Table Metadata: Headers, Augmentations and Aggregates

George Nagy*
nagy@ecse.rpi.edu

Raghav Padmanabhan*
padmar2@rpi.edu

Mukkai Krishnamoorthy*
moorthy@cs.rpi.edu

Ramana C. Jandhyala*
ramanachakradhar@gmail.com

William Silversmith*
wsilversmith@gmail.com

ABSTRACT

A sample of 200 web tables was interactively converted into layout-independent Augmented Wang Notation (AWN) using the Table Abstraction Tool (TAT). The resulting XML ground-truth files list for each table (1) cell contents, (2) relationships between the hierarchical column and row headers and the value/content/data cells, (3) designators for aggregates like totals and averages, and (4) ancillary information (*augmentations*) represented by table titles and captions, footnotes, and unit indicators. On average, these tables have 585 cells, 8.8 footnotes, and 1.4 rows of aggregates. They differ widely in number of cells, Wang dimensionality, and MHTML and AWN/XML file sizes. Even though TAT automates much of the repetitive work, interactive ground-truthing took on average four minutes per table. The collected ground truth is offered to the research community for experimentation on automated table processing and for realistic pseudo-random generation of table data.

Categories and Subject Descriptors

E.1 [Data Structures]: Tables

General Terms

Experimentation, Performance, Human Factors

Keywords

Augmentation, Aggregate, Ground Truth, Table Analysis Tool

1. INTRODUCTION

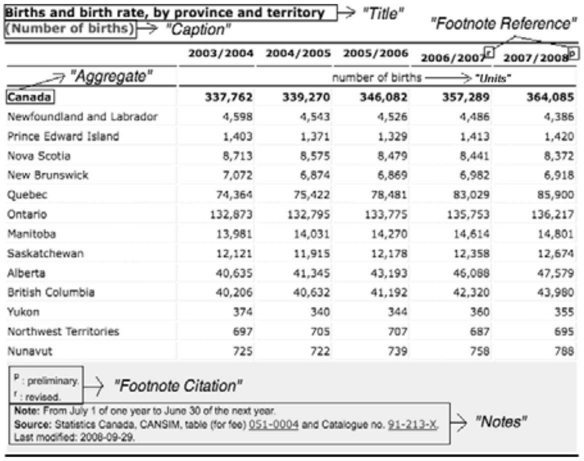
Tables are commonly used to display structured information like calendars, schedules, geopolitical characteristics, financial reports, scientific data, experimental results, and grade reports. Tables can contain words, numbers, formulae, graphics and even tables [5]. They have been adapted to word processors and page composition languages, serve as framework for spreadsheets and relational database systems [11], and share many similarities with forms [8]. They can also be used as query mechanisms [7]. Surveys of table processing contain many pointers to research on converting display tables to more computer-searchable formats [12, 1], but complete table ground-truth files are scarce.

We adopted the layout-independent *Wang Category Notation* to represent header categories and their links to the data cells [10]. Wang's motivation was to provide a flexible tool for laying out tables rather than for extracting information from them. She called the number of category trees in a table its *dimension*. Most tables have two logical dimensions, but 3-D or 4-D tables are not uncommon. Our algorithm constructs the Wang Notation from the geometry of the layout structure [2].

In the next section, we give examples of metadata encountered in tables posted on institutional web sites. In Section 3, we briefly describe TAT, our second-generation tool for interactive extraction of table structure and augmentations, and automatic conversion of this extracted information into AWN/XML. In Section 4 we report the statistical characteristics of the processed tables. In the last section we give some suggestions for possible use of our results and solicit suggestions for expanding the database.

2. AUGMENTATIONS

Augmentations are attributes of a table that are not necessarily based on either the underlying grid structure or the category structure of the table. An augmentation may apply to the entire table (e.g., Table Title, Table Caption, Note), to one or more rows or columns (Unit), or to a single cell of the table (Footnote). Aggregates often serve as proxy headers. Figure 1 illustrates augmentations and aggregates commonly found in tables. They slow down interactive processing because they require non-standard header analysis or repetitive and precise tagging.


