HOTMapper: Historical Open Data Table Mapper

Henrique Varella Ehrenfried[1], Rudolf Eckelberg[1], Hamer Iboshi[1], Eduardo Todt[1], Daniel Weingaertner[1] and Marcos Didonet Del Fabro[1]
[1]C3SL Labs, Federal University of Paraná - Curitiba, PR, Brazil
hvehrenfried@inf.ufpr.br | h.v.ehrenfried@gmail.com

Goal

Create a tool that simplifies the management of Open Data schemas and its data.

Problem been solved

The availability of Open Data raised several issues to manage its evolution and integration. Open data is often de-normalized and represents a period of time. Usually it is released periodically (i.e. annually or semesterly). Thus they can have many schemas, different representation formats and other problems. These problems make it harder to build an integrated database with historical data. Existing tools are difficult to understand and maintain. It is necessary to address three main issues:

(1) Integrated source creation: A first set of tables needs to be created to be able to execute queries considering the historical data;

(2) Schema evolution: Every year, the data sources definition changes, so it is necessary to provide schema mappings, as shown in Figure 1;

(3) Data evolution: Data has to be transformed and kept compatible along all years, requiring instance mappings (i.e: GENDER [2014] = WHEN *M* THEN *1* WHEN *F* THEN *2*; GENDER [2015] = GENDER);

Impact

The HOTMapper is used in a real world scenario. It allowed our lab to insert and manage many historical data, as it can be seen in Table 1.

Solution

We created the HOTMapper, a tool that allows to map different schemas into a single, unified one. In order to map these schemas, it is necessary to create two different files: A **mapping definition** and a **table definition**. Then when the HOTMapper runs, it will add the data into the new and unified schema. This new schema can contain transformed data that is acquired by processing the available data or created by the user.

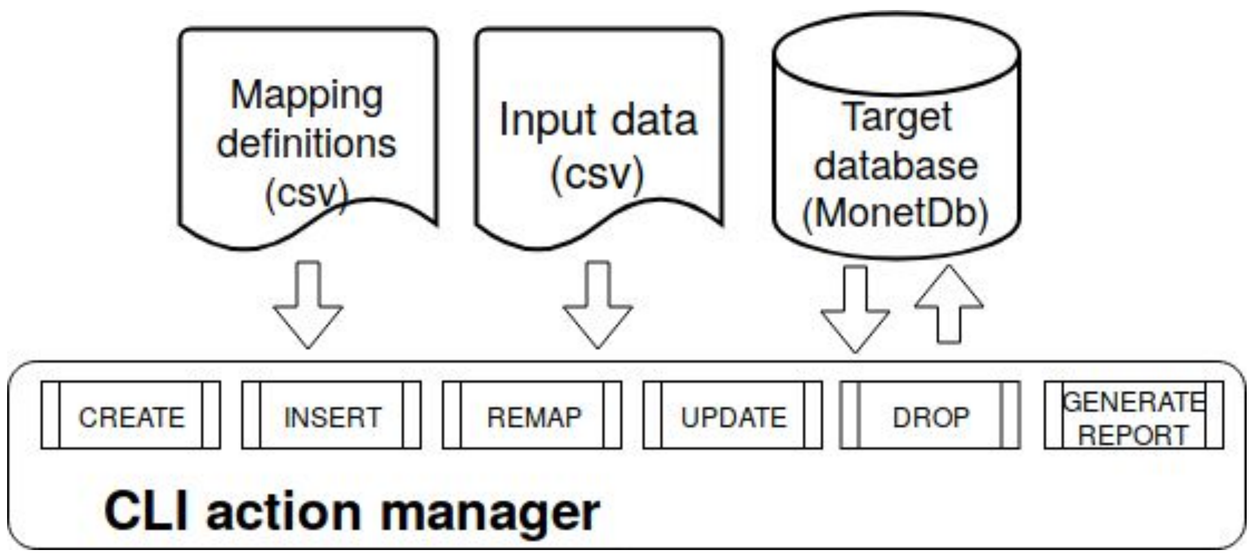

Figure 2: Overview of HOTMapper with some of its functions.**CREATE**: Create a table in the database; **DROP**: Delete the selected table; **INSERT**: Execute a bulk import in the database using the mapping protocol; **REMAP**: Modify the initial table definition; **UPDATE**: Update a table using the mapping protocol. It is recommended that the REMAP action be run first; **GENERATE REPORT:** Generates a equivalence between the inputtable and the current database.

OpenData-Mapper.csv

Lab.Var	Standard Label	New label	Temp Column	DB name	Data type	2010	2011
ID1	sg_uf	UF abbreviation	0	sigla_uf	VARCHAR(4)	SGL_UF	SG_UF
▪ ▪ ▪							
IDn-1	no_ies	IES name	0	nome_ies	VARCHAR(255)	NOM_IES	NO_IES
IDn	co_ies	IES code	0	code_ies	INTEGER	COD_IES	CO_IES

OpenData2010.csv

COD_IES	NOM_IES	…	SGL_UF
571	Univ. Fed. PR		PR
⋮	⋮	⋮	⋮
953	Univ. Fed. SC	…	SC
953	Univ. Fed. SC		SC

OpenData2011.csv

CO_IES	NO_IES	…	SG_UF
572	Univ. Fed. MG		MG
⋮	⋮	⋮	⋮
944	Univ. Fed. PE	…	PE
944	Univ. Fed. PE		PE

Figure 1: Example of open data (The tables in the bottom) and a mapping protocol (table at top)

Year	Tables	Records
2017	13	102,176,661
2016	20	116,009,013
2015	18	116,946,948
2014	17	121,115,913
2013	18	112,645,020
2012	11	036,029,271
2011	7	012,025,035
2010	7	008,768,490

Table 1: Real world usage of the HOTMapper

Some references

- D. Deng, R. C. Fernandez, Z. Abedjan, S. Wang, M. Stonebraker, A. K. Elmagarmid, I. F. Ilyas, S. Madden, M. Ouzzani, and N. Tang. 2017. The Data Civilizer System.. In CIDR.
- R. Fagin, L. M.Haas, M. Hernández, R. J. Miller, L. Popa, and Y. Velegrakis. 2009. Clio: Schema Mapping Creation and DataExchange. Springer Berlin Heidelberg, Berlin, Heidelberg, 198–236.
- X. Ling, A. Halevy, F. Wu, and C. Yu. 2013. Synthesizing union tables from the Web. In IJCAI.
- R. J. Miller. 2018. Open Data Integration.Proc. VLDB Endow.11, 12 (Aug.2018), 2130–2139. https://doi.org/10.14778/3229863.3240491

Tool Available at: https://github.com/C3SL/hotmapper