EXTRACTING DATA CUBES FROM MULTIDIMENSIONAL TABLES ON THE WEB

by

Norah Alrayes

B.Sc. (Hons., Computer and Information Sciences)
King Saud University, Saudi Arabia, 2005

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in the
School of Computing Science

SIMON FRASER UNIVERSITY

Spring 2011

Approval

Name: **Norah Alrayes**

Degree: **Master of Science**

Title of Thesis: **Extracting Data Cubes from Multidimensional Tables on the Web**

Examining Committee:

Chair: **Dr. Oliver Schulte**
Associate Professor

Dr. Wo-Shun Luk
Senior Supervisor
Professor

Dr. Jian Pei
Supervisor
Associate Professor

Dr. Qianping Gu
Internal Examiner
Professor

Date Defended/Approved: March 09, 2011

Abstract

Large organizations, such as government agencies, often distribute their information on the web in the form multidimensional tables. This thesis describes the extraction of data cubes from the tables, which can be collectively queried by decision-makers using popular OLAP tools. Those tables are also a valuable resource for answering user questions, improving faceted search, and generating ontology. Improving the quality of information extraction from multidimensional tables is mandatory, because of their inherent sophisticated design. In this thesis, algorithms are presented for assigning labels to dimensions, domain integration, identification of measure dimension, table integration, and table partitioning. Experiments were conducted on some 800 tables from Statistics Canada, and our success rate was greater than 90% for each component that was tested.

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my supervisor, Prof. Wo-Shun Luk, who has supported me throughout my thesis with patience and wisdom while allowing me the latitude to work in my own way. I attribute my success in obtaining a master's degree to his encouragement and effort. His guidance continually helped me while I was researching and writing this thesis.

Furthermore, I would like to thank the rest of my thesis committee Dr. Jian Pei, Prof. Qianping Gu and Dr. Oliver Schulte for their encouragement and insightful comments.

Finally, I would like to thank my family for all their love and encouragement. I am deeply grateful to my parents, who instilled in me a love of science and supported me in all my endeavors. Most of all, I sincerely appreciate the faithful support of my loving, supportive, encouraging, and patient husband Saad. Thank you.

Table of Contents

List of Figures

List of Tables

1. Introduction

The online realm has evolved to the state where most websites contain a combination of both simple and rich media. Moreover, the content is often used as presented on-screen, but can also be used for analysis and comparisons. Multidimensional tables, which considered as rich beneficial media, appear on the websites of various organizations are valuable resources for answering questions [1], improving faceted search [2], and generating ontology [3]. A common use for multidimensional tables is to present statistical information. In this chapter, we discuss the motivation for extracting multidimensional tables and how the results of this research can benefit different applications.

1.1 Motivation

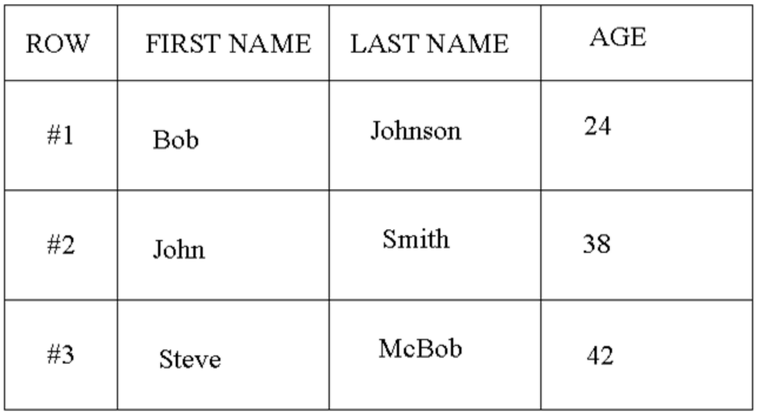
Many research studies have explored information extraction from web pages. The studies are classified by targeted input type: free text [4], semi-structured data (e.g., lists [5]), and tables [3]. Virtually no structures exist in free text, other than those arising from the natural language. Tables, on other hand, impose a two-dimensional grid structure on a number of texts, so that texts on the same row or column often, imply a special relationship among them. In between, semi-structured data, according to [6], contains "tags or other markers to separate semantic elements" in the text; for example, XML.

It does seem entirely natural that the extraction techniques that work well for one type of input may do poorly for another type of input. Indeed, the study in [7] suggest that multiple extractors should run in parallel over web pages, and the outputs should be combined into some sort of databases. It is contended that an extractor specialized for a certain type of input often does well with a certain topic, since a *natural fit* occurs between the data model and a topic area. For example, population statistics are usually published as tables. On the other hand, lists are often associated with on-line shopping (e.g., [8], [9], [10] and [11]), where some researchers, e.g., [11] assume that the data sources observe the 'Amazon effects'. Therefore, a special type of extractor should be able to focus on multidimensional tables.

1.1.1 Relational Tables vs. Multidimensional Tables

Data tables presented on the web can be classified into two different kinds: relational tables and multidimensional tables (Figure 1.1 and Table 1.1). In this research, we argue that, in extracting web tables, multidimensional tables should be distinguished from relational tables. In relational tables, the data is stored as tuples under a number of columns, with optional column labels. The tuples can be text or numbers. In contrast, a multidimensional table consists of three or more dimensions, mapped onto a multidimensional array. Each cell in the array contains a data item, called a measure. Moreover, the data items are mostly numeric values and can be addressable by their coordinate (i.e., index) in each dimension. The multidimensional array is generally called a *data cube* [12] and the coordinates for a dimension, which are called members of a dimension in this study, are organized as a linear hierarchy, where each member have only one parent. Each dimension is assigned a unique label and its members are also unique within the same dimension. As an example for the multidimensional table, Table 1.1 easily fits a cube design represented in Figure 1.2 with multidimensional information. In this example, three aspects are used to measure the number of persons: sex, marital status, and year. Each of these represents a dimension in the cube, with the measure dimension being the number of persons.

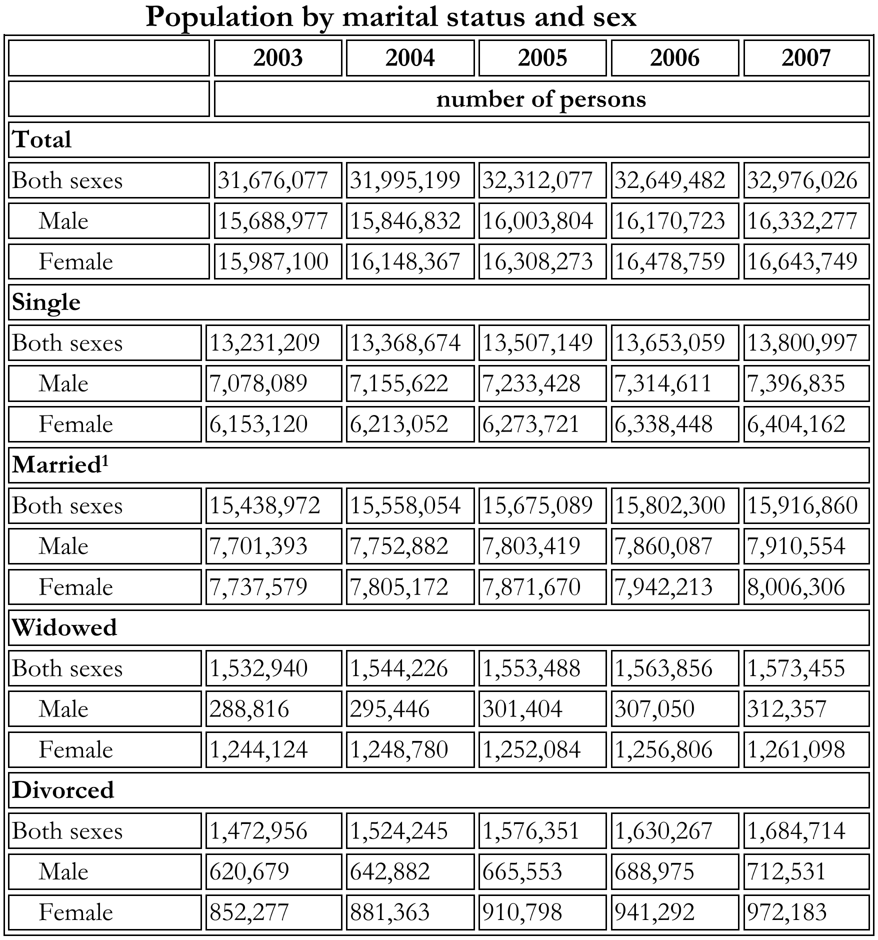
Currently, the data that is extracted from a multidimensional table is presented as attribute-value pairs, where the attribute consists of labels at every level of every dimensional hierarchy [13][14]. In those attribute-value pairs, it is not clear which are the labels from different dimensions and the labels that presenting different level with one dimension. Realizing the limitations of this approach, recent research has focused on adding more constructs to better capture the semantics of complex tables, such as partition labels and over-expanded labels ([15], [16]), and nested labels ([17]). Compared to the full-fledged OLAP schema, these approaches are not as holistic. The metadata, the data which is required to describe the structure of the data cube (such as the dimension labels), is missing, though it is just as valuable as the data itself, for all kinds of applications (as described in Section 1.3). Obviously, any table extractor considering the metadata, would be incapable of handling the numerous tables that exist on the web, as can ***WebTables*** [18] with the relational tables. Multidimensional tables; however, are more important compared to the average web table published online, since they contain more information for data analytics. Therefore, a need exists to improve the quality of the extracting.



ROW	FIRST NAME	LAST NAME	AGE
#1	Bob	Johnson	24
#2	John	Smith	38
#3	Steve	McBob	42

Figure 1.1: Relational table

Table 1.1: Sample of a statistical multidimensional table



Population by marital status and sex

	2003	2004	2005	2006	2007
	number of persons				
Total					
Both sexes	31,676,077	31,995,199	32,312,077	32,649,482	32,976,026
Male	15,688,977	15,846,832	16,003,804	16,170,723	16,332,277
Female	15,987,100	16,148,367	16,308,273	16,478,759	16,643,749
Single					
Both sexes	13,231,209	13,368,674	13,507,149	13,653,059	13,800,997
Male	7,078,089	7,155,622	7,233,428	7,314,611	7,396,835
Female	6,153,120	6,213,052	6,273,721	6,338,448	6,404,162
Married[1]					
Both sexes	15,438,972	15,558,054	15,675,089	15,802,300	15,916,860
Male	7,701,393	7,752,882	7,803,419	7,860,087	7,910,554
Female	7,737,579	7,805,172	7,871,670	7,942,213	8,006,306
Widowed					
Both sexes	1,532,940	1,544,226	1,553,488	1,563,856	1,573,455
Male	288,816	295,446	301,404	307,050	312,357
Female	1,244,124	1,248,780	1,252,084	1,256,806	1,261,098
Divorced					
Both sexes	1,472,956	1,524,245	1,576,351	1,630,267	1,684,714
Male	620,679	642,882	665,553	688,975	712,531
Female	852,277	881,363	910,798	941,292	972,183

Source: Statistics Canada: http://www40.statcan.gc.ca/l01/cst01/famil01-eng.htm

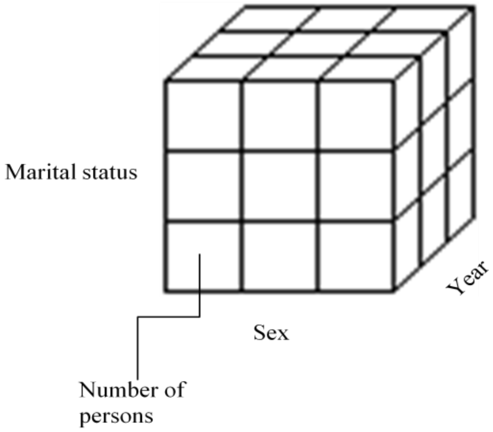


Figure 1.2: The cube representation of table 1.1

Multidimensional tables are widely used, especially among large organizations, such as national and international agencies, public institutions (e.g., universities), public corporations, etc. They often used to present statistical or economic data. As indicated by the seminal paper on OLAP data cubes [12], the concepts, such as group-by, sub-total, and cross-tab, found in most report writers, are also applicable to data cubes. Thus, multidimensional tables tend to be machine-generated, and are the primary sources of high-quality data essential to public and private policy-makers on a daily basis. Unfortunately, a tendency exists to publish the tables in the PDF format, which renders them inadmissible to most extractors, though new conversion software systems are becoming more available on the market.

1.2 Thesis Objective

This research is about the design and implementation of an information extractor for multidimensional tables. The goal is to demonstrate the feasibility of automatically constructing a database and a knowledge base from a collection of these tables from the web, presented by organizations for various kinds of applications. Given a webpage's URL that contains the tables, the extractor produces: (i) a set of data cubes and the corresponding schemas and (ii) a set of domains. An OLAP schema consists of the dimension hierarchies associated with the dimensions and the measure. A domain consists of a set of values from which a dimension draws its members. Those domains are used to identify the dimensions from the same domain, and their corresponding cubes can be joined together.

1.3 Targeted Applications

There has been a steady stream of research studies on information extraction from all types of tables, from plain-text to web tables. This is due to the on-going growth of information being published in web tables, which is dictating the need to improve the techniques of table extractors. Recent research has been mostly devoted to processing the vast number of web tables. In this thesis, the focus is on the quality of data and on the associated metadata that is extracted. In particular, we are interested in showing how the approach advocated here will affect the following applications that some of them are targeted by most information extraction systems [3].

1.3.1 Question-Answer System

As the most obvious application, with respect to a single data cube, an MDX-like OLAP query language, or data visualization tools, may be applied to the cube for OLAP-style data analytics. With an SQL-like interface that cover the collection of data cubes, a questions like: "What are the divorce rates of the top-3 most populated Canadian provinces?" may be answered by joining together two cubes on marriage/divorce and population respectively, for the common dimension of Canadian provinces. Recently, some studies have helped to answer questions from relational data-bases, and retrieve rows from the relational table that contain the results, but not the whole web page; for example

DBXplorer [19]. Nevertheless, such applications are for relational tables and are incapable of analyzing or extracting information from multidimensional tables.

1.3.2 Faceted Search

Many agencies offer a keyword search facility on their collections of published web pages. Unlike Google's search engine, most agencies do not have the resources to build a sophisticated search system. For example, the search engine for Statistics Canada is quite primitive. A search for a relevant table is more effective when the user knows the exact keywords contained in the table row and column headers. Indexing is offered for browsing, though, it is not very effective because many tables are included under multiple indices. Recently, faceted search has become an effective alternative search mechanism, as a complement to keyword search [2]. According to [20], a faceted search facility consists of two main components: faceted metadata, and faceted category interface. In the OLAP terminology, the faceted metadata is the dimension hierarchies of a data cube about some kind of entities. For example, we can classify families by their family structure; e.g., married couple families, single-male-parent families, etc. Alternatively we could classify families by their incomes, geographic locations, number of children, housing, etc. As pointed out by [2], identifying these classifications was mostly done manually or borrowed from somewhere else. In this research, we show that the classifications can be automatically generated, since we understand the structure of multidimensional tables, which can also help to avoid the challenges, as indicated by [21]. In regards to the faceted category interface, a good example can be found in [22]. In a recent study [23], the authors consider the power of combining faceted search with OLAP for analytical data. By using this research multidimensional table extractor, thus, this kind of applications can be improved.

1.3.3 Ontology Generation

According to [3], one of the applications of web tables understanding is ontology generation. Much research has been done on ontology-generation from free text; however, as pointed out by [24], much work is still needed to improve the quality of ontology generated by information extractors. Likely, the quality will be improved if the input to the information extractors has more structure, as suggested by [25]. The quality of ontology is high and it can be generated from most authoritative sources, since the tables generated from international and national agencies tend to cover many diverse areas; e.g., education, health, economics, etc. In this thesis, we experiment with the tables from Statistics Canada. We believe that the generated ontology can also help in the extraction of tables (in HTML format) published by other national statistics agencies.

1.4 Challenges

The main challenge for us is to produce a well defined cube from extracting multidimensional table, to realize the potentials of this research for the targeted applications described in Section 1.3. The cube is concerned with measurements of an entity set in different aspects. The entity refers to the aspect being measured by the cube; for example, number of people of 25 years, living in Canada. In this case, the entity is comprised of people, being measured in two aspects (age and country). To generate useful ontological information, proper labels are needed for the dimensions and for singling out the measure dimensions. Further, each dimension must be a complete classification of the entity set, to ensure summarizability [26]. In data analytics, it is common to include more than one type of entity in a table. The information extractor must be able to recognize this situation, and split the table into multiple cubes, one for each type of entity. Most research studies assume that a

1:1 correspondence exists between the set of input tables and the set of output tables; however, this is not always the case for some of the more sophisticated websites. Furthermore, due to the limited space available in a web page, especially in terms of column size, a large table may be broken into a series of tables by splitting a deep and/or bushy hierarchy, into a number of smaller ones. The information extractor must be able to reintegrate them back into a single cube.

