



# ODCleanStore

ETL Tool for RDF Data

## Tomas Knap

Charles University in Prague,  
Department of Software Engineering,  
XML and Web Engineering Research Group

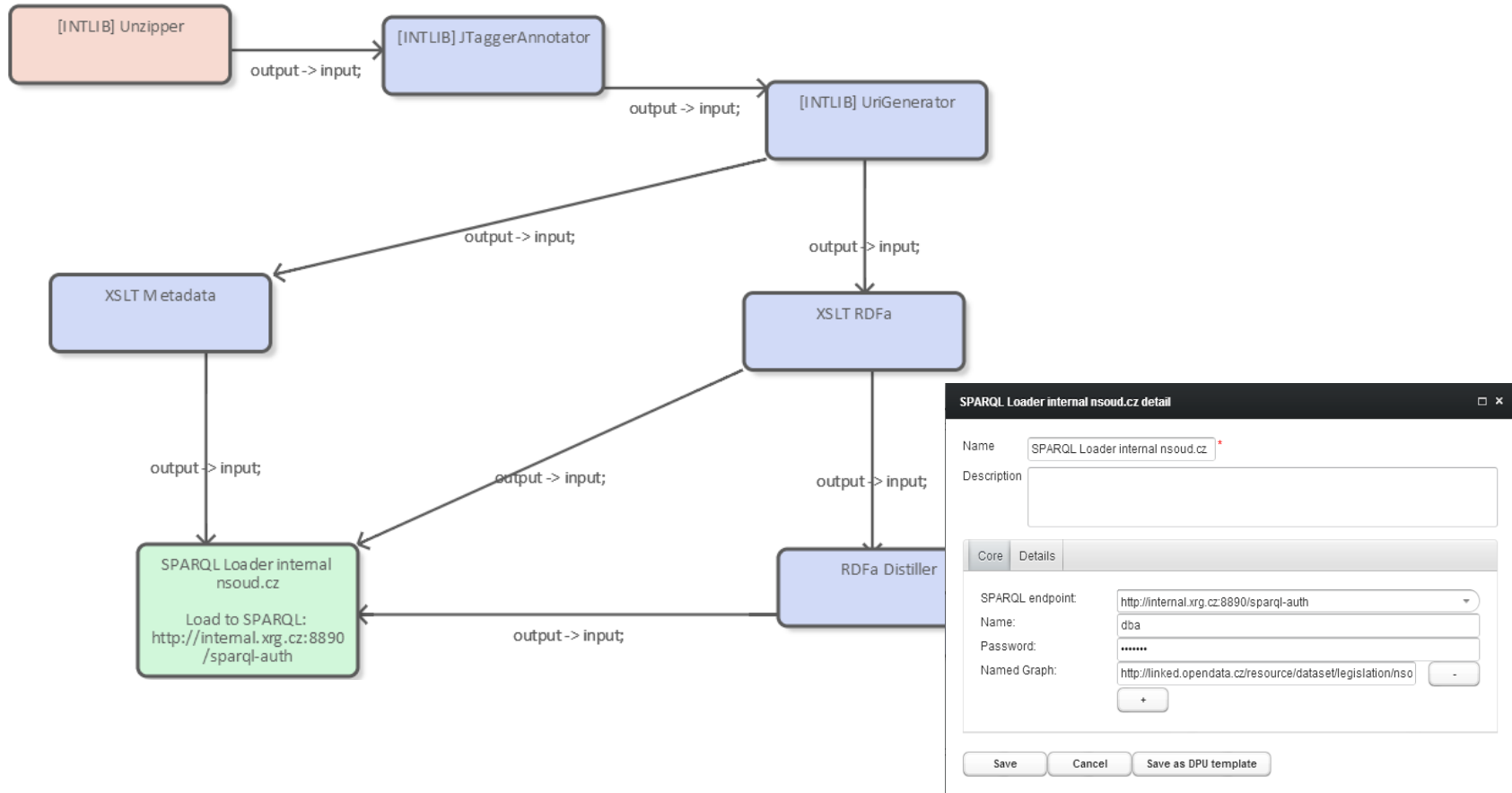
*The COMSODE project has received funding from the Seventh Framework Programme of the European Union in the grant agreement number 611358.*



# Motivation

- Tools available for RDF data extraction, enrichment, linking, transforming, ...
- No tool for flexible management of RDF data extraction/transformation **tasks**
  - task = progression of data processing units (DPUs)
    - Extract data from SPARQL Endpoint A
    - Extract data from CSV files B
    - Refine data with SPARQL queries X,Y, Z
    - Deduplicate data using Linker L
    - Publish data to SPARQL Endpoint B

# ETL Task



# ODCleanStore

- Framework for managing ETL tasks and data processing units (DPUs)
- Developed in the EU FP7 Project COMSODE. Based on the previous development effort at Charles University in Prague (with support from Semantic Web Company, Austria)

# Key Features

- Administration interface to create/configure/manage tasks, DPUs
  - Possibility to schedule tasks
    - Time based, chaining of tasks
  - Possibility to debug tasks, browse logs/events produced by DPUs
  - Multi-user environment, possibility to share pipelines, DPUs
- Robust engine running the tasks
- Core DPUs available
  - Easy way how to add your own DPUs

# Demo

- <http://odcs.xrg.cz:8080/odcleanstore/>

The screenshot displays the ODCS Execution Monitor interface. The top navigation bar includes: Pipelines, DPU Templates, Execution Monitor, Browse Data, Scheduler, and Settings. The main area is divided into two panels.

