Transforming web tables to a relational database*

David W. Embley	Sharad Seth	George Nagy
BYU	UNL	RPI
embley@cs.byu.edu	seth@cse,unl.edu	nagy@ecse.rpi.edu
DBMS	algebra & algorithms	senior programmer

* ICPR 2014, Stockholm, August 21-28, 2014

We gratefully acknowledge Mukkai Krishnamoorthy's contributions via Protegé

Summary

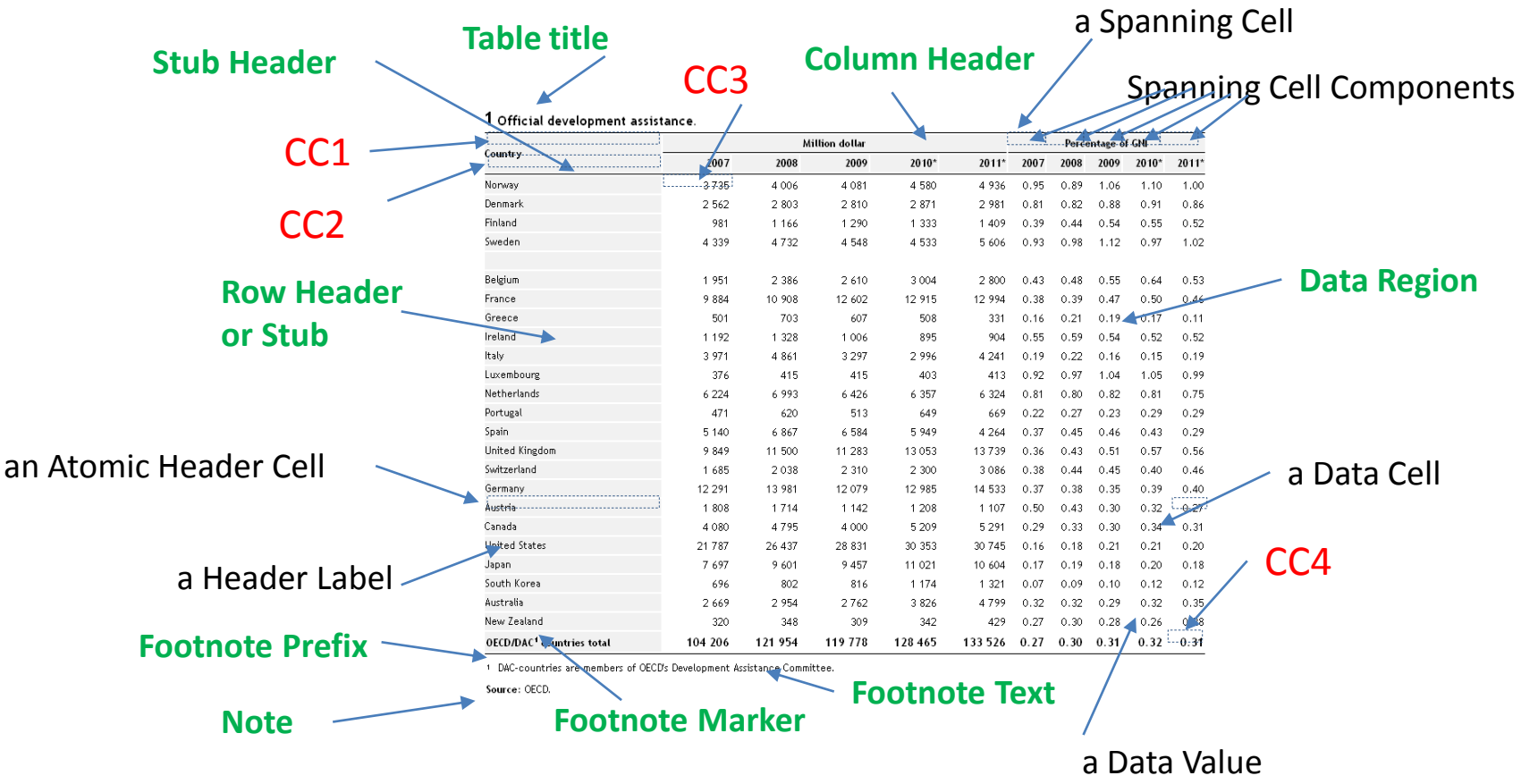
200 web tables imported as CSV files were

segmented, factored, and categorized

using the fundamental indexing property of tables headers.