Despite the challenges in this research, outlined in the previous paragraph, we show that, as far as multidimensional tables are concerned, it is feasible to extract the information for the targeted applications, by applying some novel techniques. In addition to the visual clues appearing in multidimensional tables, we rely on the table title to provide valuable information about the metadata, which, to the best of our knowledge, has not been utilized by any information extractor. A typical table title reveals the table structure in a multidimensional manner, though far from perfect, since it is generated for and by humans. We also make use of the numeric data to verify that a set of labels forms a classification scheme, or part of a multi-level classification, which we call the summation rule. Indeed, the summation rule is used to discover classification schemes that are not obvious to the human eye, because of the lack of visual clues in the table design. These techniques are discussed in detail in Chapter 4.

1.5 Organization of the Thesis

This thesis is organized into six chapters, as follows:

- Chapter 1: Introduction: an overview of the problem and how this thesis attempts to solve it.

- Chapter 2: Related work: research in related areas and how this thesis differs from it.

- Chapter 3: The Table Model: the specific challenges of extracting information from multidimensional tables.

- Chapter 4: Table Conversion Process: we present our technique for extracting the information and generating cubes.

- Chapter 5: Experimental Results: we describe the experiments and show their results for assigning labels to dimensions, domain integration, identification of measure dimension, table integration, and table partitioning.

- Chapter 6: Conclusion and Future Work: summarize this thesis and outline of how this thesis may be further modified to provide more functionality and efficiency.

2. Related Work

2.1 Table Processing

This research is about generating cubes from the extraction of multidimensional tables, as they are presented on the web. Some authors have examined this problem and proposed solutions. Leung [27], in his master's thesis, presented a model for identifying tables in an HTML page and transforming the tables into cubes. Nevertheless, he relied on the visual clues in the table design, to extract the components of the HTML table and generate the cube. For this research, we find that visual clues are insufficient to extract a schema for the data cube. In short, the previous work was mainly about the dimensionality of the data in multidimensional tables, while the focus of this research is on the metadata.

According to [28], the information in a statistical table can be modeled as an OLAP database. Thus, statistical tables can be transformed (by human observation) into OLAP cubes with; for example, three dimensions and measure function, with one of the dimensions being a multi-level hierarchy. In [29][30], the authors devised algorithms to accurately identify the dimensions inside a statistical table. In this case; however, no schemas were derived and the algorithms were not capable of handling more complex tables, as the ones described in Chapter 3.

In [26], the authors present a scheme to combine a content management system and OLAP systems, together, so that an OLAP keyword search can be used in a content management system. For example, a bibliography database in XML can be converted into a multidimensional database, illustrating the issue of summarizability. In this research, we resolve this issue with the summation rule (see Section 4.6.1).

2.2 Labeling

Recently, a flurry of research can be seen about searching on the deep web (e.g., [8], [9], [11], [31],[32], [33], [34]). Databases are often hidden from the web, but may be searched through a customized web interface. The results are usually presented on the web as lists, or relational tables. Annotations are used on the search engine to understand the output. This line of research is related to this thesis in terms of the sub-problem of label assignment.

Label assignment is concerned with finding a name that collectively describes a set of words, a process that is sometimes called data annotation. A typical problem would be to find the attribute name for a column of a table, or relation, given the set of values in the column. Many methods rely on some assumptions, such as the likely places to look for the label, or on supervised machine learning [9], [11], [32]. In our case, the label is assumed to be present in the table title, or in a nearby row. If neither of these are present, it is likely to be a common dimension that is shared by more than a few tables. Thus, our heuristic involves finding a match among the members of the unlabeled dimension with members of the other dimensions whose labels have already been discovered. A similar approach was proposed by [8], where the decision is made on the basis of frequencies of co-occurrences in the web page.

Many other assumptions involve using external resources to assign the label. The external resources contain datasets that classify each group of words into a single group. In these methods, a label is assigned to the dimensions without any further testing by matching the dimension members to the external group. Examples of this appear in Reference Match

[35], or by using OpenCalais [36] and ThingFinder [37]. The problem with these methods; however, is the limitation for the number of groups they may contain. Furthermore, they can assign a label that is not precise for the dimension that appears in the table, which can lead to a misunderstanding about the content of the table.

2.3 Table and Domain Integration

Attribute matching is another common sub-problem in schema matching and semantic integration [38], [39]. Matching two schemas often involves matching two attributes, with or without their associated instances. In this thesis, we need to determine whether or not two dimensions share the same domain. This information is important to deciding whether or not two cubes may be joined for querying or for faceted search purposes. Thus, the sets of the members of two dimensions must be ensured of having a sufficiently large overlap.

Broadly speaking, two main approaches are used for attribute matching: supervised and unsupervised learning. In [40], matching rules are derived by human trainers during the training stage. In [41], a program is deployed to investigate three 'facets': terminological relationships (e.g., synonyms), data-value characteristics (e.g., average values in the populated attributes), and target-specific, regular-expression matches of attribute values.

2.4 Ontology Extraction

Extraction of ontology from texts and tables has been a popular research topic lately. In [2], a technique is presented for automatic extraction of facets for use with browsing text databases. Nevertheless, [25] claimed that the results of automatic ontology extraction from tables are more effective than texts. In [42], the authors present a semi-automatic method for enhancing the basic database scheme in a domain, where attributes can be dimensions, as defined here, with additional semantics that include summarizability-related constraints. Schemes for extracting ontology from tables and storing it in the Semantic Web format; i.e., RDF [43], are presented in [16] and [44]. In this research, we do not describe how the ontology is stored, though a description for storing ontology in RDF, associated with an OLAP data cube, can be found in [45] and [46].

3. The Table Model

Online statistical tables are usually presented as multidimensional tables with a sophisticated design. Since they are maintained by professional organizations (i.e., Statistics Canada), they often contain extremely dense analytical information. Furthermore, multidimensional tables tend to be different from other basic relational tables that can be found online; thus, the extraction of data must be performed with meticulous care. As observed in Table 1.1 example that the multidimensional tables can be clearly mapped into data cube Figure 1.2.

To proceed, we first must understand how to read such complex tables. Some the libraries that provide directions in this regard [47], part of the multidimensional table components can be identified. In the next sections, each useful component from the HTML page is described, for the accurate design of a multidimensional table extractor.

For this research, we make use of two different table models: generic and extended. The generic model is one that is commonly assumed. With our techniques developed in this research, we show that our information extractor can work on web pages produced by statistics agencies in Canada, Austria, and Finland. The extended one includes an additional feature about table series that is specific to web pages from Statistic Canada. Since over half of the tables in our samples belong to some table series, we can't process them properly unless our model is extended to take this feature into consideration.

3.1 Generic Table Model

Statistics tables are examples of multidimensional tables. They are well constructed [47], and usually designed in a specific common layout that is standardized for readers of statistics. Figure 3.1 shows the main components of a multidimensional table, including the table title, column headers, row headers, measures and data cells. The measures are optional and may not always appear in some layouts. In the following sections, each component will be described in detail.

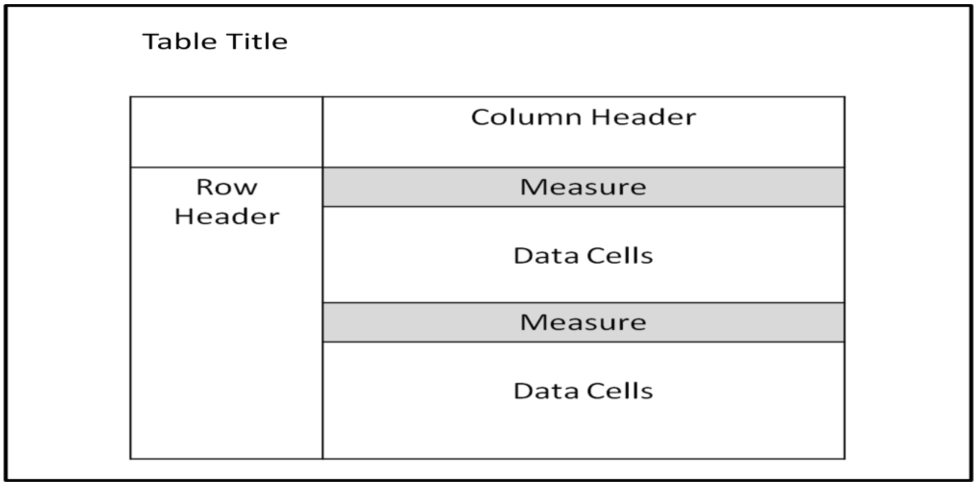


Figure 3.1: Table Model

3.1.1 Table Title

The title provides the first piece of information for readers. Typically, titles are carefully prepared by a professional organization. The title is used to inform readers of the purpose of the particular table. Articles provided by Finland Statistics [48], describes four important elements of a title: "the title, which identifies the population covered; the variables described in the table and their classification; the time period of the observations; and the units of measurement." In short, the table title should follow the general description of what a table consists of.

In some cases, table titles provide clues about the measure (e.g., population), and the labels of the dimensions (e.g., marital status and sex). In the table title for Table 1.1, nothing refers to the year dimension, and since the time dimension is almost always present in most data cubes, we can assume that one is present and we can search for it.

By observing most multidimensional table's titles, they consist of two components (see Figure 3.2): measures of the entities stored in the cells of the table, and the dimensions, connected together with the word 'by'. Most of the multidimensional tables are fairly similar to the one shown in Table 1.1, but table titles do not always fully portray the table's contents. Although omissions occur, possibly due to human error since the title is produced by human being, readers will usually be able to recognize the missing dimensions by viewing the table. The processing of the title resembles natural language processing, with rules and frequent exceptions. The following section describes some typical structures for different kinds of table titles.

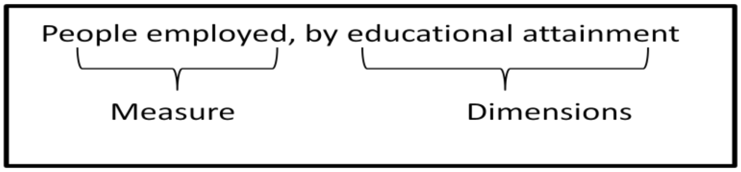


Figure 3.2:Sample of table title

3.1.1.1 Variety of Table Titles

1) Some table titles list all dimension labels that are represented in the statistics table. Although those labels are not listed in standard order, the labels are located in the title after specific prepositions, for example, "by". Thus, the appearance of this preposition will help to assign labels to the dimensions.

2) Some table titles are very brief, and do not list dimension labels that are commonly found in many tables (e.g., the year dimension).

3) Occasionally, special prepositions, such as 'by', does not appear in the table title. Those prepositions usually precede the list of dimension labels, in this case the dimension labels are listed in a different organization.

4) Some table titles list all of the members of a single dimension. In such cases, no common word can be used to represent the dimension, or the table designer may have found it more meaningful to list each member of the dimension. This case can only occur for one dimension in the table, and not for all of the dimensions.

3.1.2 Multidimensional Table Headers

The multidimensional table consists of two kinds of headers the row header and the column header. They have different representations, because of the usual table design. Each one of them has different properties listed in the following sections.

3.1.2.1 Column Header

The first few rows in the table, before the numeric data cells (i.e., the measured value), are the column header. In some cases, the header will be comprised of more than one row. The header is an alphanumeric text. Because the data in multidimensional tables can be structured as cubes, and they are attributed value pairs, one or more of the coordinates that represent the values will appear in the column header. For example, if two neighboring rows in the column header are on top of each other, we would understand that for every numeric cell within the table, two coordinates will represent it and these can be taken from the column headers. The coordinates that appear in the same row are usually part of the same dimension.

3.1.2.2 Row Header

In statistics tables, the row header appears in the leftmost column of the table, which includes the other coordinates that are used to represent the numeric values in the table cells. The information in the row header can be a single hierarchy such as a list of cities, or multi-level hierarchies such as provinces and their list of cities. They are represented with different visual clues, such as indentation, column spanning, font type, font style, font size, font color, and background color [27]. The coordinates that represent a numeric cell in a specific row are taken from the row header for the same row, and any other row that appears in the row above that is a higher level in the hierarchy (i.e., higher level in visualization). For example, if the specific row is in a regular font and the higher row is bold font, then they are both used as coordinates, but if the row has bold font and the regular font is on top, then it is clearly not a higher visual level. In the row header, different visualizations are used, since only one column can be used. This limitation is related to the screen size of a normal monitor, and because readers usually avoid scrolling horizontally.

Referring to the table shown in Table 1.1, the numeric cells in the 7th row have two coordinates from the row headers (*Male* and *Single)*. *Male* appears in the same row and *Single* appears in a different row, but at a higher presentation level. In short, the table row header could refer to different kinds of hierarchy even if the same visualization techniques were used, more detail in Section 4.3.2 and 4.6.

3.1.3 Multidimensional Table Headers Members

The data inside the row and column headers are alphanumeric text, which are usually organized as groups, with each group of data being combined when they have the same visual clues; we call those data as the header members. They are combined as a collection, since the table designer has categorized them together. For example, if a statistics table contains a group *marital status*, they would be placed together within the same dimension and with the same visual clues. The members within a single dimension can be a single word, an abbreviation, a phrase, a short sentence, or a number, and can be verbs, nouns, or adjectives. Thus, the members do not have to contain the same linguistic properties, and the dimensions would rarely have a label assigned to them.

3.1.4 Measure

The numeric data cells presented in the table are measured by specific metrics, such as number, $, or %. The metric appears in the last row of the column header, or occasionally in a spanning row in the middle of the table, when the table needs to be partitioned; Section 4.2 discussed this case in more detail. The appearance of the measure unit is optional, especially if it is a common unit of measure, such as number, or if it is already merged with the measure-related dimension. For example, in the row header, "*number of families*" appears as one of the measure-related dimension members.

Some members in the headers; i.e., row or column headers, have measure-related coordinates. For example, in some tables, we might have *revenue* in one of the headers and the measure unit, $, in the last row of the column header. Readers would understand that for each numeric cell, revenue is expressed in $. The measure-related members also need to be distinguished from the other dimensions that appear in the header. In Section 4.5, we show how to identify these.

3.1.5 Data Cells

As mentioned earlier, data cells contain numeric values. They should be extracted as attribute value pairs taken from the coordinates from the table headers. As the numbers are presented within the table body, they can reveal the structure of the table. The relationship between members of the same column or row header can be observed from the number distribution. In Table 1.1, consider the three numbers in the 2nd to 4th rows, and in the 2nd column. The numbers 15,688,977 and 15, 987,100 summate to the number in the cell above them; i.e., 31,676,077. This means that the member *Both Sexes* is verified to contain two members, *Male* and *Female*. In non-OLAP terms, the rule is that the entities (i.e., persons), may be classified into *Male* and *Female*, and the classification is a linear one; i.e., no entities can be included in both classes. This linear classification is a very powerful one, as it will help in discovering the dimensions within the table. In the table, the aggregate function is summation, though it may not always be the case, and other aggregate functions can also be used.

3.2 Extended Table Model

Some statistics websites do not include all of the dimension members for a single dimension within the same statistics tables, so that the dimension is not fully classified in the table. Instead, due to space constraints, the data is placed outside of the row and column headers. Although that this model only found in Statistics Canada, it can be a good design practice for other agencies. In such instances, the collection of members for a single dimension that exists outside the statistics table may be represented by a links, a drop-down menu, a radio button, or another HTML form feature (Figure 3.4). Thus, these links will lead to other statistics tables that contain the new collection of members. For example, table A might have "*British Columbia,*" and table B might have "*Alberta,*" and both would be members in the province dimension. The new tables are usually similar to the primary table, with the same table structure and list of dimensions, except that some of the members of the dimensions would be different. Since each table does not have the full classification for one of the dimensions, an integration of the tables would be valuable to the users.

