Data-Transformation on historical data using the RDF Data Cube Vocabulary

Sebastian Bayerl, Michael Granitzer
Department of Media Computer Science
University of Passau

SWIB15 – Semantic Web in Libraries
22.10.2015

Overview

- Motivation
- Vocabulary and Dataset
- Problem Setting and Approach
- Workflow
- Contributions

Motivation

- Statistical and historical data source
  - Statistics of the German Reich (Digitalized)
- Access the encapsulated knowledge
- Data Analytics and Recommendation
  - Using Linked Data (RDF Data Cube Vocabulary)
- But first: Data Integration
  - Data Cleaning, -Transformation and -Fusion

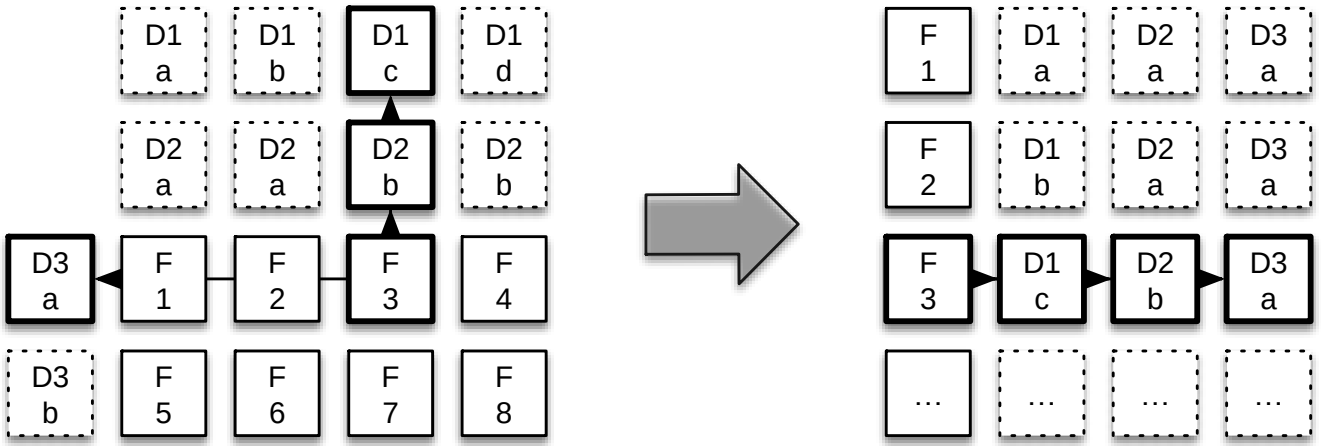
Source structure

II 8 A. Nachweisung über die Abfertigungen in Bezug auf die

Bezirke der Direktivbehörden und Hauptämter.	I. Abfertigungen im Ansageverfahren.										II. Abfertigungen mit Begleitzetteln.								
	Beim Eingange seewärts				Beim Eingange auf Landstrassen, Flüssen und Kanälen.						Ausfertigung von Begleitzetteln.			Erledigung von Begleitzetteln.					
	Zahl der mit Ansagezetteln abgegangenen Schiffe:		Zahl der mit Ansagezetteln bei Aemtern im Innern angekomm. Schiffe:		Zahl der mit Ansagezetteln abgegangenen Landfuhrwerke und Wasserfahrzeuge:				Zahl der mit Ansagezetteln bei Aemtern im Innern angekommenen Landfuhrwerke u. Wasserfahrzeuge.										
					von Ansageposten zu Grenzämtern:		von Grenzämternauf Aemter im Innern:												
	von Ansageposten zu Grenzämtern.	von Grenzämtern auf Aemter im Innern.	Hauptschiffe.	Leichterschiffe.	Landfuhrwerke.	Wasserfahrzeuge.	Landfuhrwerke.	Wasserfahrzeuge.	Landfuhrwerke.	Wasserfahrzeuge.	Zahl der ausgefertigten Begleitzettel.	Zahl der zugehörigen Ladungsverzeichnisse.	Zahl der abgefertigten Wagen.	Zahl der erledigten Begleitzettel.	Zahl der zugehörigen Ladungsverzeichnisse.	Darunter zur unmittelbaren Durchfuhr.	Zahl der abgefertigten Wagen.	Darunter zur unmittelbaren Durchfuhr.	Zahl der übergebenen speziellen Deklarationen.
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.
I. Preussen.																			
1. Provinz Preussen.																			
a) Ostpreussen:																			
Eydtkuhnen* . . .	—	—	—	—	59	—	—	—	—	—	122	122	330	462	462	434	1 693	1 688	5
Johannisburg* . . .	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Memel*	—	—	—	—	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Neidenburg* . . .	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Pillau*	—	407	—	—	—	—	—	—	—	—	115	132	777	2	2	2	3	3	—
Schmaleningken* .	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Tilsit*	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

