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

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

Big Data Analytics: Big Data management and Data Integration
i-KNOW
22.10.2015

Overview

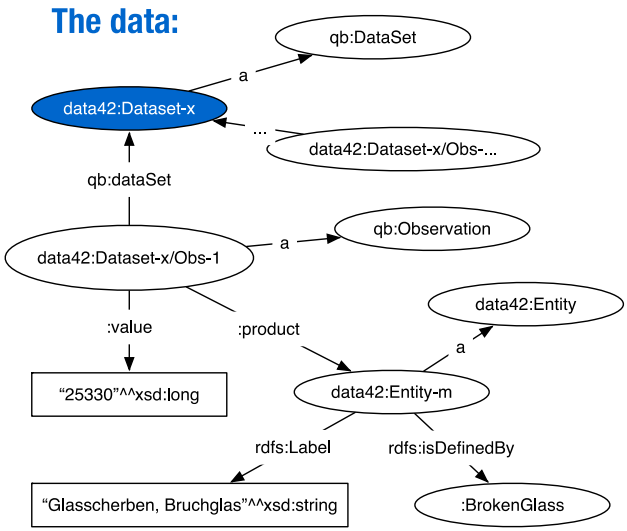
- Motivation
- Vocabulary and Dataset
- Problem Setting and Approach
- Workflow
- Contributions

Motivation

- EEXCESS Vision:
  - Unfold the treasure of cultural, educational and scientific long-tail content.
- Data Analytics!
  - Using Linked Data (RDF Data Cube Vocabulary)
  - Statistical and historical data source: German Reich statistics
- But first: Data Integration
  - Data Cleaning, -Transformation and -Fusion

The RDF Data Cube

- Cube: Multi-dimensional data structure
- Observation: measures and dimensions
  - Measure: numerical fact
  - Dimension: describes the fact(s)



-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

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

Example

Die deutschen überseeischen Auswanderer in den Jahren 1871 bis 1882.

Deutsche Auswanderer im Jahre	überhaupt (durch die amtliche deutsche Statistik nachweisbare)	davon wurden befördert über				nach Amerika, nämlich						nach Afrika	nach Asien	nach Australien
		Bremen	Hamburg	Stettin	Antwerpen	den Vereinigten Staaten v. A.	britisch Nord-A.	Mexiko u. Zentral-A.	Westindien	Brasilien	anderen Theilen v. A.			
1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
1871	75 912	45 658	30 254	.	.	73 816	9	21	37	920	263	18	11	817
1872	125 650	66 919	57 615	.	1 116	120 056	690	38	61	3 232	387	2	12	1 172
1873	103 638	48 608	51 432	.	3 598	96 641	49	32	28	5 048	496	4	9	1 331
1874	45 112	17 907	24 093	1 536	1 576	42 492	138	24	83	1 019	418	5	33	900
1875	30 773	12 613	15 826	268	2 066	27 834	38	26	47	1 387	377	1	37	1 026
1876	28 368	10 972	12 706	202	4 488	22 767	11	8	35	3 432	804	54	31	1 226
1877	21 964	9 328	10 725	75	1 836	18 240	11	25	243	1 069	289	750	31	1 306
1878	24 217	11 329	11 827	85	976	20 373	89	22	74	1 048	449	394	50	1 718
1879	33 327	15 828	13 165	245	4 089	30 808	44	17	59	1 630	441	23	31	274

Example 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 Grenzämtern auf 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 Meere	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

Seereisen Deutscher Schiffe im Jahre 1874.