They were imported into MS Access and queried via SQL.

Table Terminology



1 Official development assistance.

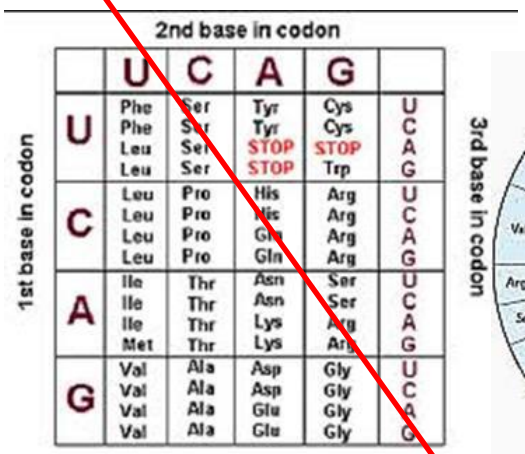
Country	Million dollar					Percentage of GNI				
	2007	2008	2009	2010*	2011*	2007	2008	2009	2010*	2011*
Norway	3 735	4 006	4 081	4 580	4 936	0.95	0.89	1.06	1.10	1.00
Denmark	2 562	2 803	2 810	2 871	2 981	0.81	0.82	0.88	0.91	0.86
Finland	981	1 166	1 290	1 333	1 409	0.39	0.44	0.54	0.55	0.52
Sweden	4 339	4 732	4 548	4 533	5 606	0.93	0.98	1.12	0.97	1.02
Belgium	1 951	2 386	2 610	3 004	2 800	0.43	0.48	0.55	0.64	0.53
France	9 884	10 908	12 602	12 915	12 994	0.38	0.39	0.47	0.50	0.46
Greece	501	703	607	508	331	0.16	0.21	0.19	0.17	0.11
Ireland	1 192	1 328	1 006	895	904	0.55	0.59	0.54	0.52	0.52
Italy	3 971	4 861	3 297	2 996	4 241	0.19	0.22	0.16	0.15	0.19
Luxembourg	376	415	415	403	413	0.92	0.97	1.04	1.05	0.99
Netherlands	6 224	6 993	6 426	6 357	6 324	0.81	0.80	0.82	0.81	0.75
Portugal	471	620	513	649	669	0.22	0.27	0.23	0.29	0.29
Spain	5 140	6 867	6 584	5 949	4 264	0.37	0.45	0.46	0.43	0.29
United Kingdom	9 849	11 500	11 283	13 053	13 739	0.36	0.43	0.51	0.57	0.56
Switzerland	1 685	2 038	2 310	2 300	3 086	0.38	0.44	0.45	0.40	0.46
Germany	12 291	13 981	12 079	12 985	14 533	0.37	0.38	0.35	0.39	0.40
Austria	1 808	1 714	1 142	1 208	1 107	0.50	0.43	0.30	0.32	0.27
Canada	4 080	4 795	4 000	5 209	5 291	0.29	0.33	0.30	0.34	0.31
United States	21 787	26 437	28 831	30 353	30 745	0.16	0.18	0.21	0.21	0.20
Japan	7 697	9 601	9 457	11 021	10 604	0.17	0.19	0.18	0.20	0.18
South Korea	696	802	816	1 174	1 321	0.07	0.09	0.10	0.12	0.12
Australia	2 669	2 954	2 762	3 826	4 799	0.32	0.32	0.29	0.32	0.35
New Zealand	320	348	309	342	429	0.27	0.30	0.28	0.26	0.28
OECD/DAC[1] countries total	**104 206**	**121 954**	**119 778**	**128 465**	**133 526**	**0.27**	**0.30**	**0.31**	**0.32**	**0.31**

