

```
<xs:complexType name="CategoryType">
```

```
<xs:sequence>
```

```
<xs:element name="description" type="xs:string" />
```

```
<xs:element name="category" type="CategoryType"  
minOccurs="0" maxOccurs="unbounded"/>
```

```
<xs:element name="books">
```

```
<xs:sequence>
```

```
<xs:element name="book" type="BookType"  
minOccurs="0" maxOccurs="unbounded"/>
```

```
</xs:sequence>
```

```
</xs:complexType>
```

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

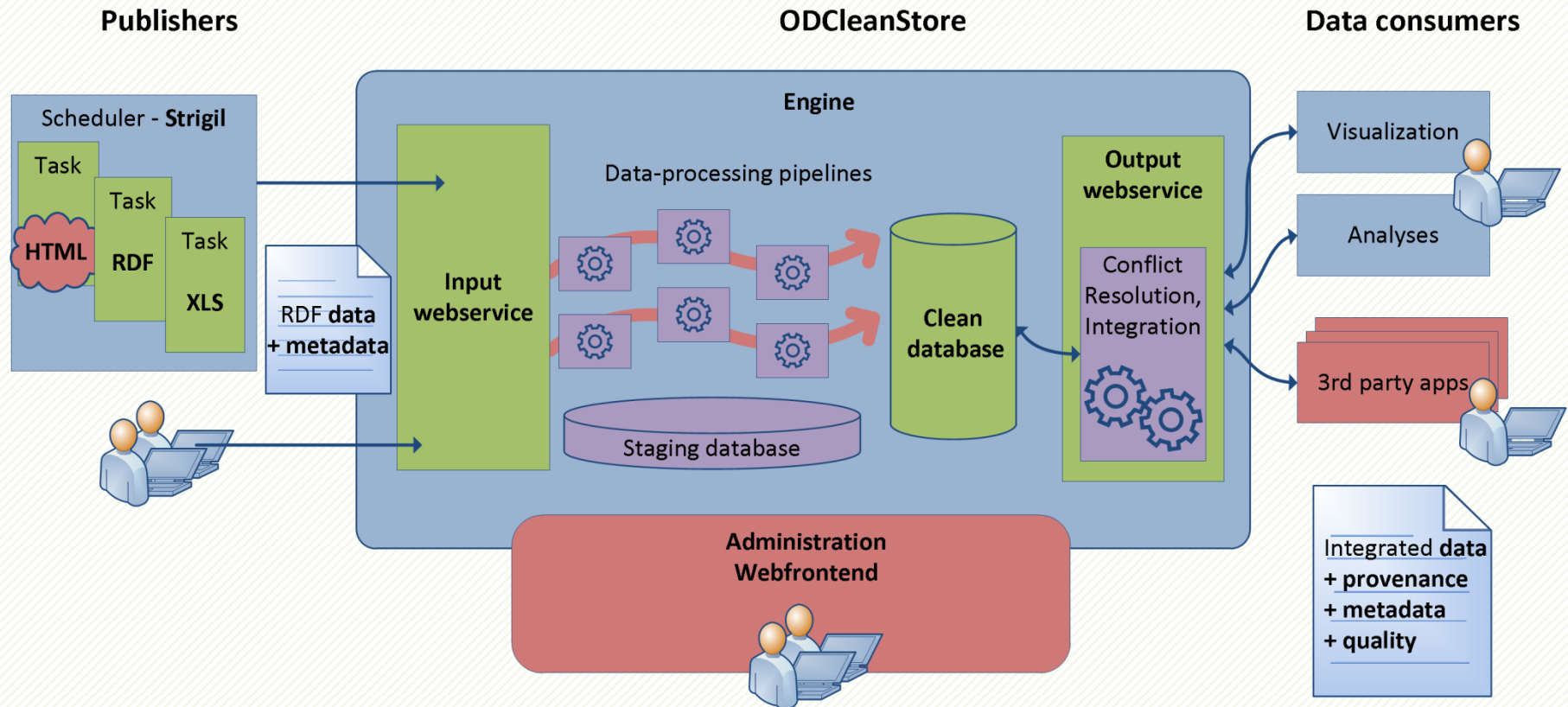
- 1. Extract
- 2. Process
- 3. Store
- 4. Publish

ODCleanStore



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

ODCleanStore & other projects



Example

ODCleanStore - Administration

Home Pipelines Rules ▾ Engine ▾ Output webservice ▾ Ontologies ▾

Home

Welcome to ODCleanStore Administration

Welcome to administration of ODCleanStore, a Linked Data management system.

