Table Abstraction Tool

by

Raghav Krishna Padmanabhan

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the degree of

MASTER OF SCIENCE

Major Subject: ELECTRICAL ENGINEERING

Approved:

_____

Dr. George Nagy, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

May, 2009

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The Table Abstraction Tool (TAT) is an interactive tool that converts tables from HTML pages imported into MS Excel to layout independent Augmented Wang XML. This XML file not only records all relationships between cells of a table in an abstract form that does not rely on the physical structure of tables but also includes information about title, caption and various augmentations like aggregates and footnotes. TAT builds on the Wang Notation Tool (WNT) by improving the interaction with the system and the quality of the XML file, which is used in building domain ontologies. Like WNT, TAT also forms intermediate category trees describing the relationships within each category. The categories constructed by TAT are shown to the user for approval or further interactive editing via Excel. TAT is robust and almost always produces the correct Wang XML representation for a table if the input table is transformed to the desired format. The XML file is based on the indented notation representation of the category trees. The evaluation of TAT was conducted on a collection of 200 tables, mainly from the geopolitical domain. TAT was able to produce the Wang XML for 193 of them. Observations on the cost (i.e., operator time) of interaction as a function of various table characteristics are reported. Detailed analysis of the results reveals a number of options for accelerating web table entry. The tables that describe the results of the experiment were themselves processed through TAT and converted to Wang Notation as a start towards the construction of a table processing ontology.

1. Introduction

The current World Wide Web (WWW) presents information in the form of pages combining natural language text and graphics. Humans can process such information, make sense of the data and create mental associations between different pages. It would be desirable, especially with the information explosion we are facing today, to realize a Semantic Web in which software agents could combine information from different sources in the web, automatically draw analogies, make sense of partial information in some cases and take decisions based on the available data [1]. This is not a trivial task as it involves issues involving semantics, mutual understanding, concept matching and interoperability. Ontologies are offered as a prospective solution to these problems [2]. However, the prevalent assumption is that building ontologies is a human intensive task. The project TANGO [3] is an effort to disprove this notion by demonstrating that almost automatic generation of ontologies is possible.

1.1 TANGO

TANGO (Table ANalysis for Generating Ontologies) is a collaborative effort between several universities and departments to generate ontologies, with little human intervention, from information contained in tables. Using concepts from ontology building, table understanding, conceptual modeling and machine learning, TANGO aims to find a potential solution to generate ontologies almost automatically. TANGO operates in four steps:

1. *Understand* a table's structure, its conceptual content and discover the constraints between the concepts in a table.
2. Match concepts obtained from *understanding* different tables and construct a mini-ontology.
3. Discover the mappings present between the mini-ontologies.
4. Merge the mini-ontologies to form a domain ontology.

To illustrate the above mentioned ideas, tables from geopolitical domain are chosen. The work described in this thesis is essentially the implementation of Step 1 in project TANGO and a record of the observations made in the process about web tables. Steps 2-4 are implemented at Brigham Young University, Provo, Utah.

The Wang Notation Tool (WNT) [4] was developed to *understand* and *interpret* web tables. WNT was able to handle a variety of tables, both in shape and size. The Table Abstraction Tool (TAT) builds on WNT by improving the interaction with the system and the quality of the output produced.

1.2 Tables

Tables are the customary means of representing structured data. They contain words, numbers, formulae, graphics and sometimes even tables. They have been adapted to word processors and page composition languages and form the underlying paradigm for spreadsheets and relational database systems. Some common examples of data usually presented in the form of tables are calendars, geopolitical data, financial data, scientific data, experimental results, and grade reports. Tables have also been shown to be successful query mechanisms [5].

1.2.1 Wang Notation

Wang proposed a layout-invariant representation of tables [6]. Wang defines an *abstract table* as an *abstract data type* and its layout structure as the *presentation form* of a table. The logical structure consists of entries and labels. The organization of the labels is called a *frame*, and the number of *categories* in the frame is the *dimension* of the abstract table. Applying a *layout specification* to an abstract table generates a *concrete table*. Tabular abstraction separates the logical structure of a table from its layout structure. The advantages of tabular abstraction are that the tables can be manipulated independently of their layout structure and we can easily alter the layout of a table by associating different topologies with the logical structure. Informally, the Wang Notation consists of two components (C, δ) where C is a finite set of labeled domains and δ is a mapping from the tree paths labels (or headers) to the possible values. Consider the simple 3-dimensional table shown in Figure 1 where the XXs represent the delta cells. The categories for this table are Year, Region and Statistics.

Year	Region	Statistics	
		Median_Total_ Income	Infant_Mortality _Rate
2002	Alberta	XX	XX
	Manitoba	XX	XX
2003	Alberta	XX	XX
	Manitoba	XX	XX

Fig. 1 Statistics for Alberta and Manitoba

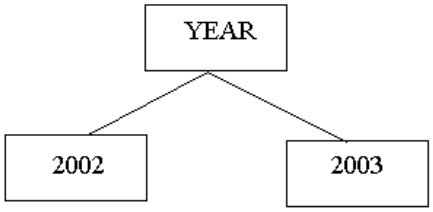


Fig. 2 Category Tree 1

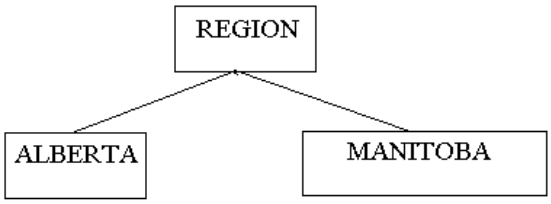


Fig. 3 Category Tree 2

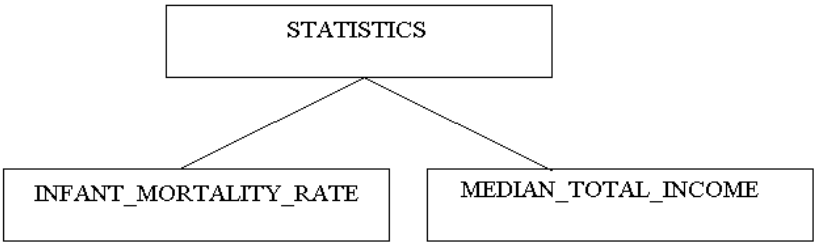


Fig. 4 Category Tree 3

Category Notation:

(Year, {(2002, phi), (2003, phi)})
(Region, {(Alberta, phi), (Manitoba, phi)})
(Statistics, {(Median_Total_Income, phi),(Infant_Mortality_Rate, phi)})

Year is the first category with 2002 and 2003 as subcategories. Region is the next category with Alberta and Manitoba as subcategories. Statistics is the last category with two subcategories: Median_Total_Income and Infant_Mortality_Rate.

Delta Notation:

delta({Statistics.Median_Total_Income, Year.2002, Region.Alberta})=XX
delta({Statistics.Infant_Mortality_rate, Year.2002, Region.Alberta})=XX
delta({Statistics.Median_Total_Income, Year.2002, Region.Manitoba})=XX
delta({Statistics.Infant_Mortality_rate, Year.2002, Region.Manitoba})=XX
delta({Statistics.Median_Total_Income, Year.2003, Region.Alberta})=XX
delta({Statistics.Infant_Mortality_rate, Year.2003, Region.Alberta})=XX
delta({Statistics.Median_Total_Income, Year.2003, Region.Manitoba})=XX
delta({Statistics.Infant_Mortality_rate, Year.2003, Region.Manitoba})=XX

However, there are tables in which spanning labels are not present. For example, the table in Figure 1 would still be understood even if the header "Statistics" were absent from the table. But the logical structure requires a root for the column header tree paths. This requires the addition of what Wang called "virtual header". Virtual headers are explicit headers added by the user or the system in order to complete the logical structure of the table to obtain its abstract notation.

1.2.2 Why Wang Notation?

Consider the table shown in Figure 5 which is a variation of the one in Figure 1.

Region	Infant_Mortality_ Rate		Median_ Total_Income	
	2002	2003	2002	2003
Alberta	XX	XX	XX	XX
Manitoba	XX	XX	XX	XX

Fig. 5 Statistics for Alberta and Manitoba-II

This table requires two virtual headers: Year (or VH1), which is a suitable header name for the subcategories 2002 and 2003 and Statistics (or VH2), which is a reasonable header name for the subcategories Infant_Mortality_Rate and Median_Total_Income. In this table one of the categories (Year) is subsumed by another category (Statistics). The category trees would be exactly the same as the ones shown in Figures 2-4 after the addition of the virtual headers. The Wang Notation becomes

Category Notation:

(Region, {(Alberta, phi), (Manitoba, phi)})
(Year/VH1, {(2002, phi), (2003, phi)})
(Statistics/VH2, {(Median_Total_Income,phi), (Infant_Mortality_Rate,phi)})

Delta Notation:

delta({Statistics/VH2.Infant_Mortality_rate, Year/VH1.2002, Region.Alberta})=XX
delta({Statistics/VH2.Infant_Mortality_rate, Year/VH1.2003, Region.Alberta})=XX
delta({Statistics/VH2.Median_Total_Income, Year/VH1.2002, Region.Alberta})=XX
delta({Statistics/VH2.Median_Total_Income, Year/VH1.2003, Region.Alberta})=XX
delta({Statistics/VH2.Median_Total_Income, Year.2002, Region.Manitoba})=XX
delta({Statistics/VH2.Median_Total_Income, Year.2003, Region.Manitoba})=XX
delta({Statistics/VH2.Infant_Mortality_rate, Year.2002, Region.Manitoba})=XX
delta({Statistics/VH2.Infant_Mortality_rate, Year.2003, Region.Manitoba})=XX

Even though the layout of the table in Figure 5 is different from the layout of the one in Figure 1, the Wang Notation preserves the underlying logical structure. Thus, it can be seen that a table's layout can be changed in many ways yet convey the same meaning with the same values. Wang's formalism can help us *understand* the table i.e., to recover the labels and the functions that map them to a set of domains. The process of understanding a table is not trivial. Experiments have shown that even human "experts" often disagree on the label-value pairs for a table [7]. Thus, instead of trying to achieve automatic understanding of tables, an interactive tool where the human expert helps to abstract the table is developed. The tool is called Table Abstraction Tool (TAT) and the Wang Notation produced using TAT is recorded in the form of an XML file whose schema encodes these aspects necessary to interpret the table. However, TAT, like WNT, displays the Wang category trees to the user in the form of *indented notation*, where nodes (i.e., categories and subcategories) at the same level appear in the same column.

1.2.3 Table Content: Other Aspects (Augmented Wang Notation)

Although Wang Notation is an efficient method to describe the logical structure of a table, it does not describe it fully. For example consider the table in Figure 6 from Canada Statistics website:

Military personnel and pay					
(Personnel)					
	2003	2004	2005	2006	2007
	Annual average number of employees[1]				
Canada and outside Canada	**83,766**	**84,059**	**85,706**	**87,730**	**89,332**
Newfoundland and Labrador	1,295	1,402	1,375	1,226	1,227
Prince Edward Island	262	266	284	213	230
Nova Scotia	10,598	10,696	10,830	10,520	10,536
New Brunswick	4,949	4,959	5,084	5,300	5,763
Quebec	15,384	15,402	16,121	17,663	18,200
Ontario	27,751	27,681	28,413	29,741	29,904
Manitoba	3,960	3,908	3,927	3,824	4,002
Saskatchewan	1,100	1,104	1,150	1,108	1,112
Alberta	9,052	9,209	9,078	9,090	9,217
British Columbia	7,741	7,776	7,793	7,298	7,305
Yukon Territory	x	x	x	x	x
Northwest Territories	148	153	150	166	174
Nunavut	x	x	x	x	x
Outside Canada	1,521	1,496	1,494	1,577	1,656
x : suppressed to meet the confidentiality requirements of the Statistics Act					
Notes:					
- Employment data are not in full-time equivalent and do not distinguish between full-time and part-time employees.					
- As at December 31.					
1. Civilian employees are excluded. Reservists are included as of January 1974.					
Source: Statistics Canada, CANSIM, table (for fee) 183-0004 and Catalogue no 68-213-XIB.					
Last modified: 2008-05-29.					

Fig. 6 Military Personnel and Pay

In the above table, one can observe additional elements like table title (“Military Personnel and Pay”), caption (“Personnel”), footnotes (“1.Civil employees are excluded. Reservists are included as of January 1974.”), aggregates (“Canada and Outside Canada”) and other details regarding the table in the form of “Notes”. The original Wang Notation does not have provisions to incorporate these details. Since these aspects of a table may be essential to its interpretation, an augmented Wang XML is proposed which incorporates all the above-mentioned information. The Table Abstraction Tool produces this augmented Wang XML as its output after accepting inputs from the user.

1.2.4 Organization of the rest of the Thesis

Section 2 defines the central ideas used in TAT along with common terms in table processing and the transformations required on tables to make them TAT-friendly. Section 3 describes the software aspects of the Table Abstraction Tool and the XML schema. Section 4 is a description of the evaluation of TAT, including the experimental design and criteria for selection of tables used in the experiment. Section 5 presents the observations and quantitative results. Section 6 discusses possible future work and Section 7 contains concluding remarks. References can be found after Section 7 and finally, there is an Appendix containing the User Manual for using the tool, the list of URLs of the tables used for evaluation and a sample log file.

2. Table Abstraction Tool: Central Ideas

There are a few concepts which are specific to TAT and need to be explained before discussing the tool further. *Row and Column categories* refer to the location of the category headers with respect to delta cells and are useful when dealing with tables having more than two categories. *Canonicalization* makes the sizes of all the cells in the table homogenous and makes processing of the table easier. *List-row notation* transforms the table categories into a VBA data structure for manipulation. The list-row notation is used to construct the *indented notation* which is the Wang category tree representation in Excel. These ideas differentiate TAT from the Wang Notation Tool. The reasons for choosing VBA and Excel as a platform for the tool and why TAT is an improvement over WNT are also explained in the subsections below.

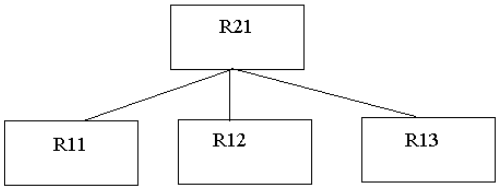
2.1 Row and Column Categories

Regardless of the dimensionality of an abstract table, its presentation form has two spatial dimensions. The headers that point to a delta cell are almost always to the left and above it (tables which do not follow this rule are not considered well formed by TAT). We call categories (*or Wang dimensions*) with headers to the left of delta cells *row-categories*, and categories with headers above delta cells *column-categories.* Row and column categories may be interchanged without altering the meaning of the table. These designations are convenient to link the categories to a specific presentation form, but have no role in the Wang Notation.

Consider the Simulated Table 1 shown in Figure 7.

STUB		C21	
		C11	C12
R21	R11	XX	XX
	R12	XX	XX
	R13	XX	XX

Fig. 7 Simulated Table 1



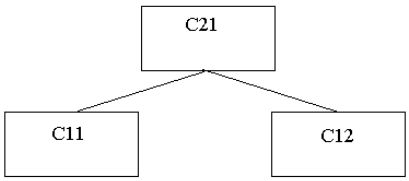


Fig. 8 Row Category Tree for table in Fig. 7

Fig. 9 Column Category Tree for table in Fig. 7

Tables with Wang dimensionality greater than two have more than one row or column category depending on the physical layout of the table. For example, Simulated Table 2 in Figure 10 has a Wang dimensionality of three with two row categories and one column category. Both row categories require the addition of virtual headers (VH and VH1).

STUB		C21	
		C11	C12
R21	R11	XX	XX
	R12	XX	XX
	R13	XX	XX
R22	R11	XX	XX
	R12	XX	XX
	R13	XX	XX

Fig. 10 Simulated Table 2

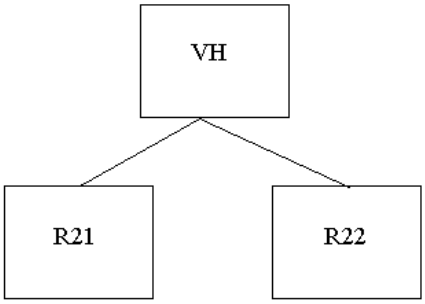


Fig. 11 Row Category 1 for Simulated Table 2

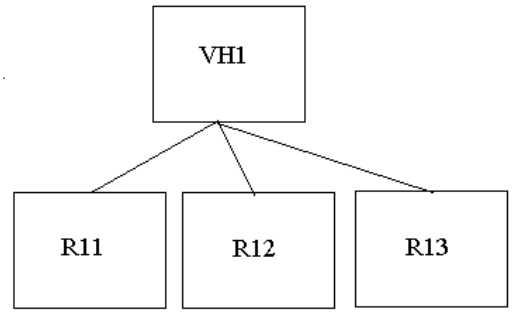


Fig. 12 Row Category 2 for Simulated Table 2

The column category tree is the same for the tables in Figures 7 and 10.

2.2 Canonicalization

As seen in Figures 7 and 10, a label (or value) might span over multiple cells which are merged. *Canonicalization* is the process of splitting the merged cells and repeating the label across all the split cells. The canonicalized version of Simulated Table 1 is shown in Figure 13.

STUB		C21	C21
		C11	C12
R21	R11	XX	XX
R21	R12	XX	XX
R21	R13	XX	XX

Fig. 13 Canonicalized version of Simulated Table 1

Canonicalization makes the size of all the cells uniform, preserves the spatial relationship between adjacent cells and helps to form the *list-rows* (next subsection) of the table.

2.3 List-Row Notation

The idea of list-rows is the most important aspect of table processing in TAT. This idea helps in checking the *validity* of tables and is used to form the category tree for each table. The list-row notation is an array representation of the category selected by the user. For Simulated Table 1, the list-row notation for the row category is the following 3X2 array:

R11 R21

R12 R21

R13 R21

The list-row notation for the column categories is the following 2X2 array:

C11 C21

C12 C21

Comparing the list-row notation for Simulated Table 1 with its Wang category tree representations (in Figures 8 and 9), it can be seen that the list-row notation simply specifies all the category tree paths in the category. Thus, using list-rows helps us to transform the table categories into a data structure which is supported by VBA and hence manipulated to get the indented notation of the categories.

2.4 Indented Notation

Wang's category trees for tables are represented using the indented notation constructed from the list-row notation. In the indented notation, nodes at the same level of the tree appear in the same column. The children of a particular node in the tree appear below the node in the next column. The indented notation is formed by looping through the list-row notation and printing the values of the rows in such a way that no paths in the tree are repeated. Figures 14 and 15 represent the indented notation for Simulated Table 1.

R21	
	R11
	R12
	R13

Fig. 14 Indented Notation for Row Category

C21	
	C11
	C12

Fig. 15 Indented Notation for Column Category

The similarities between Figures 8, 14 and 9, 15 can be observed. The indented notation is also used to create the Wang XML notation using a notation called the OID notation (represented in separate worksheets).

The OID notation for the above categories is:

C1	
	C1.1
	C1.2
	C1.3

Fig. 16 OID Notation for Row Category

C2	
	C2.1
	C2.2

Fig. 17 OID Notation for Column Category

The C21, C11 and C12 nodes are just used to represent column categories as in previous Figures. But the Cx.x notation in the OID notation is a notation for categories used for every table in the XML notation. By creating separate worksheets with OIDs, the XML notation is generated. This is explained in detail in Section 3.3.

2.5 VBA and Excel

Excel is a spreadsheet application written and distributed by Microsoft. I used Excel to build TAT because of its user-friendliness, familiarity and because it is available in most computers with Windows OS. Spreadsheet applications are perfectly suited for table processing as they can represent tables on a grid of rows and columns, each cell containing text or numeric values. Each cell in the table/grid has an *address* and the spatial relationships between the cells in a table can be captured using the cell address properties.

Visual Basic for Applications (VBA) is an implementation of the programming language Visual Basic and associated integrated development environment (IDE) which is built into most Microsoft Office applications. VBA functions within a host application rather than as a standalone programming language. It is functionally rich and flexible. VBA code is a set of macros. Since

Visual Basic is an event-driven programming language, one can utilize the same capabilities in VBA also. These capabilities have been utilized in building some features of TAT.

2.6 TAT: An Improved WNT

Since TAT is built in VBA and utilizes its user friendly features to manipulate tables, it saves time in processing them compared to the Wang Notation Tool (WNT) which uses MATLAB to convert tables into Wang XML. TAT can be regarded as an improved version of the Wang Notation Tool for the following reasons:

1. WNT used MATLAB GUI buttons to manipulate the category trees and corrections were not very easy. TAT uses Excel cells which can be easily edited to form the indented notation.
2. TAT incorporates augmentations like title, footnotes, and units in the XML notation. This increases the utility of ontologies constructed compared to WNT.
3. TAT is a complete application in itself. It requires only the source table to obtain the Wang XML. WNT on the other hand requires a textual representation (ASCII format) of the table as an input.
4. For tables with categories without roots, one needs to add a virtual header. This was a tedious procedure in WNT. TAT automatically identifies when a category root is absent and generates a unique virtual header for every rootless tree. This saves considerable time for the user.
5. TAT connects the spatial structure of the table with its indented notation by adding the address of the category/subcategory text in the original table.

2.7 Well Formed Tables

A well-formed table is a table which can be processed by TAT. The idea of a well-formed table is borrowed from [4] and it is extended. Jha describes the following requirements for a table to be deemed well-formed:

1. Every table must have n categories, where $n \geq 2$.
2. Every category must have a root (sometimes requiring the addition of virtual headers)
3. Every delta cell must be specified by n paths, one through each category tree.
4. Category trees cannot contain subcategory trees that are identical.
5. Category cells appear only in the topmost rows and leftmost columns of a table.

Rule 4 is enforced by adding the condition that the list-row for every category should have distinct rows (discussed in the next subsection). For a correct Wang representation, the categories and subcategories should be present in different rows/columns in the physical layout. If they are

expressed using any visual cues like boldface or indentation, they must be transformed (discussed in the section 2.8.2)

For TAT the first condition is not necessary. Some lists can be represented as tables as well. Even though they are not in a sense well formed, TAT does not enforce the condition for the number of dimensions to be greater than or equal to two to process them.

2.7.1 Detecting an Anomalous Table

Consider the table in Figure 18 which is obtained by adding a row to Simulated Table 2.

STUB		C21	
		C11	C12
R21	R11	XX	XX
	R12	XX	XX
	R13	XX	XX
R22	R11	XX	XX
	R12	XX	XX
	R13	XX	XX
R21	R13	XX	XX

Fig. 18 An anomalous table

The Category trees for this table are:

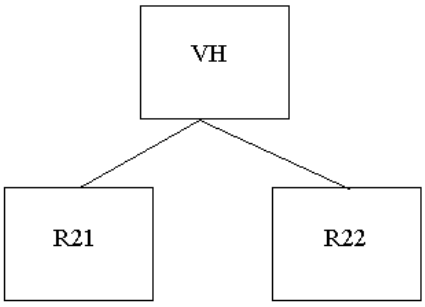


Fig. 19 Row Category 1 for table in Fig. 18

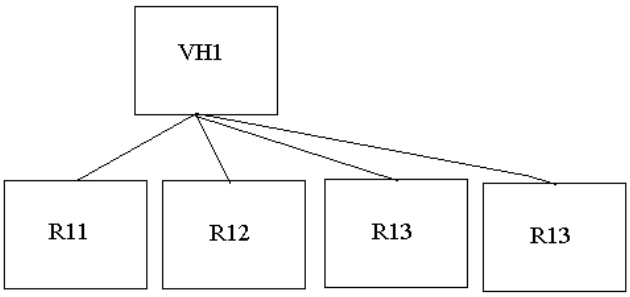


Fig. 20 Row Category 2 for table in Fig. 18

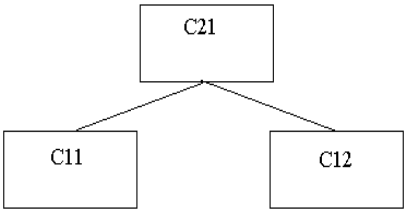


Fig. 21 Column Category 1 for table in Fig. 18

It can be seen that node *R13* repeats twice in tree shown in Figure 20 (Row Category 2). This violates rule 4 for well-formed tables. Therefore, it is not a well-formed table. The list-row notation for the row categories of the above table is:

R11 R21

R12 R21

R13 R21

R11 R22

R12 R22

R13 R22

R13 R21

It can be seen that in the list-row array for the row categories, the 3rd and 7th row are the same. This table violates the condition that the rows of the list-row notation should be distinct. Hence, TAT construes this table as an invalid table which is a valid interpretation.