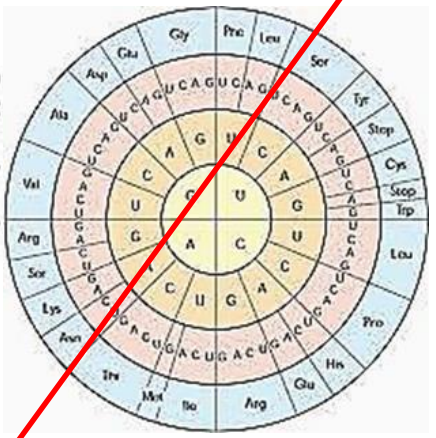
[1] DAC-countries are members of OECD's Development Assistance Committee.

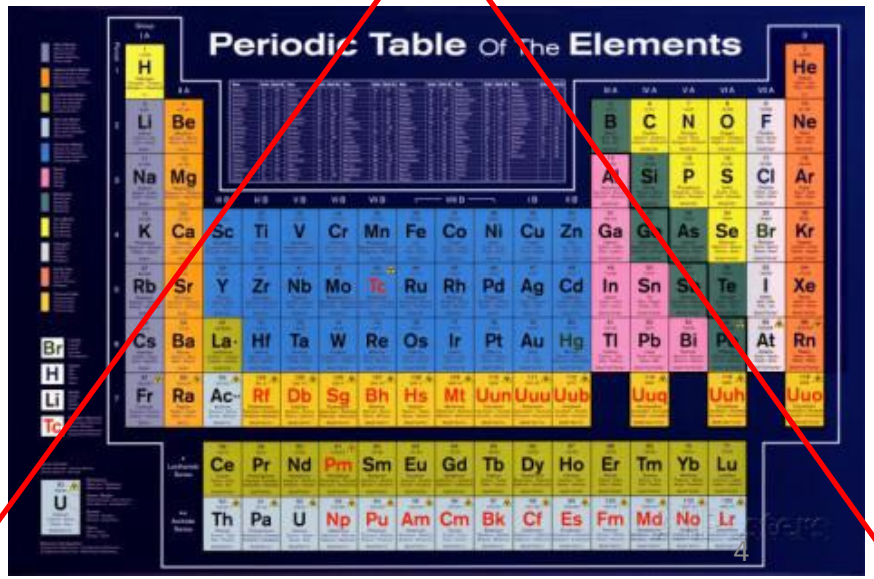
Source: OECD.

Layout of a Well-Formed Table (WFT)

TableTitle		
Notes1		
CC1 StubHeader CC2	ColHeader	ColHeader
Notes2		
RowHeader	CC3 Data	Data
RowHeader	Data	Data
RowHeader	Data	Data CC4
Footnotes		
Notes3		







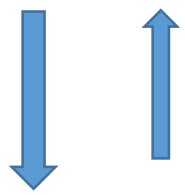
Segmentation via a fundamental indexing table property

that is seldom used in table analysis

What are the row and column headers of this table?

K	K	A	A	A	A	A	A	A	A
K	L	B	B	B	B	C	C	C	C
M	M	D	D	E	E	D	D	E	E
M	N	H	F	H	F	H	F	H	F
K	K	J	K	J	L	A	F	E	G
K	L	D	B	C	E	D	C	E	D
M	M	B	B	B	B	M	M	E	D
M	N	D	E	E	M	N	M	B	B

Column header segmentation using indexing property



K	K	A	A	A	A	A	A	A	A
K	L	B	B	B	B	C	C	C	C
M	M	D	D	E	E	D	D	E	E
M	N	H	F	H	F	H	F	H	F
K	K	J	K	J	L	A	F	E	G
K	L	D	B	C	E	D	C	E	D
M	M	B	B	B	B	M	M	E	D
M	N	D	E	E	M	N	M	B	B