ODCleanStore

- accepts, processes and stores RDF data;
- makes data processing highly customizable;
- provides predefined transformers for data processing;
- provides integrated views on stored data;
- supports data provenance tracking and quality estimation;
- uses standard technologies in order to make integration with other applications.

For more information, visit the [official website](#) or consult user manual ([draft](#)).

ODCleanStore - Administration

Home Pipelines Rules ▾ Engine ▾ Output webservice ▾ Ontologies ▾ Accounts ▾ Transformers Prefixes

Home > Backend > Pipelines > Edit

User: adm Roles: PIC, ADM, ONC [My Account](#) [Log out](#)

Edit a pipeline

[Back to the list of pipelines](#)

[Help](#)

Label:

test-pipeline *

Description:

description

[Submit](#)

Is default:

No

Assigned transformers

[Assign a transformer](#)

[View graphs in error](#)

[Help](#)

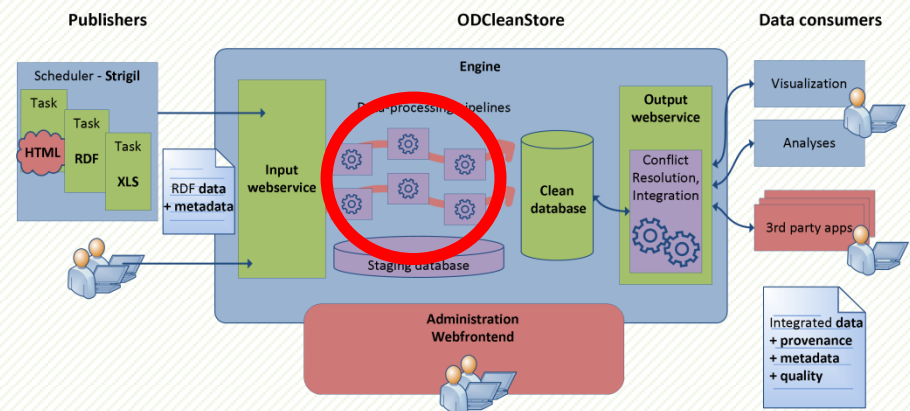
Order ▲	Label	Configuration	Allow to be run on clean DB	Action			
1	Blank node remover		No	Detail	Up	Down	Delete
2	Data Normalization		Yes	Detail	Up	Down	Delete
3	Quality Assessment		Yes	Detail	Up	Down	Delete
4	Linker	linkWithinGraph=false	Yes	Detail	Up	Down	Delete
5	Quality Aggregator		Yes	Detail	Up	Down	Delete

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)

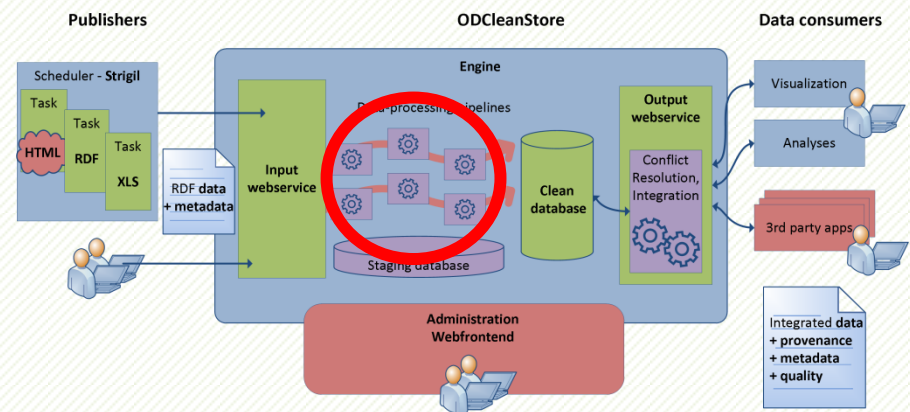
- ❑ Basic cleaning operations, e.g.
 - converting date formats
 - merging object values from two properties
 - removal of invalid values
- ❑ Configured by user-defined rules
 - SPARUL based