"Title"
"Caption"
"Footnote Reference"
"Aggregate"
"Units"
"Footnote Citation"
"Notes"

Births and birth rate, by province and territory
(Number of births)

	2003/2004	2004/2005	2005/2006	2006/2007[r]	2007/2008[p]
	number of births				
Canada	**337,762**	**339,270**	**346,082**	**357,289**	**364,085**
Newfoundland and Labrador	4,598	4,543	4,526	4,486	4,386
Prince Edward Island	1,403	1,371	1,329	1,413	1,420
Nova Scotia	8,713	8,575	8,479	8,441	8,372
New Brunswick	7,072	6,874	6,869	6,982	6,918
Quebec	74,364	75,422	78,481	83,029	85,900
Ontario	132,873	132,795	133,775	135,753	136,217
Manitoba	13,981	14,031	14,270	14,614	14,801
Saskatchewan	12,121	11,915	12,178	12,358	12,674
Alberta	40,635	41,345	43,193	46,088	47,579
British Columbia	40,206	40,632	41,192	42,320	43,980
Yukon	374	340	344	360	355
Northwest Territories	697	705	707	687	695
Nunavut	725	722	739	758	788

[p] : preliminary.
[r] : revised.

Note: From July 1 of one year to June 30 of the next year.
Source: Statistics Canada, CANSIM, table (for fee) 051-0004 and Catalogue no. 91-213-X.
Last modified: 2008-09-29.

Figure 1. Table with title, caption, footnotes, notes, units, and an aggregate serving as a proxy header

*DocLab, Rensselaer Polytechnic Institute, Troy, NY 12180

DAS '10, June 9-11, 2010, Boston, MA, USA
Copyright © 2010 by the author(s)/owner(s) 978-1-60558-773-8/10/06

3. Table Abstraction Tool

Spreadsheets are well suited for table processing because they represent tables on a grid of rows and columns. The Table Abstraction Tool (TAT) is an interactive tool (written in Visual Basic) that converts tables from HTML pages imported into Microsoft Excel to Augmented Wang Notation. TAT can process only well-formed tables (WFT). The requirements listed in [3] for well-formed tables are:

1. The table must have two or more categories;
2. Each category must have a *root* header (sometimes requiring the addition of one or more *virtual headers*);
3. Every delta (data/value/content) cell must be specified by *n* paths, one through each category tree;
4. A category tree cannot contain identical root-to-leaf paths.
5. Category cells appear only in the topmost rows and leftmost columns of the table.

Lists have no headers and only one row or column of data cells. *Linear tables* consist of a single row or column category header that indexes one column or row of data cells. TAT can process one-dimensional tables.

If a table is not accepted by TAT, the operator uses Excel commands to change it into a WFT. After verifying the syntactical validity of the edited table, TAT creates category notation for the relationship of the header hierarchies to the content cells. The edits can be visually verified by highlighting either the headers that index a designated content cell, or the content cells indexed by a designated header (a method based on studies of ambiguities in table interpretation [4] and developed for our earlier MATLAB-based Wang Notation Tool [3]). TAT also processes augmentations like footnotes and aggregates tagged by the operator. The Augmented Wang Notation (AWN), which contains both the category information and the augmentations, is rendered into an XML file for portability. All relevant user interactions are timed and logged for later analysis.

4. Web Table Characteristics

We sought to determine the factors that affect the scalability of processing web tables using TAT. We collected and processed 200 Excel and HTML tables from ten non-profit websites (c.f. Table 1). The table selection process and the experimental protocol were reported in [6].