K	K	A	A	A	A	A	A	A	A

K	K	A	A	A	A	A	A	A	A
K	L	B	B	B	B	C	C	C	C

K	K	A	A	A	A	A	A	A	A
K	L	B	B	B	B	C	C	C	C
M	M	D	D	E	E	D	D	E	E

K	K	A	A	A	A	A	A	A	A
K	L	B	B	B	B	C	C	C	C
M	M	D	D	E	E	D	D	E	E
M	N	H	F	H	F	H	F	H	F

K	L	B	B	B	B	C	C	C	C
M	M	D	D	E	E	D	D	E	E
M	N	H	F	H	F	H	F	H	F

Column header candidate with minimal unique column header paths

Column header AND row header segmentation

K	K	A	A	A	A	A	A	A	A
K	L	B	B	B	B	C	C	C	C
M	M	D	D	E	E	D	D	E	E
M	N	H	F	H	F	H	F	H	F
K	K	J	K	J	L	A	F	E	G
K	L	D	B	C	E	D	C	E	D
M	M	B	B	B	B	M	M	E	D
M	N	D	E	E	M	N	M	B	B

Original table and column header candidate

K	K	J	K	J	L	A	F	E	G
K	L	D	B	C	E	D	C	E	D
M	M	B	B	B	B	M	M	E	D
M	N	D	E	E	M	N	M	B	B

Row header candidate
(every row below column header candidate)

K	K	M	M
K	L	M	N
J	D	B	D
K	B	B	E
J	C	B	E
L	E	B	M
A	D	M	N
F	C	M	M
E	E	E	B
G	D	D	B

Transpose, then search for minimal top rows with unique columns:

K	L	M	N

Transpose again

K	B	B	B	B	C	C	C	C
L	D	D	E	E	D	D	E	E
M	H	F	H	F	H	F	H	F
N								

Final column and row headers

Prefixing (adding a row or column) to align header paths with data

Table 9. Numbers of outgoing short messages and multimedia messages from mobile phones in 2002-2008					
Year	Short messages, thousands 1)	Change, %	Short messages/ subscription	Multimedia messages, thousands	Change, %
2003	1 647 218	24,3	347	2 314	
2004	2 193 498	33,2	439	7 386	219,2
2005	2 728 230	24,4	507	15 993	116,5

Table 9. Numbers of outgoing short messages and multimedia messages from mobile phones in 2002-2008					
Year	Short messages, thousands 1)	Short messages, thousands 1)	Short messages/ subscription	Multimedia messages, thousands	Multimedia messages,
ditto	ditto	Change, %	ditto	ditto	Change, %
2003	1 647 218	24,3	347	2 314	
2004	2 193 498	33,2	439	7 386	219,2
2005	2 728 230	24,4	507	15 993	116,5