**Left Panel: Pipeline Execution Log**

DATE	NAME	USER	STATUS	DEBUG	OBSOLETE	ACTIONS
Aug 30, 2013 3:18:48 PM	DBpedia		✓	🔥		Show
Aug 30, 2013 3:07:35 PM	DBpedia		✓	🔥		Show
Aug 30, 2013 1:52:30 PM	DBpedia		✓	🔥		Show
Aug 27, 2013 12:01:43 AM	JTagger - nsoud.cz		✗	🔥		Debug
Aug 27, 2013 12:01:39 AM	DBpedia		✓	🔥		Show
Aug 26, 2013 12:58:59 PM	DBpedia		✓	🔥		Show
Aug 25, 2013 9:00:53 AM	ARES downloader 1000		✓	🔥		Show
Aug 23, 2013 9:45:32 AM	Immediate ARES downloader 1000		✓	🔥		Show
Aug 23, 2013 9:44:01 AM	ARES downloader 1000		✓	🔥		Show
Aug 22, 2013 10:10:41 PM	JTagger - nsoud		✗	🔥		Debug
Aug 22, 2013 8:20:54 PM	Buyer profiles		✓	🔥		Show
Aug 22, 2013 7:45:14 PM	DBpedia		✓	🔥		Show
Aug 22, 2013 5:34:27 PM	DBpedia		✓	🔥		Show
Aug 22, 2013 5:33:13 PM	DBpedia		✓	🔥		Show
Aug 22, 2013 5:26:57 PM	DBpedia		✓	🔥		Show
Aug 22, 2013 4:42:50 PM	DBpedia		✓	🔥		Show
Aug 22, 2013 4:42:50 PM	DBpedia		✓	🔥		Show
Aug 22, 2013 4:42:49 PM	Buyer profiles		✗	🔥		Debug
Aug 22, 2013 4:21:53 PM	DBpedia		✓	🔥		Show
Aug 22, 2013 4:20:53 PM	DBpedia		✓	🔥		Show

Page: 1 / 2 >>>

**Right Panel: Job Log Details**

DATE	TYPE	DPU INSTANCE	SHORT MESSAGE
Aug 27, 2013 12:01:44 AM	✓	JTagger one day	Extractor started.
Aug 27, 2013 12:05:00 AM	✓	JTagger one day	Extract completed.
Aug 27, 2013 12:05:02 AM	✓	SPARQL Loader nsoud	Loader started.
Aug 27, 2013 12:05:08 AM	✗	SPARQL Loader nsoud	Loader failed.
Aug 27, 2013 12:05:08 AM	✗	SPARQL Loader nsoud	Pipeline execution failed.

Page: 1 / 1 >>>

Select DPU: [Dropdown]

Buttons: Browse, Log, Query

**INFO+**

DATE	THREAD	LEVEL	SOURCE
Aug 27, 2013 12:01:44 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:45 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:45 AM	pool-2-thread-3	INFO	cz.cuni.xrg.intlib.rdf.impl.VirtuosoRDFRepc
Aug 27, 2013 12:01:45 AM	pool-2-thread-3	INFO	cz.cuni.xrg.intlib.rdf.impl.VirtuosoRDFRepc
Aug 27, 2013 12:01:45 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:45 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:47 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:47 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:47 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:47 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:47 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:47 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:50 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:50 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:50 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac
Aug 27, 2013 12:01:50 AM	pool-2-thread-3	INFO	cz.cuni.mff.xrg.intlib.extractor.jtaggerExtrac

Page: 1 / 55 >>>

Buttons: Close, Export

# COMSODE Project

- EU FP7 project
- Goals: to provide publication platform for publishing (Linked) Open Data, enable searching on top of the published data
  - It will use ODCleanStore for preparing the data
  - Proofed on datasets from various institutions
  - We are just starting, building the user group!
    - Please contact me if your institution would like to participate
  - To follow the news: FP7 COMSODE linkedIn Group
  -

# Conclusions

- ODCleanStore
  - Framework for managing ETL tasks and data processing units (DPUs)
  - Developed in COMSODE project
- Used:
  - in COMSODE Project
  - in Open Data activities in Czech Republic
  - by Semantic Web Company and their customers (planned)
  - LOD2 stack (in preparation)



# Thank You!

- Would you like to try ODCleanStore?
  - <http://www.ksi.mff.cuni.cz/~knap/odcs/>
- Would you like to know more about the tool? Would you like to participate in the user group of COMSODE?
  - [knap@xrg.cz](mailto:knap@xrg.cz)

[www.comsode.eu](http://www.comsode.eu)