There are some clues found in the webpage that shows if the special links are related to the table. Some of the members in the row or column headers appear the same in the special links, this will confirm that those links are related to the table, and will lead for the continuous information. However, it is not always the case; those tables may not be related together, thus, extra processing is needed (see Section 4.7). As each table is discussing part of a dimension and not the full classification, some table titles will show that by including supplemental information in the table title describe the sub-dimension members. This part usually appears visually different than the main table title. In Figure 3.3, will show the extended table model that includes the locations for the special links and how they related to other information in the table. The table headers in page 1 and page 2 are usually identical except for one of the dimension members.

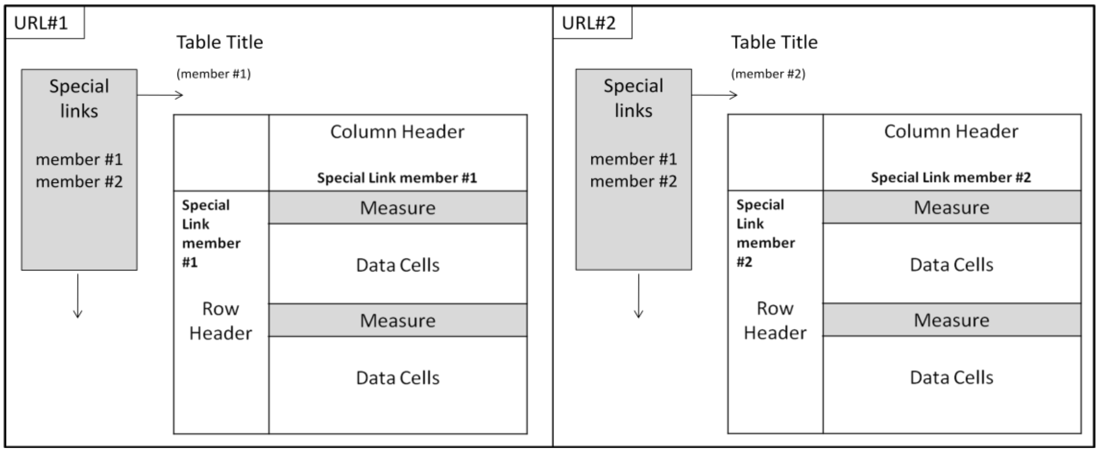
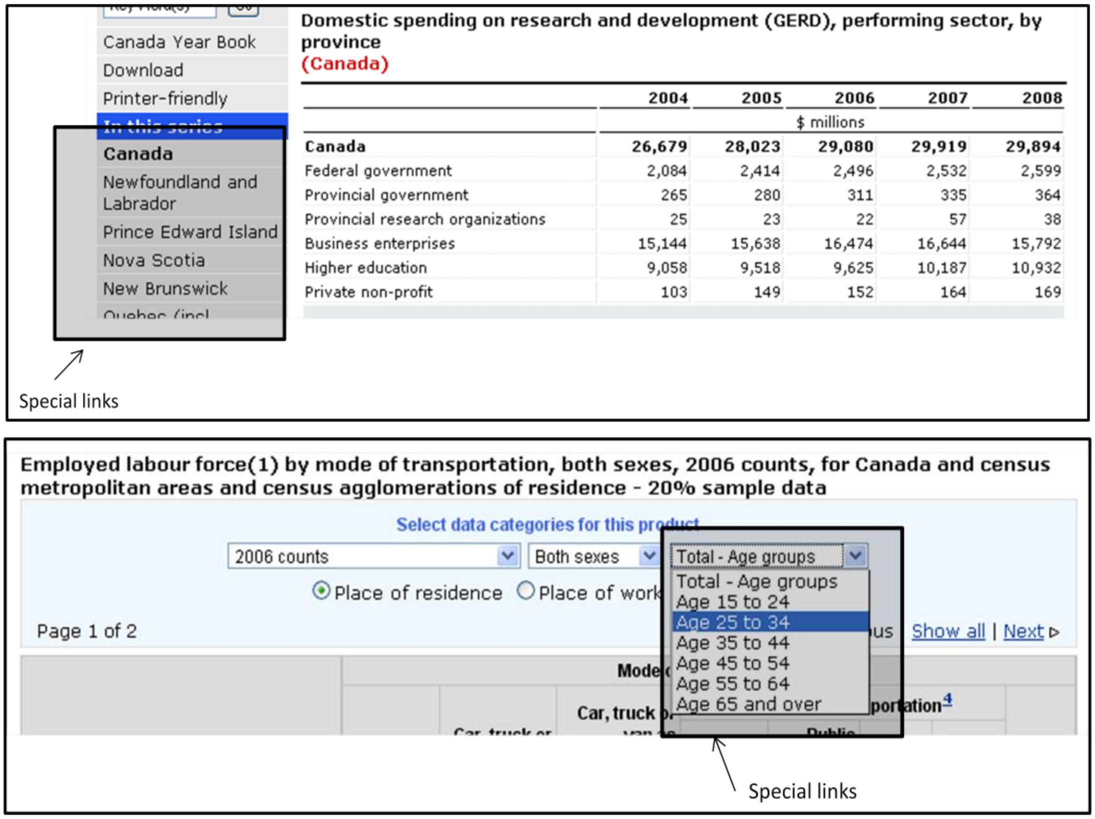


Figure 3.3: Extended table model



Domestic spending on research and development (GERD), performing sector, by province
(Canada)

	2004	2005	2006	2007	2008
			$ millions		
Canada	**26,679**	**28,023**	**29,080**	**29,919**	**29,894**
Federal government	2,084	2,414	2,496	2,532	2,599
Provincial government	265	280	311	335	364
Provincial research organizations	25	23	22	57	38
Business enterprises	15,144	15,638	16,474	16,644	15,792
Higher education	9,058	9,518	9,625	10,187	10,932
Private non-profit	103	149	152	164	169

Employed labour force(1) by mode of transportation, both sexes, 2006 counts, for Canada and census metropolitan areas and census agglomerations of residence - 20% sample data

Source Statistics Canada:
http://www40.statcan.gc.ca/l01/cst01/scte02a-eng.htm
http://www12.statcan.ca/census-recensement/2006/dp-pd/hlt/97-561/T603-eng.cfm?Lang=E&T=603&GH=8&SC=1&SO=0&O=A

Figure 3.4: Special links appears outside the table

4. Table Conversion Process

To properly transform HTML table components to cube components, we need to use heuristics approaches to best judge the representation of each component in the HTML page and to transform them to the cube components. A statistics table could represent one or more cubes, and each header in the HTML table could represent one or more dimensions. We rely on different heuristics to generate the cube, starting from the basic visual clues [27]. In our analysis of different tables, the visual clues are not sufficient for transforming the table to a cube. Thus, the generated cube cannot be used to search and extract the metadata because the dimension labels are absent. Therefore, in addition to the visual clues, we also apply linguistics and mathematical processing in analyzing the table title, the headers, the special links that appear in the webpage, and the representation of the numeric results in the statistical table.

Figure 4.1 is a flowchart showing how we combine the heuristics to extract the table. We begin by recognizing the various table components in an HTML table to disassemble the table, as described in Section 4.1. If there is a measure dimension found in spanning a row in the table, we should partition the table and process each partition alone. Then, with each components obtained from extracting and parsing the HTML table, we derive the dimensions, assign a dimension label to each, and identify the measure dimension. We carry on from these results to partition the table accurately using heuristics such as the summation rule (described in Section 4.6). After the table is partitioned, we can check to see if other tables are available in the same statistical website, and if present, determine whether or not they refer to the same study with continuous information. Next, the tables can be merged, as described in Section 4.7. Later, we can collect all of the available domains from the tables, as described in Section 4.8. The first process starts by initializing the cube storage, and each time a process is done, the cube will be updated. The domain, on the other hand, will be updated each time a dimension is derived, and it will be queried each time dimension label is needed.

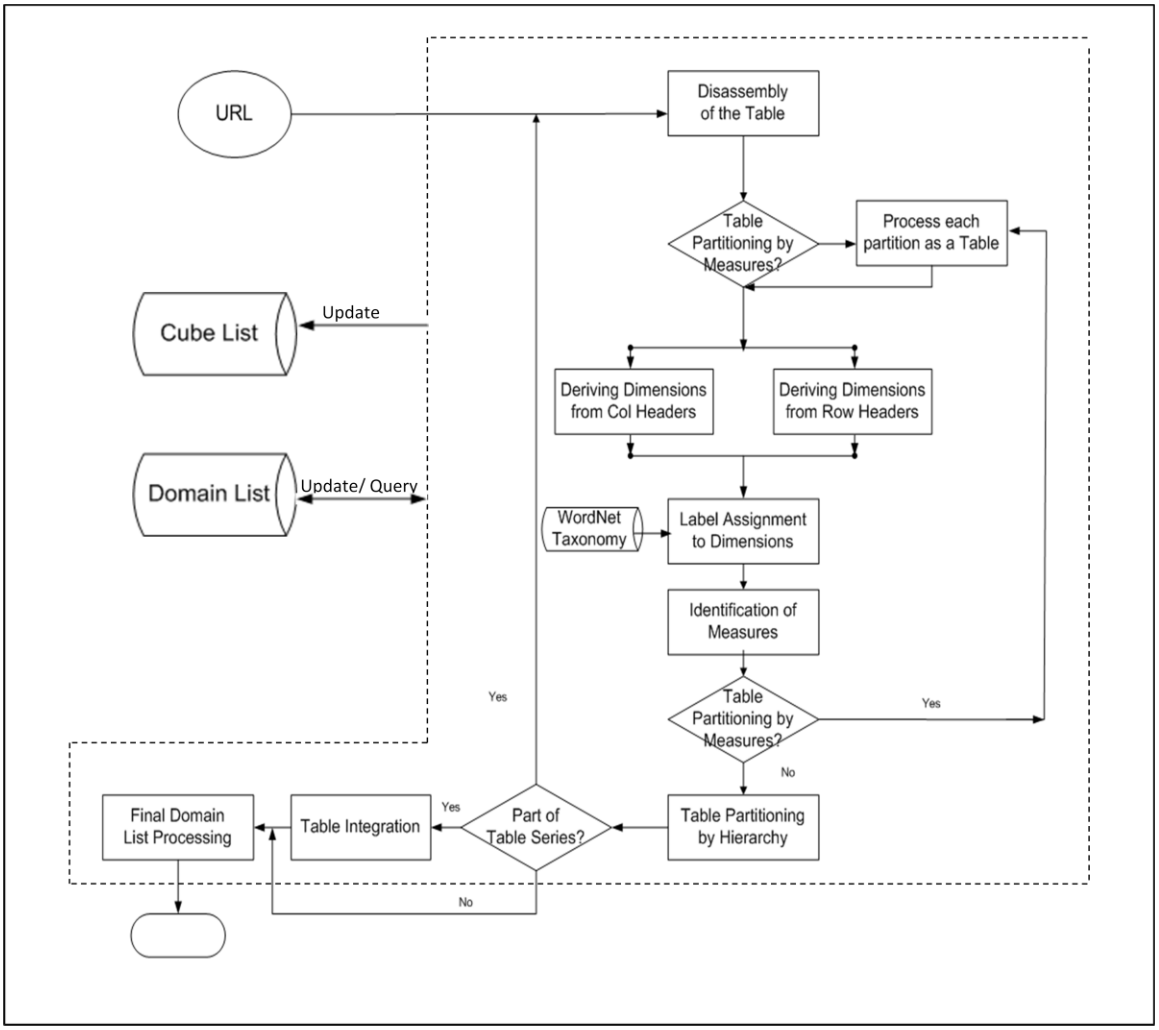

URL
Disassembly of the Table
Table Partitioning by Measures?
Process each partition as a Table
Cube List
Update
Deriving Dimensions from Col Headers
Deriving Dimensions from Row Headers
Domain List
Update/ Query
WordNet Taxonomy
Label Assignment to Dimensions
Identification of Measures
Yes
Table Partitioning by Measures?
Yes
No
Final Domain List Processing
Table Integration
Yes
Part of Table Series?
Table Partitioning by Hierarchy
No

Figure 4.1: Flowchart showing how to transform table to cube

4.1 Disassembly of the Table

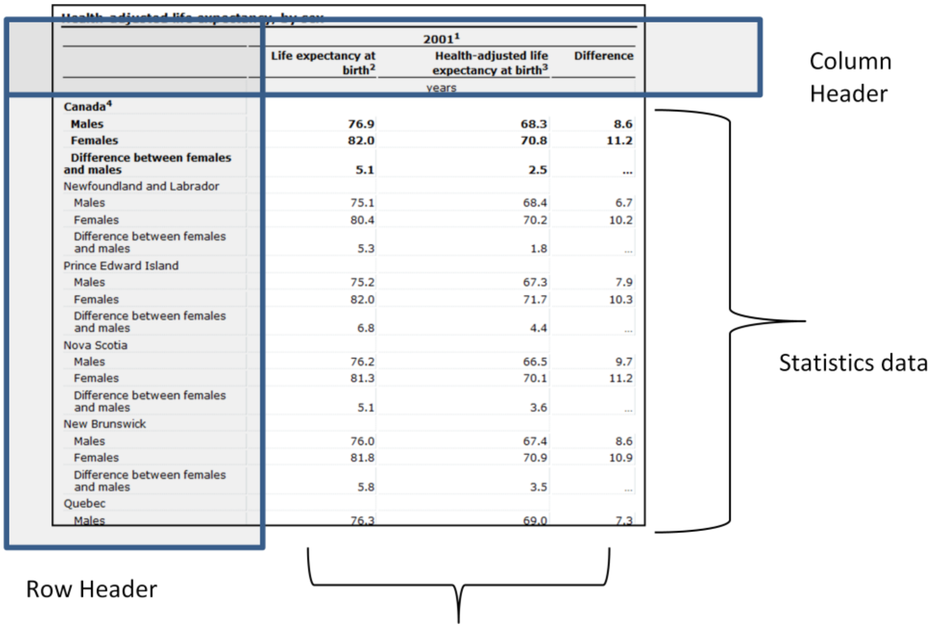
Statistics tables are usually designed according to a common layout that is standardized for readers of statistics tables. The headers are important components of the tables, with the column and row headers appearing in the first few rows, and the leftmost column, respectively. They contain the most important information representing the data cells, which are the actual numeric observations for the data studied.

To begin, we describe the solution for retrieving the components that appear in the HTML page, containing the multidimensional table. After identifying the table components, they are retrieved by disassembling the multidimensional table. In addition, we will identify the location of the table title which appears on the same webpage.

4.1.1 Identify the Headers Location

Statistics tables are designed to have row headers, column headers, and the numeric data (Figure 3.1 shows the table model, and Figure 4.2 is an example of a table). The most reliable rule used in [27] for extracting the components of the table refers to the appearance of the numeric data, since they are numbers, while the headers are alphanumeric text. The column header appears in the first few rows of the table, before the appearance of the data values, and the row header appears in the leftmost column of the table.

The headers contain important information such as the dimensions that represent the numeric data in the cube. The extraction of the dimensions from the headers will be discussed in more detail in Section 4.3.



