End-to-End Conversion of HTML Tables for Populating a Relational Database

George Nagy, RPI, Troy, NY, USA
David W. Embley, BYU, Provo, UT, USA
Sharad Seth, UNL, Lincoln, NE, USA

Motivation

There are *many* tables on the web.

A long-standing goal of table processing is to import them into a database management system where all this data can be combined and queried .

Early research focused on scanned tables, but attention has largely shifted to HTML and spreadsheet tables harvested from the web.

Work in the same spirit reported in 2013:

Adelfio & Samet; Chen & Cafarella; Lautert, Scheidt, & Dorneles.

Outline

1. Segmentation via *indexing* headers

2. Category extraction by *factorization*

3. Conversion to *canonical* form for DBMS

HTLM table: Statistics Norway Table 4

Schools, and pupils attending schools, of various sizes, 2003-2011

Statistics Norway
Statistisk sentralbyrå

Education statistics. Primary and lower secondary schools

4 Schools and pupils by school size. School years 2003/04-2010/11. Per cent

School years	Schools			Pupils		
	Less than 100 pupils	100-299 pupils	300 pupils or more	Less than 100 pupils	100-299 pupils	300 pupils or more
2003/04	36.2	39.0	24.8	8.7	39.3	52.0
2004/05	35.2	39.0	25.8	8.7	38.3	53.0
2005/06	35.2	39.0	25.8	8.8	38.3	52.9
2006/07	34.3	40.0	25.7	8.4	39.0	52.6
2007/08	34.0	39.6	26.4	8.3	38.2	53.5
2008/09	33.3	40.0	26.7	8.1	38.2	53.7
2009/10	32.0	40.7	27.3	7.7	38.2	54.1
2010/11	31.2	41.0	27.8	7.4	38.1	54.5

Explanation of symbols

http://www.ssb.no/a/english/kortnavn/utgrs_en/arkiv/tab-2010-12-09-04-en.html

NB This table has 3 categories: Years, Schools-or-pupils, Size
8 x 2 x 3 = 48 data cells

Table 4 rendered from CSV file

	A	B	C	D	E	F	G
1	4 Schools and pupils by school size. School years 2003/04-2009/10. Per cent						
2	School years	Schools			Pupils		
3		Less than 100	100-299 pupils	300 pupils	Less than 100	100-299	300 pupils
4	2003/04	36.2	39	24.8	8.7	39.3	52
5	2004/05	35.2	39	25.8	8.7	38.3	53
6	2005/06	35.2	39	25.8	8.8	38.3	52.9
7	2006/07	34.3	40	25.7	8.4	39	52.6
8	2007/08	34	39.6	26.4	8.3	38.2	53.5
9	2008/09	33.3	40	26.7	8.1	38.2	53.7
10	2009/101	32	40.7	27.3	7.7	38.2	54.1
11	1	Preliminary figures.					
12	Explanation of symbols						

(b)

Minimum Indexing Point (MIP)

	A	B	C	D	E	F	G
1	4 Schools and pupils by school size. School years 2003/04-2009/10. Per cent						
2	School years	Schools	Schools	Schools	Pupils	Pupils	Pupils
3		Less than 100	100-299 pupils	300 pupils	Less than 100	100-299	300 pupils
4	2003/04	36.2	39	24.8	8.7	39.3	52
5	2004/05	35.2	39	25.8	8.7	38.3	53
6	2005/06	35.2	39	25.8	8.8	38.3	52.9
7	2006/07	34.3	40	25.7	8.4	39	52.6
8	2007/08	34	39.6	26.4	8.3	38.2	53.5
9	2008/09	33.3	40	26.7	8.1	38.2	53.7
10	2009/101	32	40.7	27.3	7.7	38.2	54.1
11	1	Preliminary figures.					
12	Explanation of symbols						

DATA

(c)

Prefixing to ensure unique indexing

Table 32. Urban Population

As of Aug.1	Blois	Male	Female	Tours	Male	Female
1990	47,243	22677	24566	129,509	62164	67345
2000	49,318	23673	25645	133,267	63968	69299
2010	46,013	22086	23927	135,480	65030	70450

