DSNotify Evaluation Readme

Date: 27.10.2009
Author: niko.popitsch@univie.ac.at

Introduction

This document describes the evaluation data set that was used for the evaluation of DSNotify (http://dsnotify.org) in September/October 2009.

The Evaluation Strategy

The overall evaluation strategy is depicted in Figure 1. A *simulator* (e.g., the *SimpleSimulator.class*) takes a *source dataset* (SDS), a *target dataset* (TDS), an *eventset* (ES) and a configuration file (not depicted) as input.

The simulator starts a simple RDF server and loads the 3 RDF graphs (SDS, TDS, ES) and creates a new *working graph*, in the figure depicted as Observed data set (ODS). The ODS is exposed as Linked Data under a configurable HTTP URI (For the
Actual simulation this was http://localhost:18081/sim/working).

The simulator starts to continuously update the ODS according to the event information stored in the ES. The overall time of the simulation is configurable. The simulator interprets the timestamps of the events stored in the ES relative to this configured simulation-duration.

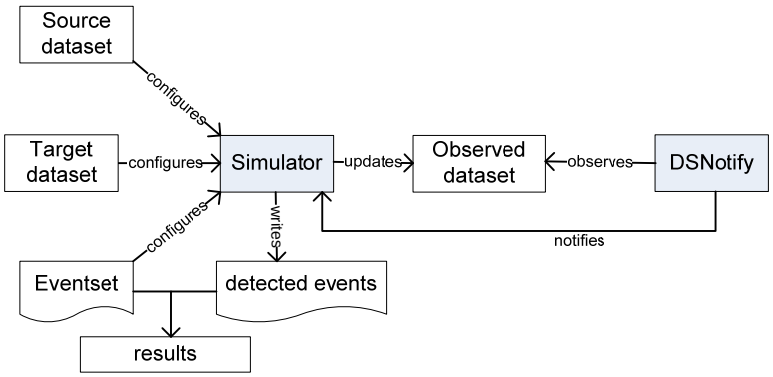

Figure 1: DSnotify evaluation strategy

Before the simulator starts with updating the ODS, DSNotify is started and configured programmatically to observe the ODS. For the evaluation we configured DSNotify to periodically retrieve all items (i.e., resources and their properties) from the ODS RDF Graph and update its indices accordingly. Immediately after this monitoring cycle, a housekeeping cycle was started that tries to reason create/remove/move events from the data stored in the DSNotify indices and reports the detected events back to the simulator.

The simulator creates a result graph by creating two new eventsets: a false-negative eventset containing all events from the ES that were not detected by DSNotify and a false positive eventset that contains all events detected by DSNotify that were not in the original ES. The result graph is then written to the configured output file.

Data Preparation

For the evaluation two different types of eventsets were created:

- IIMB eventsets derived from the ISLab Instance Matching Benchmark
- A DBpedia eventset derived from DBpedia persondata sets.

For the iimb eventsets no further data preparation was required. The eventset was created using the IIMBToDsnotifyConverter.class.

For the DBpedia eventset, the original persondata files were enriched with additional properties, with redirect information and with event dates retrieved from Wikipedia. The overall process of data preparation is depicted in Figure 2:
First, the original dbpedia 3.2 and 3.3 persondata graphs were combined with a configurable set of properties from the ontology-based infobox files (1,3). Then the enriched 3.2 persondata graph was further extended with redirect properties from the dbpedia 3.3 redirect graph (2). Then the SimpleDBPediaConverter.class combined these data to the resulting eventset. The dates for create, move and remove events were retrieved from Wikipedia (the dates are buffered in a persistent file cache to avoid accessing Wikipedia again in subsequent requests).