Health-adjusted life expectancy, by sex

	2001[1]		
	Life expectancy at birth[2]	Health-adjusted life expectancy at birth[3]	Difference
	years		
Canada[4]			
Males	**76.9**	**68.3**	**8.6**
Females	**82.0**	**70.8**	**11.2**
Difference between females and males	**5.1**	**2.5**	**...**
Newfoundland and Labrador			
Males	75.1	68.4	6.7
Females	80.4	70.2	10.2
Difference between females and males	5.3	1.8	...
Prince Edward Island			
Males	75.2	67.3	7.9
Females	82.0	71.7	10.3
Difference between females and males	6.8	4.4	...
Nova Scotia			
Males	76.2	66.5	9.7
Females	81.3	70.1	11.2
Difference between females and males	5.1	3.6	...
New Brunswick			
Males	76.0	67.4	8.6
Females	81.8	70.9	10.9
Difference between females and males	5.8	3.5	...
Quebec			
Males	76.3	69.0	7.3

Source: Statistics Canada: http://www40.statcan.gc.ca/l01/cst01/hlth67-eng.htm

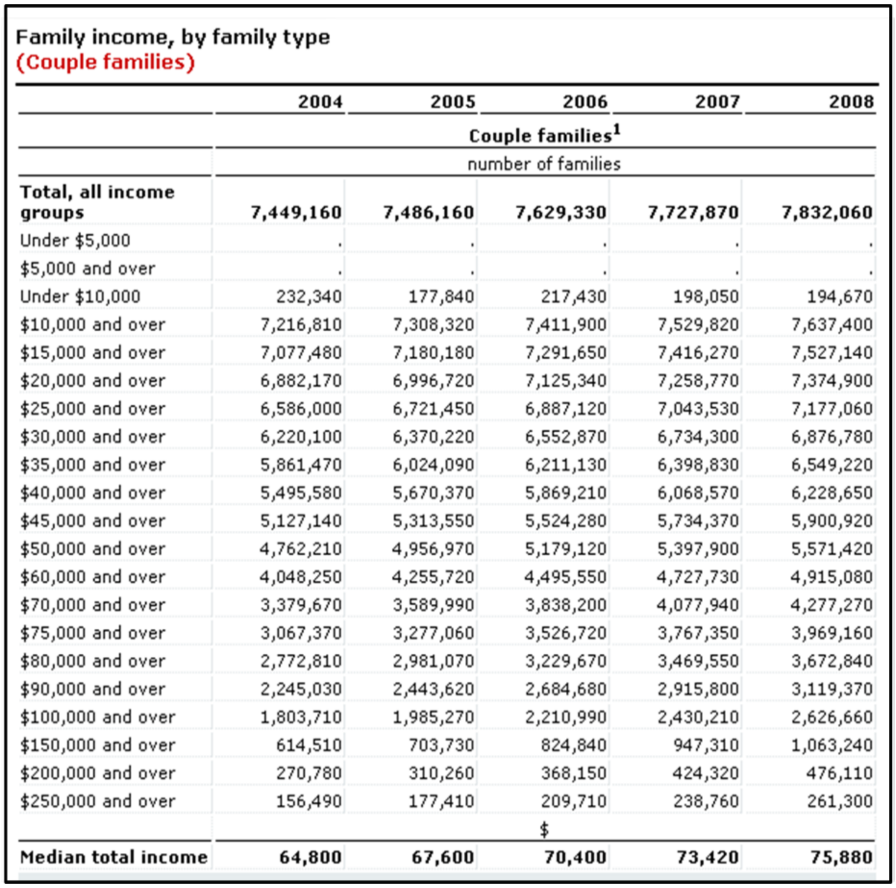
Figure 4.2: Example of a statistical table

4.1.2 Identify the Table Title

Table titles are extracted from the HTML page by identifying the appropriate HTML tags. These can be in the caption tag for the table being processed; however, if no caption tag is present, they can be found in the header tags <h> that are above the place where the statistical table is located; i.e., before the <table> tag [49]. In some rare cases, the table title could be in the <p> tag, with a larger font size. In any case, if no header or caption tag is found, the sentence that appears directly above the table will be checked.

4.2 Table Partitioning by Measure

As illustrated in Section 3.1.4, sometimes, a table may be combined from different sections. As it is shown in Figure 3.1, these sections are visibly apparent. They are separated from each other by a row which spans cross the columns of the data cells. The sections are combined within a single table for various reasons. As a rule, they are related to each other to the extent that they fall under the same table title. They are segregated primarily because they do not share the same measure, though they share other information. As an example in Figure 4.3, one section may be about the number of families in each income group, while the other section may be about the median total income for the family type. This information is included in the table as a context for viewing the data, although other tables may contain the information about median total income in more detail as seen in Figure 4.4. This is a good practice for designing the table presentation on the web, but, for our purposes, we need to segregate the data into different tables, since they are separate, though related, entities.

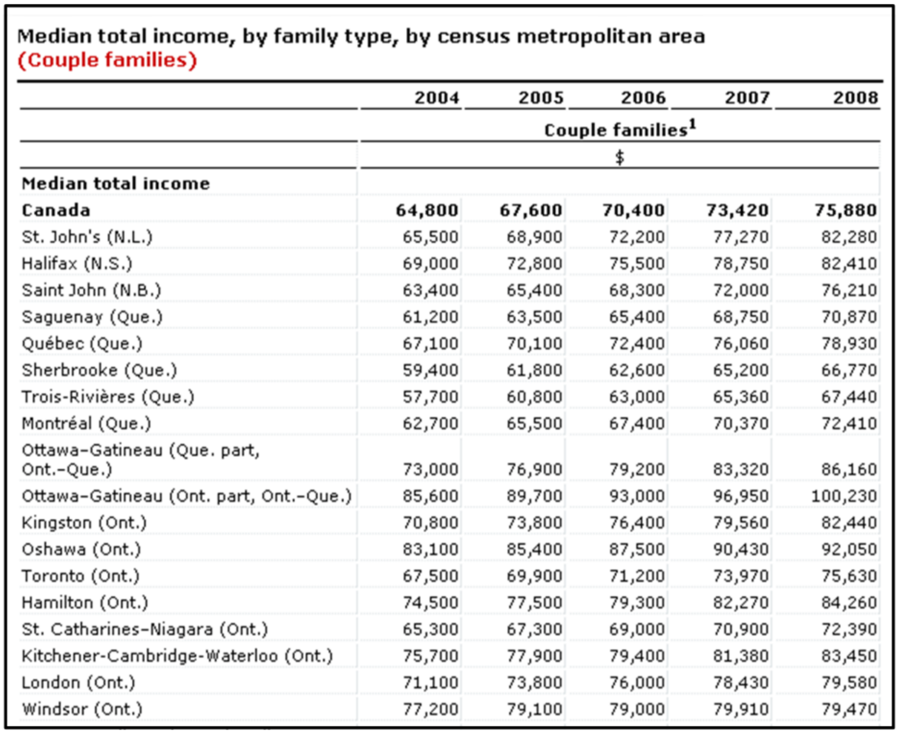


Family income, by family type
(Couple families)

	2004	2005	2006	2007	2008
			Couple families[1]		
			number of families		
Total, all income groups	**7,449,160**	**7,486,160**	**7,629,330**	**7,727,870**	**7,832,060**
Under $5,000	.	.	.	.	.
$5,000 and over	.	.	.	.	.
Under $10,000	232,340	177,840	217,430	198,050	194,670
$10,000 and over	7,216,810	7,308,320	7,411,900	7,529,820	7,637,400
$15,000 and over	7,077,480	7,180,180	7,291,650	7,416,270	7,527,140
$20,000 and over	6,882,170	6,996,720	7,125,340	7,258,770	7,374,900
$25,000 and over	6,586,000	6,721,450	6,887,120	7,043,530	7,177,060
$30,000 and over	6,220,100	6,370,220	6,552,870	6,734,300	6,876,780
$35,000 and over	5,861,470	6,024,090	6,211,130	6,398,830	6,549,220
$40,000 and over	5,495,580	5,670,370	5,869,210	6,068,570	6,228,650
$45,000 and over	5,127,140	5,313,550	5,524,280	5,734,370	5,900,920
$50,000 and over	4,762,210	4,956,970	5,179,120	5,397,900	5,571,420
$60,000 and over	4,048,250	4,255,720	4,495,550	4,727,730	4,915,080
$70,000 and over	3,379,670	3,589,990	3,838,200	4,077,940	4,277,270
$75,000 and over	3,067,370	3,277,060	3,526,720	3,767,350	3,969,160
$80,000 and over	2,772,810	2,981,070	3,229,670	3,469,550	3,672,840
$90,000 and over	2,245,030	2,443,620	2,684,680	2,915,800	3,119,370
$100,000 and over	1,803,710	1,985,270	2,210,990	2,430,210	2,626,660
$150,000 and over	614,510	703,730	824,840	947,310	1,063,240
$200,000 and over	270,780	310,260	368,150	424,320	476,110
$250,000 and over	156,490	177,410	209,710	238,760	261,300
			$		
Median total income	**64,800**	**67,600**	**70,400**	**73,420**	**75,880**

Source: Statistics Canada: http://www40.statcan.ca/l01/cst01/famil106a-eng.htm

Figure 4.3: Table has two sections, one for the number of families, and the other for the median total income in $



Median total income, by family type, by census metropolitan area
(Couple families)

	2004	2005	2006	2007	2008
	Couple families[1]				
	$				
Median total income					
Canada	**64,800**	**67,600**	**70,400**	**73,420**	**75,880**
St. John's (N.L.)	65,500	68,900	72,200	77,270	82,280
Halifax (N.S.)	69,000	72,800	75,500	78,750	82,410
Saint John (N.B.)	63,400	65,400	68,300	72,000	76,210
Saguenay (Que.)	61,200	63,500	65,400	68,750	70,870
Québec (Que.)	67,100	70,100	72,400	76,060	78,930
Sherbrooke (Que.)	59,400	61,800	62,600	65,200	66,770
Trois-Rivières (Que.)	57,700	60,800	63,000	65,360	67,440
Montréal (Que.)	62,700	65,500	67,400	70,370	72,410
Ottawa-Gatineau (Que. part, Ont.-Que.)	73,000	76,900	79,200	83,320	86,160
Ottawa-Gatineau (Ont. part, Ont.-Que.)	85,600	89,700	93,000	96,950	100,230
Kingston (Ont.)	70,800	73,800	76,400	79,560	82,440
Oshawa (Ont.)	83,100	85,400	87,500	90,430	92,050
Toronto (Ont.)	67,500	69,900	71,200	73,970	75,630
Hamilton (Ont.)	74,500	77,500	79,300	82,270	84,260
St. Catharines-Niagara (Ont.)	65,300	67,300	69,000	70,900	72,390
Kitchener-Cambridge-Waterloo (Ont.)	75,700	77,900	79,400	81,380	83,450
London (Ont.)	71,100	73,800	76,000	78,430	79,580
Windsor (Ont.)	77,200	79,100	79,000	79,910	79,470

Source: Statistics Canada: http://www40.statcan.ca/l01/cst01/famil107b-eng.htm

Figure 4.4: Median total income are shown in details with the metropolitan areas

By having the spanning row, the table is divided into two or more tables, depending on the number of spanned cells. Each sub-table inherits the column header from the main table. This is done by understanding the HTML tags that are used to build the table, and by observing where the spanned cells appear.

After this partitioning step, each section of the table is now processed as an individual table.

4.3 Deriving the Dimensions from the Headers

The table headers contain the most important information in the table; they carry the coordinates that are used to represent each numeric cell within the table. Each group of coordinates is a dimension in the cube. Different methods are used to derive the dimensions from the row header, compared to those of the column header, since, as described in Section 3.1.2, the row header and the column header have different representations.

4.3.1 Deriving the Dimensions from the Column Header

The first few rows in the table, before the numeric data, are the column headers. In some cases, the column header may be comprised of more than one row. These rows may be classified into two types: measure and dimensions. The measure row, which contains the measure unit, is segregated from the others, to be processed later (see Section 4.5). Here we are concerned with only the remaining rows.

If only one row appear in the header, they are considered to be members of a single dimension. Otherwise, we may see three different visible relationships between two neighboring rows, as shown Figure 4.5. [27] Presented three different common representations of column hierarchy. Note that (iii) is actually a special case of (ii).

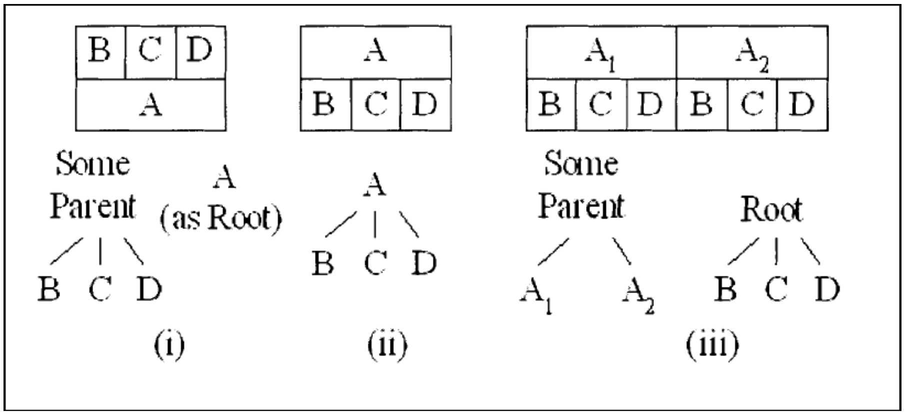


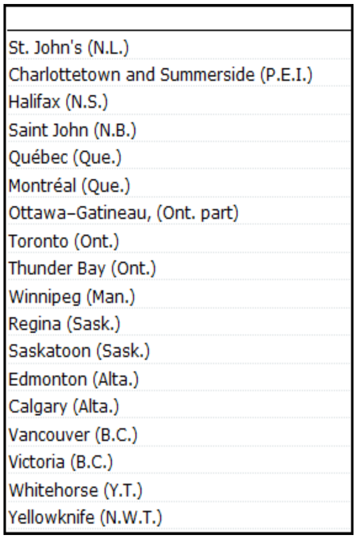
Figure 4.5: Extracting the dimensions from a column header [27]

In the case of (i), we consider the two rows to contain members from different dimensions. In (ii), the column header could represents one-dimension, but two neighboring levels. Thus, the members in the higher row act as a parent for the members in the lower row. Nevertheless, (ii) does not always represent the case of two neighboring levels for the same dimension. Additional linguistic processing is needed to determine whether or not they are the same dimension, as described in more detail in Section 4.4. In short, two rows representing two neighboring levels for the same dimension, when, the lower level members have their dimension label taken from the higher level member. This will confirm that the two levels are from the same dimension.

Each row represents a dimension, except when a relation exists between two neighboring rows. In some cases, factoring the row is used when redundancy occurs in the members of the same row, so that those members should be merged into a single dimension. For such cases, no relation can exist between the redundant group in a row and the members in the higher row, since this group of members would appear under two or more different parents (for example, see case (iii) in Figure 4.5). In case (iii), A1 and A2 have identical children nodes, i.e., B, C, and D. Then, the new dimension will consist of these three members, with unknown dimension label. A label will be assigned for the dimension, as shown in Section 4.4. We will call this dimension a cross-product dimension, as it is embedded into another relation.

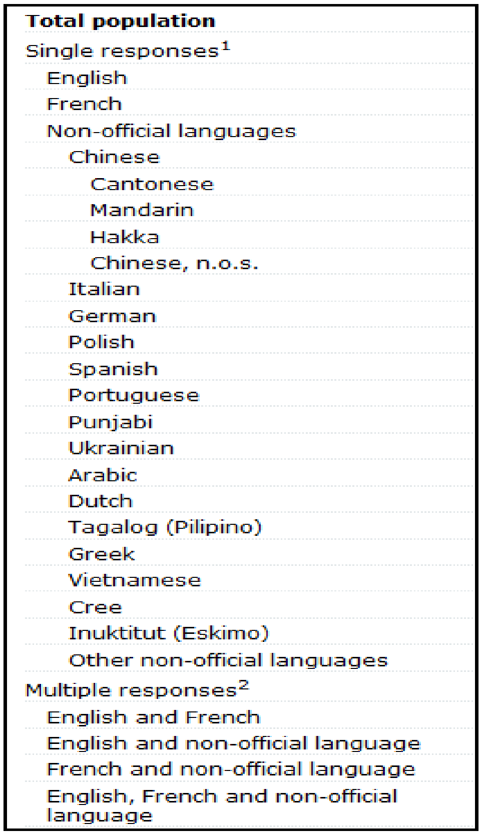
4.3.2 Deriving the Dimensions from the Row Header

In statistics tables, the row header appears in the leftmost column of the table. It includes other dimensions that are used to represent the numeric data. It may be a single-level hierarchy, where the data forms a list (Figure 4.6), or a multi-level hierarchy (Figure 4.7). The hierarchy may indicate that the data is part of the same dimension, but at different levels. When the hierarchy contains group of members with different visual clues, then the lower-level group is related to the higher-level group by the member that appears directly above the lower group, we call it the *dimension header*. This is always the case in the row header hierarchy, except when factoring is used, to be described below. To identify the relationship between two consecutive rows, we compare them visually. If they are the same, then they are at the same level and part of the same dimension. Otherwise, they could be from different neighboring levels of the same dimension or from different dimensions (cross-product dimension).



St. John's (N.L.)
Charlottetown and Summerside (P.E.I.)
Halifax (N.S.)
Saint John (N.B.)
Québec (Que.)
Montréal (Que.)
Ottawa–Gatineau, (Ont. part)
Toronto (Ont.)
Thunder Bay (Ont.)
Winnipeg (Man.)
Regina (Sask.)
Saskatoon (Sask.)
Edmonton (Alta.)
Calgary (Alta.)
Vancouver (B.C.)
Victoria (B.C.)
Whitehorse (Y.T.)
Yellowknife (N.W.T.)