Factorization and category extraction

		B	B	B	B	C	C	C	C
		D	D	E	E	D	D	E	E
		H	F	H	F	H	F	H	F
	K								
	L								
	M								
	N								

RowHeader (transposed): K+L+M+N Single Category

ColHeader: (B+C)*(D+E)*(H+F) Three categories

RowCat_1	ColCat_1	ColCat_2	ColCat_3
K	B	D	H
L	C	E	F
M			
N			

Factorization and category extraction

		B	B	B	B	C	C	C	C
		D	D	E	E	D	D	E	E
		H	F	H	F	H	F	H	F
	K								
	L								
	M								
	N								

RowHeader (transposed): K+L+M+N Single Category

ColHeader: (B+C)*(D+E)*(H+F) Three categories

RowCat_1	ColCat_1	ColCat_2	ColCat_3
K	B	D	H
L	C	E	F
M			
N			

e.g. M B D F

Mx1 Category table for export to Access

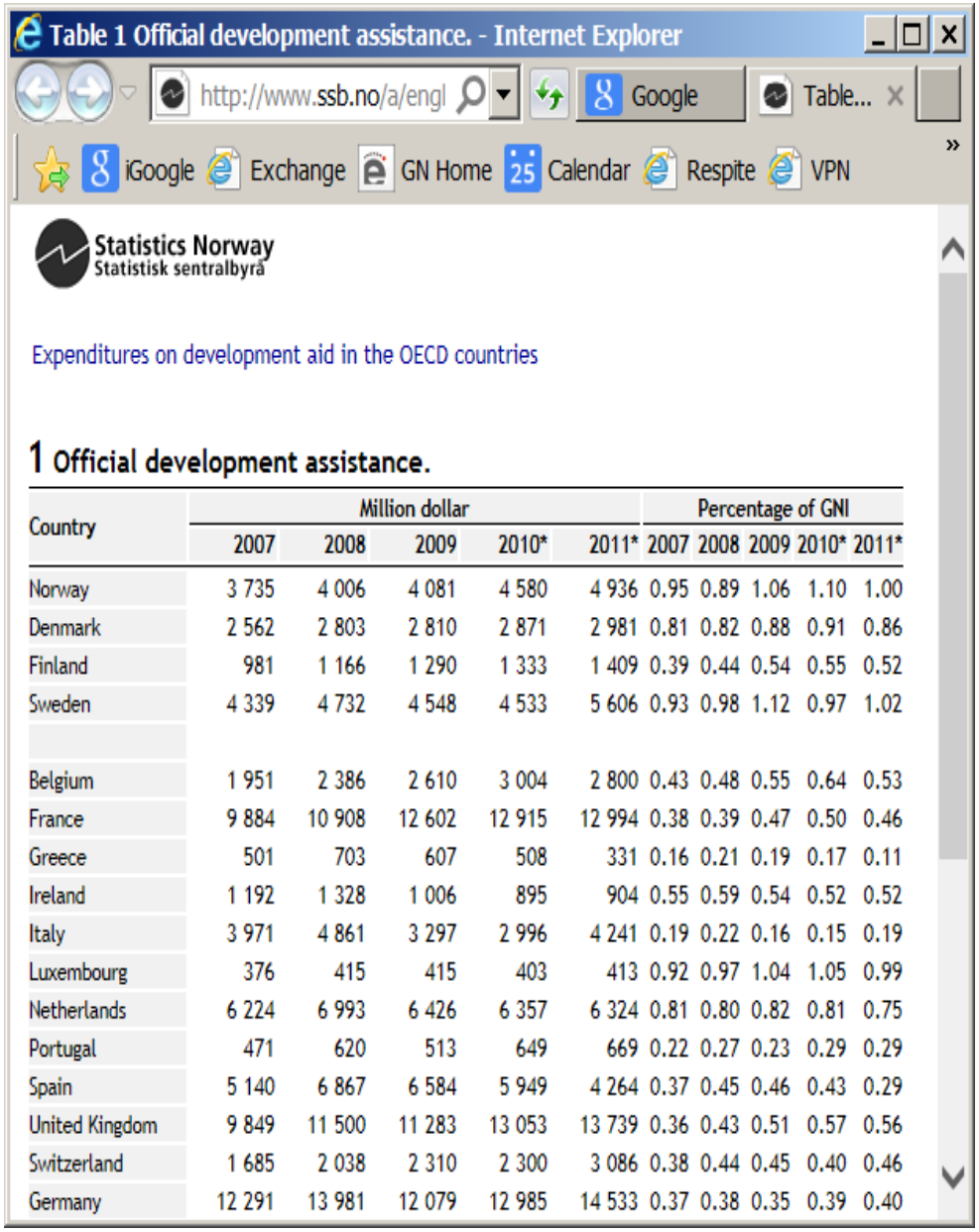


Table 1 Official development assistance. - Internet Explorer

http://www.ssb.no/a/engl

Google

Table...