2.8 Transformation of Invalid Tables into Valid Tables

As mentioned in the above section, TAT checks for the validity of a table by analyzing its list-row notation. The check for the table is based on uniqueness of cells in a column of categories. TAT decides that the table is not well formed if it finds any repeated rows in the list-row notation of the category. This check, although it detects anomalous tables, is not the best way to check the *validity* of a table. Repeating cells across a column becomes almost inevitable when there is similar data aggregated over different regions or periods which are commonly found in web documents, especially in the geopolitical domain. Hence there is an urgent need to devise methods to check the *validity* of tables. But until we find robust methods, we need to transform the table using Excel operations to process them while preserving the logical structure (Wang dimensionality and relationship between categories and subcategories). Section 2.8.1 illustrates two examples on simulated tables to give an idea of what kind of transformations are required and why. This is then followed by a discussion of transformations on real world tables, with the category trees and the list-row notations shown and compared.

2.8.1 Illustration on Simulated Tables

STUB			P	
			Q	R
A	B	C1	XX	XX
		C2	XX	XX
	D	D1	XX	XX
		D2	XX	XX

Fig. 22 Simulated Table 3

Simulated Table 3 is a two category table. The Wang category tree representation for it is:

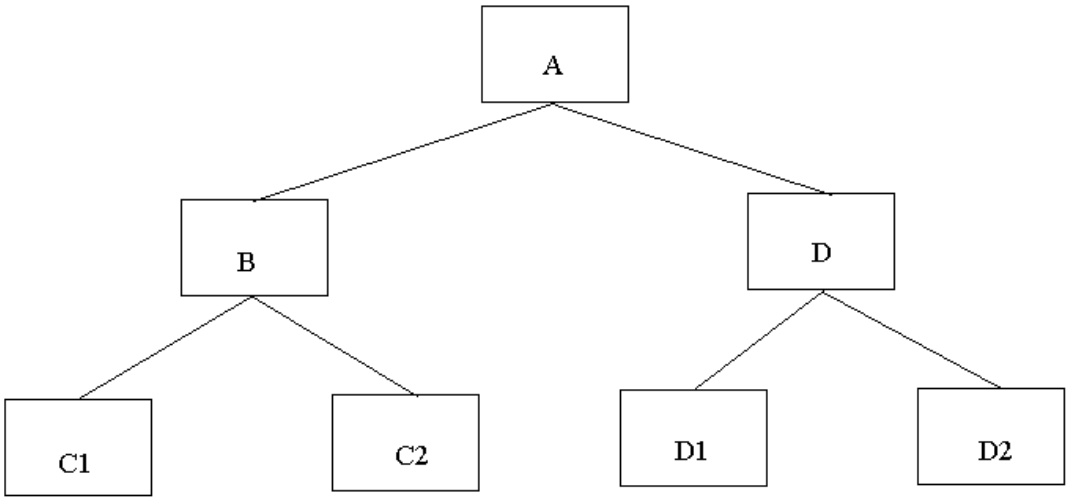


Fig. 23 Row Category for Simulated Table 3

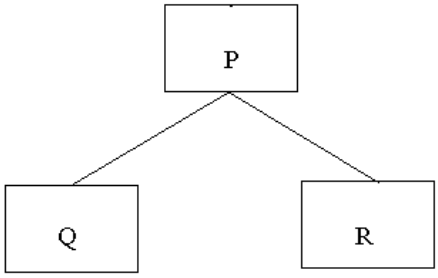


Fig. 24 Column Category for Simulated Table 3

Any combination of paths in the above trees would lead us to a delta cell and hence this is the correct Wang representation. The list-row notation for the above categories is:

Row Categories:

C1 B A

C2 B A

D1 D A

D2 D A

Column Categories:

Q P

R P

Since there are no repeating rows in either list-row, TAT considers it valid. Note that the levels of all the leaf nodes are the same.

A	P	
	Q	**R**
A1	1000	1000
B1	700	600
B2	300	400
B3	100	300
B4	200	100
A2	350	500
C1	200	200
C2	150	300
A3	500	900
C1	200	400
C2	300	500

Fig. 25 Simulated Table 4

Simulated Table 4 is an interesting table and very often found in tables on the web, especially in the geopolitical domain. Aggregates in tables are indicated in common practice using indentations and font characteristics like bold. Here A2, A3 aggregate one kind of data, represented by C1 and C2, while A1 aggregates a different kind of data, represented by B1 and B2. But B2 is an aggregate of the rows B3 and B4. This table is still a two category table. The Wang category tree representation is interesting because A1, A2, A3 and B2 have subcategories and also have delta cell values associated with them directly. The Wang category notation for this table is shown in Figures 26 and 27.

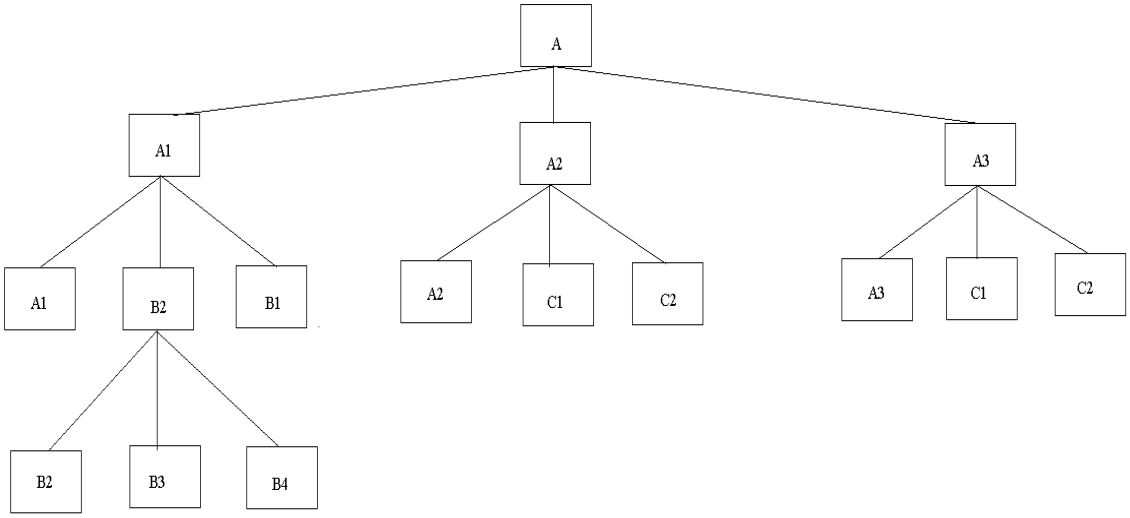


Fig. 26 Row Category for Simulated Table 4

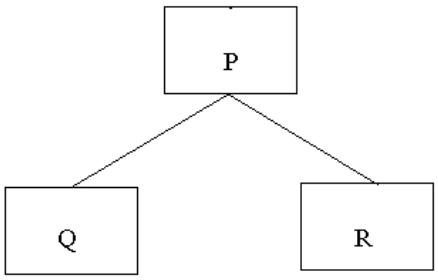


Fig. 27 Column Category for Simulated Table 4

It can be seen that the nodes A1, A2, A3 and B2 are repeated in the row category tree. This is the correct Wang tree. For the row-category, the list-row notation is just a one-dimensional array:

[A A1 B1 B2 B3 B4 A2 C1 C2 A3 C1 C2] ' where' is the transpose operator.

Since C1 and C2 repeat in the list-row notation, TAT construes this as an invalid table. This table, however, can be perfectly interpreted by humans and is valid. To make this table TAT-friendly, we need to transform the table. There are two options to make this table processable by TAT. But, one of them is more meaningful than the other and requires more transformations.

STUB			P	
			Q	R
		A1	1000	1000
		B1	700	600
		B2	300	400
		B3	100	300
	A1	B4	200	100
		A2	350	500
		C1	200	200
	A2	C2	150	300
		A3	500	900
		C1	200	400
A	A3	C2	300	500

Fig. 28 TAT-friendly (but incorrect) transformation of Simulated Table 4

The list-row notation for the row-categories of the table shown in Figure 28:

A1 A1 A

B1 A1 A

B2 A1 A

B3 A1 A

B4 A1 A

A2 A2 A

C1 A2 A

C2 A2 A

A3 A3 A

C1 A3 A

C2 A3 A

Since there are no repetitions in the list-rows for table in Figure 28, it can be processed by TAT. But in this list-row notation, cells B3 and B4 are incorrectly associated with A1. The correct

transformation should associate the cells B3 and B4 with B2. The table shown in Figure 29 is the correct transformation for this table.

STUB				P	
				Q	R
		A1	**A1**	1000	1000
		B1	B1	700	600
		B2	B2	300	400
	A1		B3	100	300
			B4	200	100
	A2	A2	**A2**	350	500
			C1	200	200
			C2	150	300
	A3	A3	**A3**	500	900
			C1	200	400
A			C2	300	500

Fig. 29 TAT-friendly and correct transformation of Simulated Table 4

Note that the indentation is removed from the table as it is no longer required to make the distinction. Every node is at the same depth. The Wang category tree notation for the row category of the table in Figure 29 is shown below:

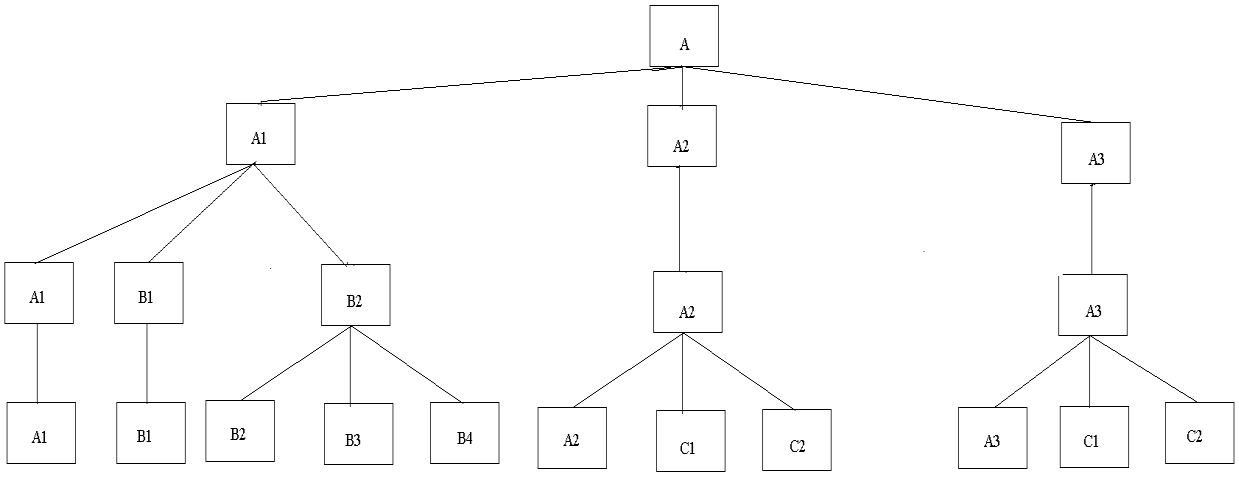


Fig. 30 Row Category for table in Fig. 29

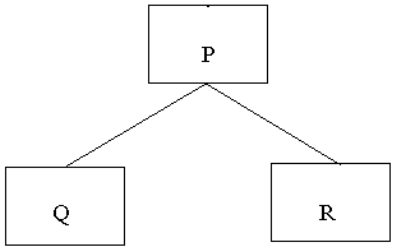


Fig. 31 Column Category for table in Fig. 29

It can be seen that the only effect the transformation has on the original table is that all the leaf nodes of the row category are at the 4th level in Figure 30 as opposed to Figure 26 in which only three leaf nodes were at the 4th level. TAT requires all the leaf nodes at the same level because TAT forms the indented notation (which is similar to the Wang category tree notation) using list-rows and hence the number of nodes required to reach a leaf node must be the same for all of them.

The list-row notation for row-category of table in Figure 30 is:

A1 A1 A1 A
B1 B1 B1 A
B2 B2 A1 A
B3 B2 A1 A
B4 B2 A1 A
A2 A2 A2 A
C1 A2 A2 A
C2 A2 A2 A
A3 A3 A3 A
C1 A3 A3 A
C2 A3 A3 A

There are no repetitions in the list-row notation of the row-categories for this table and it associates the cells B3 and B4 with B2. Hence, TAT can process it correctly.

2.8.2 Table Transformations on Real-World Tables

This section describes the transformations described above on real tables to make them TAT-friendly. Figure 32 shows a screenshot of a table requiring transformations.

Subject	Number	Percent
POPULATION 5 YEARS AND OVER BY LANGUAGE SPOKEN AT HOME AND ABILITY TO SPEAK ENGLISH		
Population 5 years and over	**262,375,152**	**100.0**
Speak only English	215,423,557	82.1
Speak a language other than English	46,951,595	17.9
Spanish	**28,101,052**	**100.0**
Speak English "very well"	14,349,796	51.1
Speak English "well"	5,819,408	20.7
Speak English "not well"	5,130,400	18.3
Speak English "not at all"	2,801,448	10.0
Other Indo-European languages	**10,017,989**	**100.0**
Speak English "very well"	6,627,688	66.2
Speak English "well"	2,091,447	20.9
Speak English "not well"	1,078,930	10.8
Speak English "not at all"	219,924	2.2
Asian and Pacific Island languages	**6,960,065**	**100.0**
Speak English "very well"	3,370,041	48.4
Speak English "well"	2,023,303	29.1
Speak English "not well"	1,260,264	18.1
Speak English "not at all"	306,457	4.4
All other languages	**1,872,489**	**100.0**
Speak English "very well"	1,283,663	68.6
Speak English "well"	399,398	21.3
Speak English "not well"	151,125	8.1
Speak English "not at all"	38,303	2.0
ABILITY TO SPEAK ENGLISH		
Population 5 years and over	**262,375,152**	**100.0**
Speaks a language other than English	46,951,595	17.9
5 to 17 years	9,779,766	3.7
18 to 64 years	32,756,989	12.5
65 years and over	4,414,840	1.7
Speak English less than "very well"	21,320,407	8.1
5 to 17 years	3,493,118	1.3
18 to 64 years	15,486,421	5.9
65 years and over	2,340,868	0.9
ABILITY TO SPEAK ENGLISH IN HOUSEHOLD		
Linguistically isolated households	4,361,638	(X)
Population 5 years and over in households	**254,620,291**	**100.0**
In linguistically isolated households	11,893,572	4.7
5 to 17 years	2,687,603	1.1
18 to 64 years	7,926,537	3.1
65 years and over	1,279,432	0.5

Fig. 32 Ability to Speak English

The table aggregates data of English speakers based on many criteria and hence there are repetitions in the "Subject" column of the table. The indicated area in the Figure represents the leaf nodes present at the lowest level of the Wang Category tree whose path would visit the following nodes:

SUBJECT > ABILITY TO SPEAK ENGLISH > POPULATION 5 YEARS AND OVER > SPEAK A LANGUAGE OTHER THAN ENGLISH > 5 TO 17 YEARS.

Thus, the row-category tree must have 5 levels with all the leaf nodes at the lowest level. We must therefore transform the table so that the delta cells start in the 6th column of the table. Figure 33 shows a screenshot of the table which has been modified in Excel into a TAT-friendly format.

SUBJECT		Asian and Pacific Island languages	Speak English "well"	Speak English "well"	2,023,303	29.1
			Speak English "not well"	Speak English "not well"	1,260,264	18.1
			Speak English "not at all"	Speak English "not at all"	306,457	4.4
		All other languages	**All other languages**	**All other languages**	**1,872,489**	**100**
		All other languages	Speak English "very well"	Speak English "very well"	1,283,663	68.6
			Speak English "well"	Speak English "well"	399,398	21.3
			Speak English "not well"	Speak English "not well"	151,125	8.1
			Speak English "not at all"	Speak English "not at all"	38,303	2
	ABILITY TO SPEAK ENGLISH	Population 5 years and over	**Population 5 years and over**	**Population 5 years and over**	**262,375,152**	**100**
		Population 5 years and over	Speak a language other than English	Speak a language other than English	46,951,595	17.9
			5 to 17 years	5 to 17 years	9,779,766	3.7
			18 to 64 years	18 to 64 years	32,756,989	12.5
			65 years and over	65 years and over	4,414,840	1.7
			Speak English less than "very well"	Speak English less than "very well"	21,320,407	8.1
			5 to 17 years	5 to 17 years	3,493,118	1.3
			18 to 64 years	18 to 64 years	15,486,421	5.9
			65 years and over	65 years and over	2,340,868	0.9
	ABILITY TO SPEAK ENGLISH IN HOUSEHOLD	Linguistically isolated	Linguistically isolated households[1]	Linguistically isolated households[1]	4,361,638	(I)
		Population 5 years and over in households	**Population 5 years and over in households**	**Population 5 years and over in households**	**254,620,291**	**100**
		Population 5 years and over in households	In linguistically isolated households[1]	In linguistically isolated households[1]	11,893,572	4.7
			5 to 17 years	5 to 17 years	2,687,603	1.1
			18 to 64 years	18 to 64 years	7,926,537	3.1
			65 years and over	65 years and over	1,279,432	0.5

Fig. 33 Ability to Speak English: Transformed table

The table in Figure 34 is another example of a real table requiring transformations.

Table 2. Mean Absolute Percentage Error in State Population Projections, By Region And Division From Series A and B, And Extrapolated Projections, 2000

Region and division	Series A	Series B	Extrapolation
United States	**2.63**	**2.44**	**2.54**
Northeast	**2.50**	**2.57**	**3.15**
New England	2.42	2.58	3.51
Middle Atlantic	2.67	2.55	2.43
Midwest	**1.58**	**1.40**	**1.11**
East North Central	1.54	1.36	1.07
West North Central	1.60	1.43	1.14
South	**2.60**	**2.58**	**2.69**
South Atlantic	3.50	3.50	3.90
East South Central	0.89	0.87	0.86
West South Central	2.29	2.21	1.79
West	**3.75**	**3.13**	**3.22**
Mountain	4.41	3.91	3.88
Pacific	2.69	1.89	2.16

Mean absolute percentage error (MAPEs) are results for 5 years-out from the 1995 population. Based on the enumerated 2000 census counts, the 2000 population for Series A, B, and Extrapolated Projections derived from the Absolute Percentage Errors calculated for the states and the District of Columbia, see text for detailed explanation. Source: U.S. Bureau of the Census.

Source: U.S. Census Bureau
Internet Release date: November 6, 2002

Fig. 34 Mean Absolute Percentage Error in State Population Projections, By Region and Division from Series A and B, and Extrapolated Projections, 2000

Figure 34 presents a well-formed, TAT-friendly table, as there are no repetitions in the list-rows. But processing the table directly will not give us the complete information. In the table, the bold format for *Northeast*, *Midwest*, *South* and *West* indicates that these headers designate aggregates of the rows below them in some form. But if we process the table using TAT this information will be lost and *South* will be treated exactly as *South Atlantic*, *East South Central* and *West South*

Central. Thus in order to exploit this information, we transform the table into the form shown in Figure 35.

Table 2. Mean Absolute Percentage Error in State Population Projections, By Region And Division From Series A and B, And Extrapolated Projections, 2000						
				Series A	Series B	Extrapolation
Region and division	United States	United States	**United States**	**2.63**	**2.44**	**2.54**
		Northeast	**Northeast**	**2.50**	**2.57**	**3.15**
			New England	2.42	2.58	3.51
			Middle Atlantic	2.67	2.55	2.43
		Midwest	**Midwest**	**1.58**	**1.40**	**1.11**
			East North Central	1.54	1.36	1.07
			West North Central	1.60	1.43	1.14
		South	**South**	**2.60**	**2.58**	**2.69**
			South Atlantic	3.50	3.50	3.90
			East South Central	0.89	0.87	0.86
			West South Central	2.29	2.21	1.79
		West	**West**	**3.75**	**3.13**	**3.22**
			Mountain	4.41	3.91	3.88
			Pacific	2.69	1.89	2.16
			Mean absolute percentage error (MAPEs) are results for 5 years-out from the			
			1995 population. Based on the enumerated 2000 census counts, the 2000			
			population for Series A, B, and Extrapolated Projections derived from the			
			Absolute Percentage Errors calculated for the states and the District of Columbia,			
			see text for detailed explanation. Source: U.S. Bureau of the Census.			
			Source: U.S. Census Bureau			
			Internet Release date: November 6, 2002			

Fig. 35 Transformed Table in Fig. 34 to capture the aggregate relationship

The transformed table in Figure 35, when processed correctly, will indicate in the category notation itself that *New England* comes under *Northeast*. This relationship can also be specified by annotating *Northeast* and others using the "Aggregate" option provided in the Augmentations action in TAT. However, this action will preserve that relationship only in the XML. If the user wants to see the relationship in the indented notation for the Wang category tree, the transformation shown in Figure 35 is necessary.

3. TAT Software Description

The Table Abstraction Tool is built in VBA as a set of macros. The coding paradigm can be described as functional programming with a set of functions interacting with one another. There are two different types of functions in TAT. The first type of functions is invoked when a button is pressed. These functions form the heart of TAT and are used for checking the *validity* of the table, annotating the augmentations, generating the indented notation for the various categories and generating the Wang XML and the log file. The second set of functions is unique to the VBA language and Excel. These functions/events are invoked by mouse clicks or cursor change. These functions are called SelectionChange events in the VBA jargon and allow the user to check the indented notation for the categories. They make the interaction with the tool easier and more user-friendly.

A large portion of the code *reacts* to the user input either by adding an annotation using the comment feature in Excel or coloring the cells to indicate that some action has been performed on the cells. However, TAT also does some processing of the user input for checking the validity of a table after the user selects the delta cells, forming the indented notation for every category after the user selects the category region, and checking the indented notation.

The following sections describe in pseudo-code the algorithms in TAT.

3.1 Pseudo-Code to Generate Indented Notation

```
Segment the table into delta cells and category cells //(based on user input)
Perform Canonicalization
Form the list-row notation for the row and column categories
Check for repeated rows in the list-row notation for both row and column categories
If repetition, display error
Else
   Repeat for number of categories
     Prompt category location, form list-rows and sort the rows
     If no category root present, generate a unique virtual header //This acts as the root
          Rearrange list-row columns with category root as the first column
          Create a new sheet to print indented notation
          Loop through modified list-row and print indented notation for category.
   End Repeat
End If
```

3.2 Wang XML Notation

The XML notation is chosen to represent the Wang notation because our BYU TANGO collaborators use XML to construct ontologies. The XML file contains 7 main sections that detail various parts of a table. This section describes the Wang XML notation using the table in Figure 36 as an example.

Region and State Information

Location	Population* (2000)	Longitude†
Northeast	3.120	
Maine	1.275	69°14.0′W
New Hampshire	1.236	71°34.3′W
Vermont	0.609	72°40.3′W
Northwest	9.315	
Washington	5.894	120°16.1′W
Oregon	3.421	120°58.7′W

*Population in Millions
†Geographic Center

Fig. 36 Region and State Information

Section1: The XML begins with:

```
<?xml version="1.0"?>
<TableOntology>
```

Fig. 37 XML Section 1

The first line of any XML document states the version and notifies the reader that it is a valid XML document. The second line informs the reader that this XML is for a table ontology.