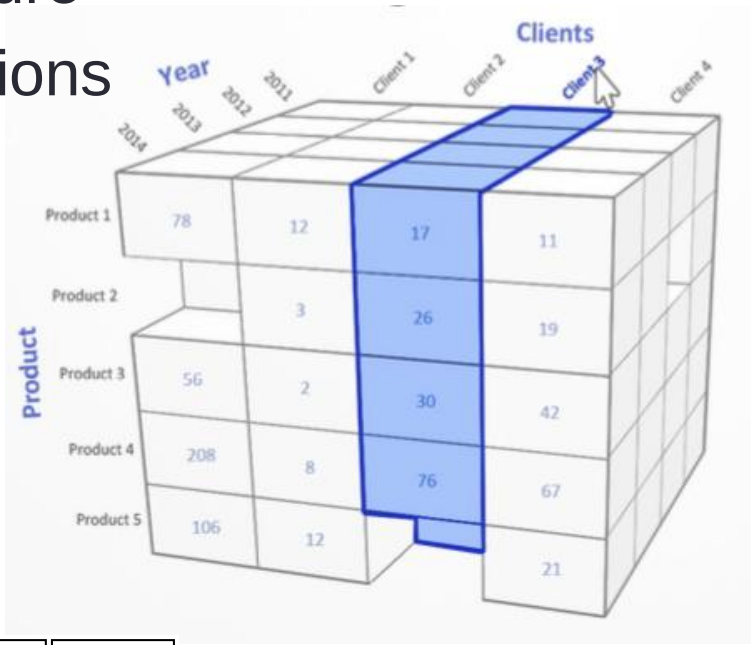
Target structure

-1	0	1	2	3	4	5
0	25330	Einfuhr	im 1. Quartal 1873.	Glasscherben, Bruchglas.	1. (132.) Pos. 1 a.	Preussen
1	21861	Einfuhr	im 1. Quartal 1872.	Glasscherben, Bruchglas.	1. (132.) Pos. 1 a.	Preussen
2	337	Einfuhr	im 1. Quartal 1873.	Glasscherben, Bruchglas.	1. (132.) Pos. 1 a.	Bayern
3	378	Einfuhr	im 1. Quartal 1872.	Glasscherben, Bruchglas.	1. (132.) Pos. 1 a.	Bayern
4	474	Einfuhr	im 1. Quartal 1873.	Glasscherben, Bruchglas.	1. (132.) Pos. 1 a.	Baden
5	120	Einfuhr	im 1. Quartal 1872.	Glasscherben, Bruchglas.	1. (132.) Pos. 1 a.	Baden



Data Cubes and OLAP

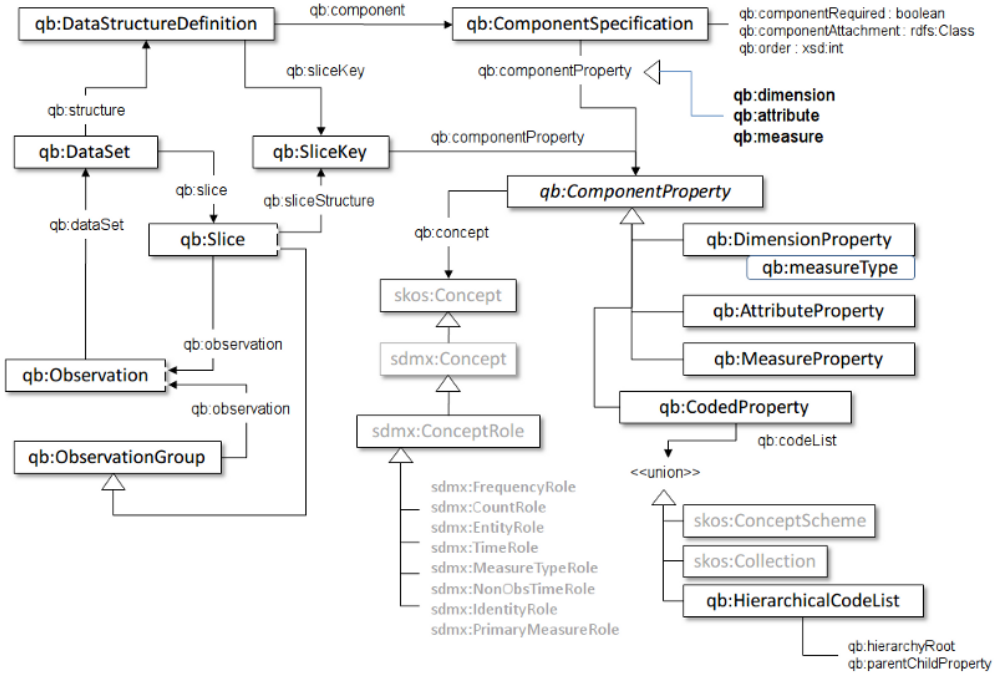
- Cube: Multi-dimensional data structure
- Observation: measures and dimensions
  - Measure: numerical fact
  - Dimension: describes the fact(s)
- Enables Data Analytics
- OLAP: Online Analytical Processing
  - Slicing, Dicing, Roll-Up,…