As of Aug.1	Blois	Blois	Blois	Tours	Tours	Tours
As of Aug.1	Blois	Male	Female	Tours	Male	Female
1990	47,243	22677	24566	129,509	62164	67345
2000	49,318	23673	25645	133,267	63968	69299
2010	46,013	22086	23927	135,480	65030	70450

NB This table has 3 categories: Year, City, Gender

Row header that requires prefixing

(~5% our tables)

Table A-3: Top 10 U.S. Land Ports for Land Trade with Canada and Mexico: 2003 and 2004									
(Thousands of current U.S. dollars)									
Excel | CSV									
U.S. Port	All land modes			Truck			Rail		
	2003	2004	Percent change	2003	2004	Percent change	2003	2004	Percent ch
U.S.-North American Trade	562,776,436	633,562,711	12.6	392,011,875	452,952,617	15.5	95,724,033	108,360,115	13.2
Top 10 ports	412,424,713	460,654,100	11.7	322,372,568	361,393,164	12.1	79,584,042	87,600,121	10.1
Detroit, MI	101,889,513	113,807,623	11.7	84,810,618	94,019,507	10.9	16,723,319	19,278,278	15.3
Laredo, TX	78,762,959	89,510,852	13.6	54,619,781	63,985,424	17.1	23,940,343	25,398,735	6.1
Buffalo-Niagara, NY	59,369,091	68,351,546	15.1	45,752,599	52,316,608	14.3	9,126,782	10,261,965	12.4
U.S.-Canada Trade	362,319,128	408,612,969	12.8	240,949,027	268,659,618	11.5	64,757,423	74,543,847	15.1
Top 10 ports	288,166,879	323,649,709	12.3	221,837,418	247,417,702	11.5	55,564,511	63,095,059	13.6
Detroit, MI	101,815,113	113,668,714	11.6	84,743,294	93,882,632	10.8	16,718,137	19,276,281	15.3
Buffalo-Niagara, NY	59,275,775	68,283,239	15.2	45,659,600	52,248,579	14.4	9,126,589	10,261,760	12.4
Port Huron, MI	62,244,347	65,879,966	5.8	35,672,586	37,704,369	5.7	22,886,271	23,959,412	4.7
Champlain-Rouses Pt., NY	14,412,634	15,945,026	10.6	12,713,518	14,147,689	11.3	898,156	1,133,615	26.2
Blaine, WA	12,005,376	14,175,533	18.1	9,881,089	11,074,258	12.1	2,098,150	3,092,083	47.4
Alexandria Bay, NY	10,035,184	11,008,768	9.7	10,025,004	11,005,130	9.8	NA	NA	NA
U.S.-Mexico Trade	200,457,309	224,949,742	12.2	163,085,879	184,292,998	13	30,966,610	33,816,269	9.2
Top 10 ports	190,980,524	211,103,066	10.5	158,942,511	179,566,108	13	30,833,262	33,587,526	8.9
Laredo, TX	78,762,959	89,510,852	13.6	54,619,781	63,985,424	17.1	23,940,343	25,398,735	6.1
El Paso, TX	39,204,331	42,779,555	9.1	35,935,405	39,531,129	10	2,472,629	2,928,668	18.4
Otay Mesa, CA	19,678,318	22,188,749	12.8	19,660,724	22,171,883	12.8	NA	NA	NA

MIP Algorithm*

Search from top-left to bottom right to find minimum number of column-header rows and row-header columns.

Backtrack when increasing the number of row-header columns decreases the number of column-header rows.

Reverse scan to eliminate redundant rows above column header.