Section2: This section contains basic information about the table:

<Table TableOID="T13967" Title="Region and State Information" Caption="Sample Table" 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="1">
<CategoryRootNodes>
<CategoryRootNode CategoryRootNodeOID="C1"/>
<CategoryRootNode CategoryRootNodeOID="C2"/>
</CategoryRootNodes>
</Table>

Fig. 38 XML Section 2

The table element contains multiple attributes like *TableOID*, *Title*, *Caption*, *DocumentCitation*, and *Number*. The table element contains the *CategoryRootNodes* element. This element contains the *CategoryRootNodeOID* attribute. The terms elements and attributes are XML jargon. Elements can be parents of other elements and/or attributes and can be repeated within the same level of an XML document. The example table has a Wang dimensionality of 2 so there are two category root nodes. The *CategoryRootNodeOID* values are obtained from the indented notation sheets and are ordered based on the location of the cells in the sheet as explained in Section 2.4.

Section3: The third section lists every category node in the table:

<CategoryNodes>
<CategoryNode CategoryNodeOID="C1"/>
<CategoryNode CategoryNodeOID="C1.1" Label="Population"/>
<CategoryNode CategoryNodeOID="C1.2" Label="Longitude"/>
<CategoryNode CategoryNodeOID="C2" Label="Location"/>
<CategoryNode CategoryNodeOID="C2.1" Label="Northeast"/>
<CategoryNode CategoryNodeOID="C2.1.1" Label="Maine"/>
<CategoryNode CategoryNodeOID="C2.1.2" Label="New Hampshire"/>
<CategoryNode CategoryNodeOID="C2.1.3" Label="Vermont"/>
<CategoryNode CategoryNodeOID="C2.2" Label="Northwest"/>
<CategoryNode CategoryNodeOID="C2.2.1" Label="Washington"/>
<CategoryNode CategoryNodeOID="C2.2.2" Label="Oregon"/>
</CategoryNodes>

Fig. 39 XML Section 3

The category nodes element contains a list of every category node in the table. Every category node element contains an attribute *CategoryNodeOID* which is the category's operational id number. From the example one can see that if a category node belongs to category root C1, it will share the same prefix in its OID. The following diagram may help explain how the id scheme works.

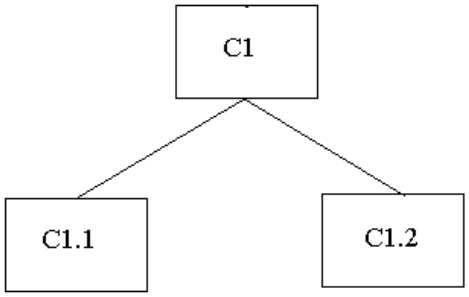


Fig. 40 Category Tree 1 for table in Fig. 36

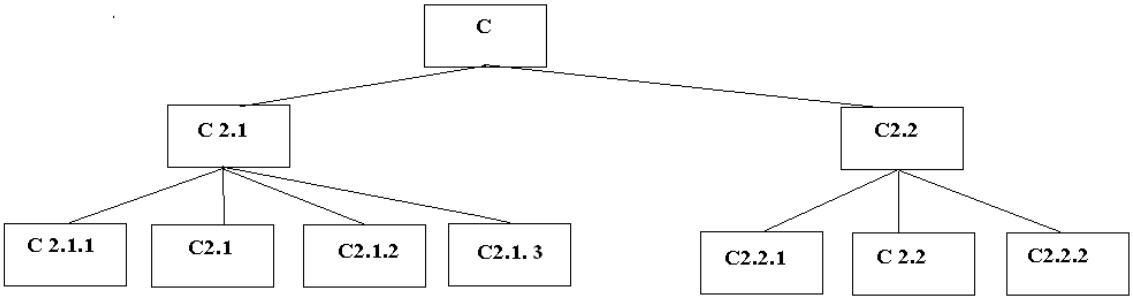


Fig. 41 Category Tree 2 for table in Fig. 36

C2.1 and C2.2 are aggregates which is why they are repeated in the tree. While they are shown twice in the tree and table they are not listed twice in the XML. If a category has a label, which should be every category node except possibly category root nodes, the label will also be added as an attribute.

Section4: The next section lists the category nodes with their children:

```
<CategoryParentNodes>
<CategoryParentNode CategoryParentNodeOID="C1">
<CategoryNodes>
<CategoryNode CategoryNodeOID="C1.1" />
<CategoryNode CategoryNodeOID="C1.2" />
</CategoryNodes>
</CategoryParentNode>
<CategoryParentNode CategoryParentNodeOID="C2">
<CategoryNodes>
<CategoryNode CategoryNodeOID="C2.1" />
<CategoryNode CategoryNodeOID="C2.2" />
</CategoryNodes>
</CategoryParentNode>
<CategoryParentNode CategoryParentNodeOID="C2.1">
<CategoryNodes>
<CategoryNode CategoryNodeOID="C2.1.1"/>
<CategoryNode CategoryNodeOID="C2.1.2"/>
<CategoryNode CategoryNodeOID="C2.1.3"/>
</CategoryNodes>
</CategoryParentNode>
<CategoryParentNode CategoryParentNodeOID="C2.2">
<CategoryNodes>
<CategoryNode CategoryNodeOID="C2.2.1"/>
<CategoryNode CategoryNodeOID="C2.2.2"/>
</CategoryNodes>
</CategoryParentNode>
</CategoryParentNodes>
```

Fig. 42 XML Section 4

This section begins with the element *CategoryParentNodes* which contains a list of all the category nodes that have children. Each element in the list is of type *CategoryParentNode* and has a *CategoryParentNodeOID* attribute whose value is the same as its *CategoryNodeOID* value. Each of these elements contains an element *CategoryNodes*. This element contains a list of the current category parent node's direct children of element type *CategoryNode*.

Section5: DataCells are listed in this section:

```
<DataCells>
<DataCell DataCellOID="D1,1" DataValue="3.120">
<HeaderNodes>
<HeaderNode HeaderNodeOID="C2.1"/>
</HeaderNodes>
<CategoryLeafNodes>
<CategoryLeafNode CategoryLeafNodeOID="C1.1" />
</CategoryLeafNodes>
</DataCell>
<DataCell DataCellOID="D1,2">
<HeaderNodes>
<HeaderNode HeaderNodeOID="C2.1"/>
</HeaderNodes>
<CategoryLeafNodes>
<CategoryLeafNode CategoryLeafNodeOID="C1.2" />
</CategoryLeafNodes>
</DataCell>
<DataCell DataCellOID="D2,1" DataValue="1.275">
<CategoryLeafNodes>
<CategoryLeafNode CategoryLeafNodeOID="C1.1" />
<CategoryLeafNode CategoryLeafNodeOID="C2.1.1" />
</CategoryLeafNodes>
</DataCell>
.
.
.
</DataCells>
```

Fig. 43 XML Section 5

This section consists of the *DataCells* element. Within this element is a list of elements that detail each data cell within the table. Each data cell has its own *DataCell* element with its *DataCellOID* attribute which is its x and y coordinates in the Excel table with (1,1) being the data cell in the top left corner. The x-coordinate increases to the right and the y-coordinate increases from top to bottom. *DataCells* usually contain a second attribute, *DataValue*, whose value is the textual or numerical content of that data cell within the table. Within each *DataCell* there may be a list of *HeaderNodes*, *CategoryLeafNodes*, or both. *HeaderNodes* are category nodes that are aggregates. For example C2.1 is an aggregate since it contains the accumulated information for Maine, New Hampshire and Vermont. It has children and it is not a leaf node, which is why it is put into the

HeaderNodes list. Each *HeaderNode* element has a *HeaderNodeOID* attribute which is its *CategoryNodeOID*. If on the other hand the category associated with the data cell is a leaf node it will be listed in the *CategoryLeafNodes* element. Each of these will be listed as an element type *CategoryLeafNode*. Similar to the *HeaderNode* elements they also have a *CategoryLeafNodeOID* attribute that is the same as their category node OID.

Section6: The next important section lists all the augmentations that occur within the table:

<Augmentations>
<Augmentation AugmentationOID="A1" AugmentationText="2000">
<CategoryNode CategoryNodeOID="C1.1"/>
</Augmentation>
<Augmentation AugmentationOID="A2" AugmentationText="Population in Millions"
FootnoteReference="%ampersand%number42;">
<CategoryNode CategoryNodeOID="C1.1"/>
</Augmentation>
<Augmentation AugmentationOID="A3" AugmentationText="Geographic Center"
FootnoteReference="%ampersand%dagger;">
<CategoryNode CategoryNodeOID="C1.2"/>
</Augmentation>
<Augmentation AugmentationOID="A4" AugmentationText="Geographic Center"
FootnoteReference="%ampersand%dagger;">
<CategoryNode CategoryNodeOID="C1.3"/>
</Augmentation>
</Augmentations>

Fig. 44 XML Section 6

In this section we have the *Augmentations* element. Within this element we see a list of *Augmentation* elements. Each *Augmentation* element has an *AugmentationOID* attribute. The numbering is sequential. If the Augmentation is a footnote it will contain the *AugmentationText* attribute and may also contain a *FootnoteReference* attribute. The *AugmentationText* is the comment about the cell and *FootnoteReference* is the type of symbol used in the cell to point to the footnote. If the *Augmentation* is a *Unit* it will have an attribute *AugmentationType* with value "Units". Lastly if the *Augmentation* is of type "Other" the element will have an attribute *AugmentationText*. Within each *Augmentation* element there is another element. This element will be of type *CategoryNode* or *DataCell* and will contain the OID of the cell(s) the augmentation corresponds to.

Section7: The last part of the XML is the closing label for the Table Ontology:

</TableOntology>

Fig. 45 XML Section 7

The following section describes in pseudo-code how TAT forms the XML for the indented notation.

3.3 Pseudo-Code to Generate XML Notation

```
Repeat for each indented notation sheet
    Create a new ID sheet
    Repeat for each cell in the indented notation sheet
        If cell value same as parent cell value, give same CategoryNodeOID
        Else generate unique CategoryNodeOIDs
    End Repeat
    Get CategoryRootNodeOID for the category
     Identify ParentNodeOIds and list category nodes for each ParentNodeOID
End Repeat
Get range of data cells (//Data Cells are Delta Cells)
Repeat for each data cell
   Get DataCellOID and DataValue
   Get corresponding CategoryLeafNodeOID / HeaderNodeOID
   If augmentations exist, get augmentation information from Excel comments
End Repeat
```

3.4 Log File

TAT records and time stamps every action the user executes to process every table. TAT also calculates the idle time between actions. The idle time between two actions is interpreted as the preparation time for the next action to be performed by the user. The log filename for each table processed is unique and reflects the time at which the table was processed. The log files were analyzed only after processing all the tables and the entries were made into a separate table which contained the detailed descriptions of the processed tables and the time taken for every action required to process each table completely. A sample log file is attached in the Appendix.

4. Evaluation of TAT

TAT can be used on any computer with Microsoft Excel 98 or higher. Preliminary pilot testing was conducted to check TAT's ability to test different kinds of tables and its user friendliness. The main experiment to evaluate TAT was conducted after fixing a few bugs which were detected in the pilot study.

4.1 Criteria for Table Selection in Evaluation

"A Tabular Survey of Automated Table Processing" by Lopresti & Nagy [8] discusses several issues in automated table processing. It also has a collection of *bizarre* tables not encountered usually. I used the paper as a guide to set the criteria for selecting tables to evaluate TAT. Apart from the characteristics of being a well-formed table in the TAT-sense, the tables I chose for evaluation also had the following characteristics:

1. Tables with rectilinear structure only.
2. Tables with text in English language only.
3. Tables that do not contain graphic symbols or figures.
4. Non recursive tables i.e., no table with a table as one of its content cells.
5. Non-concatenated tables (no tables formed by concatenating two or more tables).
6. Tables from the World Wide Web (WWW) in either HTML or Microsoft Excel format.
7. Tables from the following domains only:
   a. Geopolitical data
   b. Scientific Research data
8. Tables obtained from the following sources only:
   i. **http://www.statcan.gc.ca/**
   ii. **http://www.sciencedirect.com/**
   iii. **http://www.worldbank.org/**
   iv. **http://www.ssb.no/english/**
   v. **http://econdata.net/**
   vi. **http://www.geohive.com/**
   vii. **http://www1.lanic.utexas.edu/la/region/aid/aid98/**
   viii. **http://eia.doe.gov/**
   ix. **http://ies.ed.gov/**
   x. **http://www.census.gov/population/www/socdemo/voting/cps2006.html**
9. Tables which do not span more than one HTML page or Excel sheet. This is a matter of convenience only.

4.2 Experimental Design

The objectives of the experiment conducted using TAT are stated below:

1. To determine the percentage of tables that TAT can convert into modified Wang XML correctly.
2. Identify the major factors which affect the time taken by the users to convert tables into augmented Wang XML.

The experiment I conducted can be best described as a 'screening experiment' in which the main factors that affect the conversion time of web tables are sought. The experiment was done in two steps. Table collection from the websites mentioned above was performed first followed by table selection and processing.

Table collection:

I collected and processed 200 Excel and HTML tables from the 10 websites listed in the previous section. After pasting an HTML file into TAT, it becomes an Excel table. This action takes negligible time and hence I treated both Excel and HTML tables as one. By "collect", I mean:

a) Browsing the Tables section of the websites and saving the files containing the table in a specific location in the file system of the computer that I used to conduct the entire experiment.

b) Storing the set of tables that I looked at but rejected separately. (Number of tables rejected and the reasons for rejection are tabulated and reported in Section 4.4)

c) Make a list of all the acceptable tables and number them serially.

Table Selection and Processing:

The processing of tables took place only after all the tables were collected and was done in multiple sessions in a pseudo random order. Each table was processed according to the instructions mentioned in the TAT Manual (see Appendix). The XML and log files generated were stored separately. The logs were be analyzed only after the completion of the entire experiment. Errors in either the log file or the XML file are reported separately.

4.3 Pilot Study

I conducted an initial pilot study on a set of 15 tables from the following websites: http://www.census.gov , http://factfinder.census.gov , http://www.bls.gov and http://www.who.int . The pilot study was done to verify if the tool was indeed working fine on tables from websites

other than Canada Statistics (http://www.statcan.gc.ca/) used in developing TAT. The pilot study was a useful exercise as I discovered that TAT was able to only handle tables whose size did not exceed 100 rows. This bug was fixed after the analysis. Using the log data for each action from the tables used for the pilot study, an Excel table which includes all the information required to interpret the results was constructed. Since I had to deal with data from *15* tables only, compared to *200* in the final study, this step helped me to design the analysis table which represented the data for the entire experiment efficiently and which contained all the necessary information about every table.

4.4 Main Experiment

I collected 200 tables from the following 10 websites, 9 of which were from the geopolitical domain and 1 from the research domain (results of experiments). Table 1 presents the details:

Table 1 – URLs of table sources

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

The first six criteria for selection of the tables were the same as the ones used for the pilot study. The tables from each source were quite different from each other and tables from the same source were similar. Table 1 is the complete list of table sources used for this study with the URLs. I also rejected 12 tables in the process because they did not meet my criteria. Table 2 presents the reasons for rejection.

Table 2 – Reasons for rejection

Reason for rejection	number of tables rejected
Concatenated Table	6
Tables with graphics, figures or hyperlinks as delta cells	6

4.4.1 Method

I processed the tables only after collecting all the tables. The processing of tables was done in 15 sessions. I selected the tables using a random number generating function and processed the table corresponding to the random number generated for that trial. Each table was processed according to the instructions in the TAT Manual. The XML and log files generated were stored separately. The logs were analyzed and the results were aggregated only after the completion of the entire experiment. The session details are presented in Table 3.

Table 3 - Session Details

Session number	Number of tables processed	Time taken in minutes
1	22	247
2	10	95
3	8	49
4	4	20
5	22	175
6	19	117
7	4	20
8	3	20
9	13	87
10	5	30
11	14	95
12	3	32
13	18	151
14	16	120
15	32	223
Total	**193**	**1481**

It can be noted form the above table that only 193 tables were processed. There were 7 tables in the list which could not be processed either because I could not understand the table or because the content in the table could not be interpreted by Excel correctly. Two tables collected did not actually match the criteria mentioned, but they were collected by mistake and one of them was too large to handle (~85000 cells).

5. Results, Observations and Discussion:

First a few observations based on the experiment are reported followed by the quantitative results and their interpretation.

5.1 General Observations

The following were my observations from the experiment:

1. I realized that there is more flexibility in forming tables, which are perfectly human readable, than I had imagined.
2. Excel does not always perfectly preserve the contents of the HTML table when pasted. Hence, I had to look carefully at the Excel table before beginning to process it.
3. Excel is sometimes a little quirky and automatically converts the data in the cell into a different format. Numbers separated by "-" (to indicate ranges, for example) are automatically converted to a date format. So, care should be taken to correct these automatic conversions before processing tables.
4. Knowing all the Excel operations to manipulate cells will save the user time in processing the tables. Many preprocessing operations require merging of cells. Excel 2007 has a *Merge and Center* button provided which saves much time for the user.
5. The 'idle time' for an action is interpreted as the preparation time for the next action. This is almost always true. But in a few cases, the user might perform an action and then check the validity of the previous action. In such cases, the interpretation of idle time will be incorrect. This is rare.

5.2 Observations of Opportunities for Improvement

6. Almost every table has the "Source" and "Notes" augmentation. Currently, the provision for including this content in the XML is that the user has to push the [AUGMENTATION] button and select the "Other" Augmentation option. If there is an automatic prompt for both Source and Notes for the table, we could save some time and reduce the number of clicks.
7. A single "concept" must occupy one cell only. For example, if "From Date A to Date B" is a column subcategory associated with a column of delta cells, then we cannot have the data in this format:

From Date A
To
Date B

The entire text in the "concept" has to be present in one cell only. I observed that this rule is violated quite often in tables pasted from HTML to Excel.

8. Dealing with empty cells: Currently, whenever TAT encounters empty delta or category cells, it prompts the user to enter a value for each instance. In most of the tables from the web, the user has no idea of what that value might be, especially in the geopolitical domain as these are data collected over a period of time and not a result of some calculation. So, another change which would be helpful in reducing time is to enter default values of "NA" or "unknown" in those cells and inform the user that TAT has done so in particular cells which can be changed manually if desired by the user.
9. Tables from the source **http://ies.ed.gov/** alone were processed all at once in a pseudo random order to see if there was any significant improvement in table processing time due to learning. It was found that this was not a major factor in influencing the total processing time.

5.3 Wang XML and Highlighting

The highlighting process in TAT reflects the XML content i.e., the Highlighting action is basically a visual interpretation of the XML. While processing some tables, I noticed that the highlighting was incorrect and the XML file generated was also incorrect. This occurred usually when the table had an aggregate in the row-category.

Figure 46 is a screenshot showing an instance of a highlighting error:

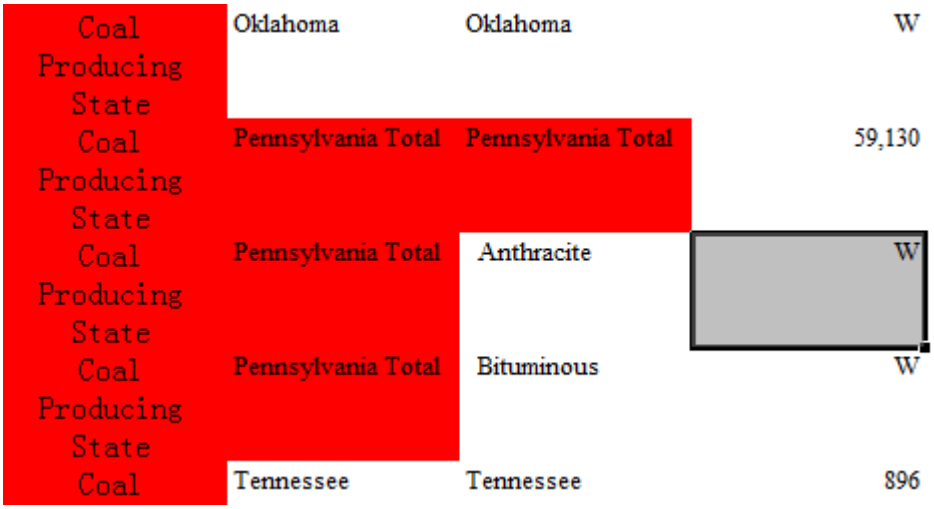


Fig. 46 Highlighting Error

As seen from Figure 46 the proper row category cell *Anthracite* is not highlighted and instead the aggregate *Pennsylvania Total* is highlighted.

Note: The "W"s are delta cells which have an associated footnote.

The XML for this entry is (only relevant portion shown)

```
.....
<CategoryRootNodes>
<CategoryRootNode CategoryRootNodeOID="C1"/>
<CategoryRootNode CategoryRootNodeOID="C2"/>
<CategoryRootNode CategoryRootNodeOID="C3"/>
</CategoryRootNodes>
.....
<CategoryNode CategoryNodeOID="C1.20.1" Label="Oklahoma "></CategoryNode>
<CategoryNode CategoryNodeOID="C1.21" Label="Pennsylvania Total "></CategoryNode>
<CategoryNode CategoryNodeOID="C1.21.1" Label="Pennsylvania Total
"></CategoryNode>
<CategoryNode CategoryNodeOID="C1.21.2" Label="  Anthracite "></CategoryNode>
<CategoryNode CategoryNodeOID="C1.21.3" Label="  Bituminous "></CategoryNode>
<CategoryNode CategoryNodeOID="C1.22" Label="Tennessee "></CategoryNode>
<CategoryNode CategoryNodeOID="C1.22.1" Label="Tennessee "></CategoryNode>
.....
<DataCell DataCellOID="D22,1" DataValue="W">
<CategoryLeafNodes>
<CategoryLeafNode CategoryLeafNodeOID="C1.21.1" />
<CategoryLeafNode CategoryLeafNodeOID="C2.1" />
<CategoryLeafNode CategoryLeafNodeOID="C3.1" />
</CategoryLeafNodes>
</DataCell>
```

Fig. 47 XML notation (part) for table in Fig. 46

Figures 48(a) and 48(b) are from another example table used in the experiment.

	Total offender	Percent of convicted offenders sentenced to incarceration\a							
	Total offender	All offenses	Felonies	Felonies	Felonies	Felonies	Felonies	Felonies	Misdemeanors
	Total offender	All offenses	Violent offens	roperty offense	roperty offense	Drug offenses	lic-order offen	lic-order offen	Misdemeanors
District	Total offender	All offenses	Violent offens	Fraudulent	Other	offenses	Regulatory	Other	Misdemeanors
All distric	53435	65.2	92.4	53.9	59.8	89.6	46.4	77.5	17.2
Alabama, Middl	292	58.6	100	69.6	27.3	94.5	100	87.9	12.6
Alabama, North	371	54.4	83.9	42.7	48.3	89.9	50	72.3	15.5
Alabama, South	424	74.1	90	40.3	61.1	87.1	35.7	93.6	...
Alaska	198	51.5	75	45.5	...	80	52.9	90	26.7
Arizona	1612	71.8	90.6	31.1	62.5	85.3	47.5	83.8	42.7

Fig. 48(a) Correct Highlighting

	Total offender	Percent of convicted offenders sentenced to incarceration\a							
	Total offender	All offenses	Felonies	Felonies	Felonies	Felonies	Felonies	Felonies	Misdemeanors
	Total offender	All offenses	Violent offens	roperty offense	roperty offense	Drug offenses	lic-order offen	lic-order offen	Misdemeanors
District	Total offender	All offenses	Violent offens	Fraudulent	Other	offenses	Regulatory	Other	Misdemeanors
All distric	53435	65.2	92.4	53.9	59.8	89.6	46.4	77.5	17.2
Alabama, Middl	292	58.6	100	69.6	27.3	94.5	100	87.9	12.6
Alabama, North	371	54.4	83.9	42.7	48.3	89.9	50	72.3	15.5
Alabama, South	424	74.1	90	40.3	61.1	87.1	35.7	93.6	...
Alaska	198	51.5	75	45.5	...	80	52.9	90	26.7
Arizona	1612	71.8	90.6	31.1	62.5	85.3	47.5	83.8	42.7

Fig. 48(b) Incorrect Highlighting

For the above table TAT did not highlight the column categories for certain delta cells and did highlight the column categories for certain other delta cells.