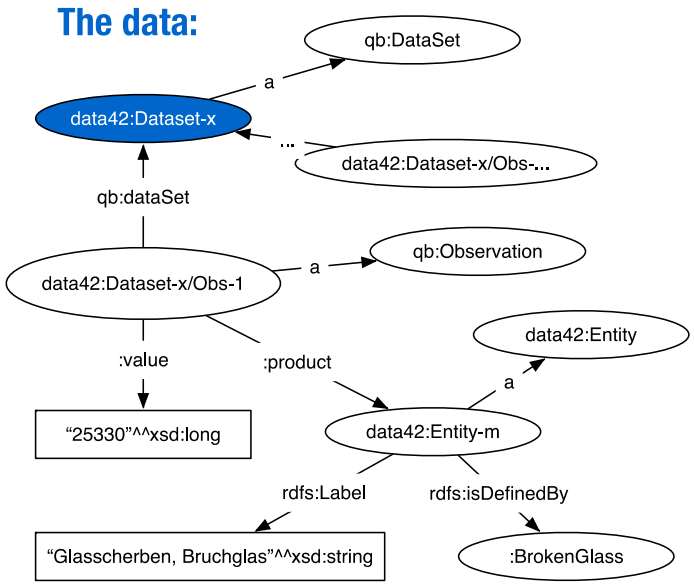


-1	0	1	2	3	4	5
0	25330	Einfuhr	im 1. Quartal 1873.	Glasscherben, Bruchglas.	1. (132.) Pos. 1 a.	Preussen
1	21861	Einfuhr	im 1. Quartal 1872.	Glasscherben, Bruchglas.	1. (132.) Pos. 1 a.	Preussen
2	337	Einfuhr	im 1. Quartal 1873.	Glasscherben, Bruchglas.	1. (132.) Pos. 1 a.	Bayern
3	378	Einfuhr	im 1. Quartal 1872.	Glasscherben, Bruchglas.	1. (132.) Pos. 1 a.	Bayern
4	474	Einfuhr	im 1. Quartal 1873.	Glasscherben, Bruchglas.	1. (132.) Pos. 1 a.	Baden
5	120	Einfuhr	im 1. Quartal 1872.	Glasscherben, Bruchglas.	1. (132.) Pos. 1 a.	Baden

The RDF Data Cube Vocabulary

- RDF based vocabulary
- Models an OLAP Data Cube
- Interlink components with existing concepts





http://www.w3.org/TR/vocab-data-cube/

Examples 1

III. Uebersicht der im Seeverkehr angekommenen und abgegangenen Schiffe nach den Flaggen und nach den Ländern (Küstenstrecken) der Herkunft und Bestimmung für das Jahr 1873.