Figure 4.6: Single level dimension

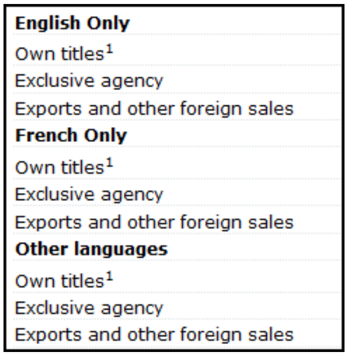


- **Total population**
- Single responses[1]
  - English
  - French
  - Non-official languages
    - Chinese
      - Cantonese
      - Mandarin
      - Hakka
      - Chinese, n.o.s.
    - Italian
    - German
    - Polish
    - Spanish
    - Portuguese
    - Punjabi
    - Ukrainian
    - Arabic
    - Dutch
    - Tagalog (Pilipino)
    - Greek
    - Vietnamese
    - Cree
    - Inuktitut (Eskimo)
    - Other non-official languages
- Multiple responses[2]
  - English and French
  - English and non-official language
  - French and non-official language
  - English, French and non-official language

Figure 4.7: Multi-level dimension

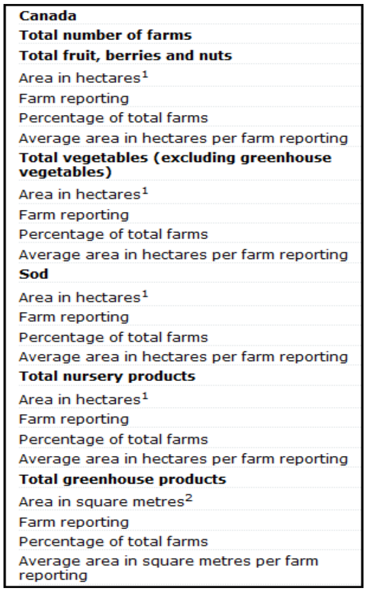
4.3.2.1 Cross-Product dimensions (Factoring)

Once a hierarchy for the row header is set up, we will look for cross-production dimensions as it is also the case for the column header (see Section 4.3.1). For example, there are two dimensions in the row header section of the table header in Figure 4.8, the set of members, i.e., Own titles, Exclusive agency, Exports and other foreign sales are the member of the cross-product dimension. However, there are cases that the sets of members are not identical for all leaf nodes of the original hierarchy, which are the nodes that appear under group of members with the same level. Consider the table header in Figure 4.9. The set of members under the leaf node 'Total greenhouse products" is slightly different from other sets under other leaf node.



English Only
Own titles[1]
Exclusive agency
Exports and other foreign sales
French Only
Own titles[1]
Exclusive agency
Exports and other foreign sales
Other languages
Own titles[1]
Exclusive agency
Exports and other foreign sales

Figure 4.8: Full redundancy hierarchy



Canada
Total number of farms
Total fruit, berries and nuts
Area in hectares[1]
Farm reporting
Percentage of total farms
Average area in hectares per farm reporting
Total vegetables (excluding greenhouse vegetables)
Area in hectares[1]
Farm reporting
Percentage of total farms
Average area in hectares per farm reporting
Sod
Area in hectares[1]
Farm reporting
Percentage of total farms
Average area in hectares per farm reporting
Total nursery products
Area in hectares[1]
Farm reporting
Percentage of total farms
Average area in hectares per farm reporting
Total greenhouse products
Area in square metres[2]
Farm reporting
Percentage of total farms
Average area in square metres per farm reporting

Figure 4.9: Partial redundancy hierarchy

To detect the redundant header along the hierarchy, we compare each group in the hierarchy with other groups at the same level. If two groups are 50% or more identical in their members, and the number of members is more than two, we join them together and identify them as a single dimension, separate from the other parts of the hierarchy.

The cross-product dimensions in the row header are clear in the above cases. Nevertheless, it is not clear when the first row in row header is representing a dimension with a single member in the table and the lower level group is representing other dimensions. Recall the extended table model discussed in Section 3.2, where there is a dimension for certain study is not full classified in this table and part of the classification for this dimension for the same study in other table. Therefore, the special links will lead to this dimension and to the other table for the integration purpose. The integration is discussed in more detail in Section 4.7. By assigning the label for each dimension placed in the header, the cross-product dimensions will be clearly distinguished. This is because each dimension will have its own label which is not related to each other. In section 4.4, labeling dimensions will be discussed in more detail.

4.3.2.2 Generalization–Specialization Dimension

Generalization–specialization indicates that the data presented in the hierarchy at different visualization levels is actually at different levels within the same category (see Figure 4.7). The upper level is the generalized version of the lower level, so that the data represented in the higher row is a summary for what is beneath. Detecting whether or not a group of members is a generalized version of another group requires linguistic processing (see Section 4.4), by recognizing whether or not the lower group gets its label from the higher group. We need to understand the meaning of the members to confirm that the higher level is truly representing the lower group with a generalized word (i.e., can the higher level member represent the lower level members as a label).

4.4 Labeling Dimensions

As mentioned above, the statistical tables and other multidimensional tables, designed to represent a certain study using different dimensions, which are defined in the tables' headers. Those dimensions in the headers are listed without any labels. The dimension labels are very important, as they ensure the cube is designed accurately. They can also help the user to search for specific information within the cube. The labels are commonly found in specific locations, as listed in the following section.

4.4.1 Dimension Label Locations

The HTML page that contains the multidimensional table has the information needed to help readers to understand the table. Therefore, the dimensions listed in the tables are usually introduced on the same webpage. Below are the common locations that have the dimension labels.

4.4.1.1 Table Title

As illustrated in Section 3.1.1, table titles usually present a list of dimension labels. Therefore, the dimensions need to match up with elements of the title. For most table titles, it is obvious where the dimension labels are listed. There are special prepositions such as: 'by', means that anything following it will be a selection of dimension labels. That said, any part of the table title cannot be assigned directly to be the dimension label because there are variety of table titles, where some of them does not include all the dimension labels, the dimension labels listed without any specific order (Section 3.1.1.1), and the table contains more than one dimension. Therefore, there is a need to compare each dimension with each part of the title in order to assign the appropriate label.

The supplemental part of the table title introduces in the extended table model, Section 3.2, will not be considered in the label assignment. This is because the labels assignment will be difficult due to the possibility of duplicating part of the table title and the supplement. Matches could exist both within the title and within the supplemental details. Therefore, only the table title will be checked without including the supplement information.

It is not always the case that we can find the appropriate dimension labels in the table title, as discussed in Section 3.1.1.1 regarding the variety of table titles. Therefore, there is a need to apply other heuristics that make it possible to find the dimension labels, for example, the dimension header.

4.4.1.2 Dimension Header

Some dimensions have a dimension header, which is assigned from an upper dimension that appears in the row header or in the column header, i.e., case (ii) in Figure 4.5. Occasionally, this dimension header can serve as the label for that dimension. Furthermore, it is also crucial to identifying the relationship between dimensions. If the dimension header is a label, then this group is in the Generalization-Specialization style. Otherwise, it might be part of other dimensions. Thus, we must confirm whether the dimension header serves as a label or not. There are five possible dimension headers cases, as shown in Table 4.1 and illustrated below, and some of these can lead to the dimension label.

1. **Other Dimension**

   As illustrated in Section 4.3.2.1, the dimension header can be part of another dimension that is not related to the current dimension, i.e., they are cross-product dimensions. In such instances, the dimensions have different labels that can be assigned using either the table title or other heuristics. This case always presents in the extended table model, where the dimension is not fully classified in a single table.

2. **Aggregate Word Only**

   The header can represent an aggregate member of the same dimension. The aggregate member could be detected by a keyword pattern. These keywords include terms such as ‘total’ or ‘all’. This type of header is not useful to the associated dimension; it will not provide any clues as to what the dimension label should be. Thus, it serves as a summary for that dimension and to show the hierarchy.

3. **Aggregate Word and Dimension Label**

   This dimension header is an aggregate word attached to the dimension label. Some statistics tables use this structure to present a summary of the information contained within the dimension, as in "Aggregate Word Only" above. One example might be *'All causes of death'*; all causes of death would contain totals pertaining to an aggregate of all the different types of causes. In the following rows, data for individual types of causes are listed separately. This is the Generalization-Specialization type. The dimension containing the dimension header is the generalized information for the dimension containing the specialized information. Furthermore, the generalized member acts as the dimension label for the lower level.

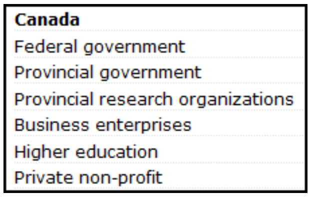
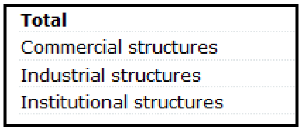
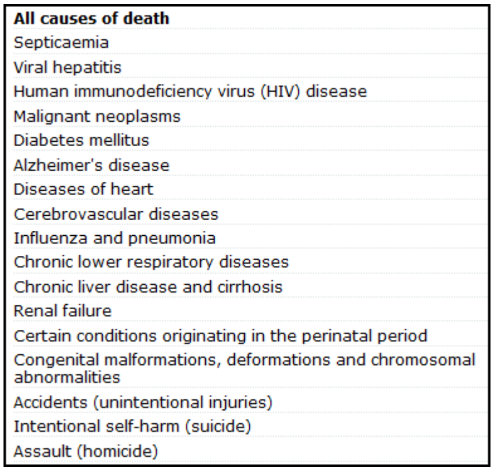
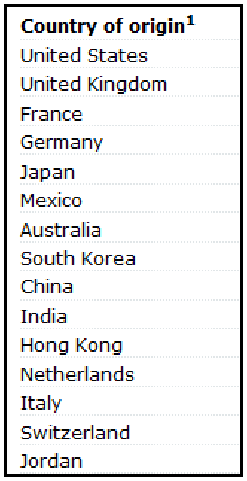
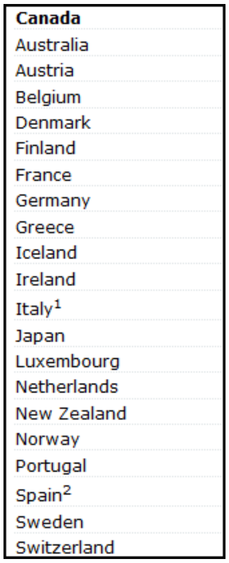
4. **Dimension Label Only**

   This dimension header contains only the dimension label for the dimension to which it belongs. In most cases, that row is not associated with any numeric data. It is only used to clarify the dimension and to make the table header look more meaningful.

5. **Member of the Same Dimension**

   In some unusual cases, the dimension header is in fact a member of the dimension it belongs to and is not the generalization member of that dimension. Therefore, the two dimensions should be merged. This case will be detected when the dimension header does not act as the label for the dimension. Thus, they both carry the same dimension label.

Table 4.1: Different kinds of row headers showing the relation between dimension header and dimension

Other Dimension	Aggregate Word Only	Aggregate Word And Dimension Label	Dimension Label Only	Member to the same dimension
 **Canada** Federal government Provincial government Provincial research organizations Business enterprises Higher education Private non-profit	 **Total** Commercial structures Industrial structures Institutional structures	 **All causes of death** Septicaemia Viral hepatitis Human immunodeficiency virus (HIV) disease Malignant neoplasms Diabetes mellitus Alzheimer's disease Diseases of heart Cerebrovascular diseases Influenza and pneumonia Chronic lower respiratory diseases Chronic liver disease and cirrhosis Renal failure Certain conditions originating in the perinatal period Congenital malformations, deformations and chromosomal abnormalities Accidents (unintentional injuries) Intentional self-harm (suicide) Assault (homicide)	 **Country of origin[1]** United States United Kingdom France Germany Japan Mexico Australia South Korea China India Hong Kong Netherlands Italy Switzerland Jordan	 **Canada** Australia Austria Belgium Denmark Finland France Germany Greece Iceland Ireland Italy[1] Japan Luxembourg Netherlands New Zealand Norway Portugal Spain[2] Sweden Switzerland

By knowing the different types of dimension headers, we can ensure that we will not assign them as dimension labels without further processing. Using this information, we need to develop a testing method so that we can properly determine the relationship between dimensions and the dimension header, and whether the dimension header should also be used as the dimension label.

4.4.1.3 Special Keywords Pattern

Some dimensions may contain special kinds of members, for example, time, measurement units and age. These members can easily be detected because we can understand what they represent directly without having to perform any further processing. These types of members may not have any linguistic meaning, as they are numbers or special characters. For this reason, we can identify them first by using a special pattern while extracting the dimension. The pattern can be a range defined by numbers, such as years or days. It can be also a set of keywords, such as a list of months. After identifying the dimensions, they can be directly assigned the appropriate label. This process should be done whenever we extract a new dimension from the table. However, after assigning the appropriate keyword, that keyword need to be compared to the table title and to the dimension header, as discussed previously, to identify a more specialized phrase that is related to the specific dimension. For example, the word *period* could appear in the table title, so this word would be assigned instead of *time*. This process also helps in singling out the measure dimension that contains the measure unit from the column header.

4.4.1.4 External Domains

All statistics tables have some common dimensions. In fact, the members of these dimensions can be exactly the same in different tables; or in other words, two tables could have the same dimension with exactly the same members. In other cases, the same dimension can appear in two tables, but their members may be written in different ways. A good example of this is two dimensions that represent *Canadian provinces*. One of the dimensions lists the complete names of each province, while the other dimension lists the abbreviations of each province. A further dimension could also list the provinces with both their full names and their abbreviations.

This type of dimension is very popular in our domain, and the majority of statistics tables have those common dimensions. Thus, these dimensions are collected in special storage each time the system identifies a new dimension, i.e. the domain storage in Figure 4.1. When there is a dimension that fails to identify its dimension label using the table title, the dimension header or keywords, we will attempt to determine the intersection between the dimension and the integrated domain in the domain storage. Afterwards, if the intersection is determined to be a 70% match of the group, then this dimension will be assigned the same label as that assigned to the integrated group. Domain integration will be discussed in more detail in Section 4.8.

4.4.1.5 Special Cases

If none of the above processes can determine the label for a dimension, then we will assign the dimension label by one of the following procedures:

- If dimensions have a dimension header, then this dimension will be assigned directly to the dimension header even if it is fail the process.
- The common word in the dimension members, if it is not a stop word and the number of dimension members is not large.
- The common ancestor retrieved according to the method described in Section 4.4.2.2, which can represent most of the members in the dimension.

4.4.2 The Process of Assigning the Dimension Label

Learning what a dimension member means is much more important than how it is presented within the table. Dimension members within the same group should always share the same dimension label. Understanding the meaning of the dimension members will help us to identify the relations between dimensions contained within the same hierarchy and therefore their labels. To find the dimension meaning or a common word that represents them, the dimension members are extracted in a preceding step. This is done by collecting the dimensions from the row or column headers that have the same visual clues, discussed in Section 4.3, and by collecting the members from the dimensions in special links, discussed in Section 4.4.2.1. The process then finds the ancestor list and compares it to the table title and the dimension header. If a match occurs with the table title, the best phrase will be assigned. If, however, no match occurs with the table title and the dimension header, the process will use the other heuristics (discussed previously), with external domains; otherwise, the special cases will be used.

4.4.2.1 Collecting Dimensions in Special Links

As illustrated previously in Section 3.2, in some cases, one part of a dimension is represented in the table, while other parts appear in separate tables. Clues are always available to locate and address these other parts. However, representing part of a dimension in one table can result inaccurate dimension label because this is insufficient information. For example, if there is a member with one province in abbreviation or word—for example, *Quebec*—it is not possible to use this information to assign a dimension label; this is because we do not know whether the table means *Quebec* as a city or as a province. However, suppose that part of the dimension contains *Quebec* and *Alberta*. In that sense, it would be possible to identify the dimension label. Therefore, finding the full list is mandatory to assign the title to this dimension.

To find these lists, the method is to search for <ul><li> tags, radio buttons, and a list box that could appear around the table in HTML format. If one of those lists is found, then we check whether a member in the table appears as a member of that list. Then, it will be possible to confirm this list is related to the table and test this conclusion to find the appropriate dimension label.

4.4.2.2 Finding the Sorted Ancestor List

Each word in a dimension member is represented within the WordNet taxonomy. “WordNet is a lexical database that is available online and provides a large repository of English lexical items” [50]. Words from the same dimension should share a common ancestor. The more members are in a dimension, the harder it is to retrieve the least common ancestor. For example, there could be a precise common ancestor that exists for 70% of all members; there could also be a common ancestor that exists for 100% of the members, but it is much vaguer. There are two important factors when choosing a common ancestor: the average distance between the common ancestor of each word within the dimension, and how many times this ancestor appears in the members list of that dimension. The combination of these factors is called the common ancestor score. The common ancestor formula is: common ancestor score = (largest distance between this ancestor and any member in the dimension + smallest distance between this ancestor and any member in the dimension)/2 * (the total number of members in the dimension - the number of members retrieving this ancestor + 1).