The XML for this entry is (only relevant portion shown)

<CategoryRootNodes>
<CategoryRootNode CategoryRootNodeOID="C1"/>
<CategoryRootNode CategoryRootNodeOID="C2"/>
</CategoryRootNodes>

.....

<CategoryNode CategoryNodeOID="C2.2" Label="Percent of convicted offenders sentenced to incarceration\a"></CategoryNode>
<CategoryNode CategoryNodeOID="C2.2.1" Label="All offenses"></CategoryNode>
<CategoryNode CategoryNodeOID="C2.2.1.1" Label="All offenses"></CategoryNode>
<CategoryNode CategoryNodeOID="C2.2.1.1.1" Label="All offenses"></CategoryNode>
<CategoryNode CategoryNodeOID="C2.2.2" Label="Felonies"></CategoryNode>
<CategoryNode CategoryNodeOID="C2.2.2.1" Label="Violent offenses"></CategoryNode>
<CategoryNode CategoryNodeOID="C2.2.2.1.1" Label="Violent offenses"></CategoryNode>

.....

<DataCell DataCellOID="D3,3" DataValue="83.9">
<CategoryLeafNodes>
<CategoryLeafNode CategoryLeafNodeOID="C1.3" />
</CategoryLeafNodes>
</DataCell>

Fig. 49 XML notation (part) for table in Fig. 48

These errors were observed during the experiment and have been fixed.

5.4 Quantitative Results

This section presents the quantitative results obtained from the main experiment. The first set of tables show the time taken to process the entire table based on different criteria like Wang Dimensionality, Number of Aggregates and Number of Footnotes.

5.4.1 Total Time Taken to Process the Table Based on Different Criteria

A. Dimensionality

Table 4 - Time taken to process entire table based on Wang dimensionality

Table Feature	Number of Tables	Time taken to process the entire table				
		Average Time (sec)	Total Time (sec)	Median Time (sec)	Average Time (min)	Median Time (min)
1-D Lists	2	150.50	301.00	150.50	2.51	2.51
2-D tables	140	224.30	31402.00	160.50	3.74	2.68
3-D tables	49	250.96	12297.04	179.00	4.18	2.98
4-D tables	2	246.99	493.98	246.99	4.12	4.12
All Tables	***193***	***230.54***	***44494.22***	***166.00***	***3.84***	***2.77***

On an average, 3-D tables take 27 seconds more than 2-D tables to process. There were two tables in the collection which were basically lists in the form of a table (1D tables).

B. Aggregates

Table 5 - Time taken to process entire table based on presence of aggregates

Table Feature	Number of Tables	Time taken to process the entire table				
		Average Time (sec)	Total Time (sec)	Median Time (sec)	Average Time (min)	Median Time (min)
Tables with 1 aggregate only	44	271.29	11936.76	231.00	4.52	3.85
Tables with 2 aggregates only	15	238.53	3577.95	188.00	3.98	3.13
Tables with > 2 aggregates	28	374.39	10482.92	303.50	6.24	5.06
Tables with aggregates	87	298.82	25997.34	258.00	4.98	4.30
Tables without aggregates	106	174.49	18495.94	140.50	2.91	2.34
All Tables	***193***	***230.54***	***44494.22***	***166.00***	***3.84***	***2.77***

Note: The max number of aggregates was 43

It can be clearly seen from Table 5 that when there are more than two aggregates, the total processing time more than doubles compared to no aggregates. Tables with aggregates require much more preprocessing than tables without aggregates. Automating the selection and processing of aggregates is definitely something to look into.

C. Footnotes

Table 6 - Time taken to process entire table based on presence of footnotes

Table Feature	Number of Tables	Time taken to process the entire table				
		Average Time (sec)	Total Time (sec)	Median Time (sec)	Average Time (min)	Median Time (min)
Tables with 1 footnote only	21	183.67	3857.07	179.00	3.06	2.98
Tables with 2 footnotes only	17	291.00	4947.00	320.00	4.85	5.33
Tables with >2 footnotes	35	297.02	10395.70	275.00	4.95	4.58
Tables with footnotes	73	263.01	19199.73	242.00	4.38	4.03
Tables without footnotes	120	210.78	25293.60	146.50	3.43	2.44
All Tables	***193***	***230.54***	***44494.22***	***166.00***	***3.84***	***2.77***

Note: The maximum number of cells with footnotes was 214

Table 6 shows that the presence of footnotes also affects the processing time. Since there is a specific format for specifying a footnote (i.e., below the delta cells using a reference), automating this process would definitely result in reducing processing time.

D. Footnotes, Aggregates and Dimensionality

Table 7 - Time taken to process entire table based on presence of footnotes, aggregates and dimensionality

Table Feature	Number of Tables	Time taken to process the entire table				
		Average Time (sec)	Total Time (sec)	Median Time (sec)	Average Time (min)	Median Time (min)
2-D Tables with aggregates	56	286.92	16067.52	247.50	4.78	4.13
3-D tables with aggregates	29	312.65	9066.85	276.00	5.21	4.60
2-D Tables with footnotes	53	246.09	13042.77	211.00	4.10	3.52
3-D tables with footnotes	20	288.65	5773.00	275.50	4.81	4.59
Tables with aggregates and footnotes	49	280.65	13751.85	276.00	4.68	4.60
Tables with no aggregates and footnotes	83	164.27	13634.41	133.00	2.74	2.22

Table 7 reiterates the point that aggregates and footnotes are the main causes for high processing time for tables.

E. Table Size

Table 8 - Time taken to process entire table based on table size (number of cells in the table)

Table Feature	Number of Tables	Time taken to process the entire table				
		Average Time (sec)	Total Time (sec)	Median Time (sec)	Average Time (min)	Median Time (min)
Tables with number of cells <=100	33	114.24	3769.92	84.00	1.90	1.40
Tables with number of cells >100 and <= 200	45	148.80	6696.00	133.00	2.48	2.22
Tables with number of cells >200 and <= 300	29	212.37	6158.73	153.00	3.54	2.55
Tables with number of cells >300 and <= 500	36	230.89	8312.04	215.00	3.85	3.58
Tables with number of cells >500 and <=800	17	384.11	6529.87	299.00	6.40	4.98
Tables with number of cells >800	33	394.76	13027.08	379.00	6.58	6.32
All Tables	***193***	***230.54***	***44494.22***	***166.00***	***3.84***	***2.77***

Note: The maximum number of cells in a table : 4247

Table 8 shows the time taken to process the entire table as a function of table size. The table size criterion presents a really strong trend with respect to the total time required to process the table. This is an expected result.

F. Table Source

Table 9 - Time taken to process entire table based on table source

Table Source	Number of tables	Time taken to process the entire table				
		Average Time (sec)	Total Time (sec)	Median Time (sec)	Average Time (min)	Median Time (sec)
http://www.statcan.gc.ca/	20	127.05	2541.00	97.50	2.12	1.63
http://www.sciencedirect.com/	15	134.40	2016.00	131.00	2.24	2.18
http://www.worldbank.org/	18	189.05	3402.90	131.00	3.15	2.18
http://www.ssb.no/english/	21	201.14	4223.94	150.00	3.35	2.50
http://www.ojp.usdoj.gov	24	357.12	8570.88	371.00	5.95	6.18
http://www.geohive.com/	26	182.11	4734.86	147.00	3.04	2.45
http://www1.lanic.utexas.edu/la/region/aid/aid98/	15	320.33	4804.95	218.00	5.34	3.63
http://eia.doe.gov/	24	298.79	7170.96	264.00	4.98	4.40
http://ies.ed.gov/	23	194.87	4482.01	155.00	3.25	2.58
http://www.census.gov/population/www/socdemo/voting/cps2006.html	7	363.71	2545.97	275.00	6.06	4.58
All Tables	***193***	***230.54***	***44494.22***	***166.00***	***3.84***	***2.77***

It can be seen that there is a large variation in the average time taken to process tables depending on the source. Some of them can be attributed to reasons seen above. Tables from http://www.ojp.usdoj.gov/ and http://www.census.gov/population/www/socdemo/voting/ cps2006.html took more time because these tables were large. Tables from http://eia.doe.gov/ took more time because these tables had many aggregates and required preprocessing. Tables from http://www1.lanic.utexas.edu/la/region/aid/aid98 were poorly constructed and took much time to analyze. In fact, three of these tables were so badly constructed that I could not understand them and hence they were not processed. Tables from http://www.statcan.gc.ca took the least amount of time, which can be attributed to my familiarity with those tables.

5.4.2 Preprocessing time

The following set of tables show how the preprocessing time changes with different criteria.

A. Dimensionality

Table 10 - Time taken to preprocess entire table based on Wang dimensionality

Table Feature	Number of Tables	Time taken to preprocess the table				
		Average Time (sec)	Total Time (sec)	Median Time (sec)	Average Time (min)	Median Time (min)
1-D Lists	2	67.50	135.00	67.50	1.13	1.13
2-D tables	140	102.77	14387.80	73.00	1.71	1.22
3-D tables	49	111.63	5469.87	85.00	1.86	1.42
4-D tables	2	51.00	102.00	51.00	0.85	0.85
All Tables	***193***	***104.12***	***20095.16***	***75.00***	***1.74***	***1.25***

3-D tables took approximately 9 seconds more on average to be made TAT-friendly through preprocessing. It is interesting to note that the two 4-D tables required less than half the average preprocessing time. It would be interesting to study these higher dimensional tables. They might be *better organized* than the lower dimensional tables. But this is just a guess and can be ascertained only if more similar data is available.

B. Aggregates

Table 11 - Time taken to preprocess table based on presence of aggregates

Table Feature	Number of Tables	Time taken to preprocess the table				
		Average Time (sec)	Total Time (sec)	Median Time (sec)	Average Time (min)	Median Time (min)
Tables with 1 aggregate only	43	122.95	5286.85	79.00	2.05	1.32
Tables with 2 aggregates only	15	106.47	1597.05	74.00	1.77	1.23
Tables with > 2 aggregates	28	177.18	4961.04	126.50	2.95	2.11
Tables with aggregates	86	137.73	11844.78	103.00	2.30	1.72
Tables without aggregates	107	77.10	8249.70	60.00	1.29	1.00
All Tables	***193***	***104.12***	***20095.16***	***75.00***	***1.74***	***1.25***

Note: The max number of aggregates was 43

Table 11 again shows the effect of aggregates on the preprocessing time. There is a more than 100 % increase in the preprocessing time for tables with more than 2 aggregates compared to tables without aggregates.

C. Table Size

Table 12 - Time taken to pre process table based on table size(number of cells in the table)

Table Feature	Number of tables	Time taken to preprocess the table				
		Average Time (sec)	Total Time (sec)	Median Time (sec)	Average Time (min)	Median Time (min)
Tables with number of cells <=100	33	64.63	2132.79	40.00	1.08	0.67
Tables with number of cells >100 and <= 200	45	53.17	2392.65	47.00	0.89	0.78
Tables with number of cells >200 and <= 300	29	90.69	2630.01	52.00	1.51	0.87
Tables with number of cells >300 and <= 500	36	115.08	4142.88	99.00	1.92	1.65
Tables with number of cells >500 and <=800	17	156.35	2657.95	118.00	2.61	1.97
Tables with number of cells >800	33	186.00	6138.00	198.00	3.10	3.30
All Tables	***193***	***104.12***	***20095.16***	***75.00***	***1.74***	***1.25***

Table 12 shows again the strong positive correlation between the preprocessing time and table size.

5.4.3 Other Results

A. Generate XML action

Table 13 - Time taken to Generate XML based on table size (number of cells in the table)

Table Feature	Number of Tables	Average Time (sec)	Median Time (sec)
Tables with number of cells <=100	33	0.33	0.00
Tables with number of cells >100 and <= 200	45	1.40	1.00
Tables with number of cells >200 and <= 300	29	6.51	2.99
Tables with number of cells >300 and <= 500	36	5.22	5.00
Tables with number of cells >500 and <=800	16	10.35	10.00
Tables with number of cells >800	33	27.09	22.00

Note: The maximum number of cells in a table was 4247

From Table 13 it can be seen that for large tables with more than 800 cells, the system takes almost 22 seconds to generate the XML on an average. Future versions of TAT could improve the XML generation process.

B. Highlight Cells Action

Table 14 -Time spent by the user in the Highlight Cells Action

Table Feature	Number of tables	Average Time (sec)
2-D tables	140	12.95
3-D tables	49	7.68
4-D tables	2	10.25
All Tables	193	11.51

An interesting observation from Table 14 is that the time spent by the user in verifying the XML through highlighting the cells is highest for 2-D tables and less for 3-D tables. This might be attributed to the XML/highlighting errors shown in the previous section which occurred in 2-D tables.

6. Future Work

As illustrated in the above sections, TAT is a robust and efficient tool to convert web/Excel tables to a layout independent representation. Additional features can be added to TAT to reduce the processing time. TAT can also be used to improve the Query By Table (QBT) mechanism which is currently implemented using WNT. Since VBA and Excel are much more suited for database interaction, TAT is a better option to implement the QBT mechanism. The current section contains two parts. The first part explains the QBT paradigm and why TAT is better equipped to handle QBT. In the second part, current research to incorporate automation and learning techniques is discussed.

6.1. Query By Table (QBT)

Query Languages are high level programming languages that allow users to make queries into databases. Examples include SMARTS, the cheminformatics standard for a substructure search, XQuery, a query language for XML data sources and SQL for relational databases. The semantics of the query are defined by a formal syntax specific to that language. In contrast, Query By Example and Query By Browsing have a simple graphical user interface. Query by Table (QBT), an intuitive mechanism based on the idea that well-formed tables can represent queries to a database. The QBT system was realized using the Wang Notation Tool. Answers to certain kinds of queries seem most naturally expressed in tabular form. Consider, for example, the query:

"*How do the volume of U.S. auto exports to Mexico compare to those to Canada for the years 2002 and 2003?*"

This question can be formulated as a query in the form shown below:

Year	U.S Exports to Canada	U.S Exports to Mexico
2002	<to be filled>	< to be filled>
2003	<to be filled>	<to be filled>

Fig. 50 Query Table 1

QBT takes inputs in the form of a table shown above and fills in the values. A table's layout can be changed in many ways yet convey the same meaning with the same values. Therefore, to use tables as queries to a database, it is important distinguish between a table's physical structure and its logical structure. Thus, Wang Notation provides the perfect platform for this paradigm. The following section explains the inner workings of QBT.

6.1.1 The QBT Mechanism

A query table is a well-formed user-constructed table that encapsulates all the information internal to its structure and uses exact attribute names for the facts to be retrieved. A query table does not use captions or titles to convey information. The Query by Table system is currently implemented in MATLAB. It accepts an MS Excel query table as an input and retrieves the values requested by the user. The process has five steps:

- Derive the Wang Notation for the query table.
- Parse the Wang Notation of the input table.
- Identify the facts and dimensions of the query.
- Form SQL queries from the facts and dimensions.
- Plug the results back into the query table.

Since WNT is built in MATLAB and the query and results are in an Excel worksheet, there is a lot of code involved in transferring the data from the format in Excel sheet to what WNT deems the correct format for a table (which is an ASCII file). This presents difficulties in practice. Since TAT is built in Excel using VBA, it can encompass the entire QBT system with some modifications, removing the need for any external data conversion routines. Also, since TAT has provisions to include augmentations, some interesting results can be achieved by coupling TAT's functionality with the power of SQL.

6.2 Visual Layer Analysis

6.2.1 Automation

To process tables using TAT, the user has to delineate the delta cells from the category cells using mouse clicks. Tables with aggregates and footnotes take much more time to process. The current implementation of selecting and annotating the aggregate cells in TAT becomes cumbersome in tables with many aggregates. If the aggregates and delta cells can be detected automatically using

visual cues in the table, it could save the user much time in processing the table. Research, which can be easily integrated with TAT, is already underway to deal with this problem of table segmentation in DocLab, RPI [9]. The proposed method relies on visual distinction between cells of a canonicalized table. The cell's features can be captured in a feature vector with both numerical and categorical attributes. By comparing the feature vectors of adjacent cells using a comparison function, a difference table can be formed, which can be used to perform orientation analysis and category-delta space segmentation.

The current check for well-formed tables is based on the repetition of list-rows as explained in Section 2.7.1. This check is done before the user selects the categories. But there is no mechanism in TAT which checks the correctness of the dimensionality of the table determined by the user. For example, a user can incorrectly interpret a well-formed two category table as a three category table. This can be prevented by adding the following requirement in addition to those described in Section 2.7:

"Every combination of paths, one through each category tree, must designate a delta cell."

6.2.2 Learning

A successful mechanism to relate the *difference table* values with the preprocessing actions that need to be performed on the tables to make them TAT-friendly can lay the platform for "learning". The preprocessing actions can be simplified if TAT can *learn* to make corrections based on past responses. TAT could compare the difference table of the current table to past difference tables and make corrections to the table structure based on their similarities.

7. Conclusion

The Table Abstraction Tool (TAT) was developed as a part of the project known as TANGO (Table ANalysis for Generating Ontologies) – a collaborative project between Rensselaer Polytechnic Institute and Brigham Young University – which aims to understand a table's structure and its conceptual content for semi-automatic ontology generation. A major step in this process is to fully interpret a table's structure and conceptual content by converting it to a layout independent form with guidance from a user. Wang Notation Tool (WNT) was one of the few attempts at complete interpretation of tables and the first to convert HTML tables to abstract Wang Notation. TAT builds upon WNT and includes additional information about tables. This improves the resulting ontology.

TAT makes use of minimal user input to convert the HTML tables, which are imported into Excel, to a layout-independent augmented Wang XML notation. This Wang XML, apart from containing the abstract Wang Notation of the table in XML form, also includes additional information like title, caption, aggregates, units and notes. After importing the table, the user can start processing the table to generate the XML with the interface coded in VBA by responding to a series of prompts using mouse-clicks. The augmented XML representation is used to generate ontologies at Brigham Young University.

TAT was evaluated by a single operator in 15 sessions that took a total of 24.7 hours. The samples were collected from ten non-profit web sites which contain thousands of tables relevant to the geopolitical domain. Two hundred tables that satisfied given criteria were processed in a pseudo-random order using TAT. Each selected sample was edited if necessary, and every editing operation was time-stamped and recorded. The time required for editing the table into the desired format along with the interaction to process title, caption, footnotes, units, and aggregates was logged. The Wang Notation for seven of the two hundred tables could not be determined.

The average total time to process a table was 231 seconds (Median: 166 seconds). The average size of the tables was 587 cells before preprocessing. The average preprocessing time to edit a table to a desired format was approximately 104 seconds (median: 75 seconds). Tables with Wang dimensionality 3 took approximately 27 seconds more than tables with Wang dimensionality 2, to

process. It was observed that tables with aggregates took much more time to process when compared to tables without them. This could be attributed to the fact that these tables required transformations to get them into the desired format (as described in Section 2.8.2). As expected, there was a strong positive correlation observed between the processing time and table size.

TAT is a fast and robust tool for generating the layout independent Wang XML notation. Evaluation reveals that a few tasks, if automated, can greatly reduce the time required to process web tables as described in Section 6. TAT can also be modified to incorporate the Query By Table mechanism.

References

1. T. Berners-Lee, J. Hendler, O. Lassila, “The Semantic Web”, Scientific American, 284(5):34-43, May 2001.

2. D. McGuinness, “Ontologies Come of Age. The Semantic Web: Why, What and How”, MIT Press, 2001.

3.Y. Tijerino, D. Embley, D. Lonsdale, Y. Ding, G. Nagy, “Toward Ontology Generation from Tables”, World Wide Web, vol. 8, no. 3, pp 261 – 285, Sept. 2005.

4. P. Jha, G. Nagy, “Wang Notation Tool: Layout Independent Representation of Tables” ICPR 08, Tampa Bay, Florida.

5. R. Padmanabhan, G. Nagy, “Query By Table”, ICPR 08, Tampa Bay, Florida

6. X. Wang, "Tabular Abstraction, Editing, and Formatting" Ph.D Dissertation, University of Waterloo, Waterloo, ON, Canada, 1996.

7. J. Hu, R. Kashi, D. Lopresti, G. Nagy, G. Wilfong, “Why table ground-truthing is hard.” In: Proceedings of the Sixth International Conference on Document Analysis and Recognition, pp. 129–133.Seattle, WA (2001).

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

9. W. Silversmith, “Visual Layer Analysis”, Technical report, DocLab, RPI.

Appendix A: TAT User Manual

The Table Abstraction Tool (TAT) is an interactive tool used to generate an abstract notation of a table in the form of an XML file that records the relationships between the cells in the table. The tool extracts the logical structure of the table from its Excel representation with some user input. Each user action is time-stamped and logged in a file. The tool is built in Microsoft Excel utilizing VBA, an event driven programming language built into Microsoft Office Applications. It is required that the user of TAT be fairly proficient with Microsoft Excel. The process consists of four steps some of which require multiple user actions.

I. IMPORTANT NOTES:

To run TAT:

The application is a set of macros written in VBA. Therefore, enable macros before running the application. In Excel 2003, the Security Settings in the menu path Tools > Macros > Security need to be set to Medium or Low to allow the macros to run. In Excel 2007, the Macro security changes can be made from Code group > Developer tab > Macro Security. If the Developer tab is not displayed then click Microsoft Office Button & click Excel options and then in the Popular category, under Top options for working with Excel, check Show Developer tab in the Ribbon.

To modify TAT:

Before editing a macro the user should be familiar with the Visual Basic Editor - an environment in which one can write new and edit existing Visual Basic for Applications code and procedures. The Visual Basic Editor contains a complete debugging toolset for finding syntax, run-time, and logical errors in the code. The Visual Basic Editor can be used to write and edit a macro that is attached to a Microsoft Office Excel workbook. To view the code, click Visual Basic. Code is present in the following code modules: Sheet1, Input_Form , Module 1 & public Class.

Other Notes:

The program works through a number of subroutines that store temporary variables. Some data whose scope is restricted to one subroutine and whose values cannot be retrieved are stored in a separate sheet called "MySheet" which is hidden. It contains data that the program uses in various subroutines. The log for the process is recorded by the system as and when the actions are performed by the user, in a hidden sheet called "Log". Once the user clicks the [GENERATE XML] button, the program creates the file in the file system. On clicking the [START] button, values in both the sheets are cleared. This manual is also embedded in the Excel Workbook itself and appears as an Adobe Acrobat icon in the worksheet in the first column. Clicking the icon opens this Help manual.