Länder bezw. Küstenstrecken der Herkunft und Bestimmung.	Angekommen					Abgegangen				
	Dampfschiffe mit schrägen Ziffern, in den Hauptzahlen mit enthalten.									
	Mit Ladung.		In Ballast oder leer.		Be-satzung.	Mit Ladung		In Ballast oder leer.		Be-satzung.
	Schiffe.	Reg.-Tons.	Schiffe.	Reg.-Tons.		Schiffe.	Reg.-Tons.	Schiffe.	Reg.-Tons.	
1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.
A. In den Preussischen Hafenplätzen.										
1. Deutsche Schiffe überhaupt.										
I. Deutsches Reich.										
Preussischer Staat.										
Preussen	1 092 *291*	94 996 *40 324*	61 *13*	8 434 *1 970*	7 003 *3 601*	761 *220*	63 652 *29 959*	520 *35*	60 574 *5 856*	7 751 *3 446*
Pommern	2 277 *598*	153 380 *84 266*	617 *62*	55 706 *10 106*	14 973 *7 740*	1 858 *582*	132 474 *80 173*	358 *51*	17 185 *3 590*	11 886 *7 155*
Schleswig-Holstein an der Ostsee . . .	4 366 *632*	105 416 *43 825*	1 982 *82*	53 638 *6 373*	16 542 *4 504*	4 563 *667*	124 958 *49 302*	1 560 *68*	27 827 *3 003*	16 155 *4 862*
" " an der Nordsee . . .	984 *4*	22 763 *220*	337 *1*	8 054 *91*	2 859 *35*	921 *4*	22 070 *218*	481 *1*	11 219 *42*	3 030 *34*
Hannover, östl. Theil	286 *23*	7 934 *2 232*	116 —	2 663 —	1 054 *220*	368 *25*	13 271 *2 452*	141 *1*	3 374 *60*	1 427 *253*
westl. Theil einschl. des Jadegeb.	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Examples 2

II 8 A. Nachweisung über die Abfertigungen in Bezug auf die

Bezirke der Direktivbehörden und Hauptämter.	I. Abfertigungen im Ansageverfahren.										II. Abfertigungen mit Begleitzetteln.								
	Beim Eingange seewärts				Beim Eingange auf Landstrassen, Flüssen und Kanälen.						Ausfertigung von Begleitzetteln.			Erledigung von Begleitzetteln.					
	Zahl der mit Ansagezetteln abgegangenen Schiffe:		Zahl der mit Ansagezetteln bei Aemtern im Innern angekomm. Schiffe:		Zahl der mit Ansagezetteln abgegangenen Landfuhrwerke und Wasserfahrzeuge:				Zahl der mit Ansagezetteln bei Aemtern im Innern angekommenen Landfuhrwerke u. Wasserfahrzeuge.										
					von Ansageposten zu Grenzämtern:		von Grenzämternauf Aemter im Innern:												
	von Ansageposten zu Grenzämtern.	von Grenzämtern auf Aemter im Innern.	Hauptschiffe.	Leichterschiffe.	Landfuhrwerke.	Wasserfahrzeuge.	Landfuhrwerke.	Wasserfahrzeuge.	Landfuhrwerke.	Wasserfahrzeuge.	Zahl der ausgefertigten Begleitzettel.	Zahl der zugehörigen Ladungsverzeichnisse.	Zahl der abgefertigten Wagen.	Zahl der erledigten Begleitzettel.	Zahl der zugehörigen Ladungsverzeichnisse.	Darunter zur unmittelbaren Durchfuhr.	Zahl der abgefertigten Wagen.	Darunter zur unmittelbaren Durchfuhr.	Zahl der übergebenen speziellen Deklarationen.
	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.	17.	18.	19.
I. Preussen.																			
1. Provinz Preussen.																			
a) Ostpreussen:																			
Eydtkuhnen* . . .	—	—	—	—	59	—	—	—	—	—	122	122	330	462	462	434	1 693	1 688	5
Johannisburg* . . .	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Memel*	—	—	—	—	2	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Neidenburg* . . .	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Pillau*	—	407	—	—	—	—	—	—	—	—	115	132	777	2	2	2	3	3	—
Schmaleningken* .	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Tilsit*	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

Seereisen Deutscher Schiffe im Jahre 1874.

Noch: III. A. Reisen zwischen ausserdeutschen Häfen, zusammengestellt nach den Küstenstrecken des Abgangs.