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

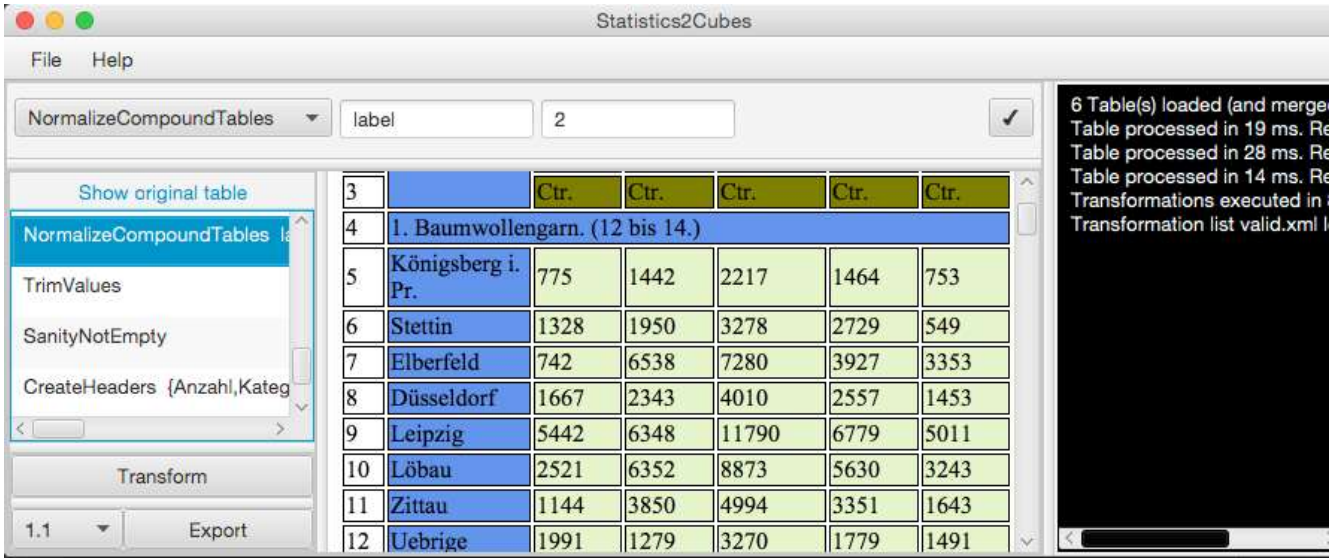
Länder bezw. Küstenstrecken des Abgangs und der Bestimmung.	Mit Ladung.		In Ballast oder leer.	
	Schiffe.	**Reg.-Tons.**	Schiffe.	**Reg.-Tons.**
1.	**2.**	**3.**	**4.**	**5.**
Noch: 7. Deutsche Schiffe überhaupt.				
Noch: Von Grossbritannien und Irland nach:				
Südamerika a. atl. Meere, südl. v. Brasilien	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 Me	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

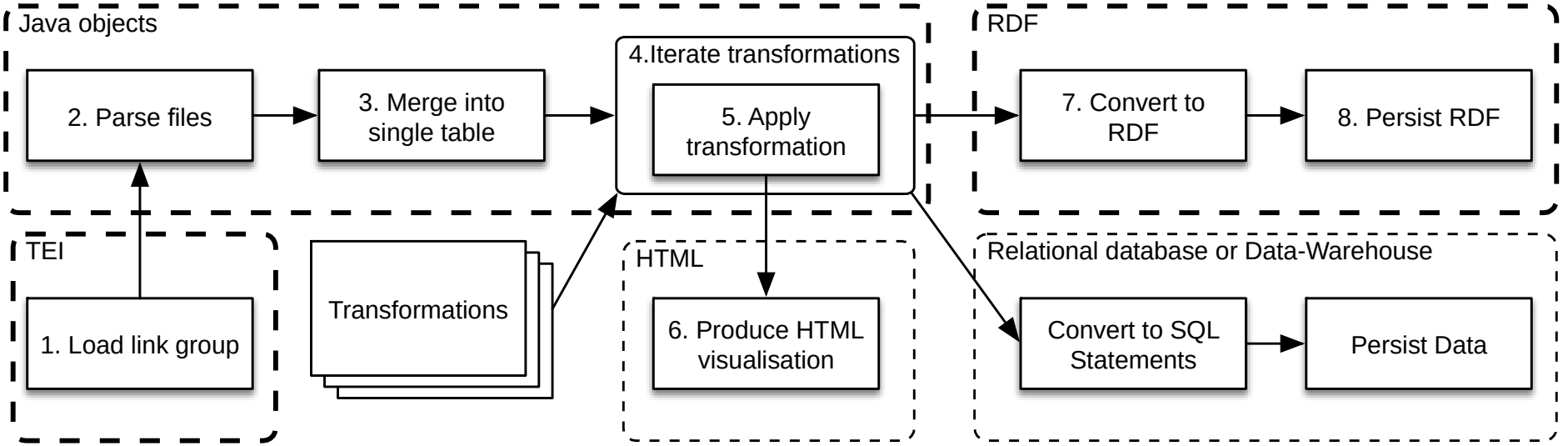
- Data is encapsulated in multiple files
- Unusable for sophisticated Data Analysis
- Normalization of complex structured data
- Dirty and faulty data, structure or annotations
- Lots of similar problems in a huge dataset