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

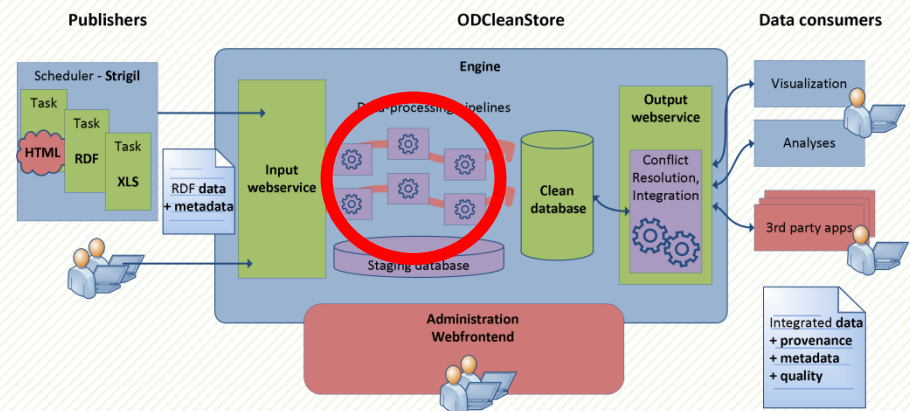
- Based on Silk



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`

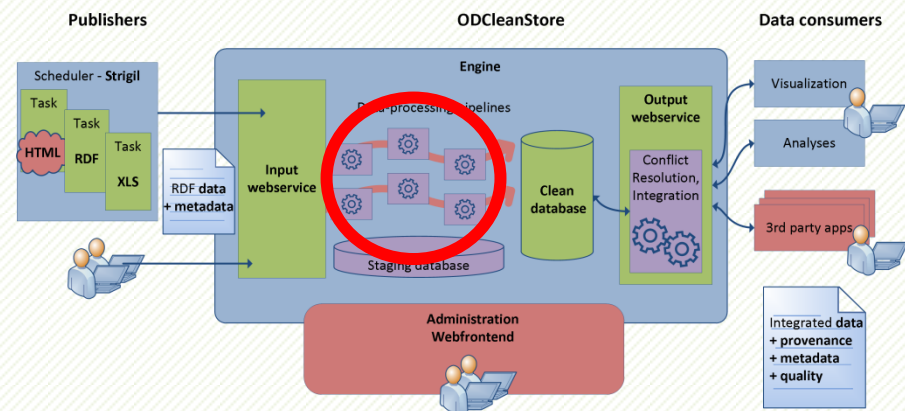
- ❑ SPARQL based



Custom transformers

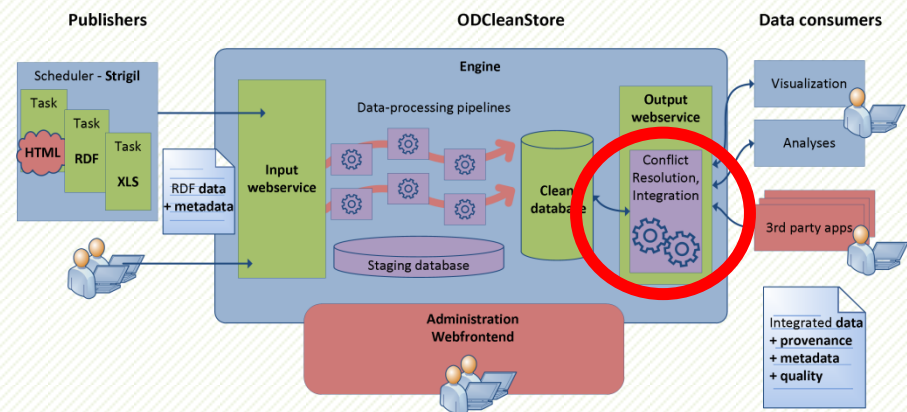
- You can write your own,

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

ODCleanStore - URI query

Server address:	<input type="text" value="localhost:8087"/>	
Searched URI:	<input type="text" value="dbpedia:Berlin"/>	
Default aggregation:	<input type="text" value="ALL"/>	
Default multivalue:	<input type="text" value="NO"/>	
Aggregation error strategy:	<input type="text" value="RETURN_ALL"/>	
Property aggregation	<input type="text" value="freebase:location.geocode.latitude"/>	<input type="text" value="AVG"/>
Property aggregation	<input type="text" value="freebase:location.geocode.longitude"/>	<input type="text" value="AVG"/>
Property aggregation	<input type="text" value="dbo:populationTotal"/>	<input type="text" value="LATEST"/>
Property multivalue	<input type="text" value="http://www.w3.org/1999/02/22-rdf-syntax-ns#type"/>	<input type="text" value="YES"/>
Property multivalue	<input type="text"/>	<input type="text" value="NO"/>
Property multivalue	<input type="text"/>	<input type="text" value="NO"/>
Output format:	<input type="text" value="HTML"/>	
	<input type="button" value="Submit"/>	

If you cannot connect to the server, make sure you have ODCleanStore Engine running.

Conflict Resolution

ODCleanStore - URI query

Server address:	<input type="text" value="localhost:8087"/>
Searched URI:	<input type="text" value="dbpedia:Berlin"/>
Default aggregation:	<input type="text" value="ALL"/>
Default multivalue:	<input type="text" value="NO"/>
Aggregation error strategy:	<input type="text" value="RETURN_ALL"/>
Property aggregation	<input type="text" value="freebase:location.geocode.latitude"/> <input type="text" value="AVG"/>
Property aggregation	<input type="text" value="freebase:location.geocode.longitude"/> <input type="text" value="AVG"/>
Property aggregation	<input type="text" value="dbo:populationTotal"/> <input type="text" value="LATEST"/>
Property multivalue	<input type="text" value="http://www.w3.org/1999/02/22-rdf-syntax-ns#type"/> <input type="text" value="YES"/>
Property multivalue	<input type="text" value=""/> <input type="text" value="NO"/>
Property multivalue	<input type="text" value=""/> <input type="text" value="NO"/>
Output format:	<input type="text" value="HTML"/>