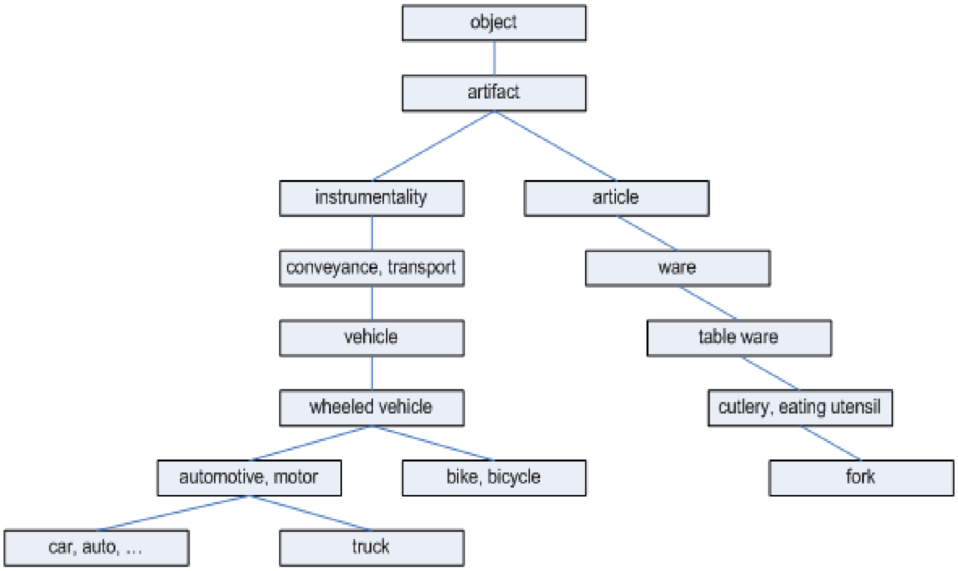
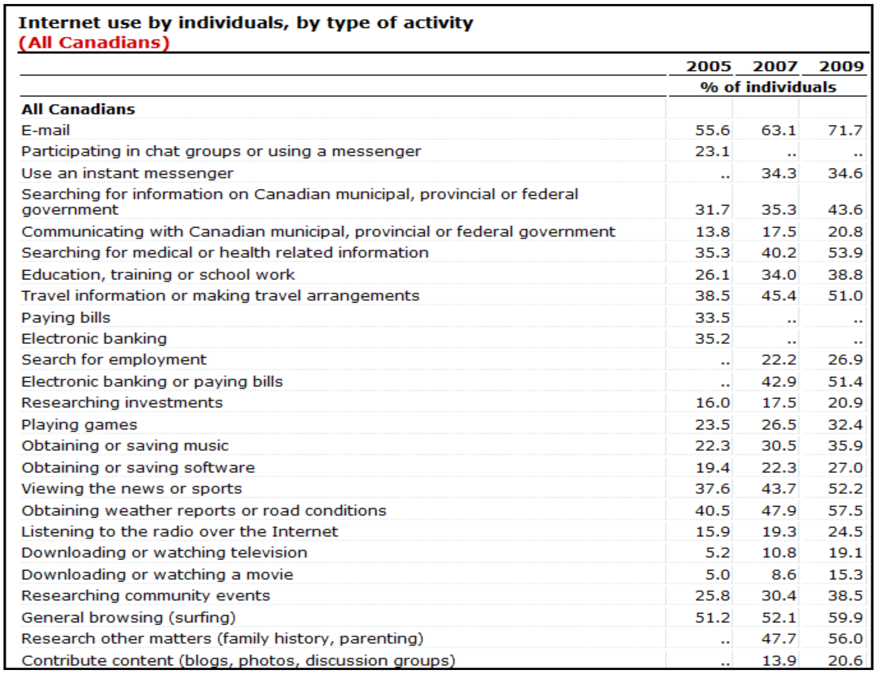


Figure 4.10: Word-Net Taxonomy [51]

Let us suppose that a dimension contains the members {car, truck, bike}; as seen in Figure 4.10, these words share the common ancestor *wheeled vehicle*. Suppose that this dimension also contained the word *fork*, which does not make sense but is chosen for the sake of this example. Then the common ancestor will be *artifact*, which is too vague. Therefore, we will consider placing the 70% ancestor match on top of the ancestor list rather than the 100% match. This type of decision is our goal when we construct the formula.

Some dimension members are phrases. Therefore, it can be difficult to collect the common ancestor for these phrases, as we cannot be sure which word in the phrase is the most closely related to this dimension. In this case, we will retrieve the ancestors for each word in the phrase. However, if two words from the same phrase retrieve the same ancestor, then the one that has the smallest distance from the word will be picked.

Table 4.2 shows the ancestors and their scores corresponding to the dimension *Type of Activity* retrieved from Figure 4.11. It also shows the best word matched from the table title by measuring the semantic similarity between the ancestor and part of the table title.



Internet use by individuals, by type of activity
(All Canadians)

	2005	2007	2009
	% of individuals		
All Canadians			
E-mail	55.6	63.1	71.7
Participating in chat groups or using a messenger	23.1	..	..
Use an instant messenger	..	34.3	34.6
Searching for information on Canadian municipal, provincial or federal government	31.7	35.3	43.6
Communicating with Canadian municipal, provincial or federal government	13.8	17.5	20.8
Searching for medical or health related information	35.3	40.2	53.9
Education, training or school work	26.1	34.0	38.8
Travel information or making travel arrangements	38.5	45.4	51.0
Paying bills	33.5	..	..
Electronic banking	35.2	..	..
Search for employment	..	22.2	26.9
Electronic banking or paying bills	..	42.9	51.4
Researching investments	16.0	17.5	20.9
Playing games	23.5	26.5	32.4
Obtaining or saving music	22.3	30.5	35.9
Obtaining or saving software	19.4	22.3	27.0
Viewing the news or sports	37.6	43.7	52.2
Obtaining weather reports or road conditions	40.5	47.9	57.5
Listening to the radio over the Internet	15.9	19.3	24.5
Downloading or watching television	5.2	10.8	19.1
Downloading or watching a movie	5.0	8.6	15.3
Researching community events	25.8	30.4	38.5
General browsing (surfing)	51.2	52.1	59.9
Research other matters (family history, parenting)	..	47.7	56.0
Contribute content (blogs, photos, discussion groups)	..	13.9	20.6

Source Statistics Canada: http://www40.statcan.gc.ca/l01/cst01/comm29b-eng.htm

Figure 4.11: Extracting best dimension label match from the table title

Table 4.2: Dimension members common ancestors and their best match from table title

Common Ancestor	Common Ancestor Score	Common Ancestor Counter	Best match from table title
Act	10.5	26	activity
Make	34.5	6	type
Use	35.5	5	Use
Activity	36	21	activity
Search	37.5	4	activity

At the end of this step, the sorted list of common ancestors and their corresponding scores for each dimension will be retrieved. However, only those ancestors that appear for 70% of the members are retained; the rest will be eliminated. Then, in the next step, this sorted list will be compared with the table title and the dimension header to decide the accurate label.

Occasionally, dimension members which belong to a single dimension are listed directly in the table title without any common word or label to represent them, case 4 in Section 3.1.1.1, therefore, if the number of members are less than 5, it will be acceptable to be listed in the title, a sliding window techniques will be used to compare each member with part of the title. If match found for majority of the members then there is no need to proceed to assign a single label and find ancestor list, the combination of the dimension members is the dimension label.

4.4.2.3 Comparing the Common Ancestor Sorted List with Testing Collection

As mentioned earlier, the dimension header and the table title need to be compared with the ancestor sorted list in order to assign the dimension label. The dimension header, if applicable, is always compared first with the ancestor list because it is nearer than the table title and is more likely to be the dimension label, especially with sub-dimensions that appear in the Generalization-Specialization case. We call the temporary collection of words that include the dimension header and the table title the *testing collection*. The testing collection contains a list for each word from the dimension header and table title. We compare semantically each word in the testing collection with the ancestor list. If we find a match between an ancestor and the dimension header, we stop. As the list of ancestors retrieved is comprised of single words, it is compared with words from the testing collection. In most cases, the highest-scoring ancestor is very similar to a word in the testing collection, but it is usually not an exact match. Thus, the comparison of the testing word with the chosen ancestor will be in a semantic, and not in a syntactic fashion.

We use the semantic score presented in "WordNet-based semantic similarity measurement" work [52] to measure the similarity in meaning between the retrieved ancestor word and the testing word. The score is usually assigned by rational numbers between 0 and 1. Along this scale, '0' means that no similarity exists between the two words, and '1' means that the two words are synonyms or exactly the same. To simplify the process, we set a threshold of 0.9 to determine whether a word in the testing collection is similar enough to an ancestor word. When a word passes this threshold, it is retained in the top 5 match list.

We must compare the ancestor-sorted list with the testing collection based on the order of highest-scoring ancestors to lower-scoring ancestors. It is not mandatory that the top ancestor should match any words within the testing collection. We may not find a match within the highest-scoring ancestor; in fact, we may instead identify a match within the middle-ranked ancestor. Therefore, we must continue comparing until we distinguish the top five matches from the testing collection. Those matches could be from different words in the testing collection.

The words that are part of the table title could be the dimension label for any dimension, unlike the dimension header, which only belongs to only one dimension because it is only related visually to one dimension. Therefore, we must confirm that we are assigning the right part of the title to the corresponding dimension. This is because, for each dimension, we will retrieve the top five matches that correlate with the ancestor and with the table title. Those matches could link different parts of table title, as seen in the example in

Table 4.2. We notice that not all the top ancestors retrieve *activity* as the dimension label. Therefore, there are two possible cases when retrieving the match.

1- Top Five Matches from the Ancestor List Match a Specific Word from the Testing Collection.

This part of the title is a 100% match to the dimension. Therefore, this word from the testing collection will be assigned as the initial dimension label.

2- Top Five Matches from the Ancestor List Match Different Words from the Table Title

After examining all the dimensions to retrieve the top 5 match list and before assigning the final label, we will check each match list result retrieved for each dimension. If one of those dimensions has its top 5 ancestor list matches different words of the table title, as seen in Table 4.2, we will assign the first match as the dimension label, after making sure that it was not assigned to any other dimension. If it was already assigned to other dimension, then we will assign the second match.

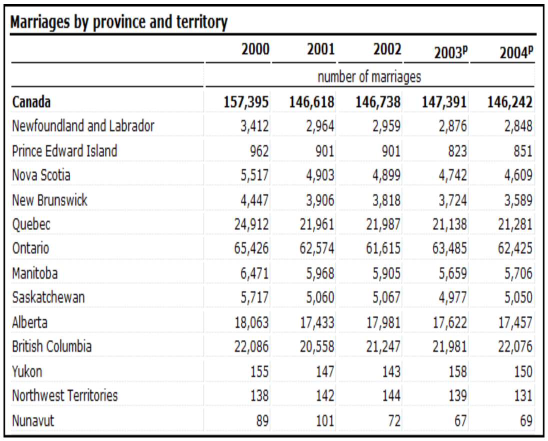
4.4.2.4 Assigning the Best Phrase from the Title

When the ancestor matches a word in the dimension header, then the dimension label will be the entire dimension header phrase. However, this is not the case when assigning a dimension from the table title. This is because it contains many phrases or words belong to different dimensions and in some cases, it is not enough to represent the dimension by only a single word. Therefore, we must retrieve a phrase from the table title that contains the matched word. This procedure is done by retrieving the part of the phrase that contains the matched word and that starts and ends with the punctuation or preposition. Then, the best possible judgment is made for the words before, and after, the preposition and they are checked for their relation, taking into account that they are not assigned to any other dimension.

4.5 Identifying the Measure Dimension

The measure dimension, which includes the measure unit, usually appears in the last row of the column header, or occasionally, in a spanning row in the middle of the table, in the case of tables that have more than one measure. Finding the measure unit dimension is done by checking those locations to find special characters; i.e., measure unit keyword pattern. Nevertheless, the measure dimension may not be sufficient for representing the measure in the table. Since it could have only a measure unit without any description of what is being measured, as in Figure 4.13. Therefore, we should find and attach the measure description to the measure unit to make the cube representation clearer.

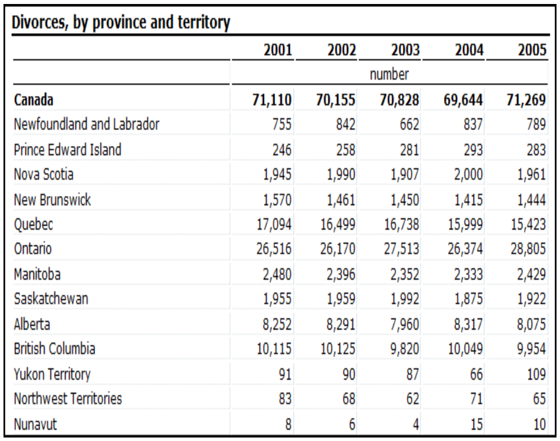
In Figure 4.12 and Figure 4.13, two statistics tables are shown. One contains the measure attached to the measure unit *number of marriages*, and the other shows only the measure unit by itself *number*.



Marriages by province and territory

	2000	2001	2002	2003^{p}	2004^{p}
	number of marriages				
Canada	**157,395**	**146,618**	**146,738**	**147,391**	**146,242**
Newfoundland and Labrador	3,412	2,964	2,959	2,876	2,848
Prince Edward Island	962	901	901	823	851
Nova Scotia	5,517	4,903	4,899	4,742	4,609
New Brunswick	4,447	3,906	3,818	3,724	3,589
Quebec	24,912	21,961	21,987	21,138	21,281
Ontario	65,426	62,574	61,615	63,485	62,425
Manitoba	6,471	5,968	5,905	5,659	5,706
Saskatchewan	5,717	5,060	5,067	4,977	5,050
Alberta	18,063	17,433	17,981	17,622	17,457
British Columbia	22,086	20,558	21,247	21,981	22,076
Yukon	155	147	143	158	150
Northwest Territories	138	142	144	139	131
Nunavut	89	101	72	67	69

Source: Statistics Canada:
http://www40.statcan.gc.ca/l01/cst01/famil04-eng.htm?sdi=marriages



Divorces, by province and territory

	2001	2002	2003	2004	2005
	number				
Canada	**71,110**	**70,155**	**70,828**	**69,644**	**71,269**
Newfoundland and Labrador	755	842	662	837	789
Prince Edward Island	246	258	281	293	283
Nova Scotia	1,945	1,990	1,907	2,000	1,961
New Brunswick	1,570	1,461	1,450	1,415	1,444
Quebec	17,094	16,499	16,738	15,999	15,423
Ontario	26,516	26,170	27,513	26,374	28,805
Manitoba	2,480	2,396	2,352	2,333	2,429
Saskatchewan	1,955	1,959	1,992	1,875	1,922
Alberta	8,252	8,291	7,960	8,317	8,075
British Columbia	10,115	10,125	9,820	10,049	9,954
Yukon Territory	91	90	87	66	109
Northwest Territories	83	68	62	71	65
Nunavut	8	6	4	15	10

Source: Statistics Canada:
http://www40.statcan.gc.ca/l01/cst01/famil02-eng.htm?sdi=divorces

Figure 4.12: "Number of marriages" shows the measure unit with the measure

Figure 4.13: "Number" appears in the table without showing what we are measuring. It should contain the word "divorces"

Because the cube contains both dimensions and measures, we need to identify the appropriate measure-related name to be attached to the measure unit, in cases where the measure does not contain a description. Having an understanding the table titles and how they are formed helps us to identify this measure-related dimension. By observing hundreds of table titles, we found that the first section in the table title usually represents the measure, as described in Section 3.1.1.

4.5.1 Cases for finding the measure dimension

The measure-related words or dimensions are found within the table and/or the table title. Below, are the different cases for deciding the identity of the measure dimension.

1) The measure-related word is already attached to the measure unit; no need for further processing (Figure 4.12):

- This case can be detected by syntactically comparing the words in the table title with the words attached to the measure unit. If a word, excluding a stop word, is found in both the table title and attached to the measure unit, then it most likely represents what the table is measuring.