iGoogle Exchange GN Home Calendar Respite VPN

Statistics Norway
Statistisk sentralbyrå

Expenditures on development aid in the OECD countries

1 Official development assistance.

Country	Million dollar					Percentage of GNI				
	2007	2008	2009	2010*	2011*	2007	2008	2009	2010*	2011*
Norway	3 735	4 006	4 081	4 580	4 936	0.95	0.89	1.06	1.10	1.00
Denmark	2 562	2 803	2 810	2 871	2 981	0.81	0.82	0.88	0.91	0.86
Finland	981	1 166	1 290	1 333	1 409	0.39	0.44	0.54	0.55	0.52
Sweden	4 339	4 732	4 548	4 533	5 606	0.93	0.98	1.12	0.97	1.02
Belgium	1 951	2 386	2 610	3 004	2 800	0.43	0.48	0.55	0.64	0.53
France	9 884	10 908	12 602	12 915	12 994	0.38	0.39	0.47	0.50	0.46
Greece	501	703	607	508	331	0.16	0.21	0.19	0.17	0.11
Ireland	1 192	1 328	1 006	895	904	0.55	0.59	0.54	0.52	0.52
Italy	3 971	4 861	3 297	2 996	4 241	0.19	0.22	0.16	0.15	0.19
Luxembourg	376	415	415	403	413	0.92	0.97	1.04	1.05	0.99
Netherlands	6 224	6 993	6 426	6 357	6 324	0.81	0.80	0.82	0.81	0.75
Portugal	471	620	513	649	669	0.22	0.27	0.23	0.29	0.29
Spain	5 140	6 867	6 584	5 949	4 264	0.37	0.45	0.46	0.43	0.29
United Kingdom	9 849	11 500	11 283	13 053	13 739	0.36	0.43	0.51	0.57	0.56
Switzerland	1 685	2 038	2 310	2 300	3 086	0.38	0.44	0.45	0.40	0.46
Germany	12 291	13 981	12 079	12 985	14 533	0.37	0.38	0.35	0.39	0.40

RowCat_1.1	ColCat_1	ColCat_2.1	DATA
Norway	Million dollar	2007	3 735
Norway	Million dollar	2008	4 006
Norway	Million dollar	2009	4 081
Norway	Million dollar	2010*	4 580
Norway	Million dollar	2011*	4 936
Norway	Percentage of GNI	2007	0.95
Norway	Percentage of GNI	2008	0.89
Norway	Percentage of GNI	2009	1.06
Norway	Percentage of GNI	2010*	1.1
Norway	Percentage of GNI	2011*	1
Denmark	Million dollar	2007	2 562
Denmark	Million dollar	2008	2 803
Denmark	Million dollar	2009	2 810
Denmark	Million dollar	2010*	2 871
Denmark	Million dollar	2011*	2 981
Denmark	Percentage of GNI	2007	0.81
Denmark	Percentage of GNI	2008	0.82
Denmark	Percentage of GNI	2009	0.88
Denmark	Percentage of GNI	2010*	0.91
Denmark	Percentage of GNI	2011*	0.86
Finland	Million dollar	2007	981
Finland	Million dollar	2008	1 166
Finland	Million dollar	2009	1 290

Results of segmentation, category extraction and export to Access

200 heterogeneous web tables from 10 international sites were imported as CSV files

--most format information (geometry, fonts, colors, text positioning) was lost

2 tables not segmented because they had duplicate columns (errors in the table)

197 tables were segmented and categorized correctly

The header of one table omitted one row compared to GT (output useable)

The run time to process all the tables and write the output files was 4 seconds on a

2 GHz Windows 7 laptop running Python 2.7 under IDLE

All 198 Mx1 Category tables that were generated were automatically imported

into MS Access via a short VBA script

The Mx1 Category tables were also converted to RDF triples
and imported and queried in Protegé

MS Access SQL query on imported Mx1 table

Query:

How far and in which direction does Canada's percentage of GNI differs from the overall average for all countries?

Query1 | DevAssistanceTable

```
SELECT t1.RowCat_11 as Country, t1.ColCat_21 as Year, t1.DATA as PercentGNI
   t2.DATA as OverallAvePercentGNI, t1.DATA - t2.DATA as DiffFromAve
FROM DevAssistanceTable t1, DevAssistanceTable t2
WHERE t1.RowCat_11 = "Canada"
   and t1.ColCat_11 = "Percentage of GNI"
   and t2.RowCat_11 = "OECD/DAC1 countries total"
   and t2.ColCat_11 = "Percentage of GNI"
   and t1.ColCat_21 = t2.ColCat_21;
```

Query Result:

Query1 | DevAssistanceTable

Country	Year	PercentGNI	OverallAvePercentGNI	DiffFromAve
Canada	2005	0.34	0.32	0.02
Canada	2006	0.29	0.3	-0.01
Canada	2007	0.29	0.27	0.02
Canada	2008	0.33	0.3	0.03
Canada	2009*	0.3	0.31	-0.01

