

Data transformations

Miroslav Blaško

November 9, 2017

1 Background

The goal of this seminar is get familiar with a tool for data transformations – *OpenRefine*.

2 Task – Data Transformations with OpenRefine

Within this task we will transform two sheets from a XLSX document into RDF which we will use for querying. Concrete steps of the task are:

- Login into GraphDB at <http://onto.fel.cvut.cz:7300>.
- Import sheet “event type x factor” provided by XLSX document¹ into OntoRefine within GraphDB².
- Fix issues with spelling/capitalization of ”Eccairs event description” using cluster & merge method³.
- Analyze “Source of model description (if relevant)” column using Text facet and filtering⁴.
- Add source type column based on different values of the analyzed column⁵. You can use OpenRefine Expression Language to define new column in the following way:

```
value.replace('^http[s]?:../, "").split("/") [0].replace(/www./, '').split('.')  
[0].replace(/$/, ' documentation')
```

- Remove all irrelevant rows⁶.

¹It is one of the resources provided for this tutorial.

²Hint: /GraphDB/Import/Tabular (OntoRefine)

³Hint: ECCAIRSEvent description/Edit cells ../Cluster and edit ...

⁴Hint: Source of model description (if relevant)/Facet/Text facet

⁵Hint: Source of model description (if relevant)/Edit column/Add column based on this column ...

⁶Hint: Source of model description (if relevant)/Facet/Text facet, pick blank to include only in filtering, then All/Edit rows/Remove all matching rows

- Export the project into a SPARQL endpoint⁷.
- Create new OpenRefine project by importing sheet “uniset factors” provided by XLSX, transform it appropriately and export it as a second SPARQL endpoint.
- Use both SPARQL endpoints to query and create data⁸ compliant with schema provided in Fig. 1.

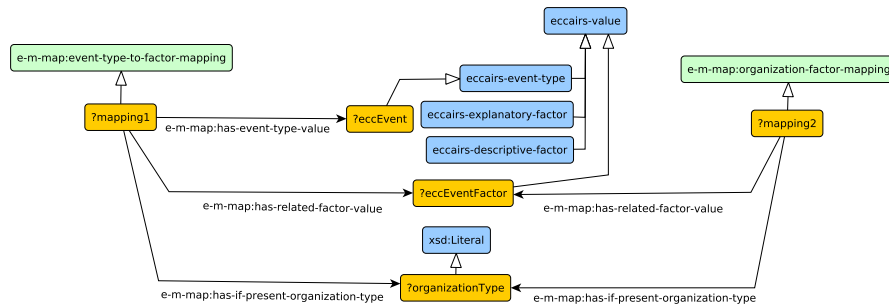


Figure 1: RDF schema for the output of transformation task

3 Other related tools

- **RDFpro** – available at <http://rdfpro.fbk.eu/>.
- **ETL LinkedPipes** – available at <https://etl.linkedpipes.com/>.

⁷RDF button + Data/Get SPARQL endpoint...

⁸Hint: Use SPARQL CONSTRUCT query with two SERVICE clauses pointing to exported SPARQL endpoints.