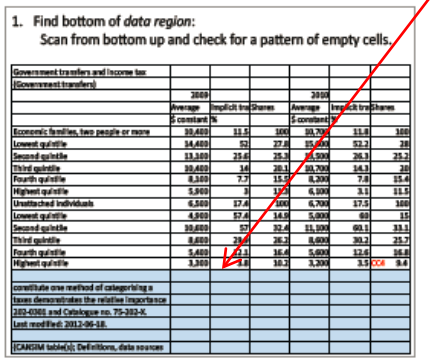
* cf. ICDAR 2013

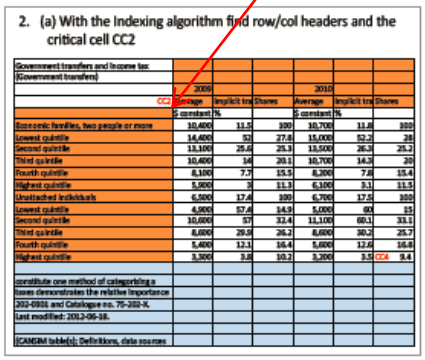
Segmentation with MIP

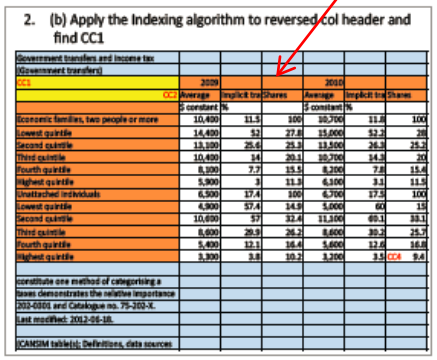
1. Determine bottom of data region
2. Find MIP
3. Eliminate redundant rows above column header
4. Determine top of data region

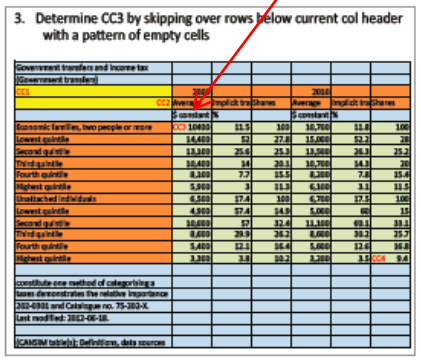
1 2 3 4

From ICDAR 2013:









Categories

The Wang categories determine equivalent layouts.

A 3-category table can be laid out in 8 ways:

- 1 row of data
- 1 column category and two row categories (3 ways)
- 1 row category and two column categories (3 ways)
- 1 column of data

Furthermore rows/columns within a category can be permuted.

A different layout for Table 4

	A	B	C	D	E
			Less than 100 pupils	100-299 pupils	300 pupils or more
1	Schools	2003/04	36.2	39	24.8
2		2004/05	35.2	39	25.8
3		2005/06	35.2	39	25.8
4		2006/07	34.3	40	25.7
5		2007/08	34	39.6	26.4
6		2008/09	33.3	40	26.7
7		2009/10	32	40.7	27.3
8	Pupils	2003/04	8.7	39.3	52
9		2004/05	8.7	38.3	53
10		2005/06	8.8	38.3	52.9
11		2006/07	8.4	39	52.6
12		2007/08	8.3	38.2	53.5
13		2008/09	8.1	38.2	53.7
14		2009/10	7.7	38.2	54.1

“Ordinary” vs. Relational tables

Ordinary:

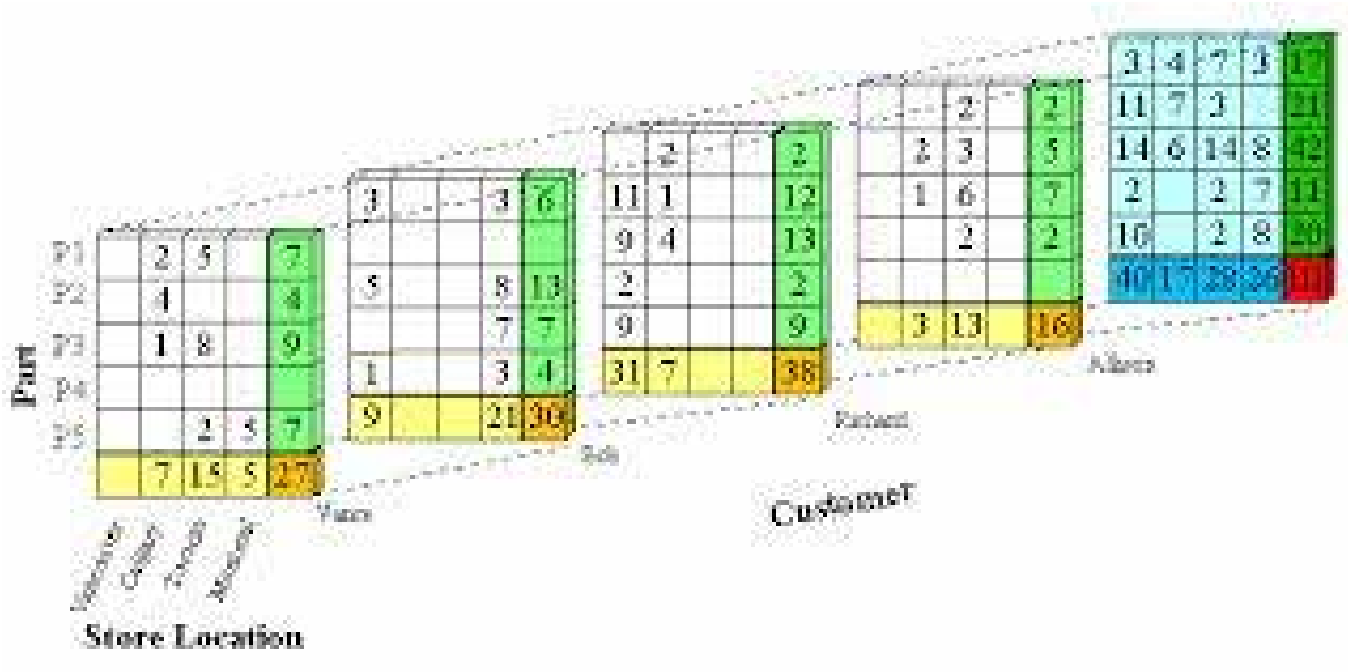
- Rows and columns logically *symmetric*
- Row/column layout is the designer’s choice

Relational:

- Rows are *records* (or *tuples)*
- Columns are *fields (*or *entities)*
- Some fields are *keys*

Data Cube View

A table can be visualized as the values of the cross-product of two or more categories.



Any category may be considered as a key field

Factorization of vertical header

	A	B	C	D	E	F	G
1	4 Schools and pupils by school size. School years 2003/04-2009/10. Per cent						
2	School years	Schools	Schools	Schools	Pupils	Pupils	Pupils
3		Less than 100	100-299 pupils	300 pupils	Less than 100	100-299	300 pupils
4	2003/04	36.2	39	24.8	8.7	39.3	52
5	2004/05	35.2	39	25.8	8.7	38.3	53
6	2005/06	35.2	39	25.8	8.8	38.3	52.9
7	2006/07	34.3	40	25.7	8.4	39	52.6
8	2007/08	34	39.6	26.4	8.3	38.2	53.5
9	2008/09	33.3	40	26.7	8.1	38.2	53.7
10	2009/101	32	40.7	27.3	7.7	38.2	54.1
11	1	Preliminary figures.					
12	Explanation of symbols						

(c)

Sch*<100 + Sch*100-299 + Sch*>300 + Pup*<100+ Pup*100-299 + Pup*>300

Factorized: (Sch + Pup)*(<100 + 100-299 + >300)

* vertical concatenation
+ horizontal concatenation

The three categories of Table 4

RowCat 1	ColCat 1	ColCat 2
2003/04	Schools	Less than 100 pupils
2004/05	Pupils	100-299 pupils
2005/06		300 pupils or more
2006/07		
2007/08		
2008/09		
2009/101		

Importing into DBMS (MS-Access)

Every table is converted to an M x 1 canonical form.

There is one row per data cell.

The row header has as many columns as there are categories.

The column header contains the names of the categories.