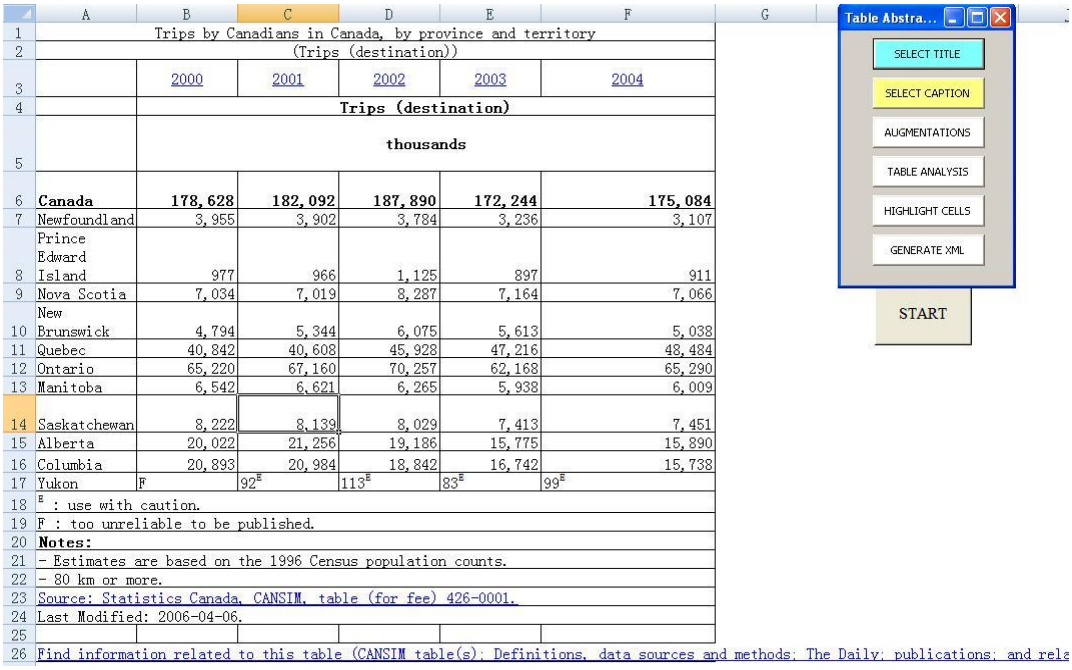
The process of *abstracting* a table is described below:

II. TABLE ABSTRACTION:

The first step is to open the excel workbook – "TAT.xls", which has all the code in the code modules (This can be verified using the Visual Basic Editor). Then, copy the original HTML table into Microsoft Excel. To preserve the table structure in Excel, the contents in the web page outside the table (some content which is not related to the table) should also be copied and pasted directly. The user can later remove the extraneous content although it is not necessary. Then click the [START] button in the Excel sheet (which should be integrated into Eclipse framework eventually). This draws the borders for the cells occupied by the table and pops up a User Form with the following options (displayed as buttons) for the user:

1. **SELECT TITLE**
2. **SELECT CAPTION**
3. **AUGMENTATIONS**
4. **TABLE ANALYSIS**
5. **HIGHLIGHT CELLS**
6. **GENERATE XML**

Figure 51 displays an example table:



	A	B	C	D	E	F
1	Trips by Canadians in Canada, by province and territory					
2	(Trips (destination))					
3		2000	2001	2002	2003	2004
4		Trips (destination)				
5		thousands				
6	**Canada**	**178,628**	**182,092**	**187,890**	**172,244**	**175,084**
7	Newfoundland	3,955	3,902	3,784	3,236	3,107
8	Prince Edward Island	977	966	1,125	897	911
9	Nova Scotia	7,034	7,019	8,287	7,164	7,066
10	New Brunswick	4,794	5,344	6,075	5,613	5,038
11	Quebec	40,842	40,608	45,928	47,216	48,484
12	Ontario	65,220	67,160	70,257	62,168	65,290
13	Manitoba	6,542	6,621	6,265	5,938	6,009
14	Saskatchewan	8,222	8,139	8,029	7,413	7,451
15	Alberta	20,022	21,256	19,186	15,775	15,890
16	Columbia	20,893	20,984	18,842	16,742	15,738
17	Yukon	F	92[E]	113[E]	83[E]	99[E]
18	[E] : use with caution.					
19	F : too unreliable to be published.					
20	**Notes:**					
21	- Estimates are based on the 1996 Census population counts.					
22	- 80 km or more.					
23	Source: Statistics Canada, CANSIM, table (for fee) 426-0001.					
24	Last Modified: 2006-04-06.					
25						
26	Find information related to this table (CANSIM table(s); Definitions, data sources and methods; The Daily; publications; and rela					

Fig. 51 TAT

Note: In Figure 51, the information in row 26 – "Find information related to this table ……." is not part of the table. But for a faithful representation of the table in Excel, that line from the HTML is also copy-pasted in the Excel. This is the way Excel works.

Figure 52 shows the Excel representation of the table if text outside the <table>… </table> tags are not copied.

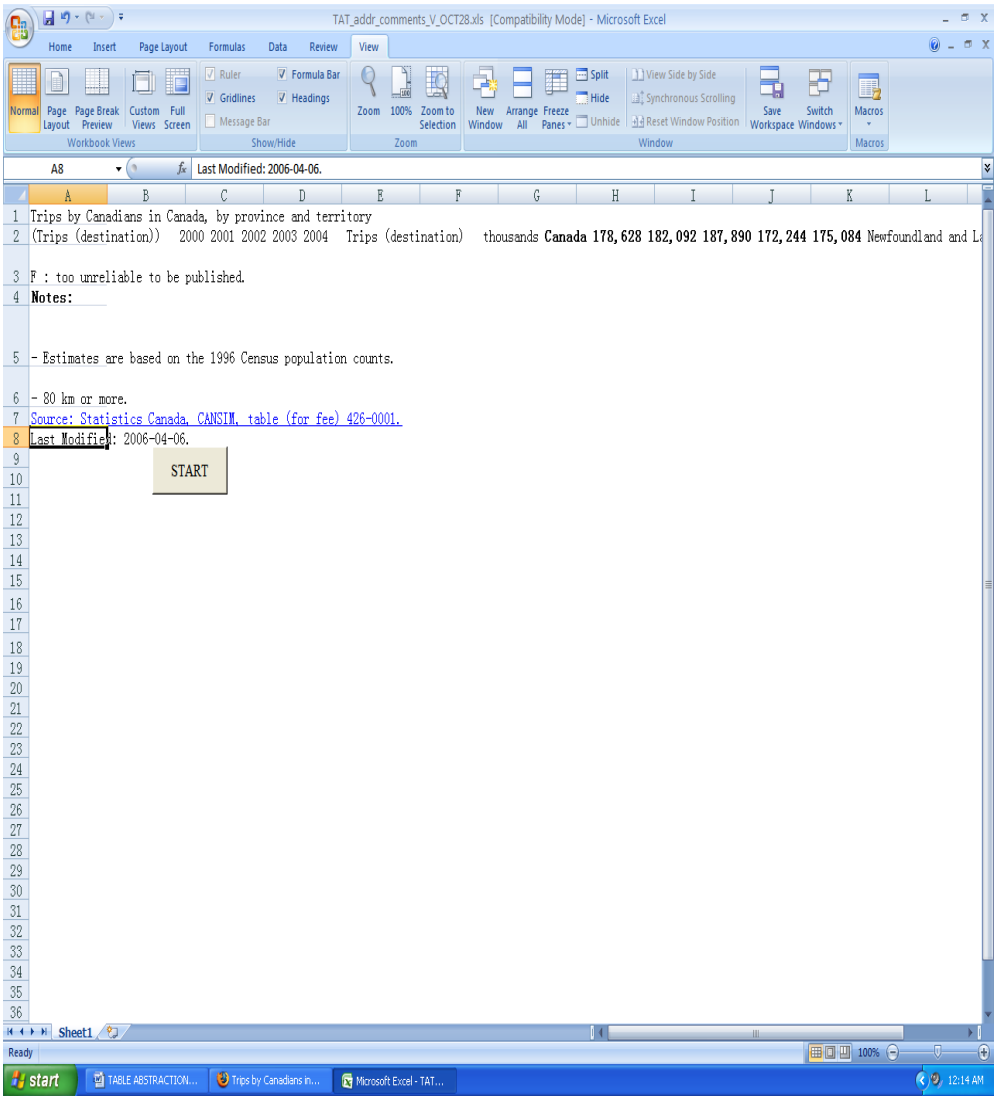


Fig. 52 Paste Error

1. SELECT TITLE:

The blue [SELECT TITLE] button displays a pop up that asks the user to select the cells containing the title. The cells selected by the user as the title are also colored. In the above example, the title would be “Trips by Canadians in Canada, by province and territory”.

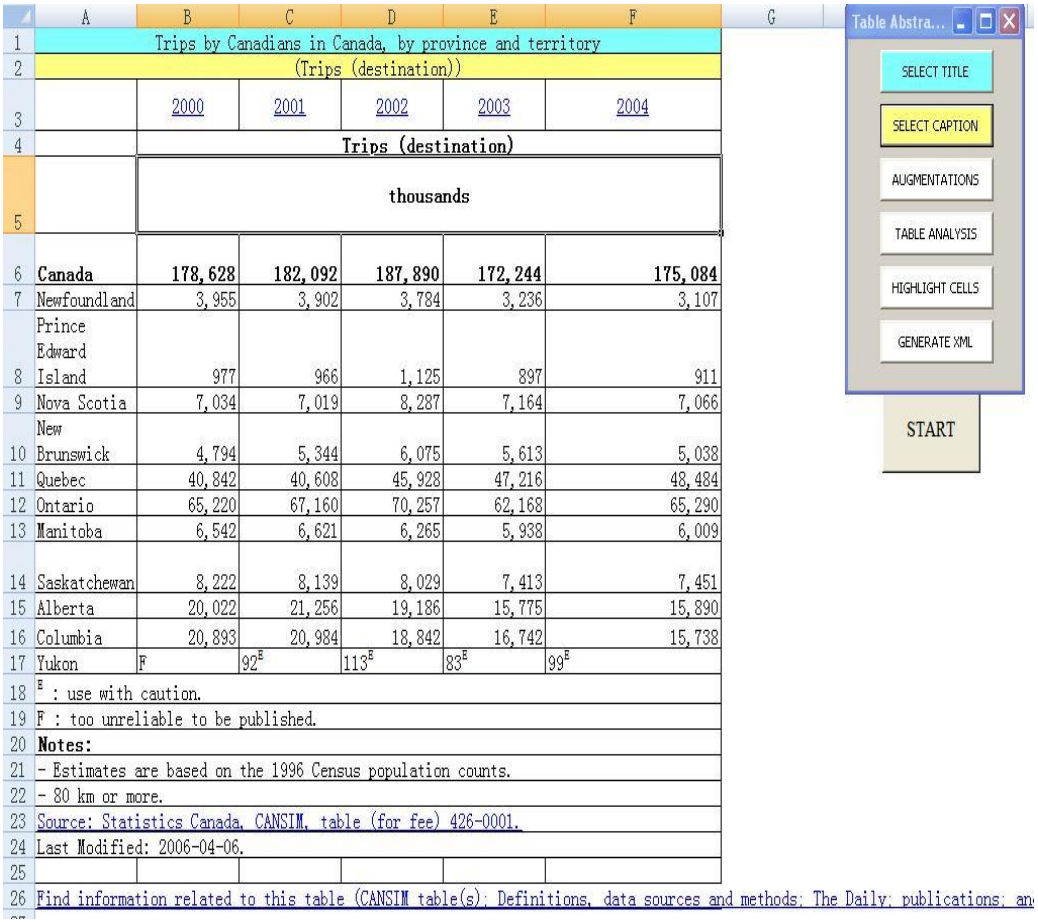
2. SELECT CAPTION:

The yellow [SELECT CAPTION] button, when clicked, displays a pop up that asks the user to select the cells representing the table caption. The cells selected by the user as caption are also colored. The caption in this case would be “(Trips (destination))”.

For tables without title and caption, the user can directly press the [TABLE ANALYSIS] button. However, if title and caption are present and the user does not select them, the XML representation will not contain those details and these cells will also be canonicalized in the "Table Analysis" action. The order of selecting the title and caption does not matter. The title can also be selected after selecting the caption.

Note: Please refer to the Table Analysis section for Canonicalization process.

After selecting the title and caption, the table is highlighted with the title and caption (Fig. 53). TAT is now ready to process "Augmentations".



	A	B	C	D	E	F	G
1	Trips by Canadians in Canada, by province and territory						
2	(Trips (destination))						
3		2000	2001	2002	2003	2004	
4		Trips (destination)					
5		thousands					
6	**Canada**	**178,628**	**182,092**	**187,890**	**172,244**	**175,084**	
7	Newfoundland	3,955	3,902	3,784	3,236	3,107	
8	Prince Edward Island	977	966	1,125	897	911	
9	Nova Scotia	7,034	7,019	8,287	7,164	7,066	
10	New Brunswick	4,794	5,344	6,075	5,613	5,038	
11	Quebec	40,842	40,608	45,928	47,216	48,484	
12	Ontario	65,220	67,160	70,257	62,168	65,290	
13	Manitoba	6,542	6,621	6,265	5,938	6,009	
14	Saskatchewan	8,222	8,139	8,029	7,413	7,451	
15	Alberta	20,022	21,256	19,186	15,775	15,890	
16	Columbia	20,893	20,984	18,842	16,742	15,738	
17	Yukon	F	92[E]	113[E]	83[E]	99[E]	
18	[E] : use with caution.						
19	F : too unreliable to be published.						
20	**Notes:**						
21	- Estimates are based on the 1996 Census population counts.						
22	- 80 km or more.						
23	Source: Statistics Canada, CANSIM, table (for fee) 426-0001.						
24	Last Modified: 2006-04-06.						
25							
26	Find information related to this table (CANSIM table(s); Definitions, data sources and methods; The Daily; publications; an						

Fig. 53 Highlighted Title and Caption

3. AUGMENTATIONS:

The [AUGMENTATIONS] button, when clicked, displays a pop up which helps the user to select the augmentations for the cells in the table. There are four kinds of augmentations that a user can currently choose (using radio buttons):

1. Footnotes
2. Aggregates
3. Units
4. Other

3 (a). Footnotes:

To specify a footnote, the user must make three selections:
(1) Cells to which the footnote applies. (These cells also contain the footnote reference.) (2) Footnote Citation- which specifies the footnote text.
(3) The symbol used as the footnote reference.

When the user clicks selects the "Footnote" option (after clicking the [AUGMENTATIONS] button) and clicks the [OK] button, the system displays a pop up that asks the user to select the cells with the footnote reference (Fig. 54).

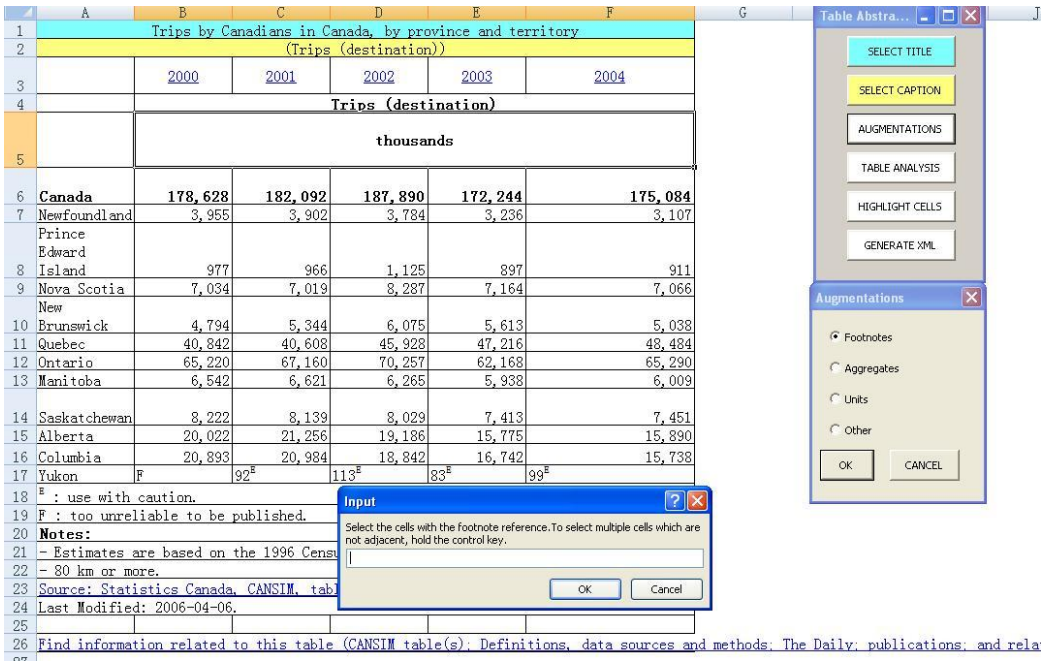


Fig. 54 Footnote Cell Selection

In the above table there are two footnotes. The first one corresponds to the cell with address “B17” i.e. the value of number of trips to Yukon territory in the year 2000 is “F” which is „too unreliable to be published’. Thus, the user selects cell “B17” for the first pop up where the cells with the footnote reference are requested and clicks the [OK] button (Fig. 54). This causes another pop up to be displayed which asks the user to select the cells with the corresponding footnote citation (Fig. 55).

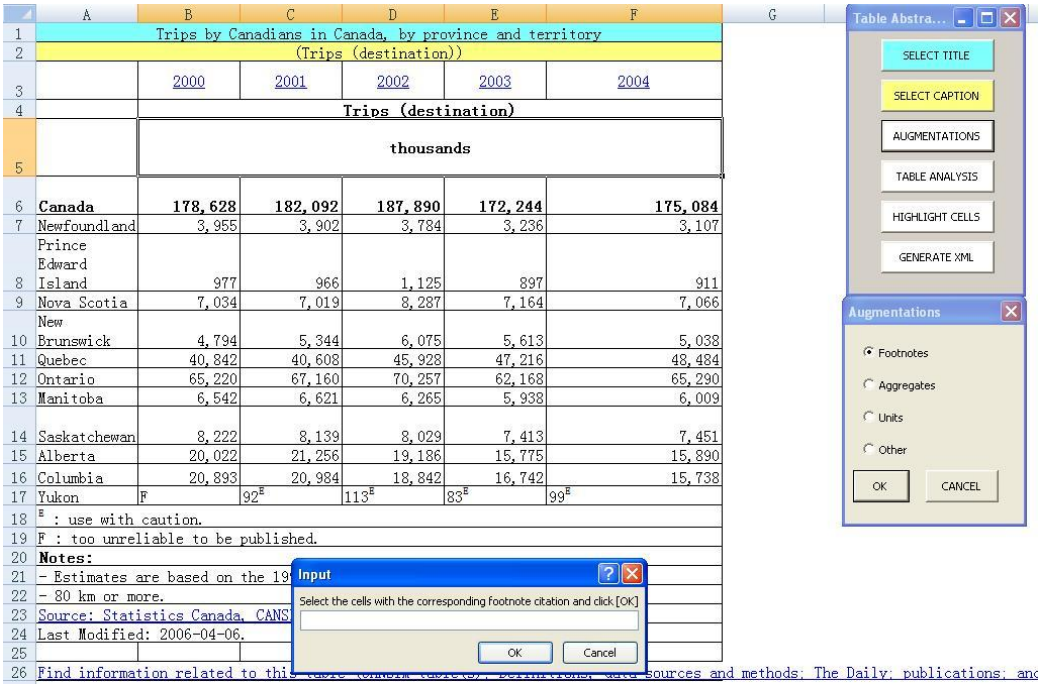


Fig. 55 Footnote Citation

The user then selects cell "A19" – with the text „too unreliable to be published", as the footnote citation, and clicks the [OK] button. This causes another pop up to be displayed which requests the user to specify the footnote reference (which is "F" in this case). After finishing this process of specifying one footnote, the following text appears as a comment for the cell "B17"–

(footnote)F : too unreliable to be published.(/footnote)(f.reference)F(/f.reference)

The string within the (footnote) and (/footnote) is the footnote text and the string between (f.reference) and (/f.reference) is the footnote reference. This allows the user to verify the actions performed. The third popup asks the user for the symbol used for footnote reference (Fig. 56).

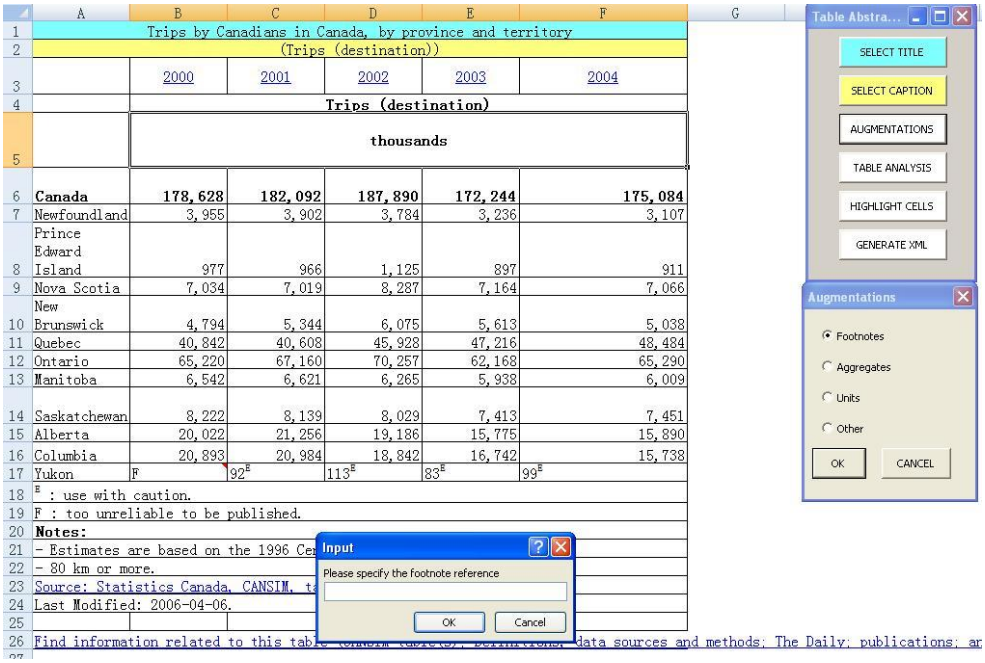


Fig. 56 Footnote Reference

To specify the values for the second footnote - cells “C17: F17”, the user has to make sure the “Footnote“ radio button in the menu is selected and click the [OK] button·

The following are the values for the second footnote:

1. Cells with the footnote reference: C17:F17
2. Cells with the footnote citation: A18 (“E: use with caution”)
3. Footnote reference: E

3(b). Aggregates:

TAT allows the users to specify the aggregate cells in the table. The aggregate cells are the category/sub-category cells whose corresponding delta cells are an aggregate (like sum or average) of delta cells of other sub-categories. In the example table, the delta cells corresponding to “Canada” i.e., the delta cells in the same row as Canada are actually a summation of all the delta cells below them. Thus, Canada is an aggregate of the sub-category cells – Newfoundland and Labrador, Prince Edward Island, Nova Scotia, etc.

When the user selects the „Aggregate" option and clicks the [OK] button, the system asks the user to select the aggregate cells (Fig. 57).

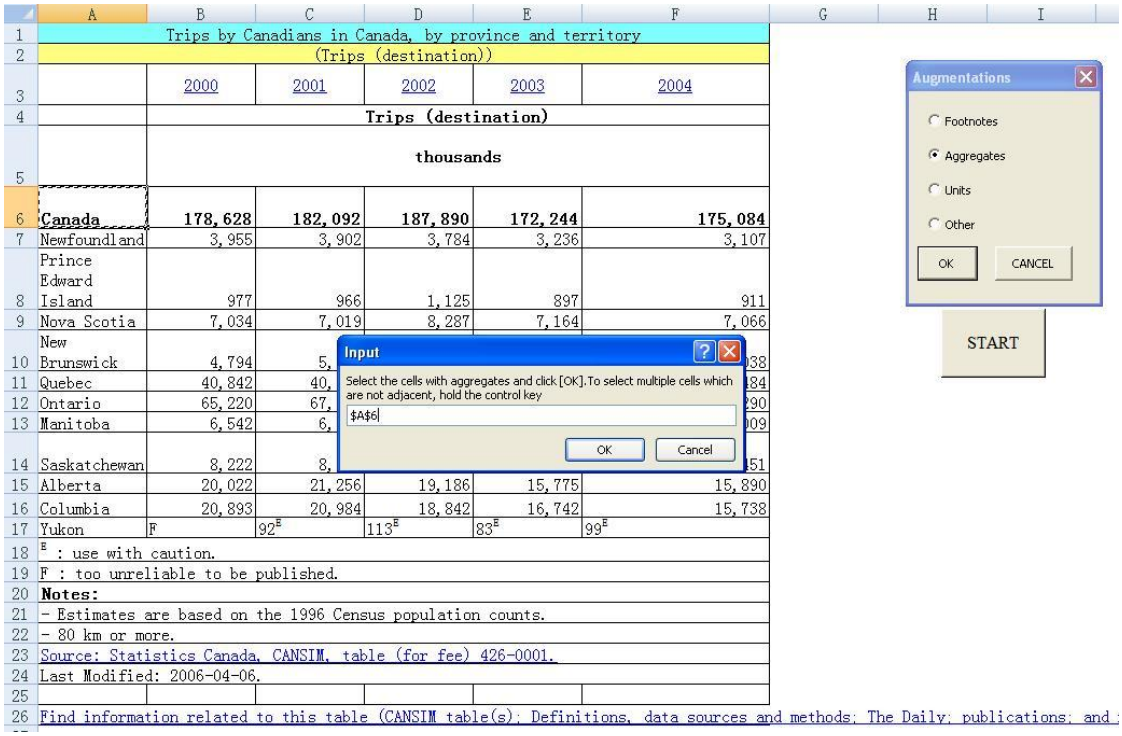


Fig. 57 Aggregate Cell Selection - 1

The user selects the cell – "Canada" and clicks the [OK] button. The system displays another pop up asking the user which subcategory cells are aggregated by the previously selected cell (Fig. 58). The user selects all the cells from "Newfoundland" to "Yukon", as the values for the delta cells corresponding to "Canada" are the sum of the delta cells corresponding to these subcategories.