Footnote prefixes, footnotes, and footnote markers

	2011				
	Hamilton (Ont.)	St. Catharines–Niagara (Ont.)	Kitchener-Cambridge-Waterloo (Ont.)	Brantford (Ont.)	Guelph (Ont.)
	number				
Total households	**282,185**	**160,455**	**181,495**	**52,725**	**54,865**
Total persons in households	708,175	383,970	469,930	133,250	139,675
Average number of persons in household	2.5	2.4	2.6	2.5	2.5
Single-detached house	161,020	108,965	101,215	36,470	32,265
Total persons in households	461,280	282,815	297,110	99,960	92,795
Average number of persons in household	2.9	2.6	2.9	2.7	2.9
Apartment, five or more storeys	44,000	9,305	18,530	3,750	5,160
Total persons in households	74,455	14,890	32,215	5,890	9,035
Average number of persons in household	1.7	1.6	1.7	1.6	1.7
Movable dwelling[1]	385	340	325	80	330
Total persons in households	750	615	570	165	570
Average number of persons in household	1.9	1.8	1.8	2.1	1.7
Other dwelling[2]	76,780	41,845	61,420	12,430	17,110
Total persons in households	171,685	85,645	140,035	27,235	37,270
Average number of persons in household	2.2	2.0	2.3	2.2	2.2

1. Includes mobile homes and other movable dwellings such as houseboats and railroad cars.
2. The category 'Other dwelling' is a subtotal of the following categories: semi-detached house, row house, apartment or flat in a duplex, apartment in a building that has fewer than five storeys and other single-attached house.

Source: Statistics Canada, 2011 Census of Population and Statistics Canada catalogue no. 98-313-XCB.

100%

Category – aggregate ineraction

	number				
Total households	**282,185**	**160,455**	**181,495**	**52,725**	**54,865**
Total persons in households	708,175	383,970	469,930	133,250	139,675
Average number of persons in household	2.5	2.4	2.6	2.5	2.5
Single-detached house	161,020	108,965	101,215	36,470	32,265
Total persons in households	461,280	282,815	297,110	99,960	92,795
Average number of persons in household	2.9	2.6	2.9	2.7	2.9
Apartment, five or more storeys	44,000	9,305	18,530	3,750	5,160
Total persons in households	74,455	14,890	32,215	5,890	9,035
Average number of persons in household	1.7	1.6	1.7	1.6	1.7
Movable dwelling[1]	385	340	325	80	330
Total persons in households	750	615	570	165	570
Average number of persons in household	1.9	1.8	1.8	2.1	1.7
Other dwelling[2]	76,780	41,845	61,420	12,430	17,110
Total persons in households	171,685	85,645	140,035	27,235	37,270
Average number of persons in household	2.2	2.0	2.3	2.2	2.2

1. Includes mobile homes and other movable dwellings such as houseboats and railroad cars.
2. The category 'Other dwelling' is a subtotal of the following categories: semi-detached house, row house, apartment or flat in a duplex, apartment in a building that has fewer than five storeys and other single-attached house.

Source: Statistics Canada, 2011 Census of Population and Statistics Canada catalogue no. 98-313-XCB.

A 5 x 3 two-category single-row row-header

A single-row column header with two categories

Search

Data collections | Products and services | News | Statistics Finland

y use in manufacturing > 2008 > Table 2. Energy use in manufacturing by industry 2008