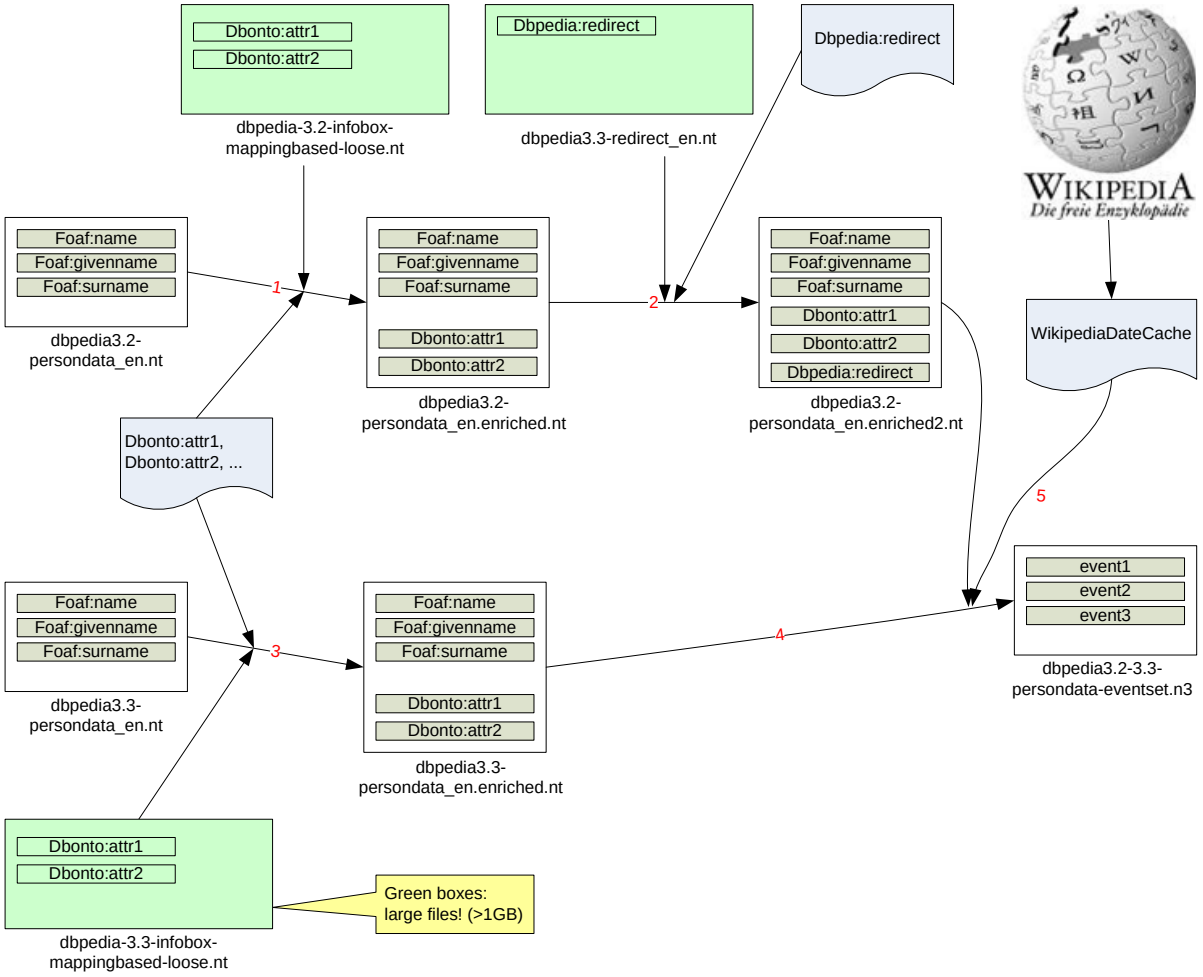

Figure 2: DBpedia eventset preparation

The Evaluation process

The actual evaluation process of the DBpedia eventset is depicted in Figure 3 and explained in the following. The iimb eventset evaluation was done analogously.

1) The enriched 3.2 and 3.3 persondata graphs are imported into a simple RDFServer. The 3.2 graph is used as the „working model“ (ODS)

2) Using the eventset and the 3.3 persondata, this working graph is updated corresponding to a „compressed“ timeline by the simulator.

3) This changing RDF graph is monitored by a DSNotify instance

4) DSNotify is configured to monitor the attributes added by the enrichment steps done in eventset preparation

5) The events reported by DSNotify are used to create a result set model

6) When the simulation is finished, this result set model is serialized to a file

7) The SimpleAnalyzer is used to calculate precision/recall/f-measure/etc. from this result set