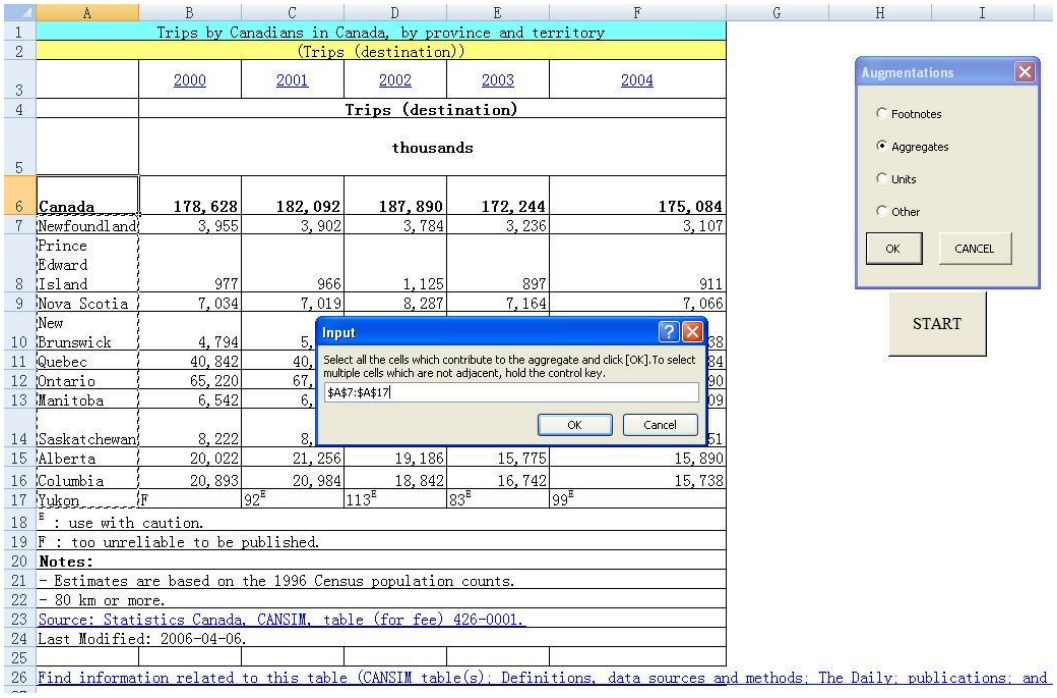


Fig. 58 Aggregate Cell Selection - 2

On selecting the subcategory cells and clicking the [OK] button, the system displays the following text as a comment for the cell "A6" containing information about the cells it aggregates.

(aggregate)$A$7:$A$17(/aggregate)

The string between the (aggregate) and /(aggregate) identifiers specifies the cell addresses of the cells it aggregates.

3.(c). Units:

The units of the delta cells are important in interpreting the table contents. TAT allows the users to specify the cells in the table which are units. In the above table, the spanning cell "thousands" specifies the units of the Number of trips by Canadians. In the Augmentations menu, if the user selects the „Units" radio button and clicks the [OK] button, the system displays a pop up asking the user to select the cells which denote units in the table (Fig. 59).

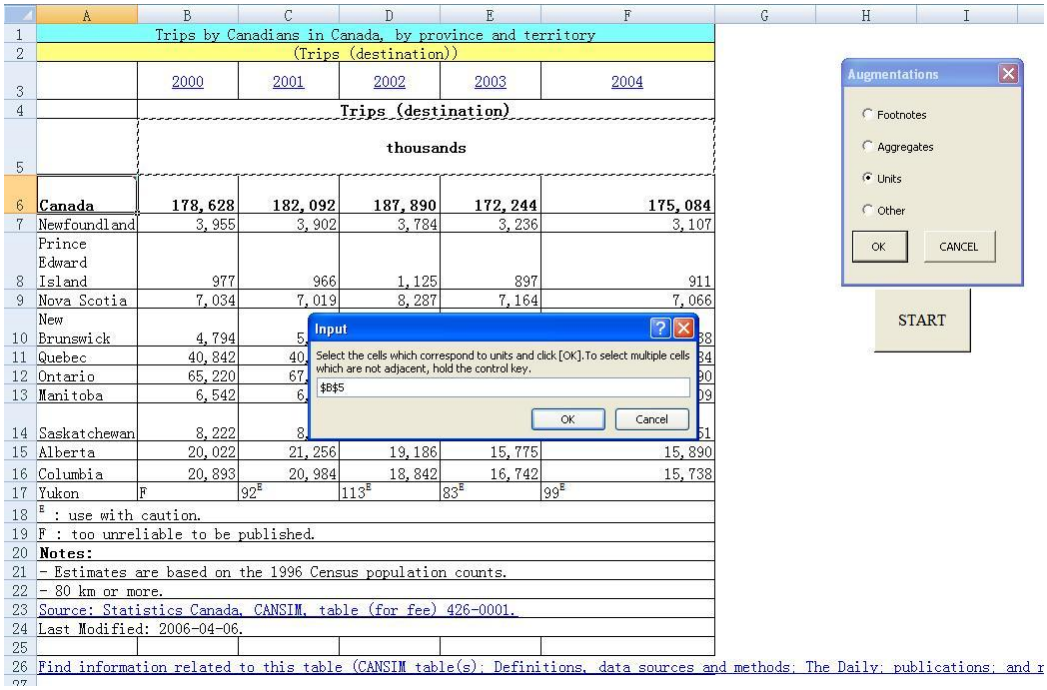


Fig. 59 Units

The user selects the cell “thousands” and clicks the [OK] button. This causes the text (Units) to appear as a comment for the cell.

3.(d). Other:

The user can also specify any other kind of augmentations for any cells in the table. For example, the cells in the rows 21 & 22 with the text– “Estimates are based on the 1996 Census population counts” and “80 km or more” are special notes for the table. TAT allows users to specify these kind of additional notes as well. In the Augmentations menu, the user selects the “Other” option and clicks the [OK] button. The system then displays a pop up to the user and asks the user to select the cells with any other augmentation (Fig. 60).

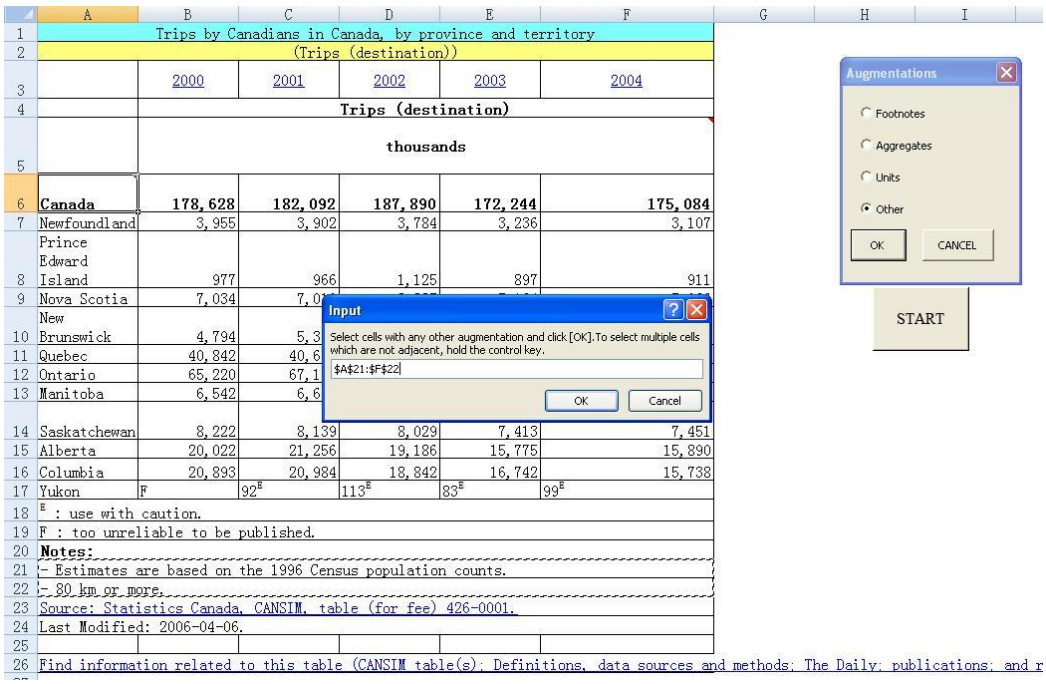
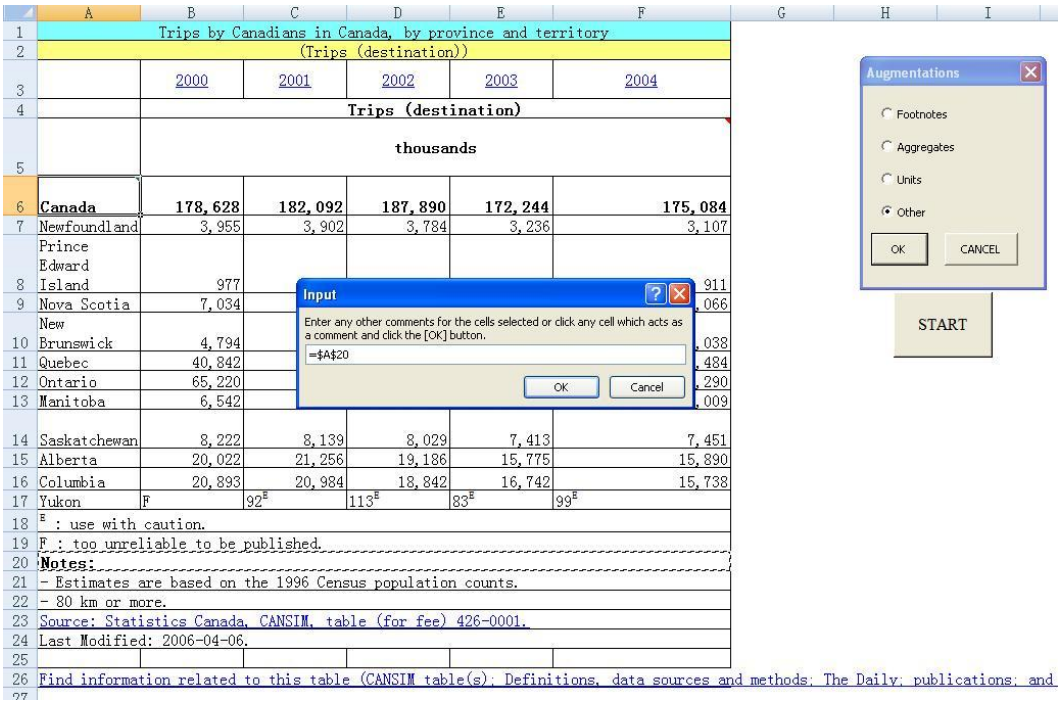


Fig. 60 Other Augmentations - 1

The user selects the cells in the rows 21 and 22 and clicks the [OK] button. The system displays another pop up to the user asking to enter any comments for these cells or select any cells which serve as comments. In the above case, these augmentations appear below the heading “Notes” in row 20, so the user selects the merged cell “A20”. The “Other” option allows the user to enter any arbitrary comments that pertain to the whole table by selecting the entire table for the domain of “Other”.



	2000	2001	2002	2003	2004
Canada	178,628	182,092	187,890	172,244	175,084
Newfoundland	3,955	3,902	3,784	3,236	3,107
Prince Edward Island	977				911
Nova Scotia	7,034				7,066
New Brunswick	4,794				7,038
Quebec	40,842				7,484
Ontario	65,220				7,290
Manitoba	6,542				7,009
Saskatchewan	8,222	8,139	8,029	7,413	7,451
Alberta	20,022	21,256	19,186	15,775	15,890
Columbia	20,893	20,984	18,842	16,742	15,738
Yukon	F	92[E]	113[E]	83[E]	99[E]

[E] : use with caution.
F : too unreliable to be published.
Notes:
- Estimates are based on the 1996 Census population counts.
- 80 km or more.
Source: Statistics Canada, CANSIM, table (for fee) 426-0001.
Last Modified: 2006-04-06.
Find information related to this table (CANSIM table(s); Definitions, data sources and methods; The Daily; publications; and

Fig. 61 Other Augmentations - 2

The text: *(Other) Notes :(/Other)* appears as a comment over the cells. It should be noted that in all the above cases, if the augmentations need to be overwritten, they have to be deleted first before adding the new augmentation (footnote, aggregate, units or other) as new augmentations are just appended to the existing ones. This can be done by right-clicking on the cell and selecting the "Delete Comment" option from the popup menu.

4. TABLE ANALYSIS:

This is the most important part of the program that actually abstracts the table. It consists of two stages: Verification (Steps 1-5) and Analysis (Steps 6-11).

4 (a) Verification:

In the first part, TAT checks if the table is well-formed or not. For this the user is first prompted to click the top-leftmost and the bottom-rightmost delta cells (which will be done automatically in future versions). The system colors the cells selected by the user in orange and does the following actions:

1. Deactivates any hyperlinks present in the cells (because of a direct copy-paste from the web page) that might cause unnecessary actions when the user clicks those cells.

2. Checks for any empty delta cell in that particular range of delta cells selected by the user. If there is an empty delta cell, the system colors the empty cell, enters "D?" in the cell and prompts the user to enter a value for that cell. If the user does not enter any value, the delta cell is given the value "D?"

3. The system performs the „Canonicalization' process, which splits all the merged cells in the table and repeats the text across the split cells. The canonicalization process is
restricted to the category and delta cells. For the above table, the process splits the merged cells with the text "Trips (destination)" & "thousands" and repeats the value/text over the entire span. The title and caption cells are not split as they were correctly assigned before performing the table analysis. (Fig. 62)

4. The next action is to form the list-rows for the column and row headers by looping through the cells above and to the left of the delta cells – categories + sub-categories. The list-row notation is a simple one- or two-dimensional array notation of the cells.

The list-rows for the current table would be

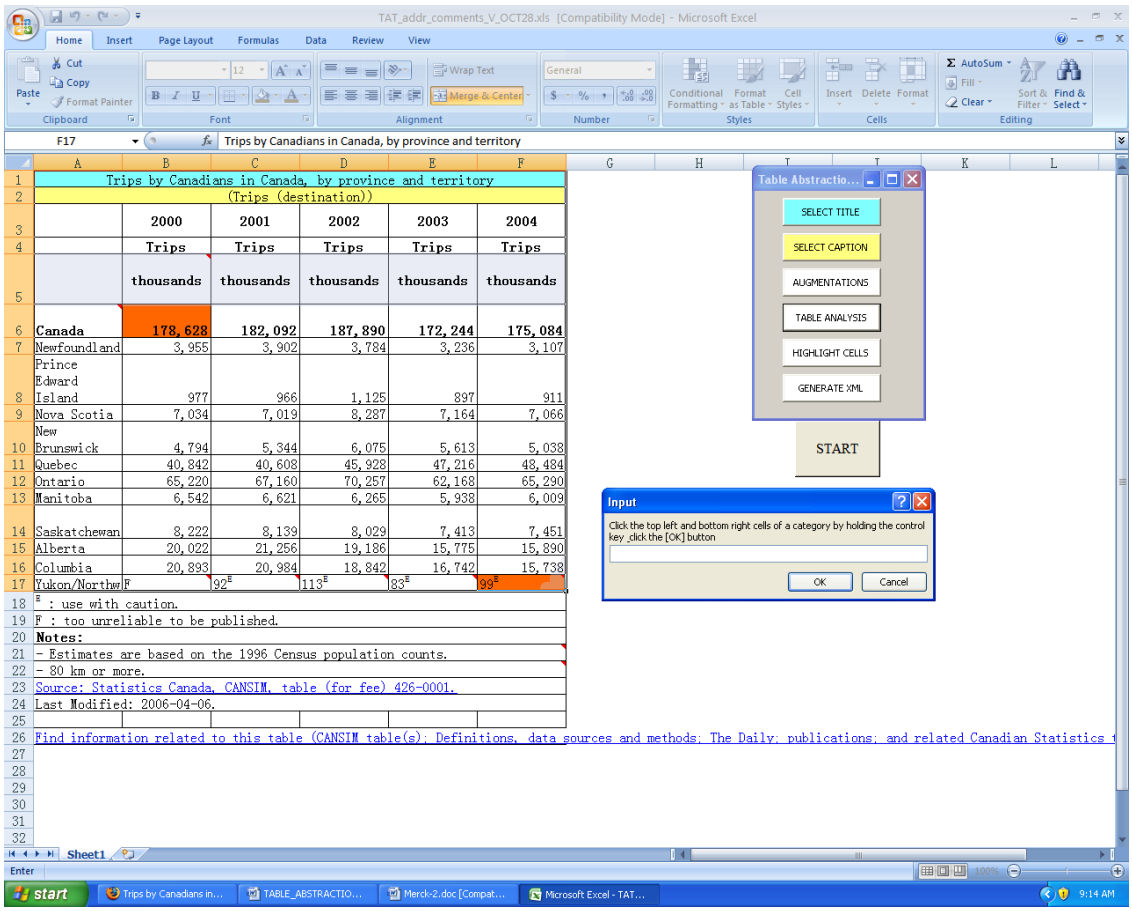
List-row notation for the column category -

Thousands − Trips (Destination) − 2000
Thousands − Trips (Destination) − 2001
Thousands − Trips (Destination) − 2002
Thousands − Trips (Destination) − 2003
Thousands − Trips (Destination) − 2004

List-row notation for the row category -

Canada
Newfoundland and Labrador
Prince Edward Island
Nova Scotia
New Brunswick
Quebec
Ontario
Manitoba
Saskatchewan
Alberta
British Columbia

Yukon Territory



	2000	2001	2002	2003	2004
	Trips	Trips	Trips	Trips	Trips
	thousands	thousands	thousands	thousands	thousands
Canada	178,628	182,092	187,890	172,244	175,084
Newfoundland	3,955	3,902	3,784	3,236	3,107
Prince Edward Island	977	966	1,125	897	911
Nova Scotia	7,034	7,019	8,287	7,164	7,066
New Brunswick	4,794	5,344	6,075	5,613	5,038
Quebec	40,842	40,608	45,928	47,216	48,484
Ontario	65,220	67,160	70,257	62,168	65,290
Manitoba	6,542	6,621	6,265	5,938	6,009
Saskatchewan	8,222	8,139	8,029	7,413	7,451
Alberta	20,022	21,256	19,186	15,775	15,890
Columbia	20,893	20,984	18,842	16,742	15,738
Yukon/Northw	F	92[E]	113[E]	83[E]	99[E]

Find information related to this table (CANSIM table(s); Definitions, data sources and methods; The Daily; publications; and related Canadian Statistics

Fig. 62 Canonicalized Table

5. The list-rows are then checked to determine if the table is well-formed or not. If the system determines that the table is not well-formed because of repetitions in its headers, it highlights them in red and gives an appropriate error message. In such a case, the user must correct the table by manipulating it with Excel operations and then start over from the beginning by clicking the [START] button. In the current example, the table is well- formed as there are no repetitions of rows in the list-row arrays. Future versions of TAT will incorporate a more robust check of the conditions for a well formed table.

4(b) Analysis:

The second stage consists of forming the indented notation for the categories in the table based on user input.

6. If the table is well-formed, the system prompts the user to select the top-leftmost and bottom-rightmost cells of a category. Depending on the location of the clicked category cells, the system determines if the category has a column-based header (header above the delta cells) or row-based header (header to the left of delta cells). The system colors the category cells selected in green. The system then forms the list-row notation for that category alone and checks for the root of the category tree. The root is repeated through all the rows of the list-row array as it is canonicalized.

7. If no root is found for the category, then a virtual header is added for the category. For the above table, since in the list row notation for categories above the delta cells, "Trips (Destination)" and "thousands" are repeated, either of the two can be the root for that category. For the category to the left of the delta cells, there is no repetition in the columns. Hence, it requires a virtual header. The virtual header is a unique string for every table processed and is "VHxxx" where xxx represents a unique number.

8. The indented notation for that category is formed in a separate sheet in the workbook. (Fig. 63). Each cell in the indented notation has a comment which refers to the address of that cell in the original table.

9. The system then prompts asking the user if there are more categories present in the table with a Yes − No message box. If the user responds Yes then steps 6-8 are repeated.

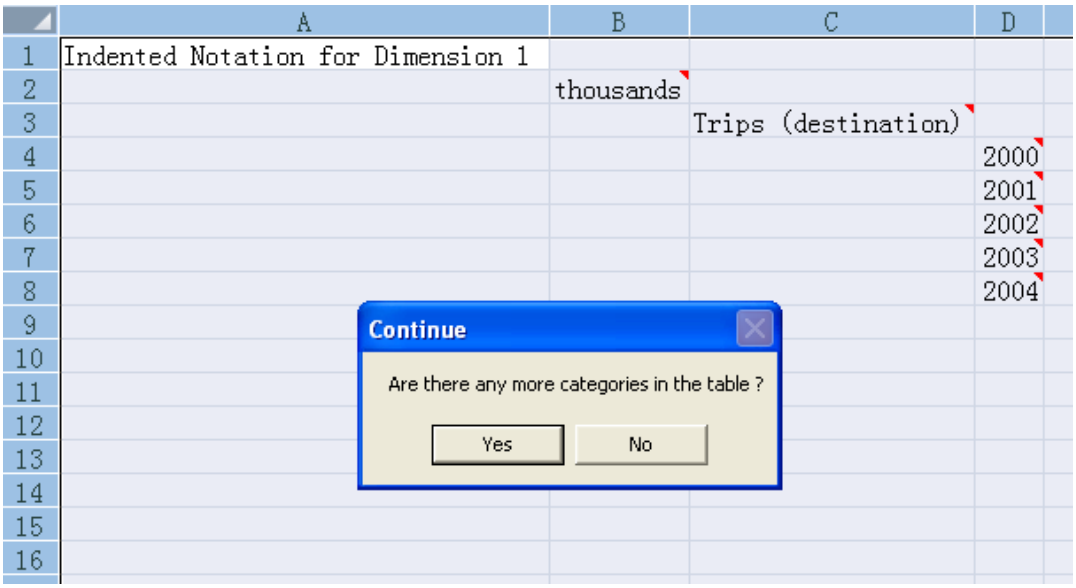


Fig. 63 Category Prompt

10. If the user clicks the No button, the system displays the table along with all the indented notation sheets arranged in tiles. The user has the chance to correct the indented notation if necessary.

11. The user can now check if the table has been interpreted correctly by looking for the category cells associated with a single delta cell and the delta cells associated with a single category or sub-category by clicking the [HIGHLIGHT CELLS] button.