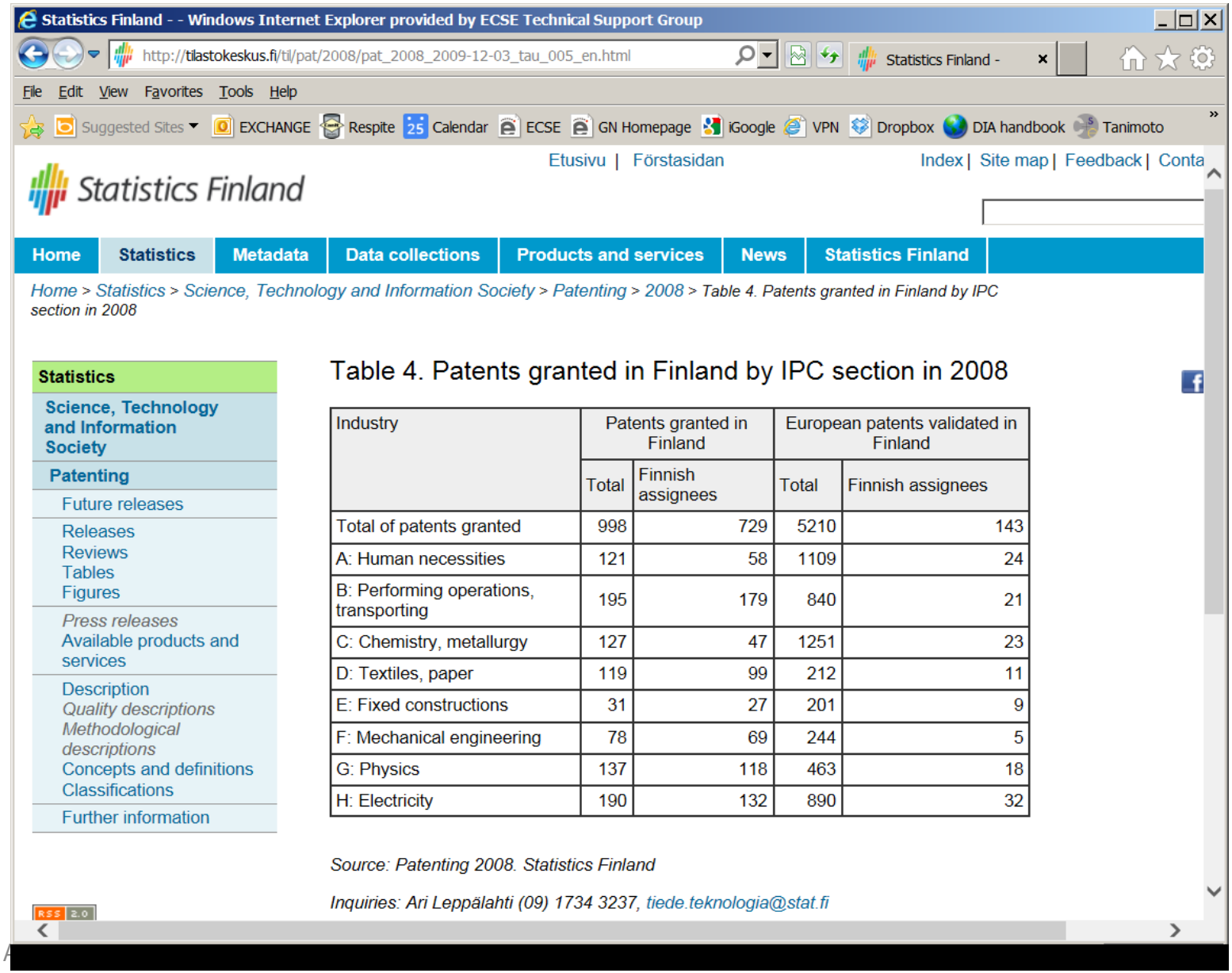
The imported canonical table is manipulated in Access using SQL

