DBpedia Data Processing and Integration Tasks in UnifiedViews

Tomas Knap
Semantic Web Company

Markus Freudenberg
Leipzig University

Kay Müller
Leipzig University
Introduction

Agenda, Team
Agenda

Team & Goal

UnifiedViews
  ▶ “An ETL tool for RDF data”

DBpedia
  ▶ Current situation
  ▶ Target solution
  ▶ Role of UnifiedViews
Tomas Knap, PhD

Architect & Researcher
Semantic Web Company

Research interests:
- Linked Data integration and quality
- Linked Data management

Contact:
- tomas.knap@semantic-web.com
Markus Freudenberg
Researcher
AKSW/KILT - Leipzig University
Release Manager DBpedia

Research interests:
- Linked Data as Big Data
- Dataset Metadata
  (Member of W3C Dataset Exchange WG)
- Knowledge Graphs

Contact:
- freudenberg@informatik.uni-leipzig.de
Kay Müller
Researcher
AKSW/KILT - Leipzig University

Research interests:
- Knowledge Graphs
- Information Retrieval
- Big Data Systems

Contact:
- kay.mueller@informatik.uni-leipzig.de
Goal

▸ Improve the process of preparing DBpedia dataset
  ▸ Extraction tasks
  ▸ Transformation tasks
  ▸ Data enrichment tasks

▸ Use UnifiedViews
  ▸ “An ETL tool for RDF data”
UnifiedViews

Introduction
Introduction

- UnifiedViews is an ETL tool for RDF data
  - It differs from other ETL tools by natively supporting RDF data format
- UnifiedViews allows users to manage RDF data processing tasks
  - Extract data from SPARQL Endpoint
  - Download CSV file, convert CSV file to RDF data
  - Refine data with series of SPARQL queries
  - Link/Fuse the data
  - Publish data to a file
Web administration interface
- Define and maintain tasks
- Validate, execute, monitor tasks
- Possibility to schedule tasks
  - Notifications
- Possibility to debug tasks
- Possibility to share tasks and plugins
- Define and maintain plugins
- Multi-user environment, SSO support

Robust engine running the tasks
API to work with tasks, executions, scheduled events
Set of Core plugins available

- **Extractors**
  - Obtaining external sources (CSV, DBF, XLS, XML files, RDF data, or relational tables)

- **Transformers**
  - Transforming them between various formats (e.g. CSV files to RDF data, relational tables to RDF data)
  - Executing typical transformations such as SPARQL Update queries, or XSL transformations

- **Loaders**
  - Loading the transformed and curated data to external systems, repositories

- **35+ plugins**
Easy way to extend UnifiedViews with your own plugins

- Guide for creating new plugins
- Tutorials
UnifiedViews is part of PoolParty Semantic Integrator

A semantic technology suite

- Organize and maintain company’s knowledge base
- Annotate documents with concepts from that knowledge base
- Provide focused search on top of the annotated document space

https://www.poolparty.biz/
UnifiedViews
Availability

▸ Available under an open source license (GPL + LGPL v3)
  ▸ Commercial license also available as part of PoolParty Semantic Integrator

▸ Hosted on GitHub
  ▸ https://github.com/UnifiedViews
UnifiedViews Demo, Resources

- UnifiedViews in 10 minutes
  - [https://www.youtube.com/watch?v=YtF31FyHkQQ](https://www.youtube.com/watch?v=YtF31FyHkQQ)
  - “UnifiedViews”
    - “Data integration with UnifiedViews”

- [http://unifiedviews.eu](http://unifiedviews.eu)
DBpedia
Current Situation and Expected Improvements using UnifiedViews
Knowledge base, which extracts structured information from Wikipedia and makes it available in machine readable form (RDF)

One of the early members (2008) and a major ‘Link Hub’ of the LOD cloud

English version describes over 6 million things (e.g. persons, places, companies, etc.)

Localized versions in 130 languages
Current Situation

Extraction tasks

- Generating RDF triples from Wikipedia’s XML data dumps
- Depends on a specialized extraction framework (in Java/Scala)
  - Needs lots of time and supervision for new releases, lots of manual effort

Transformation tasks

- Canonicalize object URIs
  - Replacing language dependant URIs
- Type consistency check
  - Domains/ranges met
1. A configurable workflow shall replace the current manual tasks
2. Support for data enrichment/fusion tasks
3. Allowing for exact reproductions of a given dataset when using the same input data (needs suitable metadata)
4. Data extractions / transformations must scale
   ▪ up to 10 Billion triples
   ▪ for 130 language editions
UnifiedViews to Address the Requirements

1. UnifiedViews provides a configurable workflow definition
2. UnifiedViews has support for data enrichment/fusion - provides schema mapping/entity linking/data fusion DPUs
3. UnifiedViews needs to produce standardized metadata for tasks
4. UnifiedViews needs to scale for tens of millions of triples
   ▶ Introducing Apache Spark into the UnifiedViews environment
Scalability - Apache Spark

- “A fast and general engine for large scale data processing”
- https://spark.apache.org/
- Identified use case used as a PoC
  - Existing UnifiedViews pipeline
  - Series of SPARQL update queries
- To be able to execute UnifiedViews pipelines/pipeline fragments on top of Apache Spark
  - Allows us to scale and stream data processing tasks executions
We were able to manually prepare the needed Apache Spark transformers for the given pipeline fragment.

Integrated into UnifiedViews - there is a DPU for now, where you can select:
- Spark pipeline to be executed
- Configure Spark environment

SANSA
- [https://github.com/SANSA-Stack](https://github.com/SANSA-Stack)
- Querying arbitrary SPARQL queries using Apache Spark
- Not working smoothly
Simple evaluation - executing SPARQL Update query

CONSTRUCT {?s ?p ?o} WHERE {
    ?s a ?type;
    ?p ?o.
FILTER(?p IN (dc:title, dc:description) &&
    ?type IN (dio:DesignRequirement, dio:DesignIntent, ....) && STRLEN(?o) > 0) }
Summary

▸ UnifiedViews
  ▹ ETL tool for RDF data management
  ▹ Open Source/Part of PoolParty Semantic Integrator
  ▹ UnifiedViews in 10 minutes
    ■ https://www.youtube.com/watch?v=YtF31FyHkQQ
    ▹ http://unifiedviews.eu

▸ DBpedia
  ▹ Plans for preparation of DBpedia in UnifiedViews
Tomas Knap, PhD
Architect & Researcher
Semantic Web Company

Research interests:
- Linked Data integration and quality
- Linked Data management

Contact:
- tomas.knap@semantic-web.com