Approach

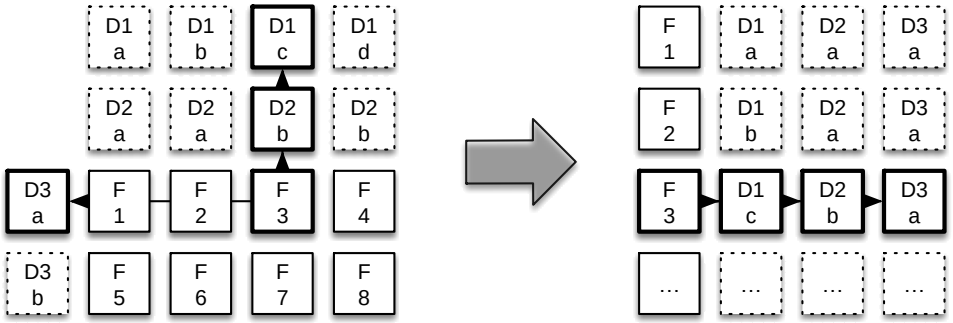
- Use the RDF Data Cube Vocabulary
  - Enables: Interlinking, merging and analytics
- Use an incremental workflow
  - Identify fine-granular transformations
- Implement the research prototype with GUI
  - Select, configure and chain transformations (save/load)
  - HTML preview



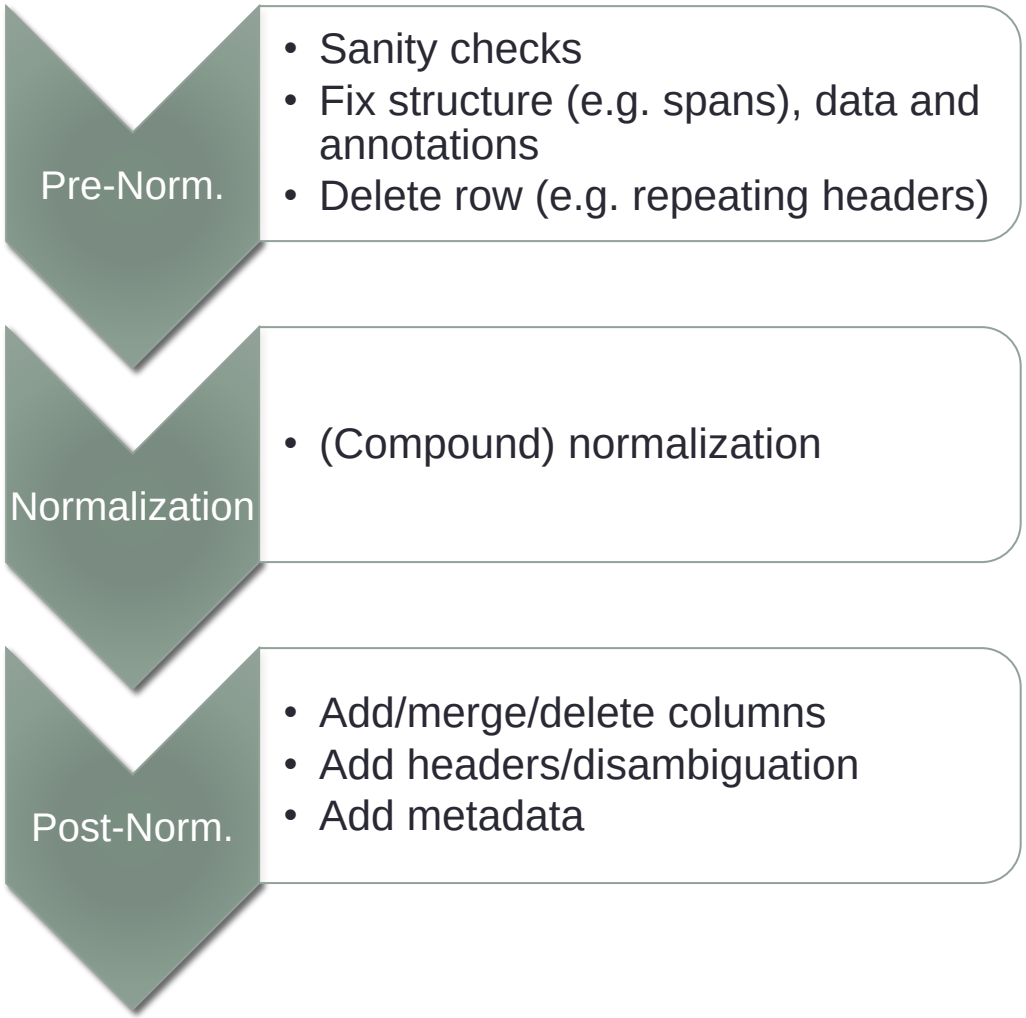
Workflow



Normalize:



Tansformations



- > 35 transformations
- Lots of Data Cleaning
- Generic- and structure dependent implementations

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 publication and Data Analytics
- Current datset contains 32169 files
  - > 10% converted
  - 10 conversion chains

https://github.com/bayerls/statistics2cubes

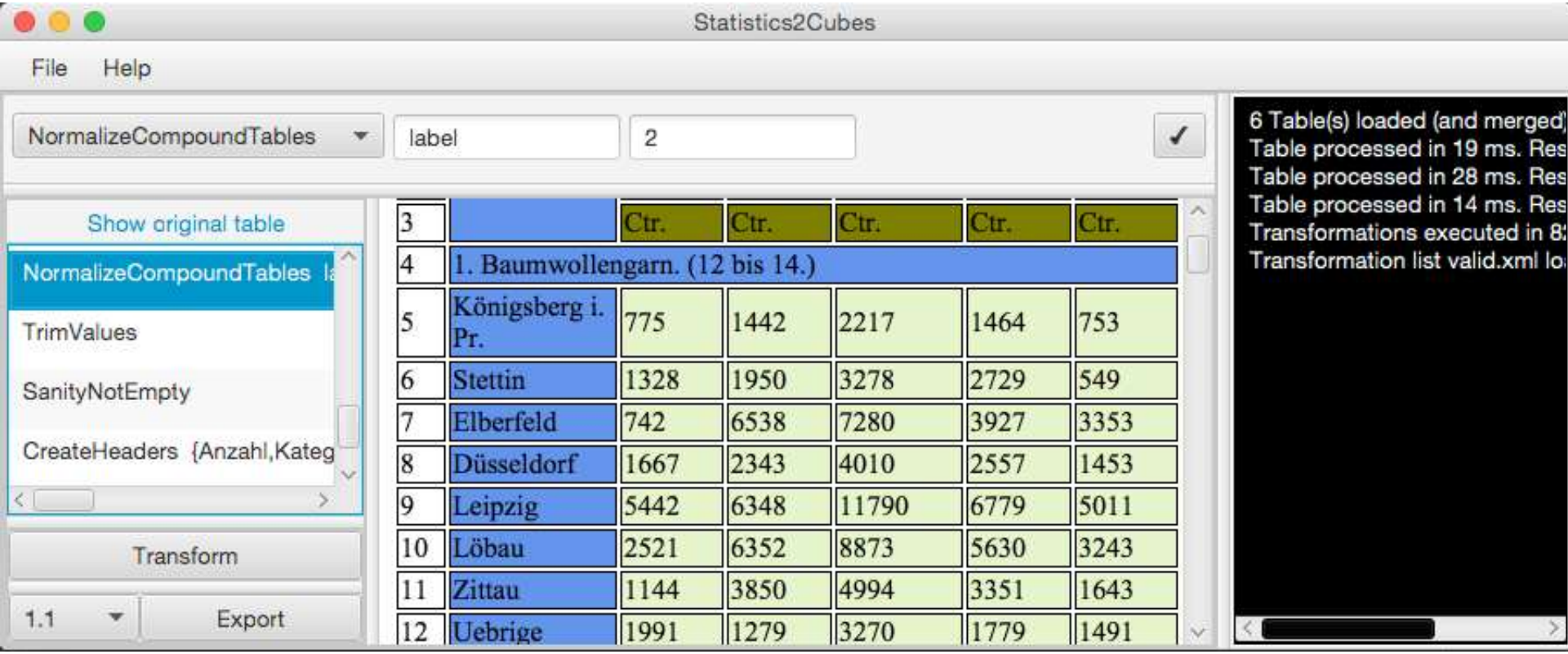
Thank you for your attention!

Question?