Part of the canonical table for Table 4

RowCat_1.1	ColCat_1.1	ColCat_2.1	DATA
2003/04	Schools	Less than 100 pupils	36.2
2003/04	Schools	100-299 pupils	39
2003/04	Schools	300 pupils or more	24.8
2003/04	Pupils	Less than 100 pupils	8.7
2003/04	Pupils	100-299 pupils	39.3
2003/04	Pupils	300 pupils or more	52
2004/05	Schools	Less than 100 pupils	35.2
2004/05	Schools	100-299 pupils	39
2004/05	Schools	300 pupils or more	25.8
2004/05	Pupils	Less than 100 pupils	8.7
2004/05	Pupils	100-299 pupils	38.3
2004/05	Pupils	300 pupils or more	53
2005/06	Schools	Less than 100 pupils	35.2
2005/06	Schools	100-299 pupils	39
2005/06	Schools	300 pupils or more	25.8
2005/06	Pupils	Less than 100 pupils	8.8
2005/06	Pupils	100-299 pupils	38.3
2005/06	Pupils	300 pupils or more	52.9
2006/07	Schools	Less than 100 pupils	34.3
2006/07	Schools	100-299 pupils	40

Some examples from our collection

A two-category column header



CSV version

Source: Statistics Canada, 2006 Census of Population.						
Last modified: 2007-10-30						
T51						
Table 4. Patents granted in Finland by IPC section in 2008						
Industry	Patents granted in Fin		European patents validated in Finland			
	Total	Finnish ass	Total	Finnish assignees		
Total of patents granted	998	729	5210	143		
A: Human necessities	121	58	1109	24		
B: Performing operations,	195	179	840	21		
C: Chemistry, metallurgy	127	47	1251	23		
D: Textiles, paper	119	99	212	11		
E: Fixed constructions	31	27	201	9		
F: Mechanical engineering	78	69	244	5		
G: Physics	137	118	463	18		
H: Electricity	190	132	890	32		

Rare three-column row header

Table 1.9 Net Summer Capacity of Plants Cofiring Biomass and Coal, 2007						
(Megawatts)						
State	Company Name	Plant I.D.	Plant Name	County	Biomass/ (	Total Plant
AL	DTE Energy Services	50407	Mobile Energy Services LLC	Mobile	91	91
AL	Georgia-Pacific Corp	10699	Georgia Pacific Naheola Mill	Choctaw	31	78
AL	International Paper Co	52140	International Paper Prattville Mill	Autauga	49	90
AR	Domtar Industries Inc	54104	Ashdown	Little River	157	157
AZ	Tucson Electric Power Co	126	H Wilson Sundt Generating Statior	Pima	173	559
DE	Conectiv Delmarva Gen Inc	593	Edge Moor	New Castl	252	710
FL	International Paper Co-Pensacola	50250	International Paper Pensacola	Escambia	83	83
FL	Jefferson Smurfit Corp	10202	Jefferson Smurfit Fernandina Bea	Nassau	74	128
FL	Stone Container Corp-Panama Ci	50807	Stone Container Panama City Mill	Bay	20	34
GA	Georgia Pacific CSO LLC	54101	Georgia Pacific Cedar Springs	Early	101	101
GA	International Paper Co-Augusta	54358	International Paper Augusta Mill	Richmond	85	85
GA	SP Newsprint Company	54004	SP Newsprint	Laurens	45	82
HI	Hawaiian Com & Sugar Co Ltd	10604	Hawaiian Comm & Sugar Puunene	Maui	46	62
IA	Ames City of	1122	Ames Electric Services Power Plar	Story	109	109
IA	Archer Daniels Midland Co	10860	Archer Daniels Midland Clinton	Clinton	180	211
IA	University of Iowa	54775	University of Iowa Main Power Pl	Johnson	21	23
KY	East Kentucky Power Coop, Inc	6041	H L Spurlock	Mason	659	1,609
LA	International Paper Co	54090	International Paper Louisiana Mill	Morehous	59	59
MD	NewPage Corporation	50282	Luke Mill	Allegany	65	65
ME	NewPage Corporation	10495	Rumford Cogeneration	Oxford	103	103
ME	S D Warren Co.- Westbrook	50447	S D Warren Westbrook	Cumberlar	15	81
MI	Decorative Panels International, Ir	10149	Decorative Panels Intl	Alpena	8	8
MI	NewPage Corporation	10208	Escanaba Paper Company	Delta	81	103
MI	S D Warren Co	50438	S D Warren Muskegon	Muskegon	51	51
MI	TES Filer City Station LP	50835	TES Filer City Station	Manistee	70	70
MN	Minnesota Power Inc	10686	Rapids Energy Center	Itasca	27	28
MN	Minnesota Power Inc	1897	M L Hibbard	St Louis	73	123
MO	University of Missouri-Columba	50969	University of Missouri Columbia	Boone	6	91
MS	Weyerhaeuser Co	50184	Weyerhaeuser Columbus MS	Lowndes	123	123
NC	Carlyle/Riverstone Renewable Ene	10381	Coastal Carolina Clean Power	Duplin	44	44
NC	Corn Products Intl Inc	54618	Corn Products Winston Salem	Forsyth	8	8
NC	Domtar Paper Company LLC	50189	Domtar Paper Co LLC Plymouth N	Martin	162	162
NC	Primary Energy of North Carolina I	10379	Primary Energy Roxboro	Person	68	68
NY	AES Greenidge	2527	AES Greenidge LLC	Yates	113	163
NY	AES Hickling LLC	2529	AES Hickling LLC	Steuben	70	70
NY	AES Jennison LLC	2531	AES Jennison LLC	Chenango	60	60
NY	Black River Generation LLC	10464	Black River Generation	Jefferson	56	56
NY	Niagara Generation LLC	50202	WPS Power Niagara	Niagara	56	56