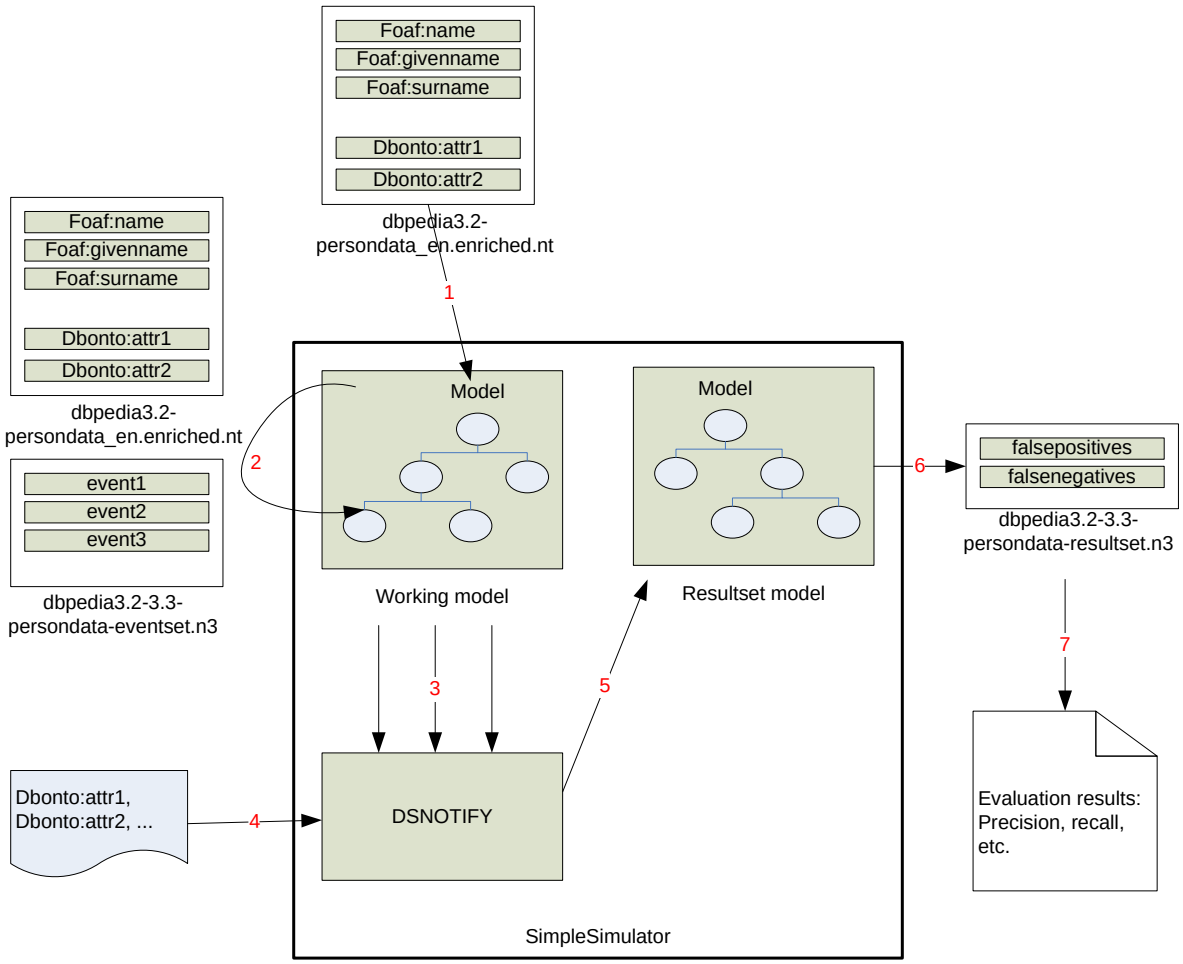

Figure 3: DSNotify evaluation process

The Evaluation Data

The created evaluation data is structured as follows:

```
.
|-- 2009_DSNotify_Evaluation.jar
Snapshot of  the DSNotify version used for the evaluation.
|-- dbpedia
Directory containing all DBpedia evaluation data.
|   |-- enricheddata
The modified (enriched) input datasets
|   |   |-- dbpedia3.2-3.3-persondata-eventset.histogram.txt
Histogram of the resulting eventset
|   |   |-- dbpedia3.2-3.3-persondata-eventset.log
|   |   |-- dbpedia3.2-3.3-persondata-eventset.n3
The resulting eventset
|   |   |-- dbpedia3.2-persondata_en.enriched.nt
The 3.2 persondata set enriched with additional infobox properties
|   |   |-- dbpedia3.2-persondata_en.enriched2.nt
The 3.2 persondata set enriched with additional infobox properties and redirect properties
|   |   `-- dbpedia3.3-persondata_en.enriched.nt
The 332 persondata set enriched with additional infobox properties
|   |-- experiments
The experiments.
|   |   |-- dbpedia-experiment1
|   |   |-- dbpedia-experiment2
|   |   |-- dbpedia-experiment3
|   |   |-- dbpedia-experiment4
|   |   |-- dbpedia-experiment5
|   |   |-- dbpedia-experiment6
|   |   |-- dbpedia-experiment7
|   |   |-- dbpedia-experiment8
|   |   `-- dbpedia-experiment9
|   `-- rawdata
The original snapshots downloaded from http://wiki.dbpedia.org/
|       |-- dbpedia-3.2-infobox-mappingbased-loose.nt
|       |-- dbpedia-3.3-infobox-mappingbased-loose.nt
|       |-- dbpedia3.2-persondata_en.nt
|       |-- dbpedia3.3-persondata_en.nt
|       `-- dbpedia3.3-redirect_en.nt
|-- iimb
Directory containing all iimb evaluation data.

|   |-- enricheddata
The modified (enriched) input datasets.
|   |   `-- iimb-converted
Directory containing asub directory for each iimb sub dataset. These contain the created eventsets. Not that the timestamps are randomly created.
|   |-- experiments
|   |   |-- iimb-experiment
|   `-- rawdata
Directory containing the original iimb datasets from http://islab.dico.unimi.it/iimb/
|       |-- 001
|       |-- …
|       |-- 037
|       |-- README.txt
Description of the IIMB dataset. Note that only the first 10 iimb sub datasets were considered in the evaluation.
|       |-- abox.owl
|       `-- tbox.owl
|-- readme
General information about the evaluation data
|   `-- readme.pdf
`-- wikipediadata
Cached article creation/deletion dates that were automatically retrieved from Wikipedia. These caches were created to avoid accessing Wikipedia every time a DBpedia eventset is created.
    |-- wikipedia-article-creationdates.n3
    `-- wikipedia-article-removedates.n3
```