2) The measure-related words are members in a dimension:

- Occasionally, some dimension members are special words that are measure-related. For example, if terms such as revenue and expenditure are found in a dimension, then that dimension represents the measure. It is not mandatory for all of the words to be measure-related keywords, providing that some of the members contain keywords to determine if they are a measure.
- By extracting the appropriate dimension label for each dimension, one dimension could be assigned to the first part of the table title as the dimension label. Thus, the members are representing a measure-related dimension, and each member should be attached to the corresponding measure unit that has the same row or column coordinates. Nevertheless, we may have the option to flatten them, depending on the number of members appearing.

3) No measure description is found with the unit and no measure-related dimension is found in the table:

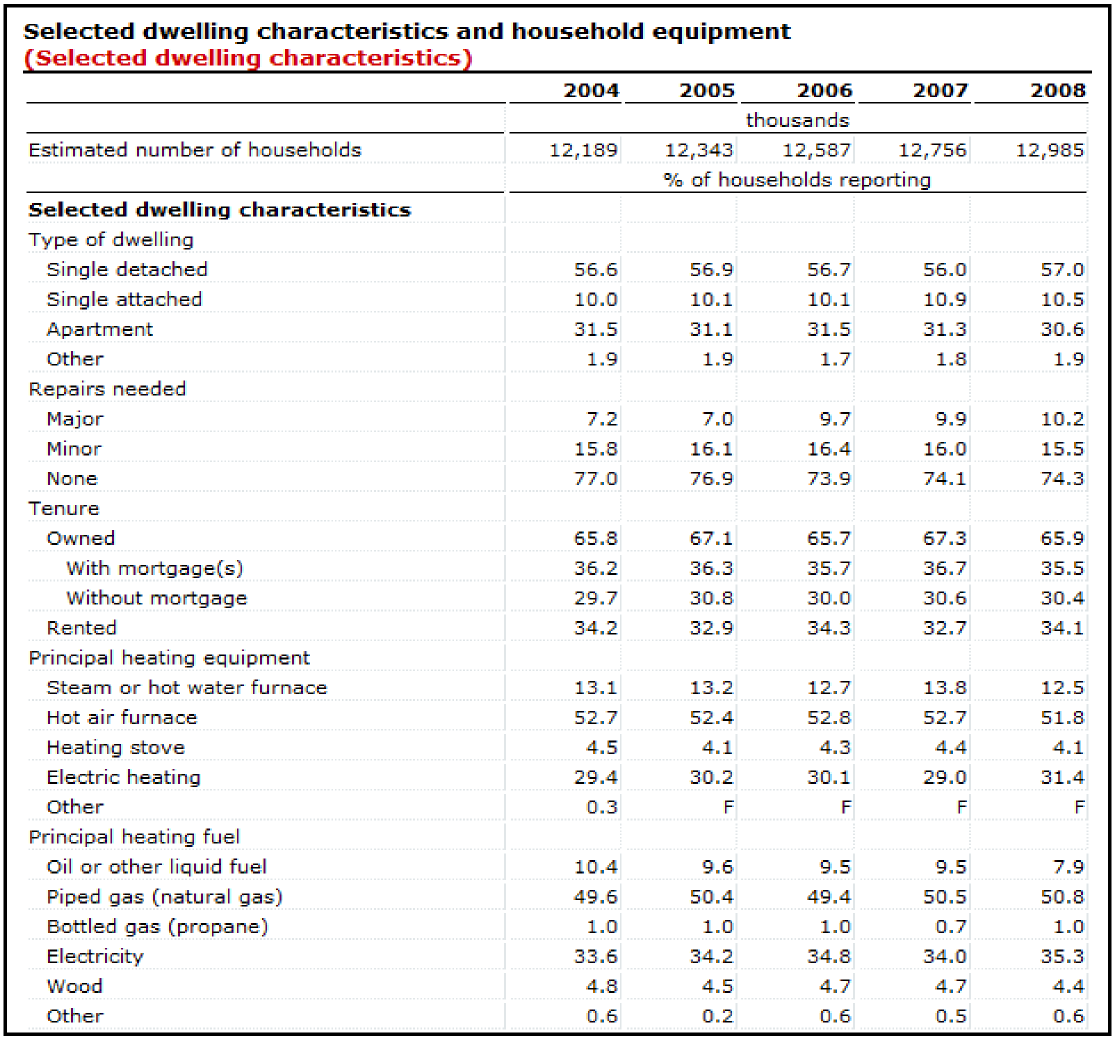
- In this case, as described earlier, the first part of the table title represents the measure, and that part should be attached to all of the measure units that are in the statistics table. (See Figure 4.13); *Divorces* should be merged with number, to give *divorces in number*.

4) No measure unit is found in the table and no measure-related dimension is present:

- In this case, the table designer would have assumed that it is obvious for readers that the unit is number. Therefore, in our method, we will assume that the measure unit is "*number*" and we attach it to the first part of the title.

4.6 Table Partitioning by Dimension Hierarchy

A table may be partitioned horizontally into a number of sections. Each section is, by itself, a table, which inherits the table title and column headers of the parent table. Detecting table sections is usually fairly easy, either visually or by a program, by the presence of thick horizontal lines (i.e., Section 4.2). Partitioning a table into a number of tables is sometimes justified even when no such lines are present. Consider the row header of the table in Figure 4.14, it is obviously not a multidimensional cube with 5 dimensions (plus year dimension), i.e., "*type of dwellings*", "*repair needed*", etc., in the same sense that the table in Figure 4.8 is a table with 2 dimensional cube. The former one cannot find out the percentage of dwellings which are single-detached and need minor-repair, wherein in the latter, one can easily find out the data about a member of one dimension and another member of another dimension, i.e., "French Only" and "Exports and other foreign sales". This is the essence of summarizability of OLAP cubes and statistical tables [53]. The table in Figure 4.14 shows a non-linear hierarchy, i.e., a member can roll-up to different parents. In this case, a dwelling can be classified in 5 different ways. The summarizability cannot be maintained by non-linear dimension hierarchies. Thus the table should be partitioned into 5 different, two-dimensional tables.



Selected dwelling characteristics and household equipment
(Selected dwelling characteristics)

	2004	2005	2006	2007	2008
			thousands		
Estimated number of households	12,189	12,343	12,587	12,756	12,985
			% of households reporting		
Selected dwelling characteristics					
Type of dwelling					
Single detached	56.6	56.9	56.7	56.0	57.0
Single attached	10.0	10.1	10.1	10.9	10.5
Apartment	31.5	31.1	31.5	31.3	30.6
Other	1.9	1.9	1.7	1.8	1.9
Repairs needed					
Major	7.2	7.0	9.7	9.9	10.2
Minor	15.8	16.1	16.4	16.0	15.5
None	77.0	76.9	73.9	74.1	74.3
Tenure					
Owned	65.8	67.1	65.7	67.3	65.9
With mortgage(s)	36.2	36.3	35.7	36.7	35.5
Without mortgage	29.7	30.8	30.0	30.6	30.4
Rented	34.2	32.9	34.3	32.7	34.1
Principal heating equipment					
Steam or hot water furnace	13.1	13.2	12.7	13.8	12.5
Hot air furnace	52.7	52.4	52.8	52.7	51.8
Heating stove	4.5	4.1	4.3	4.4	4.1
Electric heating	29.4	30.2	30.1	29.0	31.4
Other	0.3	F	F	F	F
Principal heating fuel					
Oil or other liquid fuel	10.4	9.6	9.5	9.5	7.9
Piped gas (natural gas)	49.6	50.4	49.4	50.5	50.8
Bottled gas (propane)	1.0	1.0	1.0	0.7	1.0
Electricity	33.6	34.2	34.8	34.0	35.3
Wood	4.8	4.5	4.7	4.7	4.4
Other	0.6	0.2	0.6	0.5	0.6

Source: Statistics Canada: http://www40.statcan.gc.ca/l01/cst01/famil09a-eng.htm?sdi=dwelling

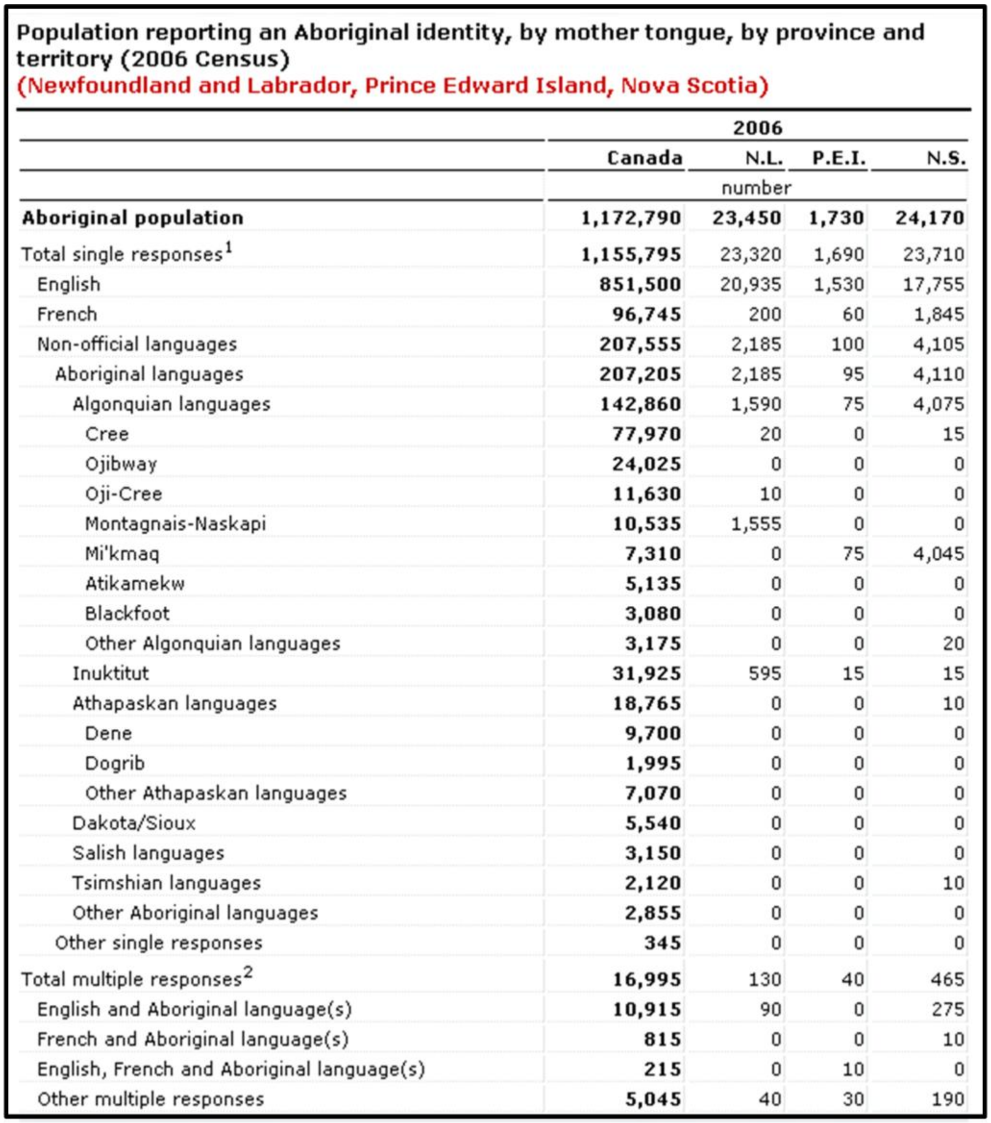
Figure 4.14: Table needs to be partitioned

There are some clues that can detect this kind of hierarchy and determine if it need to be partitioned. First rule that we rely on is the summation rule. However, if the summation rule was not sufficient, the dimension member ancestor will be tested, although the results are not accurate as the summation rule decision. In the next sections, we discuss how the decision can be made numerically and linguistically.

4.6.1 Summation Rule

The data presented in a statistics table is in numbers; they could be population, percentage, or average, etc. The numbers are useful for determining the relationships among the different levels in the headers. In most cases, the table represents the data, and at the same time, has special rows that summarize the results for a group of members. Usually, the summary does not help us to assign the appropriate dimension label, though it could help us to decide whether or not the group members are part of the same dimension. Furthermore, if all dimensions in the same header are summed up and their result is equal, the table should be partitioned to ensure summarizability for each cube.

Let us separate each group of lines in Figure 4.14, beginning at the leftmost position, followed by a number of indented lines; e.g., the line 'type of dwelling' followed by four other indented lines. The sum of the percentages under the column '2004' (associated with the four lines) is 100%. As a result, they should be segregated into separate tables. On the other hand, in Figure 4.15, the single response and the multiple responses should not be divided. As it is clear by the values in the first numeric data column that the total for single response plus multiple response are equal to the total population. This means this is a liner hierarchy and each single cell is counted only once.



Population reporting an Aboriginal identity, by mother tongue, by province and territory (2006 Census)
(Newfoundland and Labrador, Prince Edward Island, Nova Scotia)

	2006			
	Canada	N.L.	P.E.I.	N.S.
	number			
Aboriginal population	**1,172,790**	**23,450**	**1,730**	**24,170**
Total single responses[1]	**1,155,795**	23,320	1,690	23,710
English	**851,500**	20,935	1,530	17,755
French	**96,745**	200	60	1,845
Non-official languages	**207,555**	2,185	100	4,105
Aboriginal languages	**207,205**	2,185	95	4,110
Algonquian languages	**142,860**	1,590	75	4,075
Cree	**77,970**	20	0	15
Ojibway	**24,025**	0	0	0
Oji-Cree	**11,630**	10	0	0
Montagnais-Naskapi	**10,535**	1,555	0	0
Mi'kmaq	**7,310**	0	75	4,045
Atikamekw	**5,135**	0	0	0
Blackfoot	**3,080**	0	0	0
Other Algonquian languages	**3,175**	0	0	20
Inuktitut	**31,925**	595	15	15
Athapaskan languages	**18,765**	0	0	10
Dene	**9,700**	0	0	0
Dogrib	**1,995**	0	0	0
Other Athapaskan languages	**7,070**	0	0	0
Dakota/Sioux	**5,540**	0	0	0
Salish languages	**3,150**	0	0	0
Tsimshian languages	**2,120**	0	0	10
Other Aboriginal languages	**2,855**	0	0	0
Other single responses	**345**	0	0	0
Total multiple responses[2]	**16,995**	130	40	465
English and Aboriginal language(s)	**10,915**	90	0	275
French and Aboriginal language(s)	**815**	0	0	10
English, French and Aboriginal language(s)	**215**	0	10	0
Other multiple responses	**5,045**	40	30	190

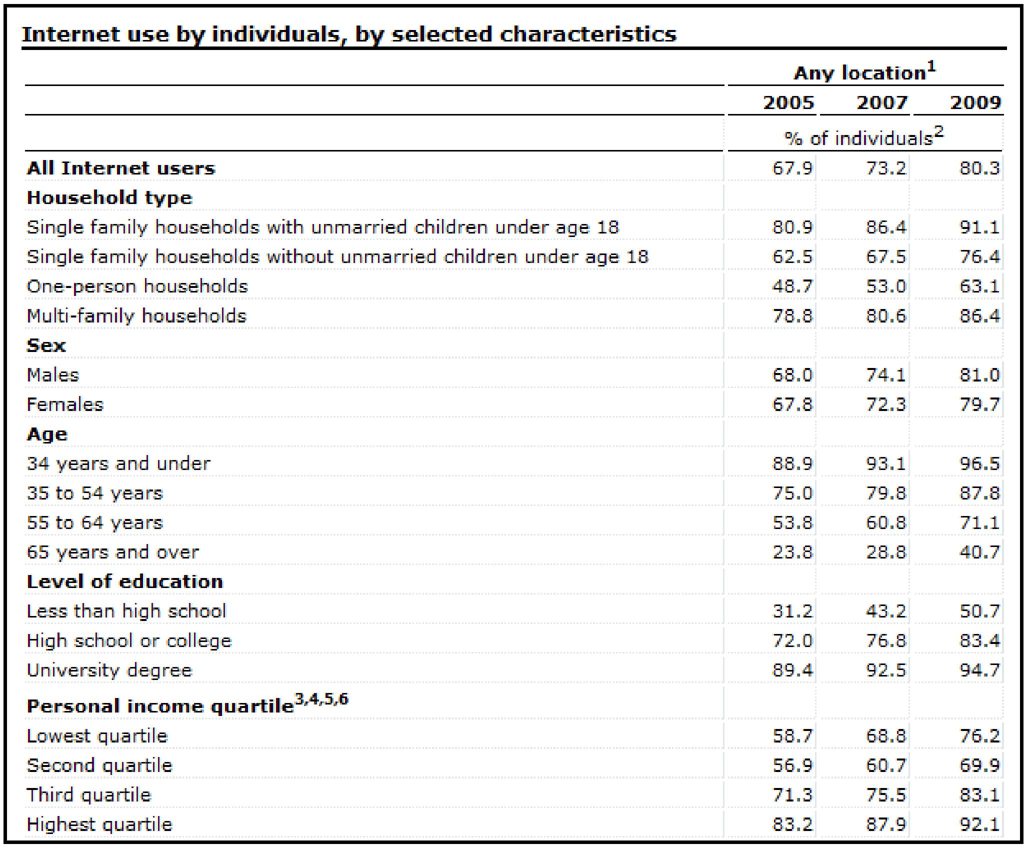
Source: Statistics Canada: http://www40.statcan.gc.ca/l01/cst01/demo38a-eng.htm

Figure 4.15: Statistics table where the summation rule work

When applying the summation rule, it is sufficient to use the first column that holds the numeric values, since the other columns should have the same distribution of numeric values. Nevertheless, the summation rule does not always provide for the partitioning decision, since the relation between the numbers in the summation functions may not hold. The numbers could be in other equations not easily detected. Therefore, a linguistics decision is needed by comparing between the dimensions members ancestors in the following section. However, we cannot guarantee the non-liner hierarchy in this case.