Contributions

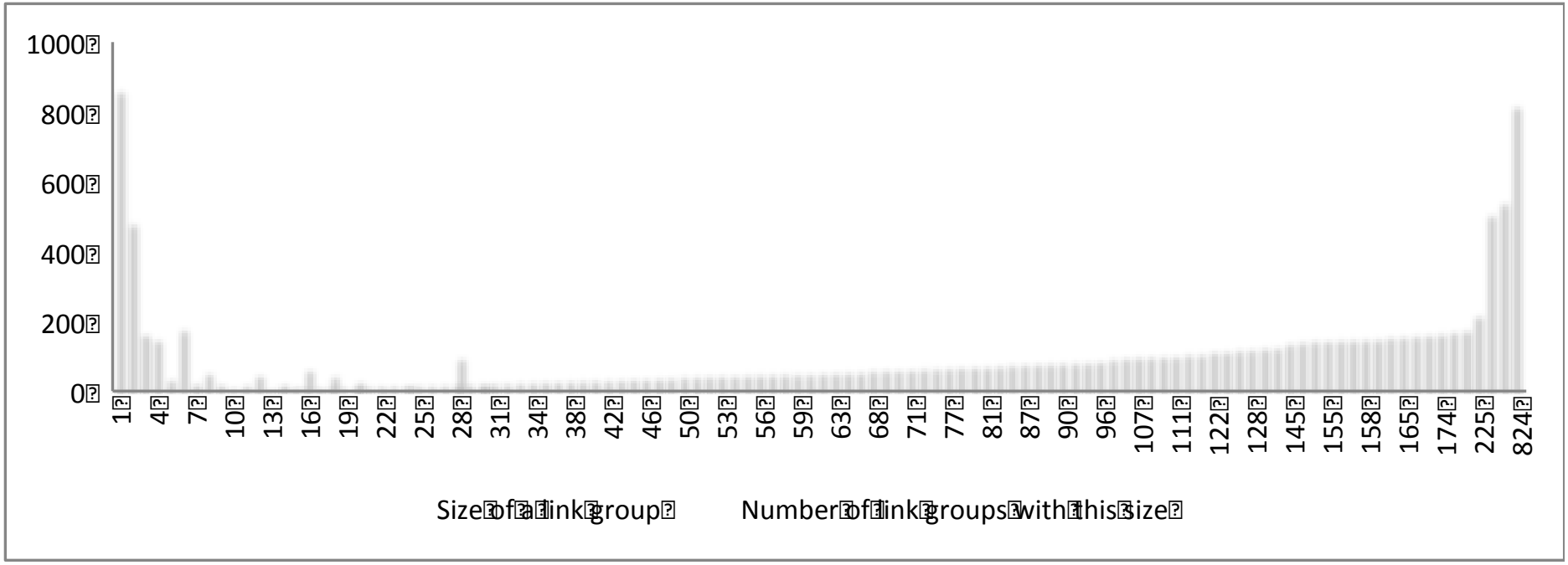
- Define a modular and extendible conversion workflow
- Implemented as an open source research prototype
- Apply Semantic Web Standard to historical statistical data
- Enrich and lift the statistics of the German Reich

GUI



3		Ctr.	Ctr.	Ctr.	Ctr.	Ctr.
4	1. Baumwollengarn. (12 bis 14.)					
5	Königsberg i. Pr.	775	1442	2217	1464	753
6	Stettin	1328	1950	3278	2729	549
7	Elberfeld	742	6538	7280	3927	3353
8	Düsseldorf	1667	2343	4010	2557	1453
9	Leipzig	5442	6348	11790	6779	5011
10	Löbau	2521	6352	8873	5630	3243
11	Zittau	1144	3850	4994	3351	1643
12	Uebrige	1991	1279	3270	1779	1491

Backup


1000
800
600
400
200
0
1
4
7
10
13
16
19
22
25
28
31
34
38
42
46
50
53
56
59
63
68
71
77
81
87
90
96
107
111
122
128
145
155
158
165
174
225
824
Size of a link group
Number of link groups with this size

Example 3

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]

The dataset

- Statistics of the German Reich around the year 1880
- Provided by the ZBW
- Data is available as TEI files
- One page = one table → Link groups
- Highly complex structured
- Hard to interpret, even for humans
- Unclean or even invalid structure or data

-1	0	1	2	3	4
0	I. In den freien Verkehr des Deutschen Zollgebiets getretene Waaren.				
1	Nummern der Waarenverzeichnisse resp. Tarifposition*)	Gebietstheile.	Einfuhr		
2			im 1. Quartal 1873.	im 1. Quartal 1872.	im 1. Quartal 1873 mehr (+) weniger (-).
3	1. (132.) Pos. 1 a.	Glasscherben, Bruchglas. — Ctr. brutto, zollfrei.			
4		Preussen	25330	21861	+3469
5		Bayern	337	378	-41
6		Baden	474	120	+354
7		Oldenburg	2252	2746	-494
8		Elsass-Lothringen	3106	1527	+1579
9		Uebriges Zollgebiet	204	256	-52