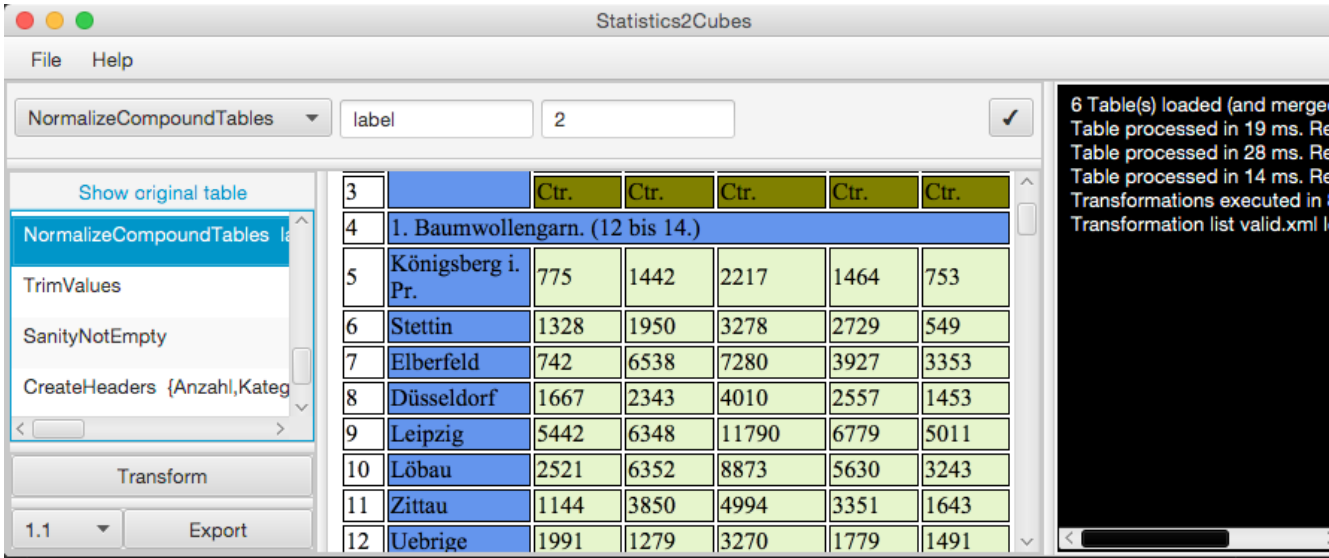
Länder bezw. Küstenstrecken des Abgangs und der Bestimmung.	Mit Ladung.		In Ballast oder leer.	
	Schiffe.	**Reg.-Tons.**	Schiffe.	**Reg.-Tons.**
1.	**2.**	**3.**	**4.**	**5.**
Noch: 7. Deutsche Schiffe überhaupt.				
Noch: Von Grossbritannien und Irland nach:				
Südamerika a. atl. Meere, südl. v. Brasilien	56	20033	—	—
Chile	12	6853	—	—
dem übrigen Südamerika am stillen Meere	13	7749	—	—
Egypten am mittelländischen Meere	11	5598	—	—
Kapland mit Natal	26	6888	—	—
Afrika am atlantischen M[illegible]	29	8163	9	1038
" " indischen und rothen Meere	2	1116	—	—
Asien a. mittell. u. schwarz. Meere (Levante)	1	918	—	—
dem übrigen Vorderasien bis Ostindien	2	1538	—	—
Ostindien mit den indischen Inseln	35	25566	1	1030
China	11	5962	—	—
Australien mit den Inseln im stillen Meere	3	1278	1	1119
Ueberhaupt	1497	540494	711	208348
Von den Niederlanden nach:				
dem Europ. Russland a. weiss. M. u. Eismeere	—	—	10	1838
" " " an der Ostsee	22	22925	37	6377
Schweden	8	1043	3	550