Experimental results

Data: 200 tables from large international sites

2	had duplicate rows or columns
197	correctly segmented with minimal headers
1	error (with arguable ground truth)

One-row column headers:	69.0%
Two-row column headers:	26.5%
Three-row headers:	4.5%
Multi-category row or column header	10.5%

Footnotes detected in all 66 tables that had them

Runtime for processing 198 tables was 27s on a 2Ghz laptop (< 0.2s/table)

Algorithms coded in Python.

The single segmentation error

National Center for Education Statistics										
Table SA-3. Percentage distribution of degree-granting institutions, by enrollment size, control and type of institution, and										
Control an	All	Enrollment size								
		Under 200	200– 499	500– 999	1,000– 2,4	2,500– 4,9	5,000– 9,9	10,000– 19	20,000– 29	30,000 or r
Total	100	12	14	15	21	15	12	7	3	1
2-year institutions										
Public	100	1	4	7	23	26	23	12	4	1
4-year institutions										
Public	100	#	2	4	15	17	23	21	12	6
Private nc	100	15	13	17	29	15	6	3	1	#
Private fc	100	15	29	26	18	9	2	1	1	#
Public 2-year institutions										
City	100	#	1	3	8	23	31	23	9	3
Suburban	100	1	2	1	9	17	42	22	6	2

The program accepted the blank under “All” as a valid header.

The Ground Truth did not, and specified a two-row column header.

Note the footnote reference marker.

To-do list

- Table title, units, unreferenced notes
- Footnote prefix, text, reference marker*
- Detect aggregates (total, change, %, ...)
- Clean data: e.g., ‘ 236 000’ à ‘236000’
- Exploit non-indexing parts of headers and stub
- Examples of queries on HTML tables executed in MS-Access submitted to ICPR 2014

* The footnote reference marker is red, the prefix is “*”, and this is the text

Summary

- Segmentation via indexing is independent of script and language, and more reliable than segmentation based on appearance features. Header indexing is the fundamental property of tables.

- Logical table layout analysis into categories is necessary prior to conversion to relational form. We accomplish this by *factoring*.

- The “single-column” canonical table bridges the significant differences between the organization of “ordinary” and relational tables.

Thank you