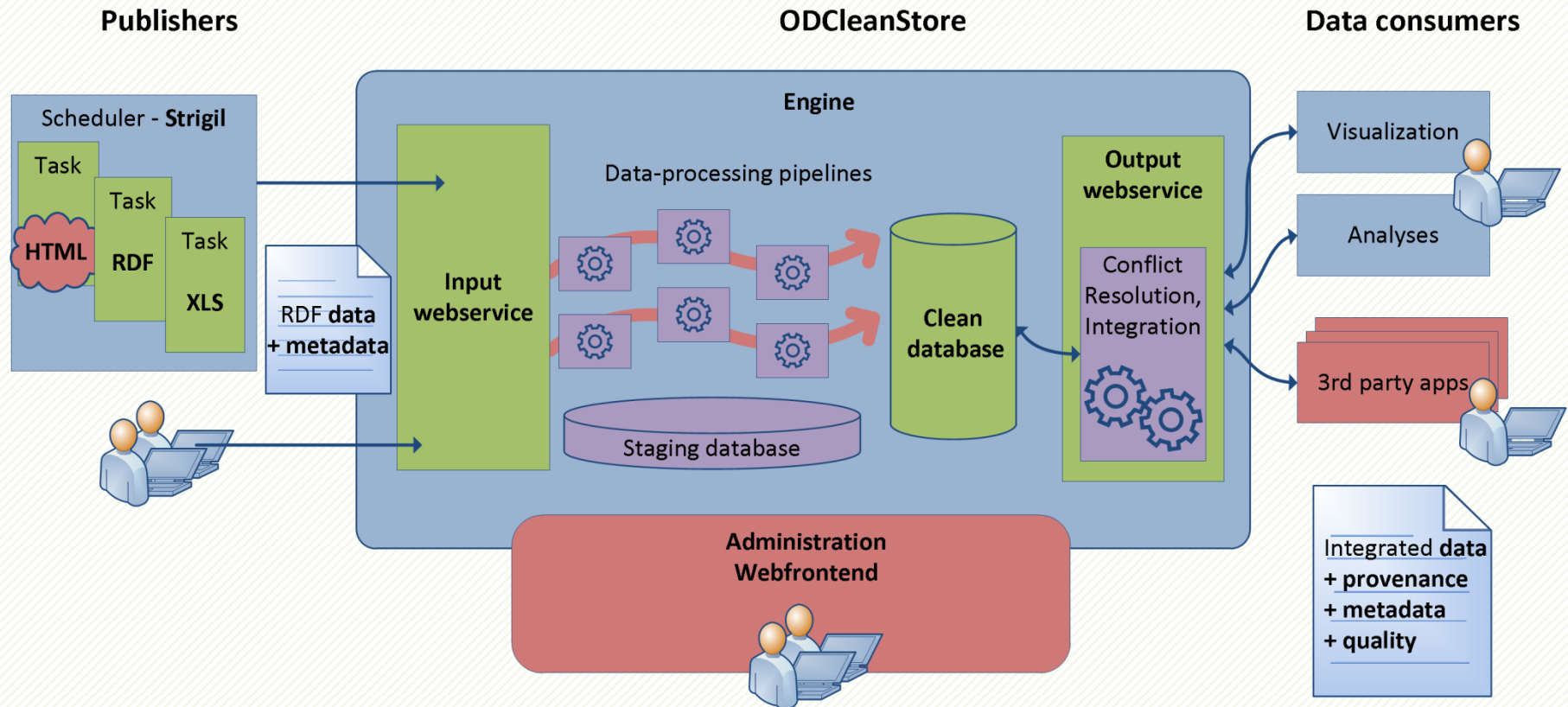   - Public procurement use case

1. Extract
2. Process
3. Store
4. Publish

ODCleanStore

- Tool for management of Linked Data
- Server application, Open Source (Java)
- Data
  - processing
    - customizable, mutliple data sources
  - storage
  - access, integrated views on data
    - standard technologies, REST, RDF, SPARQL endpoint
- Web administration interface

ODCleanStore & other projects

Example

Example

1. Define extraction scripts
2. Create pipelines
   I. Cleaner (Data Normalization)
   II. Quality Assessment
   III. Linker
   IV. … custom transformers
3. Start sending data
4. Clean database
5. Browse the data

Transformer: Cleaner (Normalization)