Table 1. URLs of table sources

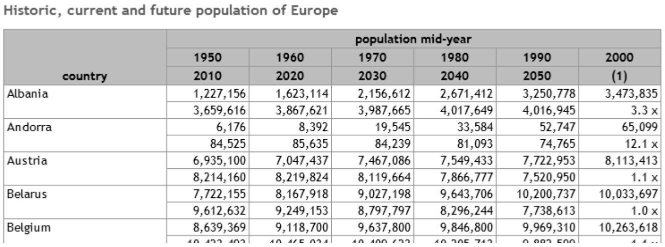
Site	URL
1	http://www.statcan.gc.ca/
2	http://www.sciencedirect.com/
3	http://www.worldbank.org/
4	http://www.ssb.no/english/
5	http://www.ojp.usdoj.gov
6	http://www.geohive.com/
7	http://www1.lanic.utexas.edu/la/region/aid/aid98/
8	http://eia.doe.gov/
9	http://ies.ed.gov/
10	http://www.census.gov/population/www/socdemo/voting/cps2006.html

The average number of cells differs considerably between sources (Table 2). Sources #5 and #10 have large tables which increased their processing time. Sources #6 and #7 contain some tables that took extra time because they were constructed poorly or laid out unconventionally.

Table 2. Summary of table features for each source.

Source	Number of tables	Avg. no. of cells	Median no. of cells	Avg. Proc. Time (sec)	Source Specific ρ
1	21	125	126	127	0.78
2	15	125	100	134	0.85
3	17	344	240	173	0.97
4	19	363	375	194	0.47
5	21	2525	2394	363	0.77
6	24	317	204	179	0.43
7	16	517	473	345	0.45
8	23	364	315	302	0.87
9	23	263	180	195	0.72
10	8	1103	656	377	0.99
Total	**187**	**585**	**264**	**231**	0.71

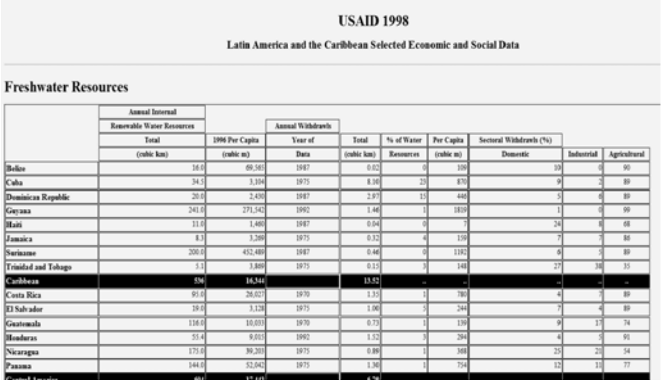
Examples of irregular tables are shown below. In Fig. 1, the aggregate cell is located at the top of the subcategories that it aggregates while in Fig. 3 "Caribbean" is at the bottom. These aggregates can be identified only with external knowledge: the table in Fig. 3 might be processed incorrectly by human operators who do not know that Cuba is in the Caribbean.



Historic, current and future population of Europe

country	population mid-year					
	1950	1960	1970	1980	1990	2000
	2010	2020	2030	2040	2050	(1)
Albania	1,227,156	1,623,114	2,156,612	2,671,412	3,250,778	3,473,835
	3,659,616	3,867,621	3,987,665	4,017,649	4,016,945	3.3 x
Andorra	6,176	8,392	19,545	33,584	52,747	65,099
	84,525	85,635	84,239	81,093	74,765	12.1 x
Austria	6,935,100	7,047,437	7,467,086	7,549,433	7,722,953	8,113,413
	8,214,160	8,219,824	8,119,664	7,866,777	7,520,950	1.1 x
Belarus	7,722,155	8,167,918	9,027,198	9,643,706	10,200,737	10,033,697
	9,612,632	9,249,153	8,797,797	8,296,244	7,738,613	1.0 x
Belgium	8,639,369	9,118,700	9,637,800	9,846,800	9,969,310	10,263,618

Figure 2. A table (cropped) with an unconventional layout. Geohive: http://www.geohive.com/earth/his_proj_europe.aspx