5. HIGHLIGHT CELLS & INDENTED NOTATION

5 (a). Highlight Cells

After the table has been abstracted, the system displays the indented notation of the trees along with the table (Fig. 64). If the user clicks the [HIGHLIGHT CELLS] button and clicks a delta cell, the system highlights all the category cells associated with it in red. The button text changes to "STOP HIGHLIGHTING".

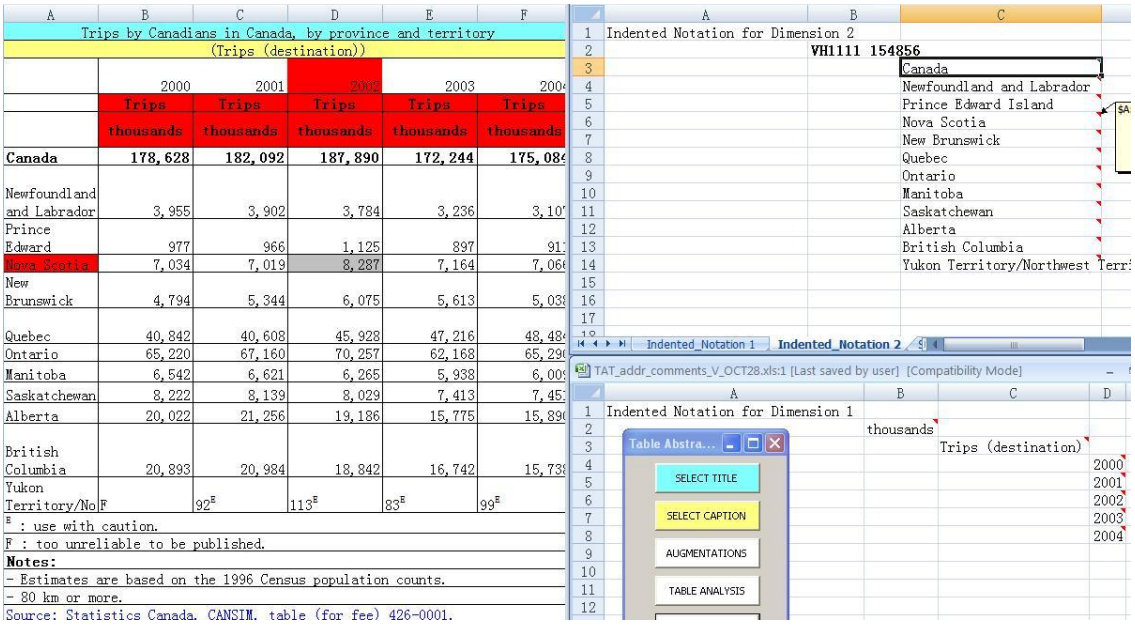


Fig. 64 TAT Highlighting Category Cells

In Figure 64, the system highlights all the category cells associated with a delta cell in red and highlights the delta cell selected itself in gray. The Indented Notations for the above table are also arranged in the form of tiles as shown in the above Figure. On clicking a category/sub-category cell, the system highlights the entire set of delta cells associated with it (Fig. 65). In Figure 65, the user selected sub-category 2001.

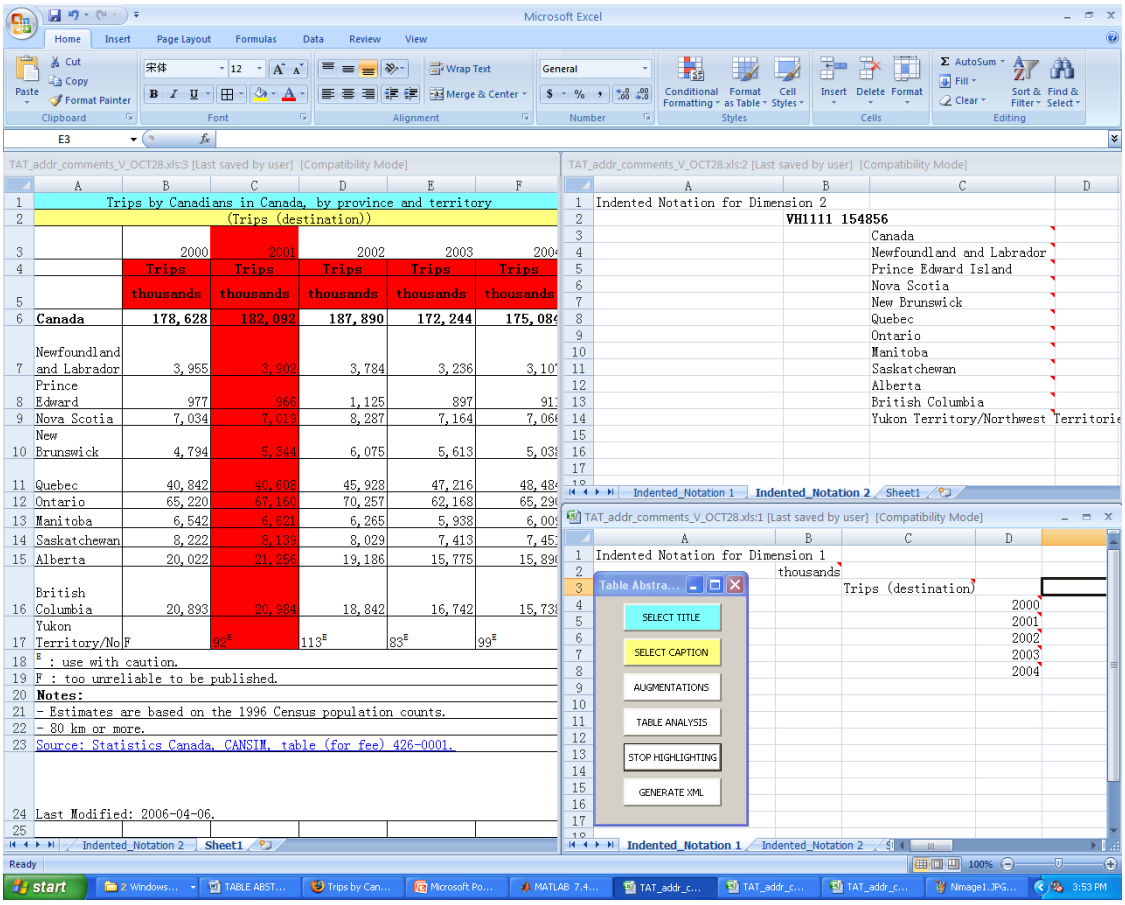


Fig. 65 TAT Highlighting Delta Cells associated with a category

The system can also highlight a specific delta cell covered by two categories. If the user clicks a column-based category/sub-category cell followed by a row-based category/sub-category cell, the system highlights a specific delta cell covered by the categories (Fig. 66). The same cell is highlighted even if the user selects the row-based category/sub-category cell first followed by the column-based category/sub-category cell.

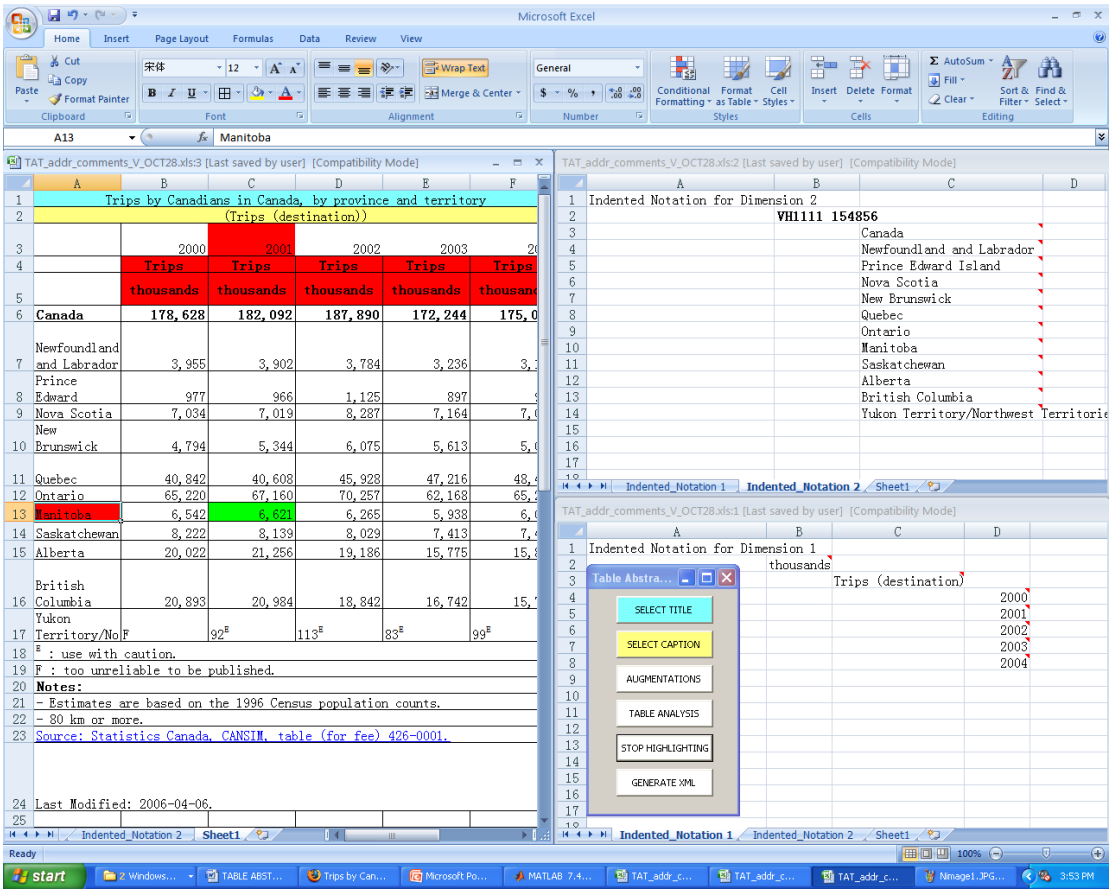


Fig. 66 TAT Highlighting one delta cell associated with two categories

In the above case, the user chose the column-based sub-category "2001" and the horizontal sub-category "Manitoba". The delta cell associated with these sub-categories is highlighted in green. To stop the highlighting process, the user should click the [STOP HIGHLIGHTING] button.

5 (b) Indented Notation

The advantage of TAT over WNT is that the corrections to the indented notation can be made by directly using Excel operations instead of MATLAB. The user can make all the changes which were allowed in WNT. The user can re-arrange the cells or delete any empty rows or columns between the cells. An important point is that the highlighting is done looking at the indented notation. So, if the user makes any changes to the contents of the indented notation with text that is not present in the table, then the system displays an error message (**Fig. 67)**.

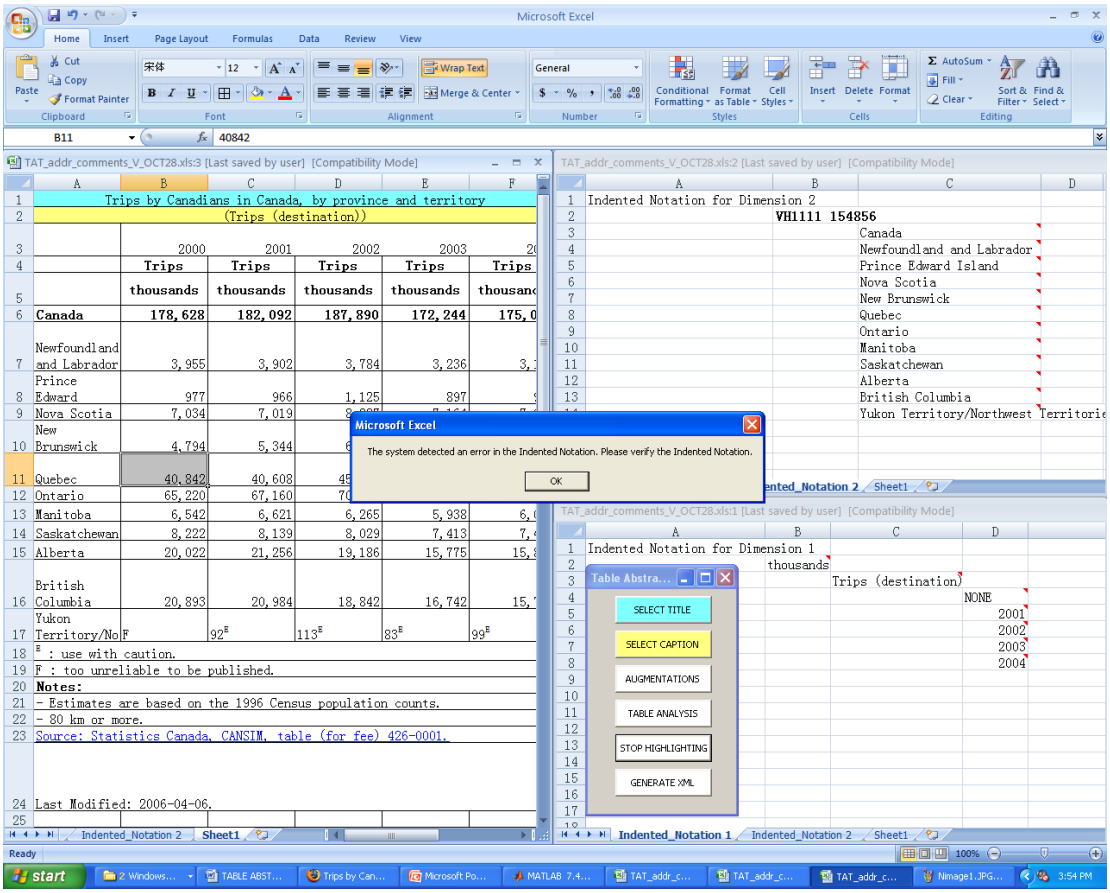


Fig. 67 TAT Indented Notation Error

In the above screenshot it can be seen in the bottom right corner i.e. Indented Notation for Dimension 1 that the year "2000" has been changed to "NONE". When a delta cell –"B11" corresponding to the subcategory "2000" is selected to view the highlighting, the system displays an appropriate error message. The error can be corrected by changing the corresponding text in the table i.e., "2000" in cell "B3" to "NONE" or by changing the text in cell "D4" in the sheet "Indented_Notation_1" to "2000" again.

6. GENERATE XML:

Clicking this button generates the XML notation for the table based on its indented notation. The XML filename contains the time stamp of the process. TAT creates the XML file in the same folder as the Excel Workbook. TAT also creates a log file with the name containing the time stamp of the process in the same folder. It records the time taken by the user for each action and the idle

time between successive actions. To change the path, please look at the createxml & Writelog procedures in the source code, which can be viewed in the Visual Basic Editor.

III. TRANSFORMATIONS ON TABLES FOR PROCESSING:

There are three most commonly occurring templates of table layouts which TAT cannot process. They are shown below:

	A1		A2	
Stub	B1	B2	B1	B2
X				
C1	XX	XX	XX	XX
C2	XX	XX	XX	XX
Y				
C1	XX	XX	XX	XX
C2	XX	XX	XX	XX

Fig. 68 Template Table 1

Template Table 1 is a 4 category table:

Category 1- Virtual header1 (root) – A1, A2

Category 2- Virtual header2 (root) – B1, B2

Category 3- Virtual header3 (root) – X, Y

Category 3- Virtual header4 (root) – C1, C2

This template cannot be processed by TAT as X and Y which are sub-categories of the same category are not contiguous and have sub-categories of another category between them.

	A1		A2	
Stub	B1	B2	B1	B2
X	XX	XX	XX	XX
C1	XX	XX	XX	XX
C2	XX	XX	XX	XX
Y	XX	XX	XX	XX
C3	XX	XX	XX	XX
C4	XX	XX	XX	XX

Fig. 69 Template Table 2

Template Table 2 is a 3 category table.

Category 1- Virtual header1 (root) : A1, A2

Category 2- Virtual header2 (root) : B1, B2

Category 3 -Virtual header3 (root) : X: X, C1, C2; Y:Y,C3,C4

Note: In category 3, X and Y are sub-categories which have their own sub-categories. Also, generally, X and Y are aggregates.

In Template Table 2, we need a virtual header for X and Y. Also, X is the parent node of C1 and C2 while Y is the parent node of C3 and C4. This relationship between the sub-categories is expressed in the form of format changes like bold and italics which are not preserved by Microsoft Excel cannot be interpreted by TAT correctly.

	A1				
Stub	B1	B2	B3	B4	
X					
C1	XX	XX	XX	XX	
C2	XX	XX	XX	XX	Table 1
Y					
C3	XX	XX	XX	XX	
C4	XX	XX	XX	XX	Table 2

Fig. 70 Template Table 3

Template Table 3 is actually a concatenation of two tables. This is also a very commonly found template for tables where related tables are concatenated for presenting a complete picture. However, logically these tables should be defined as two (or more) separate tables.

In order to make these tables "TAT-friendly", a few transformations on the tables need to be performed. These transformations preserve the logical structure of the table. For transformations on real tables, please refer to section 2.8.2 in the thesis.

Appendix B: Sample Log File

Figure 71 shows a sample log file.

Action	Start Time	End Time	Time Elapsed (sec)	Idle Time (sec)
Pre-Process Table	5:46:44 PM	5:48:02 PM	78	0
Title	5:48:02 PM	5:48:04 PM	2	0
Caption	5:48:04 PM	5:48:06 PM	2	0
Augmentation	5:48:06 PM	5:48:09 PM	3	0
Units	5:48:09 PM	5:48:10 PM	1	3
Cells with Augmentation reference	5:48:13 PM	5:48:17 PM	4	0
Cells with Augmentation citation	5:48:17 PM	5:48:18 PM	1	5
Aggregate selection	5:48:23 PM	5:48:24 PM	1	0
Aggregate cells selection	5:48:24 PM	5:48:28 PM	4	3
Delta Cells selection + Check for Well Defined table	5:48:31 PM	5:48:33 PM	2	3
Category- 1	5:48:36 PM	5:48:37 PM	1	3
Category- 2	5:48:40 PM	5:48:46 PM	6	2
Category- 3	5:48:48 PM	5:48:50 PM	2	1
Highlighting	5:48:51 PM	5:48:54 PM	3	0
Generate XML	5:48:54 PM	5:48:55 PM	1	

Fig. 71 Sample Log File

Figure 71 shows an example log file. Each row in the table corresponds to a particular action described in the first column. The ‘Start Time’ for the action is displayed in the second column. The third column displays the ‘End Time’ for the action. The difference between End Time and Start Time is the total time for the action indicated by ‘Time Elapsed (sec)’ in the fourth column. ‘Idle Time (sec)’ is the time between two consecutive actions when the user does not interact with the system. This is interpreted as the preparation time for the next action to be performed by the user. For example, the ninth entry in the log indicates that after selecting the aggregate cells in the table, the user was idle for about three seconds before selecting the delta cells. The original display

format for Start Time and End Time is *mm/dd/yyyy hh:mm*, which has been modified in the Figure to show that the calculation of Time Elapsed and Idle time is indeed correct.

Excel stores all dates as integers and all times as decimal fractions. With this system, Excel can add, subtract, or compare dates and times just like any other numbers, and all dates are manipulated by using this system. In this system, the serial number 1 represents 1/1/1900 12:00:00 a.m. Times are stored as decimal numbers between .0 and .99999, where .0 is 00:00:00 and .99999 is 23:59:59. The date integers and time decimal fractions can be combined to create numbers that have a decimal and an integer portion. For example, the number 39883.740787037 corresponds to the date 3/11/2009 17:46 which is the entry for Start Time in the table for the Pre-Process Table action. It is modified and displayed as 5:46:44 PM. Similarly, for the same action, the End Time 3/11/2009 17:48 is displayed as 5:48:02 PM and it corresponds to 39883.7416898148. To obtain the Time Elapsed in seconds, the decimal point value of these numbers is multiplied by 86400 and the difference between these values is computed (86400 x (0.7416898148 - 0.740787037) = 78).

Appendix C: List of Table URLs

The following table presents the list of URLs of tables used in the experiment

Table 15 - List of Table URLs

Sno.	Table	Table URL
1	Average incarceration sentence length imposed, by offense, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
2	Persons entitled to vote, votes cast and percentage voter turnout. Storting elections 1945-2005	http://www.ssb.no/stortingsvalg_en/tab-2005-10-27-01-en.html
3	World Agriculture	http://web.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/0,,contentMDK:20390020~menuPK:1192714~pagePK:64133150~piPK:64133175~theSitePK:239419~isCURL:Y,00.html
4	Marriages by province and territory	http://www40.statcan.ca/l01/cst01/famil04-eng.htm
5	Parole or supervised release terminating with a new crime or technical violation, by original offense, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
6	Capacity Utilization of Coal Mines by State, 2007, 2006	http://www.eia.doe.gov/cneaf/coal/page/acr/table12.html
7	Deaths by transport accidents. 2006	http://www.ssb.no/dodsarsak_en/tab-2008-06-27-16-en.html
8	Statistical Project Database Essential Public Health Functions	http://ddp-ext.worldbank.org/ext/CSIDB/getProjectInfoXML?id=90993&format=project
9	Largest Seaports in the world	http://www.geohive.com/charts/ec_seaport1.aspx
10	Global Economy: Exports	http://www.geohive.com/charts/ec_exim1.aspx
11	Under-Five Mortality Rate	http://www1.lanic.utexas.edu/la/region/aid/aid98/health/tab2.html
12	Table 3. Reported Voting and Registration, by Race, Hispanic Origin, and Age, for the United States, Regions, and Divisions: November 2006	http://www.census.gov/population/www/socdemo/voting/cps2006.html

13	U.S. Coal Production by Coal-Producing Region and State, 2006-2007	http://www.eia.doe.gov/cneaf/coal/page/acr/tables2.html
14	Probation supervisions terminating with a new crime or technical violation, by offender characteristics, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
15	World : Barley Production	http://www.geohive.com/earth/ag_barley.aspx
16	Court, youth cases by decision, by province and territory	http://www40.statcan.ca/l01/cst01/legal24d-eng.htm
17	Table S3: State Homicide Victimization Rates, Ages 14-17	http://www.ojp.usdoj.gov/bjs/dtdata.htm#State
18	Middle East and North Africa: World Bank Commitments, Disbursements, and Net Transfers l Fiscal 2003–2008	http://web.worldbank.org/WBSITE/EXTERNAL/EXTABOUTUS/EXTANNREP/EXTANNREP2K8/0,,contentMDK:21919785~menuPK:5405451~pagePK:64168445~piPK:64168309~theSitePK:5164354,00.html
19	Producer price index, services	http://www40.statcan.ca/l01/cst01/econ145f-eng.htm
20	Underground Coal Production by State and Mining Method, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table3.html
21	Infant Mortality Rate - Latin America	http://www1.lanic.utexas.edu/la/region/aid/aid98/health/tab1.html
22	Storting Election 2005. Valid votes, seats and mean valid votes per seats. Storting elections 1989-2005	http://www.ssb.no/stortingsvalg_en/tab-2005-10-27-21-en.html
23	top20 cities	http://www.geohive.com/charts/cy_notagg.aspx
24	Poultry	http://www40.statcan.ca/l01/cst01/prim55d-eng.htm
25	Recoverable Coal Reserves and Average Recovery Percentage at Producing U.S. Mines by Mine Production Range and Mine Type, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table17.html
26	Live births and late foetal deaths. 1951-2007	http://www.ssb.no/fodte_en/tab-2008-04-09-01-en.html
27	Deaths of underlying cause of death, by place of death. Per cent. 20061 (New table 24 September 2008)	http://www.ssb.no/dodsarsak_en/tab-2008-06-27-19-en.html