- Basic cleaning operations, e.g.
  - converting date formats
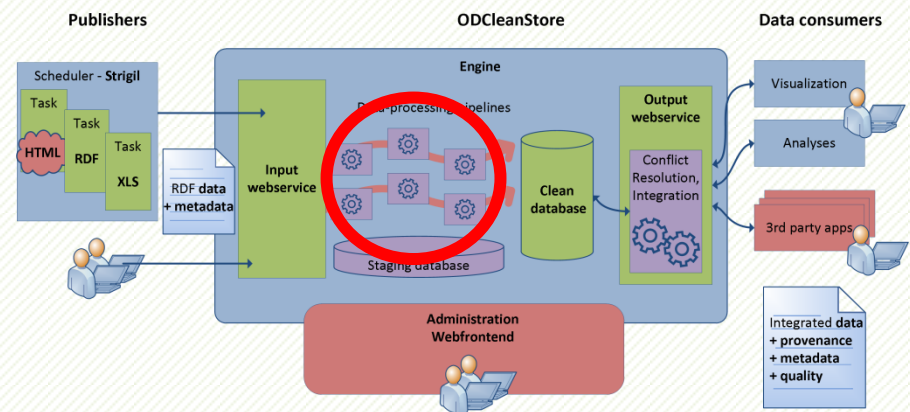 - merging object values from two properties
  - removal of invalid values
- Configured by user-defined rules
  - SPARUL based
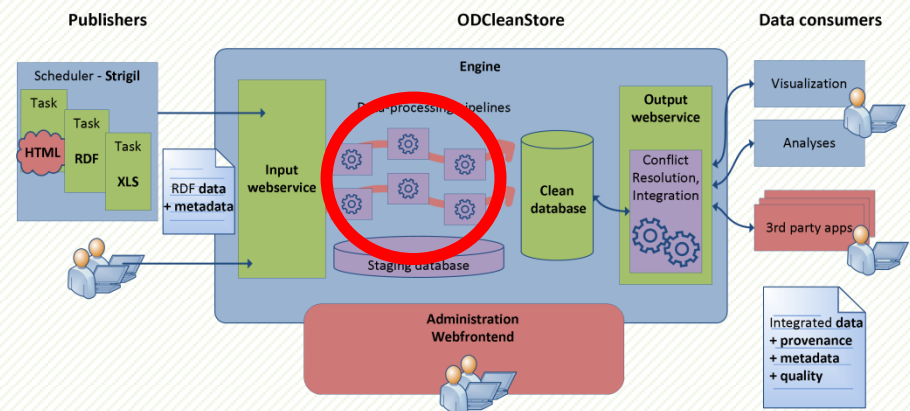
Transformer: Linker

- Creation of links between entities
  - e.g. `owl:sameAs` link between business entities based on their identification number
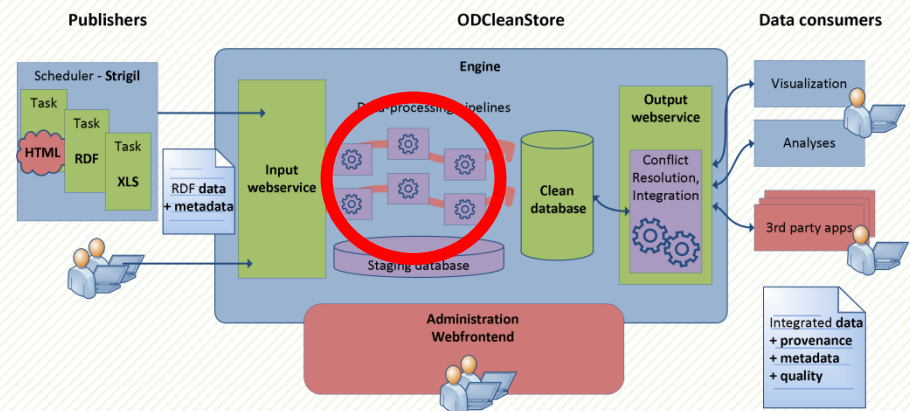 - link cities based on their name and location

- Based on Silk

Transformer: Quality Assessment

- Computes data score based on user policies, e.g.
    - email value has a correct format
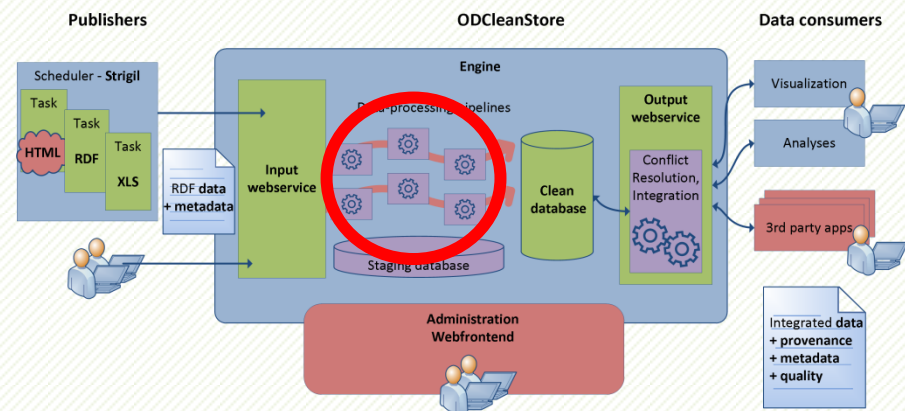   - value of `pc:awardDate` is before `pc:publicationDate`