USAID 1998

Latin America and the Caribbean Selected Economic and Social Data

Freshwater Resources

	Annual Internal		Annual Withdrawls						
	Renewable Water Resources								
	Total	1996 Per Capita	Year of	Total	% of Water	Per Capita	Sectoral Withdrawls (%)		
	(cubic km)	(cubic m)	Data	(cubic km)	Resources	(cubic m)	Domestic	Industrial	Agricultural
Belize	16.0	69,565	1987	0.02	0	109	10	0	90
Cuba	34.5	3,104	1975	8.10	23	870	9	2	89
Dominican Republic	20.0	2,430	1987	2.97	15	446	5	6	89
Guyana	241.0	271,542	1992	1.46	1	1819	1	0	99
Haiti	11.0	1,460	1987	0.04	0	7	24	8	68
Jamaica	8.3	3,269	1975	0.32	4	159	7	7	86
Suriname	200.0	452,489	1987	0.46	0	1192	6	5	89
Trinidad and Tobago	5.1	3,869	1975	0.15	3	148	27	38	35
Caribbean	536	16,344		13.52	..	..	..	..	..
Costa Rica	95.0	26,027	1970	1.35	1	780	4	7	89
El Salvador	19.0	3,128	1975	1.00	5	244	7	4	89
Guatemala	116.0	10,033	1970	0.73	1	139	9	17	74
Honduras	55.4	9,015	1992	1.52	3	294	4	5	91
Nicaragua	175.0	39,203	1975	0.89	1	368	25	21	54
Panama	144.0	52,042	1975	1.30	1	754	12	11	77

Figure 3. A poorly constructed table (cropped). USAID 1998, Latin America and the Caribbean Selected Economic and Social Data: http://www1.lanic.utexas.edu/la/region/aid/aid98/environment/tab10.html

The source-specific correlation coefficient ρ in Table 2 indicates the extent to which table processing time is predicted by the Wang dimensionality and number of cells, footnotes and aggregates of the table. Some of the sites with many tables show surprisingly high correlation.

Table 3. Footnotes and Aggregates by source. The numbers in parentheses indicate the number of tables with footnotes or aggregates in each source

Source	Average # Footnotes	Median # Footnotes	Average # Aggregates	Median # Aggregates
1	5.5 (11)	2	1.4 (8)	1
2	8.5 (2)	8	1.0 (1)	1
3	2.0 (1)	2	2.4 (5)	1
4	3.7 (11)	1	2.6 (14)	1
5	27.8 (12)	2	1.2 (10)	1
6	6.3 (3)	1	1.5 (6)	1
7	60.0 (1)	60	2.8 (5)	3
8	51.6 (14)	22	6.2 (18)	4
9	23.1 (14)	2	1.6 (12)	1
10	21.0 (3)	18	5.8 (5)	4
Total	**22.8 (72)**	**2**	**3.0 (84)**	**1**

There is considerable difference between sources in the number of tables with footnotes or aggregates (Table 3). Table 3 also shows a significant difference between the average and median number of footnotes for most of the sources, indicating some tables with many footnotes. However, the same cannot be inferred about aggregates, because the tables in these websites are organized into groups based on their subject matter (similar layouts across tables in a group reveal relationships between related data).

Table 4. Size of different table representations. Italics indicate weighted averages

sour-ce	# of tables	# HT-ML	# Excel +CSV	Avg. size (KB)	Avg. # cells	Size XML (KB)	Ratio original / XML
1	21	21	0	117	125	20	5.8
2	15	3	0	874	125	17	50.6
3	17	4	13	139	344	48	2.9
4	19	19	0	44	363	93	0.5
5	21	1	20	20	2525	239	0.1
6	24	19	0	35	317	71	0.5
7	16	14	0	100	517	90	1.1
8	23	0	23	20	364	67	0.3
9	23	0	23	18	263	37	0.5
10	8	0	8	25	1102	189	0.1
Tot.	**187**	**81**	**87**	***121***	***585***	***82***	***1.5***