28	Reported Voting and Registration, by Race, Hispanic Origin, Sex, and Age, for the United States: November 2006	http://www.census.gov/population/www/socdemo/voting/cps2006.html
29	Forest fires and forest land burned, by province and territory	http://www40.statcan.ca/l01/cst01/envir02a-eng.htm
30	cntyfed.xls	http://www.ojp.usdoj.gov/bjs/dtdata.htm#County
31	Virus- Disease	http://www.cdc.gov/ncidod/dvrd/spb/mnpages/dispages/arena.htm
32	Europe and Central Asia: World Bank Commitments, Disbursements, and Net Transfers l Fiscal 2003–2008	http://web.worldbank.org/WBSITE/EXTERNAL/EXTABOUTUS/EXTANNREP/EXTANNREP2K8/0,,contentMDK:21919785~menuPK:5405451~pagePK:64168445~piPK:64168309~theSitePK:5164354,00.html
33	Coal Production and Coalbed Thickness by Major Coalbeds and Mine Type, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table5.html
34	Private and public capital expenditures	http://www40.statcan.ca/l01/cst01/econ17-eng.htm
35	Reported Voting and Registration, by Sex and Single Years of Age: November 2006	http://www.census.gov/population/www/socdemo/voting/cps2006.html
36	Local governments, financial assets and liabilities, by province and territory	http://www40.statcan.ca/l01/cst01/govt40a-eng.htm
37	Trips by Canadians in Canada, by province and territory.mht	http://www40.statcan.ca/l01/cst01/arts26a-eng.htm
38	Deaths by accidents. 2006 1	http://www.ssb.no/dodsarsak_en/tab-2008-06-27-15-en.html
39	Global internet and PC use	http://www.geohive.com/earth/ec_inet.aspx
40	Poverty Indicators: El Salvador	http://www1.lanic.utexas.edu/la/region/aid/aid98/poverty/tab6.html
41	Poverty Indicators: Honduras	http://www1.lanic.utexas.edu/la/region/aid/aid98/poverty/tab6.html
42	Poverty Indicators: Peru	http://www1.lanic.utexas.edu/la/region/aid/aid98/poverty/tab6.html
43	Restaurants. Sales1. Period. Million kroner	http://www.ssb.no/servering_en/tab-2001-10-08-01-en.html
44	Infant deaths. County. 1991-2006	http://www.ssb.no/dodsarsak_en/tab-2008-06-27-05-en.html

45	Reported Voting and Registration, by Race, Hispanic Origin, Duration of Residence, and Tenure: November 2006	http://www.census.gov/population/www/socdemo/voting/cps2006.html
46	Data Collection	Query Result from- www.worldbank.org
47	Adopters and non-adopters of e-procurement in Singapore: An empirical study … Factor Analysis	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VC4-4V17CW7-1&_user=659639&_coverDate=10%2F31%2F2009&_alid=879602068&_rdoc=2&_fmt=high&_orig=search&_cdi=5944&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=63b5e4dddf7b939456dc26b0342fc9a0
48	Correlation Matrix	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VC4-4V17CW7-1&_user=659639&_coverDate=10%2F31%2F2009&_alid=879602068&_rdoc=2&_fmt=high&_orig=search&_cdi=5944&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=63b5e4dddf7b939456dc26b0342fc9a0
49	Demographic profile of the respondent companies	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VC4-4V17CW7-1&_user=659639&_coverDate=10%2F31%2F2009&_alid=879602068&_rdoc=2&_fmt=high&_orig=search&_cdi=5944&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=63b5e4dddf7b939456dc26b0342fc9a0
50	Characteristics of e-procurement activities	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VC4-4V17CW7-1&_user=659639&_coverDate=10%2F31%2F2009&_alid=879602068&_rdoc=2&_fmt=high&_orig=search&_cdi=5944&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=63b5e4dddf7b939456dc26b0342fc9a0

51	Logistic regression analysis	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VC4-4V17CW7-1&_user=659639&_coverDate=10%2F31%2F2009&_alid=879602068&_rdoc=2&_fmt=high&_orig=search&_cdi=5944&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=63b5e4dddf7b939456dc26b0342fc9a0
52	PLS analysis for factors associated with extent of usage on e-procurement	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VC4-4V17CW7-1&_user=659639&_coverDate=10%2F31%2F2009&_alid=879602068&_rdoc=2&_fmt=high&_orig=search&_cdi=5944&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=63b5e4dddf7b939456dc26b0342fc9a0
53	Legal Strategic Framework	http://ddp-ext.worldbank.org/ext/CSIDB/getCountryByCategory
54	Demographic status of the world	http://www.geohive.com/earth/world.aspx
55	Africa: World Bank Commitments, Disbursements, and Net Transfers l Fiscal 2003–2008	http://web.worldbank.org/WBSITE/EXTERNAL/EXTABOUTUS/EXTANNREP/EXTANNREP2K8/0,,contentMDK:21919785~menuPK:5405451~pagePK:64168445~piPK:64168309~theSitePK:5164354,00.html
56	Milestones	http://www.geohive.com/earth/his_history2.aspx
57	Poverty indicators - Latin America	http://www1.lanic.utexas.edu/la/region/aid/aid98/poverty/tab3.html
58	Education Price Index	http://www40.statcan.ca/l01/cst01/educ46-eng.htm
59	East Asia and Pacific: World Bank Commitments, Disbursements, and Net Transfers l Fiscal 2003–2008	http://web.worldbank.org/WBSITE/EXTERNAL/EXTABOUTUS/EXTANNREP/EXTANNREP2K8/0,,contentMDK:21919785~menuPK:5405451~pagePK:64168445~piPK:64168309~theSitePK:5164354,00.html
60	Table 7. Coal Production by State, Mine Type, and Union Status, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table7.html

61	Coal Disposition by State	http://www.eia.doe.gov/cneaf/coal/page/acr/table8.html
62	Population Youngest 2000	http://www.geohive.com/charts/pop_age.aspx
63	Population: oldest 2000	http://www.geohive.com/charts/pop_age.aspx
64	Population Youngest 2050	http://www.geohive.com/charts/pop_age.aspx
65	Population: oldest 2050	http://www.geohive.com/charts/pop_age.aspx
66	Probation supervisions terminating with a new crime, by original offense, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
67	The 50 largest (population) countries in the world	http://www.geohive.com/earth/pop_top50.aspx
68	Components of population growth, by province and territory	http://www40.statcan.ca/l01/cst01/demo33a-eng.htm
69	Historic, current and future population of Europe	http://www.geohive.com/earth/his_proj_europe.aspx
70	Millenium Development Goals	http://ddp-ext.worldbank.org/ext/ddpreports/ViewSharedReport?&CF=1&REPORT_ID=1336&REQUEST_TYPE=VIEWADVANCED&HF=N
71	Freshwater Resources	http://www1.lanic.utexas.edu/la/region/aid/aid98/environment/tab10.html
72	Storting Election 2005. Elected representatives, by party/electoral list, sex and county	http://www.ssb.no/stortingsvalg_en/tab-2005-10-27-05-en.html
73	Global Economy: Exports	http://www.geohive.com/charts/ec_exim1.aspx
74	Food Security	http://www1.lanic.utexas.edu/la/region/aid/aid98/environment/tab8.html
75	World bank-1	http://web.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/0,,contentMDK:20535285~menuPK:1192694~pagePK:64133150~piPK:64133175~theSitePK:239419,00.html
76	Threatened Species - 1996	http://www1.lanic.utexas.edu/la/region/aid/aid98/environment/tab2.html
77	Total fertility rate1, by county. 1968-2007	http://www.ssb.no/fodte_en/tab-2008-04-09-05-en.html

78	Probation supervisions terminating with a new crime, by offender characteristics, 1993	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6W87-4TX186H-1&_user=659639&_coverDate=07%2F31%2F2009&_alid=879602068&_rdoc=28&_fmt=high&_orig=search&_cdi=6647&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=baa17926612607da4ba267c69b44aa27
79	Study locations (abbreviation), number of streams, and number of sites (30-m reaches) across permanence categories (E: ephemeral, I: intermittent, and P: perennial) within each location.	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6W87-4TX186H-1&_user=659639&_coverDate=07%2F31%2F2009&_alid=879602068&_rdoc=28&_fmt=high&_orig=search&_cdi=6647&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=baa17926612607da4ba267c69b44aa27
80	Significant ($p < 0.05$) Pearson correlations between environmental variables and nonmetric multi-dimensional scaling axes for species- and family-level ordinations.	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6W87-4TX186H-1&_user=659639&_coverDate=07%2F31%2F2009&_alid=879602068&_rdoc=28&_fmt=high&_orig=search&_cdi=6647&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=baa17926612607da4ba267c69b44aa27
81	Frequencies of phyla, reproductive and colonial growth forms and literature-derived moisture associations for bryophytes among hydrologic permanence classes.	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6W87-4TX186H-1&_user=659639&_coverDate=07%2F31%2F2009&_alid=879602068&_rdoc=28&_fmt=high&_orig=search&_cdi=6647&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=baa17926612607da4ba267c69b44aa27

82	ANOVA results comparing species richness across forests, streams (nested within forests), and permanence classes.	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6W87-4TX186H-1&_user=659639&_coverDate=07%2F31%2F2009&_alid=879602068&_rdoc=28&_fmt=high&_orig=search&_cdi=6647&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=baa17926612607da4ba267c69b44aa27
83	Reported crime in Alabama	http://bjsdata.ojp.usdoj.gov/dataonline/Search/Crime/State/statebystaterun.cfm?stateid=1
84	Population by year, by province and territory	http://www40.statcan.ca/l01/cst01/demo02a-eng.htm
85	Europe and Central Asia - Data And Statistics	http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/ECAEXT/0,,menuPK:258604~pagePK:158889~piPK:146815~theSitePK:258599,00.html
86	State Estimated Homicide Offending Rates, Ages 14-24	http://www.ojp.usdoj.gov/bjs/dtdata.htm#State
87	Value of building permits, by province and territory (monthly)	http://www40.statcan.ca/l01/cst01/econ67a-eng.htm
88	Parole or supervised release terminating with a new crime or technical violation, by offender characteristics, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
89	Retail trade, operating statistics, by province and territory	http://www40.statcan.ca/l01/cst01/trad38d-eng.htm
90	Recoverable Coal Reserves	http://www.eia.doe.gov/cneaf/coal/page/acr/table17.html
91	Conviction rate, by most serious offense charged, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
92	Dependencies with their governing/controlling country	http://www.geohive.com/earth/gen_dependencies.aspx
93	U.S. Direct-Investment Position in LAC	http://www1.lanic.utexas.edu/la/region/aid/aid98/investment/tab1.html
94	Coal Production by Coalbed Thickness and Mine Type, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table4.html
95	Recoverable Coal Reserves and Average Recovery Percentage at Producing Underground Coal Mines by State and Mining Method, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table16.html

96	Turks and Caicos Islands	http://www.geohive.com/cntry/turkscaicos.aspx
97	Military personnel and pay	http://www40.statcan.ca/l01/cst01/govt16a-eng.htm
98	Sheep inventories, by province	http://www40.statcan.ca/l01/cst01/prim52d-eng.htm
99	Net farm income, by province	http://www40.statcan.ca/l01/cst01/agri02c-eng.htm
100	Deaths by transport accidents. 2007	http://www.ssb.no/english/subjects/03/01/10/dodsarsak_en/tab-2009-04-07-16-en.html
101	Table D-2.2. Behavior of defendants released prior to trial, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
102	LD09_CumLendingbyCountry.xls	siteresources.worldbank.org/EXTANNREP2K8/Resources/LD09_CumLendingbyCountry.xls
103	Rough set based hybrid algorithm for text classification table 1	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V03-4V74VB4-H&_user=659639&_coverDate=07%2F31%2F2009&_alid=879602068&_rdoc=44&_fmt=high&_orig=search&_cdi=5635&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=fd8297a23ad3df7efc8454db4d6aafa9
104	Rough set based hybrid algorithm for text classification table 2	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V03-4V74VB4-H&_user=659639&_coverDate=07%2F31%2F2009&_alid=879602068&_rdoc=44&_fmt=high&_orig=search&_cdi=5635&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=fd8297a23ad3df7efc8454db4d6aafa9
105	Rough set based hybrid algorithm for text classification table 3	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V03-4V74VB4-H&_user=659639&_coverDate=07%2F31%2F2009&_alid=879602068&_rdoc=44&_fmt=high&_orig=search&_cdi=5635&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=fd8297a23ad3df7efc8454db4d6aafa9

106	Rough set based hybrid algorithm for text classification table 4	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V03-4V74VB4-H&_user=659639&_coverDate=07%2F31%2F2009&_alid=879602068&_rdoc=44&_fmt=high&_orig=search&_cdi=5635&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=fd8297a23ad3df7efc8454db4d6aafa9
107	Rough set based hybrid algorithm for text classification table 5	http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6V03-4V74VB4-H&_user=659639&_coverDate=07%2F31%2F2009&_alid=879602068&_rdoc=44&_fmt=high&_orig=search&_cdi=5635&_sort=d&_docanchor=&view=c&_ct=250268&_acct=C000035878&_version=1&_urlVersion=0&_userid=659639&md5=fd8297a23ad3df7efc8454db4d6aafa9
108	Waste disposal, by source, by province	http://www40.statcan.ca/l01/cst01/envir25a-eng.htm
109	Coal Mining Productivity by State, Mine Type, and Union Status, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table24.html
110	Computer systems design and related services, by province	http://www40.statcan.ca/l01/cst01/serv15d-eng.htm
111	lwenfemp.xls	http://www.ojp.usdoj.gov/bjs/dtdata.htm#justice
112	Average Number of Employees by State and Mine Type, 2007, 2006	http://www.eia.doe.gov/cneaf/coal/page/acr/table18.html
113	Convicted offenders, by most serious offense charged, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
114	U.S. Coal Supply, Disposition, and Prices, 2006-2007	http://www.eia.doe.gov/cneaf/coal/page/acr/tables1.html
115	State Homicide Victimization Rates, Ages 14-24	http://www.ojp.usdoj.gov/bjs/dtdata.htm#State
116	Productive Capacity and Capacity Utilization of Underground Coal Mines by State and Mining Method, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table13.html
117	Probation rate, by offense, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal

118	Recoverable Coal Reserves at Producing Mines, Estimated Recoverable Reserves, and Demonstrated Reserve Base by Mining Method, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table15.html
119	World Maize Production	http://www.geohive.com/earth/ag_maize.aspx
120	Gender and Education	http://www1.lanic.utexas.edu/la/region/aid/aid98/education/tab4.html
121	U.S. Coal Production, 2002-2008	http://www.eia.doe.gov/cneaf/coal/quarterly/html/t1p01p1.html
122	Coal Consumers in the Manufacturing and Coke Sectors, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table25.html
123	Suicide by method. 1976-2006	http://www.ssb.no/dodsarsak_en/tab-2008-06-27-09-en.html
124	Yemen Administrative units	http://www.geohive.com/cntry/yemen.aspx
125	Main cities	http://www.geohive.com/cntry/yemen.aspx
126	Deaths by sex, age and underlying cause of death. The whole country. 2006	http://www.ssb.no/dodsarsak_en/tab-2008-06-27-03-en.html
127	Profile of Education	http://lanic.utexas.edu/la/region/aid/aid98/education/tab5.html
128	High income Mill Dev Goals.xls	http://ddp-ext.worldbank.org/ext/ddpreports/ViewSharedReport?&CF=1&REPORT_ID=1336&REQUEST_TYPE=VIEWADVANCED&HF=N
129	Confinements of single and multiple births, by sex1. 1991-2007	http://www.ssb.no/fodte_en/tab-2008-04-09-02-en.html
130	Reported crime in Illinois	http://bjsdata.ojp.usdoj.gov/dataonline/Search/Crime/State/statebystaterun.cfm?stateid=14
131	Household waste by county. Percentage, waste sent for recovery, including energy recovery. 1995-2007	http://www.ssb.no/avfkomm_en/tab-2008-06-20-01-en.html
132	Average Sales Price of U.S. Coal by State and Disposition, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table33.html
133	National Protection Systems - 1997	http://www1.lanic.utexas.edu/la/region/aid/aid98/environment/tab1.html
134	Charitable Donors	http://www40.statcan.ca/l01/cst01/famil90-eng.htm

135	Incidence of Poverty	http://www1.lanic.utexas.edu/la/region/aid/aid98/poverty/tab4.html
136	ciies.csv	http://www.ojp.usdoj.gov/bjs/dtdata.htm#City
137	Percentage distribution of valid votes, by party/electoral list. Redistribution of votes cast for joint lists. Storting elections 1957-2005	http://www.ssb.no/stortingsvalg_en/tab-2005-10-27-02-en.html
138	Latin America and the Caribbean Selected Economic and Social Data	http://www1.lanic.utexas.edu/la/region/aid/aid98/poverty/tab2.html
139	Total Fertility rate	http://www1.lanic.utexas.edu/la/region/aid/aid98/health/tab4.html
140	LD03c_Commitments_SAR.xls	Query Result from- www.worldbank.org
141	Coal Production and Number of Mines by State and Mine Type, 2007-2006	http://www.eia.doe.gov/cneaf/coal/page/acr/table1.html
142	Electric Power Sector Net Generation, 2006-2007	http://www.eia.doe.gov/cneaf/coal/page/acr/tables3.html
143	Reported Voting and Registration, by Race, Hispanic Origin, Sex, and Age, for the United States: November 2006	http://www.census.gov/population/www/socdemo/voting.html
144	Characteristics of convicted offenders, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
145	Storting Election 2005. Changes in percentages of valid votes for parties/electoral lists from 2001-2005, by county. Percentage points	http://www.ssb.no/stortingsvalg_en/tab-2005-10-27-03-en.html
146	Age Structure	http://www.geohive.com/charts/pop_agestruc.aspx
147	Coal Disposition by State, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table8.html
148	Carbon Dioxide Emissions	http://www1.lanic.utexas.edu/la/region/aid/aid98/environment/tab13.html
149	Regional population from 1750 to 2050	http://www.geohive.com/earth/his_history1.aspx
150	Storting Election 2005.Elected representatives by party/electoral list and sex. Storting elections 1945-2005	http://www.ssb.no/stortingsvalg_en/tab-2005-10-27-06-en.html
151	Parole or supervised release terminating with a new crime, by	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal

	original offense, 1993	
152	Projects - Algeria Second Rural Employment Project.mht	http://web.worldbank.org/external/projects/main?pagePK=64283627&piPK=73230&theSitePK=40941&menuPK=228424&Projectid=P076784
153	Freedom Ratings	http://www1.lanic.utexas.edu/la/region/aid/aid98/democracy/tab1.html
154	Probation supervisions terminating with a new crime or technical violation, by original offense, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
155	GeoHive 50 largest (area) countries in the world.mht	http://www.geohive.com/earth/area_top50.aspx
156	Major U.S. Coal Producers, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table10.html
157	Storting Election 2005. Elected representatives, by party /electoral list, sex and county	http://www.ssb.no/stortingsvalg_en/tab-2005-10-27-05-en.html
158	Reported Voting and Registration, by Veteran Status and Race and Hispanic Origin: November 2006.	http://www.census.gov/population/www/socdemo/voting.html
159	Percentage of inhabitants living in municipalities offering collection at home of sorted waste fractions, by county. 2002-2007 (Corrected 20 June 2008 at 1130)	http://www.ssb.no/avfkomm_en/tab-2008-06-20-04-en.html
160	Niue	http://www.geohive.com/cntry/niue.aspx
161	Reported Voting and Registration, by Race, Hispanic Origin, Sex, and Age, for the United States: November 2006	http://www.census.gov/population/www/socdemo/voting.html
162	Emissions to air of NOX, SO2, NH3 and NMVOC. 1973-2007*	http://www.ssb.no/agassn_en/tab-2009-02-09-04-en.html
163	Infant mortality rates, by province and territory	http://www40.statcan.ca/l01/cst01/health21a-eng.htm
164	Production of eggs, by province	http://www40.statcan.ca/l01/cst01/prim53d-eng.htm
165	State Estimated Homicide Offending Rates, Ages 14-17	http://www.ojp.usdoj.gov/bjs/dtdata.htm#State

166	Public Investment in Education	http://www1.lanic.utexas.edu/la/region/aid/aid98/education/tab2.html
167	Projects Approved for IBRD and IDA Assistance by Region and Country l Fiscal 2008	http://web.worldbank.org/WBSITE/EXTERNAL/EXTABOUTUS/EXTANNREP/EXTANNREP2K8/0,,contentMDK:21919785~menuPK:5405451~pagePK:64168445~piPK:64168309~theSitePK:5164354,00.html
168	coe_table2_1	http://nces.ed.gov/programs/coe/2007/section1/table.asp?tableID=662
169	Placement examinations coe_table3	http://nces.ed.gov/programs/coe/2007/analysis/sa_table.asp?tableID=853
170	coe_table3_2	http://nces.ed.gov/programs/coe/2005/section1/table.asp?tableID=228
171	coe_table5_1	http://nces.ed.gov/programs/coe/2007/section1/table.asp?tableID=667
172	coe_table5_2	http://nces.ed.gov/programs/coe/2007/section1/table.asp?tableID=669
173	coe_table_8_1	http://nces.ed.gov/programs/coe/2008/section1/table.asp?tableID=867
174	coe_table_17_1	http://nces.ed.gov/programs/coe/2004/section3/table.asp?tableID=58
175	coe_table_18_1	http://nces.ed.gov/programs/coe/2007/section2/table.asp?tableID=692
176	coe_table_21_1	http://nces.ed.gov/programs/coe/2007/section3/table.asp?tableID=697
177	coe_table_31_1	http://nces.ed.gov/programs/coe/2007/section4/table.asp?tableID=717
178	coe_table_36_1	http://nces.ed.gov/programs/coe/2008/section4/table.asp?tableID=930
179	coe_table_37_3	http://nces.ed.gov/programs/coe/2008/section4/table.asp?tableID=933
180	coe_table_47_1	http://nces.ed.gov/programs/coe/2007/section5/table.asp?tableID=749
181	coe_table_47_2	http://nces.ed.gov/programs/coe/2007/section5/table.asp?tableID=750
182	coe_table_s8	http://nces.ed.gov/programs/coe/2004/section2/table.asp?tableID=124
183	coe_table_s8_1	http://nces.ed.gov/programs/coe/2004/section2/table.asp?tableID=158

184	coe_table_s18_2	http://nces.ed.gov/programs/coe/2004/section3/table.asp?tableID=180
185	coe_table_s21_1	http://nces.ed.gov/programs/coe/2007/section3/table.asp?tableID=787
186	coe_table_S23.xls	http://nces.ed.gov/programs/coe/2006/section3/table.asp?tableID=637
187	coe_table_S24_3.xls	http://nces.ed.gov/programs/coe/2008/section3/table.asp?tableID=979
188	das_table_1	nces.ed.gov/das/library/tables_listings/tableXLS.asp?tableID=3661
189	das_table_12	http://nces.ed.gov/das/epubs/2007165/showTable2007.asp?rt=p&tableID=3683&b=tables_figures.asp%23p1
190	affil_2004_30	http://nces.ed.gov/Surveys/SASS/tables_affil.asp
191	Table D-5-7. Parole or supervised release terminating with a new crime, by offender characteristics, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
192	Incarceration rate, by offense, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
193	Coal Production and Number of Mines by State and Coal Rank, 2007	http://www.eia.doe.gov/cneaf/coal/page/acr/table6.html
194	SUDAN - Administrative units	http://www.geohive.com/cntry/sudan.aspx
195	Sudan - Main cities	http://www.geohive.com/cntry/sudan.aspx
196	Engineering Services	http://www40.statcan.ca/l01/cst01/serv12d-eng.htm
197	Reported Voting and Registration, by Race, Hispanic Origin, and Age, for the United States, Regions, and Divisions: November 2006	http://www.census.gov/population/www/socdemo/voting.html
198	Major US Coal Mines	http://www.eia.doe.gov/cneaf/coal/page/acr/table9.html
199	Table D-4.2. Incarceration rate by offender characteristics, 1993	http://www.ojp.usdoj.gov/bjs/dtdata.htm#Federal
200	Albania Stats	http://web.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/0,,contentMDK:20535285~menuPK:1192694~pagePK:64133150~piPK:64133175~theSitePK:239419,00.html