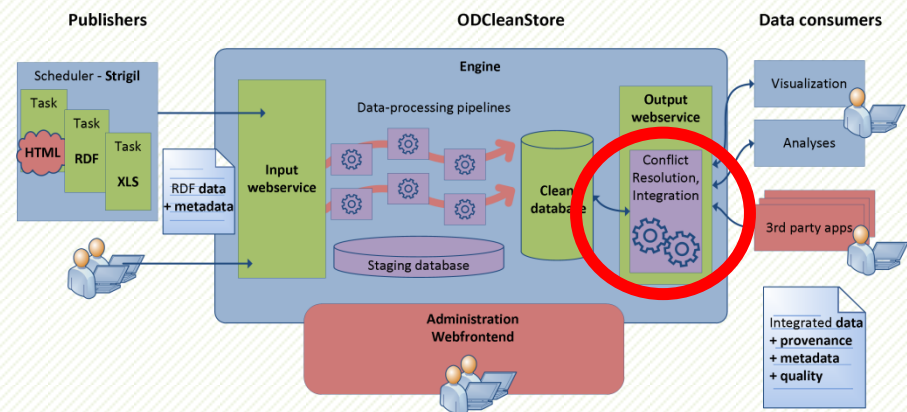- SPARQL based

Custom transformers

□ You can write your own,

e.g. checker of identification number in the Business Register

Conflict Resolution

- Resolution of conflicts at query time
  - based on user policies
  - aggregation – e.g. average geographic location
- Transparently resolves sameAs links and vocabulary mappings
- Provenance tracking, quality scores

Conflict Resolution

Conflict Resolution

Conflict Resolution

Conflict Resolution

URI query for <http://dbpedia.org/resource/Berlin>. Query executed in 0.569 s.

Subject	Predicate	Object	Quality	Source named graphs
dbpedia:Berlin	dbo:country	dbpedia:Germany	0.90000	http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/dbpedia
dbpedia:Berlin	dbo:populationTotal	"3420768"^^xsd:integer	0.79598	http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/linkedgeodata
dbpedia:Berlin	http://linkedgeodata.org/property/capital	"yes"	0.80000	http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/linkedgeodata
dbpedia:Berlin	freebase:location.geocode.latitude	"44.695598392734375"^^xsd:double	0.55381	http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/error, http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/dbpedia, http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/linkedgeodata, http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/freebase, http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/geonames
dbpedia:Berlin	freebase:location.geocode.longtitude	"13.402740096987914"^^xsd:double	0.82446	http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/linkedgeodata, http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/dbpedia, http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/geonames, http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/freebase
dbpedia:Berlin	rdf:type	http://schema.org/City	0.92000	http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/dbpedia, http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/freebase
dbpedia:Berlin	rdf:type	http://schema.org/Place	0.90000	http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/dbpedia
dbpedia:Berlin	rdfs:label	"Berlin"	0.94252	http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/geonames, http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/dbpedia, http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/freebase, http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/geonames, http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/linkedgeodata
dbpedia:Berlin	rdfs:label	"Berlino"@it	0.56970	http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/linkedgeodata

Source graphs:

Named graph	Data source	Inserted at	Graph score	License	Update tag
http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/dbpedia	http://dbpedia.org/page/Berlin	2012-04-01 12:34:56.0	0.9		
http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/error	http://example.com		0.8		
http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/freebase	http://www.freebase.com/view/en/berlin	2012-04-02 12:34:56.0	0.8		
http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/geonames	http://www.geonames.org/2950159/berlin.html	2012-04-03 12:34:56.0	0.8		
http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/linkedgeodata	http://linkedgeodata.org/page/node240109189	2012-04-04 12:34:56.0	0.8		

Conclusion

- ❑ ODCleanStore

- ❑ Related projects

- ❑ Public procurement use case

- ❑ Why ODCleanStore

  - ▪ standard technologies, Linked Data

  - ▪ Open Source server application, administration from your web browser

  - ▪ quality & provenance tracking

  - ▪ conflict resolution customized to your query

Thank you

... Questions?

Download:
http://goo.gl/49V1L
http://sourceforge.net/p/odcleanstore/

odcleanstore-user@lists.sourceforge.net

Faculty of Mathematics and Physics
Charles University in Prague