Table 4 lists the sizes of different representations of the tables. Some sources contain both MHTML and Excel/CSV tables. The discrepancies in sources #2, #6 and #7 between the number of tables (column two) and the sum of the number of different file types (columns three and four) is due to files with multiple tables. The size of the AWN/XML ground truth (Fig. 5) depends mainly on the number of cells (both header and data), the number of footnotes and the number of aggregates. However, the number of cells is a good predictor of XML file size even though some notes and footnotes contain lengthy text.

```
<TableOntology>
 <Table TableOID="tableOID" Title="AGRICULTURE" DocumentCitation="Lynn, S. and Embley, D.W., Semantically Conceptualizing
and Annotating Tables, Technical Report, Brigham Young University, July 2008, www.deg.byu.edu/papers/TableConceptualization.pdf"
Number=>
   <CategoryRootNodes>    <CategoryRootNode CategoryRootNodeOID="C1"/>…    </CategoryRootNodes>
   </Table>   <CategoryNodes>    <CategoryNode CategoryNodeOID="C1" Label="Year"></CategoryNode>…
    </CategoryNodes>   <CategoryParentNodes>   <CategoryParentNode      CategoryParentNodeOID="C1">
     <CategoryNodes>      <CategoryNode CategoryNodeOID="C1.1"></CategoryNode>…     </CategoryNodes>
      </CategoryParentNode>…    </CategoryParentNodes>
    <DataCells>
       <DataCell DataCellOID="D1,1" DataValue="3254">         <CategoryLeafNodes>
        <CategoryLeafNode CategoryLeafNodeOID="C1.1.1.1" />
        <CategoryLeafNode CategoryLeafNodeOID="C2.1" />     </CategoryLeafNodes>    </DataCell>…</DataCells>
 <Augmentations>
   <Augmentation AugmentationOID="A1"AugmentationType="Units">
    <CategoryNode CategoryNodeOID=C1.1.1.1/>  …  </Augmentation></Augmentations>
</Table Ontology>
```

Figure 4. A fragment of the verbose AWN/XML notation for a table. Ellipses indicate the many missing entries.

5. Discussion

We prepared ground-truthed test data and reported the statistical profile of tables randomly selected from large institutional web sites. Although X. Wang already observed the importance of metadata more than a decade ago [10], we believe that this collection is the first to represent it in the ground truth. Over one third of the tables contain footnotes and aggregates. They simply cannot be ignored if the objective is to transform tables meant for visual inspection into a form suitable for any type of automated search or for populating a knowledge base.

The metadata was extracted interactively. After analyzing the header regions, TAT automatically assigns each delta cell to the appropriate category path, processes annotation and aggregate tags, and generates the XML file. Analysis of the log of operator interventions reveals that the time consuming aspects of the interaction are (1) transforming row headers into a format accepted by TAT, (2) tagging the augmentations, and (3) verifying connections between headers and delta cells.

The time for item (1) could be reduced by extending TAT's automatic coverage to the most common row header layouts. Row-oriented Western scripts and the customary page and display formats result in table layouts that are distinctly asymmetric in orientation. This only increases the benefits of the layout-independent Wang representation.

The necessary human time is affected as much by the number and kind of augmentations as by the size of the table and the complexity of its header layout. Some of the augmentations like units and footnotes are generally consistent in format within each source, so it should not be difficult to detect and extract them automatically. Others, like aggregates, will be harder to detect and classify automatically.

Verification of the structural interpretation by highlighting logically connected header and delta cells works well but is dependent on the operator's skill. Furthermore, it does not verify the whole process. In our experiment the most common errors were in the table collection process where the same tables were collected more than once. In retrospect we should have checked each table for duplication. Future experiments of this kind would benefit from automation of the table harvesting process, but we are well aware of the difficulty of reliable automated table detection [9]. HTML table constructs are often used to format text or figures.

Although the preparation and checking of the ground truth for even 200 tables was time-consuming, we recognize that this dataset is too small for conclusive experimentation on automated table processing. Our next target is 2000 tables from the same websites. In order to avoid excessive human labor, we shall first automate the detection of additional common header formats, and of units and footnotes. We have also conducted, following the footsteps of others, pilot experiments on 50 tables to extract appearance features (mainly text formatting within cells). Most of these features are preserved when the web tables are imported into Excel. Indentations, for instance, often provide a clue to the presence of aggregates. Another item on our agenda is to have a subset of tables processed by several operators to determine their consistency and individual ramp-up times.