Suomeksi
På svenska
Print version

Table 2. Energy use in manufacturing by industry 2008

Industries	Fuels TJ	confidence interval, ± %	Electricity TJ 1)	confidence interval, ± %	Heat TJ 1)	confidence interval, ± %	Total TJ	confidence interval, ± %
05 Mining of coal and lignite	.	.	.	.	.	.	.	.
06 Extraction of crude petroleum and natural gas	.	.	.	.	.	.	.	.
07 Mining of metal ores	521,7	58,5	1 051,8	37,8	48,9	5,6	1 622,4	0,0
08 Other mining and quarrying	3 064,7	70,6	831,0	8,5	226,1	12,0	4 121,8	51,7
09 Mining support service activities	.	.	0,1*	138,6	.	.	0,1*	138,6
10 Manufacture of food products	3 454,8	20,0	4 395,4	21,3	4 186,4	57,8	12 036,6	20,8
11 Manufacture of beverages	675,5	11,1	618,8	21,3	876,1	16,6	2 170,3	9,6
12 Manufacture of tobacco products	.	.	.	.	.	.	.	.
13 Manufacture of textiles	619,0	61,7	573,3	37,2	367,6	10,5	1 559,8	26,0
14 Manufacture of wearing apparel	36,6	50,6	119,5	63,9	66,8*	98,1	222,9	41,6
15 Manufacture of leather and related products	33,8	77,8	59,4	67,1	34,8	72,7	128,0	20,6
trailers								
30 Manufacture of other transport equipment	454,8	26,7	862,6	12,1	579,0	9,4	1 896,5	7,9
31 Manufacture of furniture	496,2	73,5	928,9	66,4	282,7	62,5	1 707,7	39,9
32 Other manufacturing	70,1*	89,3	147,6*	80,9	22,5*	82,9	240,2	54,8
33 Repair and installation of machinery and equipment	114,0	31,9	565,3	65,7	156,5	35,4	835,9	43,5
Total	385 186,4	0,8	128 708,6	1,5	59 746,3	5,5	573 641,3	0,8

* Data in table cell are unreliable, as the variation coefficient exceeds the value 40.

. No data available (the sample contained no establishments)

Includes industries (TOL 2008) B Mining and quarrying and C Manufacturing (including autoproducer plants). Does not include the energy use of industry D Electricity, gas and steam supply.Includes the fuel consumption of industrial establishments (including autoproducer establishments) as well as the external net purchases of electricity and heat. Summing these together gives the total energy use in the industry.

1) Net purchases

Previous work

Decades of research on printed, ASCII, and HTML tables have drawn attention to the immense variety of table layouts necessary for quick human comprehension, in contrast to the uniform structure of relational tables.

Research on table analysis appears to be shifting from the PR/ML/CV/DIA community to web and database groups:

W.J. Cafarella,A. Halevy, D.Z. Wang, E. Wu, Y. Zhang, WebTables: Exploring the Power of Tables on the Web, VLDB '08

G. Limaye, S. Sarawagi, S. Chakrabarti, "Annotating and searching web tables, using entities, types, and relationships," Procs. VLDB Endowment, vol. 3, nrs. 1–2, pp. 1338–1347, 2010.

H. Gonzalez et al. Google Fusion Tables: Web-Centered Data Management and Collaboration, SIGMOD'10, 2010

P. Venetis et al., "Recovering semantics of tables on the web," Procs. VLDB Endowment, vol. 4, nr. 9, 2011

M. D. Adelfio and H. Samet, "Schema extraction for tabular data on the web," Procs. VLDB Endowment, vol. 6, nr. 6, 2013.

Z. Chen and M. Cafarella, "Automatic web spreadsheet extraction," Procs. 3rd Wks. Semantic Search Over the Web, 2013.

L. Lautert, M.M. Scheidt, C.F. Dorneles, "Web table taxonomy and formalization," SIGMOD Record, vol. 42, nr. 3, 2013.

Proposed work

Detect and annotate aggregates using factoring results
Improve footnote marker detection (currently only ~70%)
Complete cell classification
Extend headers, using cell format information if necessary
Queries that exploit header hierarchies not just header paths
Cluster category headers to find related tables
Reasoning over OWL ontologies
Extension to printed table images and to PDF tables
Cross the rickety bridge to semantic analysis of table contents

Summary

Segmented, factored, and categorized200 web tables
based on the fundamental *indexing property of tables.*

Demonstrated *algorithmic* rather than heuristic table analysis,
using no cell layout or content information other than testing
whether the symbol strings in two cells are the same.

Transformed web tables with heterogeneous layout into a uniform
category-preserving and database-compatible relational form.

Imported 198 tables into MS Access and *queried* them via SQL.

Thank you

See you at the banquet tables tonight!