<input type="button" value="Submit"/>	

If you cannot connect to the server, make sure you have ODCleanStore Engine running.

Conflict Resolution

ODCleanStore - URI query

Server address:	<input type="text" value="localhost:8087"/>	
Searched URI:	<input type="text" value="dbpedia:Berlin"/>	
Default aggregation:	<input type="text" value="ALL"/>	
Default multivalue:	<input type="text" value="NO"/>	
Aggregation error strategy:	<input type="text" value="RETURN_ALL"/>	
Property aggregation	<input type="text" value="freebase:location.geocode.latitude"/>	<input type="text" value="AVG"/>
Property aggregation	<input type="text" value="freebase:location.geocode.longitude"/>	<input type="text" value="AVG"/>
Property aggregation	<input type="text" value="dbo:populationTotal"/>	<input type="text" value="LATEST"/>
Property multivalue	<input type="text" value="http://www.w3.org/1999/02/22-rdf-syntax-ns#type"/>	<input type="text" value="YES"/>
Property multivalue	<input type="text"/>	<input type="text" value="NO"/>
Property multivalue	<input type="text"/>	<input type="text" value="NO"/>
Output format:	<input type="text" value="HTML"/>	
	<input type="button" value="Submit"/>	

If you cannot connect to the server, make sure you have ODCleanStore Engine running.

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/linkedgedata
dbpedia:Berlin	http://linkedgedata.org/property/capital	"yes"	0.80000	http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/linkedgedata
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/linkedgedata , 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.longitude	"13.402740096987914" ^^xsd:double	0.82446	http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/linkedgedata , 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/linkedgedata
dbpedia:Berlin	rdfs:label	"Berlino"@it	0.56970	http://odcs.mff.cuni.cz/namedGraph/qe-test/berlin/linkedgedata

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/linkedgedata	http://linkedgedata.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

```
<xs:complexType name="CategoryType">
```

```
<xs:sequence>
```

```
<xs:element name="description" type="xs:string" />
```

```
<xs:element name="category" type="CategoryType"
```

```
minOccurs="0" maxOccurs="unbounded"/>
```

```
<xs:element name="books">
```

```
<xs:complexType>
```

```
<xs:sequence>
```

```
<xs:element name="book" type="BookType"
```

```
minOccurs="0" maxOccurs="unbounded"/>
```

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