We are open to suggestions for improving the usefulness of such data to the research community. So far we have avoided any script-specific approaches, but are aware of the richness added to table layout by bidirectional Oriental scripts. We are most interested in pointers to large collections of tables and tools that may serve to build domain-specific ontologies.

Another possible application of the statistics reported above is the generation of simulated tables. Competitions for processing other types of documents have long been a valued feature of ICDAR and GREC conferences. Using the information we collected, it would be possible to set realistic parameters for a table synthesis program. While simulated data avoids tedious and error-prone ground-truth generation, it contributes to progress only to the extent that it reflects the characteristics of real data.

We will shortly post at least the AWN/XML files, the edited Excel tables, the log files, and the date-stamped URLs of all the web tables that we have processed so far on the IAPR TC-11 website (http://www.iapr-tc11.org).

6. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grants # 044114854 and 0414644. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation. We have also benefited from the support of the Rensselaer Center for Open Software (RCOS). Max Muthalathu (RPI student) programmed the XML conversion. We gratefully acknowledge the influence of two decades of discussions about tables and ontologies with Prof. D.W. Embley of Brigham Young University.

7. REFERENCES

[1] D.W, Embley, M. Hurst, D. Lopresti, G. Nagy: Table Processing Paradigms: A Research Survey. *Int. J. Doc. Anal. Recognit. 8* (2-3), 66-86, 2006.

[2] R.C. Jandhyala, G. Nagy, S. Seth, W. Silversmith, M. Krishnamoorthy, R. Padmanabhan: From tessellations to table interpretation, L. Dixon et al. (Eds.): *Calculemus/MKM 2009*, Springer-Verlag, Berlin, vol. 5625 of Lecture Notes in Artificial Intelligence, pp. 422-437, 2009.

[3] P. Jha, G. Nagy: Wang Notation Tool: Layout Independent Representation of Tables, *Proceedings of the Nineteenth International Conference on Pattern Recognition* (ICPR'08),Tampa, 2008.

[4] J. Hu, R. Kashi, D. Lopresti, G. Nagy, G. Wilfong: Why table ground-truthing is hard, *Proceedings of the Sixth International Conference on Document Analysis and Recognition,* pp. 129–133.Seattle, WA, 2001.

[5] D. Lopresti, G. Nagy: A Tabular Survey of Automated Table Processing, *Graphics Recognition: Recent Advances*, Springer-Verlag, Berlin, vol. 1941 of Lecture Notes in Computer Science, pp. 93-120, 2000.

[6] R. Padmanabhan, R. C. Jandhyala, M. Krishnamoorthy, G. Nagy, S. Seth, W. Silversmith: Interactive Conversion of Large Web Tables, *Proceedings of Eighth International Workshop on Graphics Recognition*, GREC 2009, Published by City University of La Rochelle, La Rochelle, France, July 22-23, 2009.

[7] R. Padmanabhan, G. Nagy: Query By Table, *Proceedings of the Nineteenth International Conference on Pattern Recognition* (ICPR'08),Tampa. 2008.

[8] F. Rahman, B. Klein (editors): Special Issue on detection and understanding forms for document processing applications, *Int. J. Doc. Anal. Recognit.* 8(2-3): 2006.

[9] Y. Wang, J. Hu: Automatic Table Detection in HTML Documents, *Web Document Analysis: Challenges and Opportunities*, pp. 135-154, October 2003.

[10] X. Wang.: Tabular Abstraction, Editing, and Formatting, Ph.D Dissertation, University of Waterloo, Waterloo, ON, Canada, 1996.

[11] G. Weikum, G. Kasneci, M. Ramanath, F. Suchanek: Database and Information-Retrieval Methods for Knowledge Discovery, *Communications of the ACM*, 52(4): 56-64, April 2009.

[12] R. Zanibbi, C. Blostein, J.R. Cordy: A survey of table recognition: Models, observations, transformations, and inferences, *Int. J. Doc. Anal. Recognit.* 7(1):1–16, 2004