Problem Setting

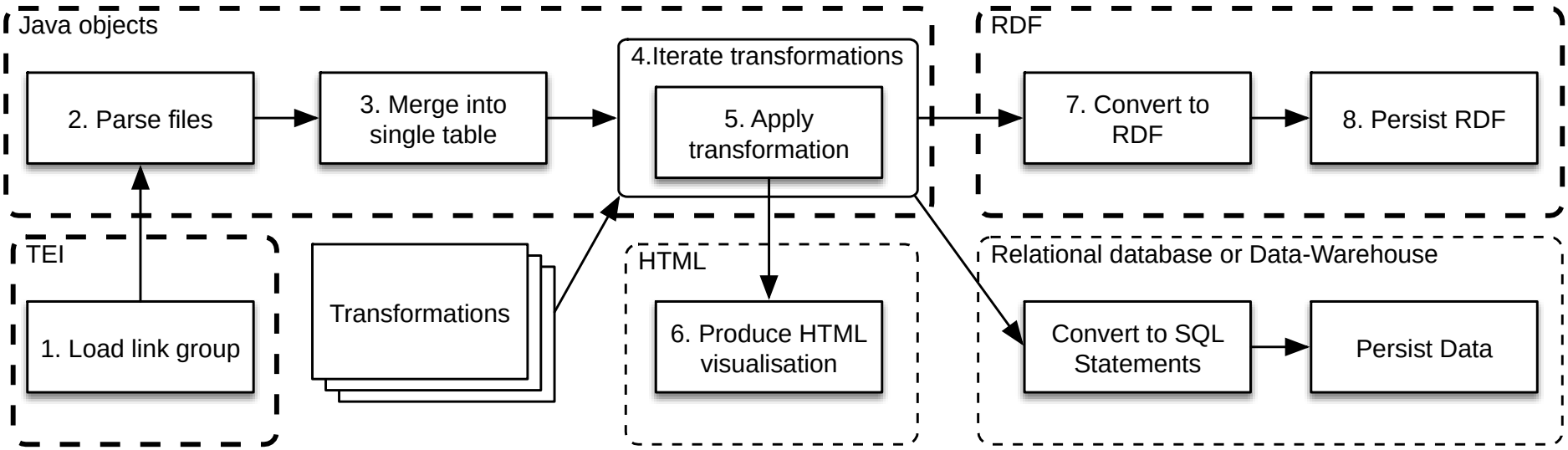
- Data is encapsulated in multiple files
- Unusable for sophisticated Data Analysis
- Normalization of complex structured data
- Dirty and faulty data, structure or annotations
- Lots of similar problems in a huge dataset

Approach

- Use the RDF Data Cube Vocabulary
  - Enables: Interlinking, merging and analytics
- Use an incremental workflow
  - Identify fine-granular transformations
- Implement the research prototype with GUI
  - Select, configure and chain transformations (save/load)
  - HTML preview



Workflow


Java objects
2. Parse files
3. Merge into single table
4.Iterate transformations
5. Apply transformation
RDF
7. Convert to RDF
8. Persist RDF
TEI
1. Load link group
Transformations
HTML
6. Produce HTML visualisation
Relational database or Data-Warehouse
Convert to SQL Statements
Persist Data

Transformations

1. Pre-Normalization

- Sanity checks
- Data Cleaning
- Fix structure (e.g. spans), data and annotations
- Delete row (e.g. repeating headers)
- 30 more…

2. Normalization

- Normalization
- Compound normalization: Horizontal or vertical partitions

3. Post-Normalization

- Add/merge/delete columns
- Add headers/disambiguation
- Add metadata
- …

Advanced transformations

- Compound transformations
  - Combine multiple transformation
  - Fix more complex problems
  - E.g. find problematic cells and fix with existing transformation
- Transformation suggestions
  - Find common problems: Repeat symbol, annotation patterns
  - A step towards automation

Contributions

- Modular workflow for the Data Integration process
  - Definition of fine granular transformation steps
  - Reusable within the same or for other data sources
- Lift and enrich historical statistical data
  - Ready for Data Analytics
- Current datset contains 32169 files
  - > 10% converted
  - 10 conversion chains

https://github.com/bayerls/statistics2cubes

Thank you for your attention!

RDF Data Cube Vocabulary:
http://www.w3.org/TR/vocab-data-cube/

https://github.com/bayerls/statistics2cubes

Sebastian Bayerl
Department of Media Computer Science
University of Passau
bayerl@dimis.fim.uni-passau.de

BACKUP

Publication

- Bayerl, Sebastian, and Michael Granitzer. "Data-transformation on historical data using the RDF data cube vocabulary." Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business. ACM, 2015.

Abstract

This work describes how XML-based TEI documents, containing statistical data, can be normalized, converted and enriched using the RDF Data Cube Vocabulary. In particular we focus on a statistical real world data set, namely the statistics of the German Reich around the year 1880, which are available in the TEI format. The data is embedded in complex structured tables, which are relatively easy to understand for humans but they are not suitable for automated processing and data analysis, without heavy pre-processing, due to their varying structural properties and differing table layouts. Therefore, the complex structured tables must be validated, modified and transformed, until they are suitable for the standardized multi-dimensional data structure - the data cube. This work especially focuses on the transformations necessary to normalize the structure of the tables. Performing validation- and cleaning-steps, resolving row- and column-spans and reordering slices are available transformations among multiple others. By combining exiting transformations, compound operators are implemented, which can handle specific and complex problems. The identification of structural similarities or properties can be used to automatically suggest sequences of transformations. A second focus is on the advantages, which come by using the RDF Data Cube Vocabulary. Also, a research prototype was implemented to execute the workflow and convert the statistical data into data cubes.