4.6.2 Dimension Members Ancestors

As indicated in the previous section, it may not always be possible for the summation in all tables to provide us with the decision for partitioning the table. Other undetermined equations may be in use, and the summation may not always work. Let us consider the statistical table in Figure 4.16, which shows the percentage of individuals using the Internet by different selected characteristics. Each individual can be in all the categories, as they cover nearly all individual characteristics. Having all of the characteristics for one individual in one cube would be insufficient because of the double count. Nevertheless, detecting the double count by the summation rule is not possible in this case, other equation is used of which we are not aware; for example, the summation of *males* and *females* in the table should be 100%, but in this case, it is 135.8%, and this is obviously not equal to the age group members.



Internet use by individuals, by selected characteristics

	Any location[1]		
	2005	2007	2009
	% of individuals[2]		
All Internet users	67.9	73.2	80.3
Household type			
Single family households with unmarried children under age 18	80.9	86.4	91.1
Single family households without unmarried children under age 18	62.5	67.5	76.4
One-person households	48.7	53.0	63.1
Multi-family households	78.8	80.6	86.4
Sex			
Males	68.0	74.1	81.0
Females	67.8	72.3	79.7
Age			
34 years and under	88.9	93.1	96.5
35 to 54 years	75.0	79.8	87.8
55 to 64 years	53.8	60.8	71.1
65 years and over	23.8	28.8	40.7
Level of education			
Less than high school	31.2	43.2	50.7
High school or college	72.0	76.8	83.4
University degree	89.4	92.5	94.7
Personal income quartile[3,4,5,6]			
Lowest quartile	58.7	68.8	76.2
Second quartile	56.9	60.7	69.9
Third quartile	71.3	75.5	83.1
Highest quartile	83.2	87.9	92.1

Source: statistics Canada: http://www40.statcan.gc.ca/l01/cst01/comm35a-eng.htm

Figure 4.16: A Statistics table should be partitioned where summation rule does not work

For cases such as the one above, the ancestors heuristic must be applied. This heuristic is based on linguistic processing. As we discussed in Section 4.4, about finding the dimension label, we retrieve the ancestors for the members in each dimension. Having the ancestors will help us to make the decision to assign the appropriate dimension label. Continuing with the example from Figure 4.16, we find that each dimension extracted from the row header is not related in meaning to each other. Our solution for assigning dimensions labels, however, will assign "*characteristics*" taken from the table title to the dimensions that contain: "*Household type, Sex, Age, Level of education,* and *Personal income quartile*". Therefore, we will compare the group of members in the lower level of the header if they share ancestors with other dimensions in the same level. For example, the ancestors of the "*Level of Education*" members will differ greatly from the ancestors for the "*Household type*," so that the group ancestors will not intersect. Therefore, this table should be partitioned into sections that each inherits the column header and the table title. This kind of partitioning may not guarantee that the table is divided to ensure summarizability, but it would guarantee that the cube members in the same dimension are related to each other and that they do not discuss different topics.

4.7 Table Integrator

Some statistics web sites do not include all possible classification dimension members for a single dimension within the same statistics table, (see Section 3.2). The data is placed outside the row and column headers due to space constraints, we will call this data as special links. In these instances, the collection of members for a single dimension existing outside the statistical table is represented by links. The links lead to other statistics tables that contain the new collection of members. In our particular model, we will collect the external dimension members to help us identify the dimension label for the dimensions that belong within the table header, since it is more difficult to assign a dimension label for one member, than to a group of members (see Section 4.4.2.1). In any case, collecting the members will not guarantee that the other table will be merged, since we did not confirm whether or not the other dimensions for the other table can be used in the integration. Our main rule for integrating two tables is that all of the dimensions in each table should be the same, before the tables can be merged.

For each table, we check each dimension, and for each dimension, we check the dimension label or the dimension members. The dimension label is checked first, because, if one group of the members in that dimension appears in one table and the other group appears in the other table; for example, if we know that each group contains provinces but their members are different, we have to merge them. Furthermore, the members are checked because, in some tables, a false result could occur in assigning the dimension label in one of the tables. Therefore, we will make a syntactical comparison between the members to determine whether or not they are the same. By checking all of the dimensions, we can make a clear decision about integrating the different statistical tables.

4.8 Domain Integrator

A domain consists of a set of values from which a dimension draws its members. The purpose of domain matching is two-fold: to identify the dimensions that share the same domain, and to find a proper label of a dimension which is not yet labeled (see Section 4.4.1.4). A domain of a dimension, as defined in this thesis, is similar to a domain for an attribute in the relational model. The members of matched dimensions that share the same domain are drawn from the same collection of members, but they do not necessarily contain the exact same set of members. In this sense, dimension-matching is similar to attribute-matching, as part of schema-matching. We opt for an unsupervised learning approach, but our rules for matching are quite different. First of all, some dimensions may not have a name to begin with, since the program cannot infer the name from the metadata within the table, especially from the table title. In fact, one of the main reasons for dimension-matching is to locate the label of a dimension, as discussed in Section 4.4.1.4. On the other hand, two matching dimensions will most likely have the same dimension labels, and vice versa. In our matching process, we keep a list of dimensions, with or without labels. Whenever a dimension is derived (with or without a label), an attempt is made to match the dimension with one of the dimensions on the list. If a match occurs, the union of the members of the two dimensions replaces the members of the dimension already on the list. If either of the matched dimensions has been labeled, the other one will be given the same label. Each domain will contain a group of words that are related to each other. For example, in one domain list, we will have the names of provinces. The list may change based on the different table being tested. In some cases, for instance, the list might contain the province *Quebec,* but instead of simply *Quebec*, it is replaced by *Quebec, Francophone* and *Quebec, Anglophone*. These sub-categories would not appear in the same domain list with *Quebec*, but they would appear within another domain containing provinces without *Quebec*.

To define the matching process, we syntactically compare the two lists of members. We check the dimension label first, and if the two lists carry the same dimension label, we can determine syntactically whether or not the two groups of members intersect with a low threshold. We will not apply the union of these groups directly without confirming the label, because we cannot guarantee that the dimension label was assigned correctly. On the other hand, if the dimension labels are different, then we will make the decision to merge by syntactically intersecting the dimension members with a high threshold.

5. Experimental Results

In Chapter 4, we described the algorithms that allow for the extraction of data and metadata from multidimensional tables to create cubes. The cubes are designed to help users to query specific statistical data. In this chapter, show the effectiveness of our information extractor as described in Chapter 4. Section 5.1 provides an overview of the system; and Section 5.2 presents our experiments and the results by using tables form Statistics Canada [54]. To prove that our multidimensional table extractor is accurate for other statistical agencies, in Section 5.3, we apply it with table from Statistics Austria [55] and Statistics Finland [56].

5.1 Overview of the System

The system is designed by implementing the algorithms discussed in Chapter 4, so that users can choose the statistics tables to be retrieved and stored in a cube. Users do this by submitting the URL containing the statistics table. Then, based on the information presented for the retrieved HTML page, the cube is generated and integrated with other available tables from the website. With regards to the cube, the measure should be identified and each dimension within the cube should have a label, and the members can be organized as one level or as a hierarchy. In the future direction of this work, the system will be automated so that it will retrieve all of the HTML tables for an organization and process them at the same time.

5.2 Performance assessment

To test the performance of our system we need to use real data, which is available online for statistics readers and researchers. At the UN data website [57], a list of statistics web sites can be found for all countries. From the list, we analyzed the tables and determined that most of them are presented in a standardized way. We tested the summary tables in Statistics Canada website because it has a wide range of HTML multidimensional tables on various subjects; the dataset contains 800 randomly selected tables. They are domain-independent and cover such topics as: education, construction, household, travel, etc. Some may be part of a series, with the same table subject, but with different dimension members, which is important in our table integration test methodology, i.e., the extended table model.

In Table 5.1, we show experimental results for our system process. Each process had a different number of components tested, which was different from the number of tables. The main system processes are:

1) Deriving the dimension and extracting the dimension label

   The number of components in this process depends on the number of dimensions found in each table. The total number of components is the total number of dimensions extracted from all tables processed. Successful components are extracted with a correct full dimension and appropriately assigned label.

2) Identifying the measure dimension

   The number of components for testing the measure dimension is determined by the measure dimensions found in the table. Each correct component should have a measure unit attached to a measure-related word or dimension. If a measure unit is retrieved without attaching the appropriate measure-related word, it is considered as a failure for the process.

3) Partitioning the table

The number of components tested for the partition depends on how many sections the tables can be partitioned into. The successful result occurs when the table is partitioned into the correct number of sections.

4) Table integration

This depends on how many series of tables are found in the dataset. The series should be integrated to form one complete cube. Each series of table is considered as one component and each failure to integrate all of the tables in the same series is considered as a failure for one component.

5) Domain integration

The number of integrated domains depends on how many domains are to be extracted by integrating all of the dimensions. A domain is successfully integrated if all of the similar dimensions are integrated into one group.

Table 5.1 : Experimental results

Process name	Number of components	Success rate
Deriving the dimension and assigning the dimension label	2446	91%
Identifying the measure dimension	877	90.3%
Partitioning the table	889	93.7%
Table integration	119	92%
Domain integration	50	96%

5.3 Real examples from different statistics agencies

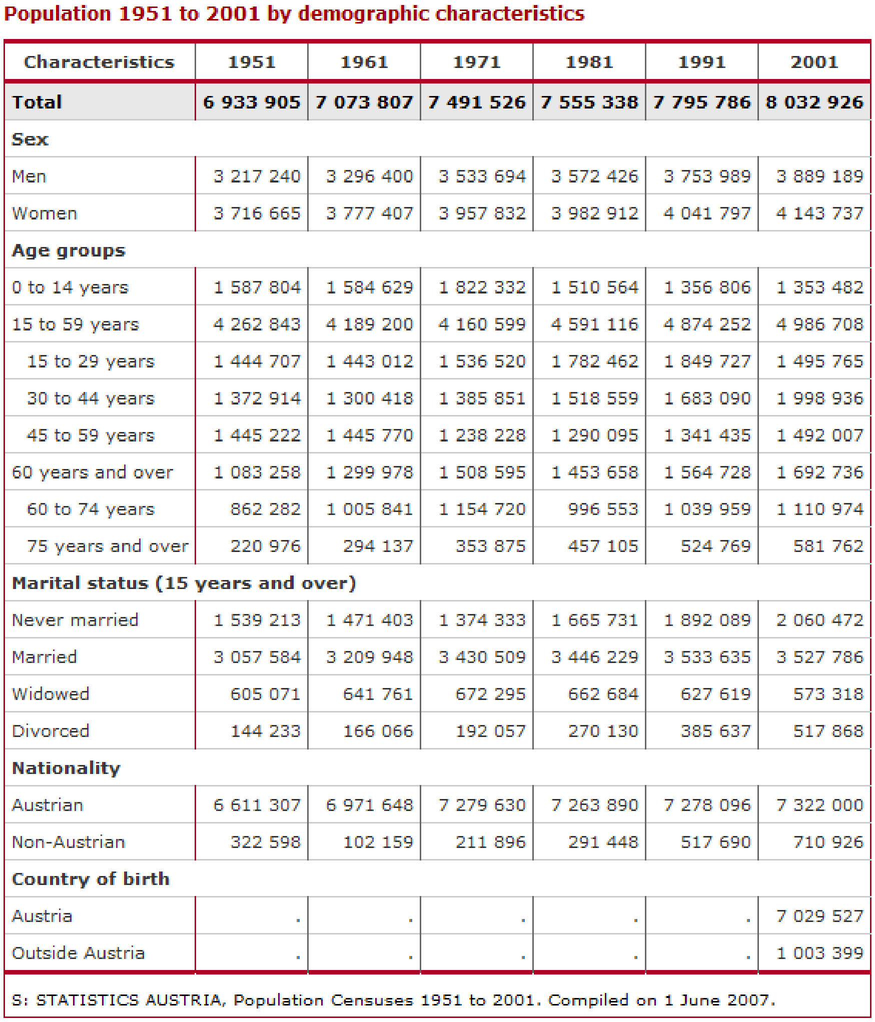
To confirm that our method works for different government agencies that have HTML multidimensional tables (in English), we present our results for the following tables.

5.3.1 Statistics Austria

Figure 5.1 shows an example of a table taken from Statistics Austria [55]. The data is presented in different formats, with one of them being HTML. This table was successfully tested in four components, and no table integration was needed.

In this table, the summation rule works perfectly; for example, in the sum of the sex section (3rd and 4th rows) and the 1st column (3 217 240+ 3 716 665= 6 933 905). The sum for the marital status section (15th to 18th rows) and the 1st column are also equal to 6 933 905, and the rule also works for the other sections (age groups, nationality, and country of birth). Because of the equality in the summation results for all sections, the table needs to be partitioned.

Regarding the assignment of labels for the dimensions, each dimension in the row header gets its label from the dimension header that appears above. For example, *Austrian* and *non-Austrian* have nationality as their labels, and the same applies for the other dimensions in the same header. The time dimension, on the other hand, is indicated by a keyword pattern indicating the year range. Finally, the table has no measure unit or measure-related dimension. Thus, our system will assign *Population in number* as the measure dimension.



Population 1951 to 2001 by demographic characteristics

Characteristics	1951	1961	1971	1981	1991	2001
Total	**6 933 905**	**7 073 807**	**7 491 526**	**7 555 338**	**7 795 786**	**8 032 926**
Sex						
Men	3 217 240	3 296 400	3 533 694	3 572 426	3 753 989	3 889 189
Women	3 716 665	3 777 407	3 957 832	3 982 912	4 041 797	4 143 737
Age groups						
0 to 14 years	1 587 804	1 584 629	1 822 332	1 510 564	1 356 806	1 353 482
15 to 59 years	4 262 843	4 189 200	4 160 599	4 591 116	4 874 252	4 986 708
15 to 29 years	1 444 707	1 443 012	1 536 520	1 782 462	1 849 727	1 495 765
30 to 44 years	1 372 914	1 300 418	1 385 851	1 518 559	1 683 090	1 998 936
45 to 59 years	1 445 222	1 445 770	1 238 228	1 290 095	1 341 435	1 492 007
60 years and over	1 083 258	1 299 978	1 508 595	1 453 658	1 564 728	1 692 736
60 to 74 years	862 282	1 005 841	1 154 720	996 553	1 039 959	1 110 974
75 years and over	220 976	294 137	353 875	457 105	524 769	581 762
Marital status (15 years and over)						
Never married	1 539 213	1 471 403	1 374 333	1 665 731	1 892 089	2 060 472
Married	3 057 584	3 209 948	3 430 509	3 446 229	3 533 635	3 527 786
Widowed	605 071	641 761	672 295	662 684	627 619	573 318
Divorced	144 233	166 066	192 057	270 130	385 637	517 868
Nationality						
Austrian	6 611 307	6 971 648	7 279 630	7 263 890	7 278 096	7 322 000
Non-Austrian	322 598	102 159	211 896	291 448	517 690	710 926
Country of birth						
Austria	.	.	.	.	.	7 029 527
Outside Austria	.	.	.	.	.	1 003 399

S: STATISTICS AUSTRIA, Population Censuses 1951 to 2001. Compiled on 1 June 2007.

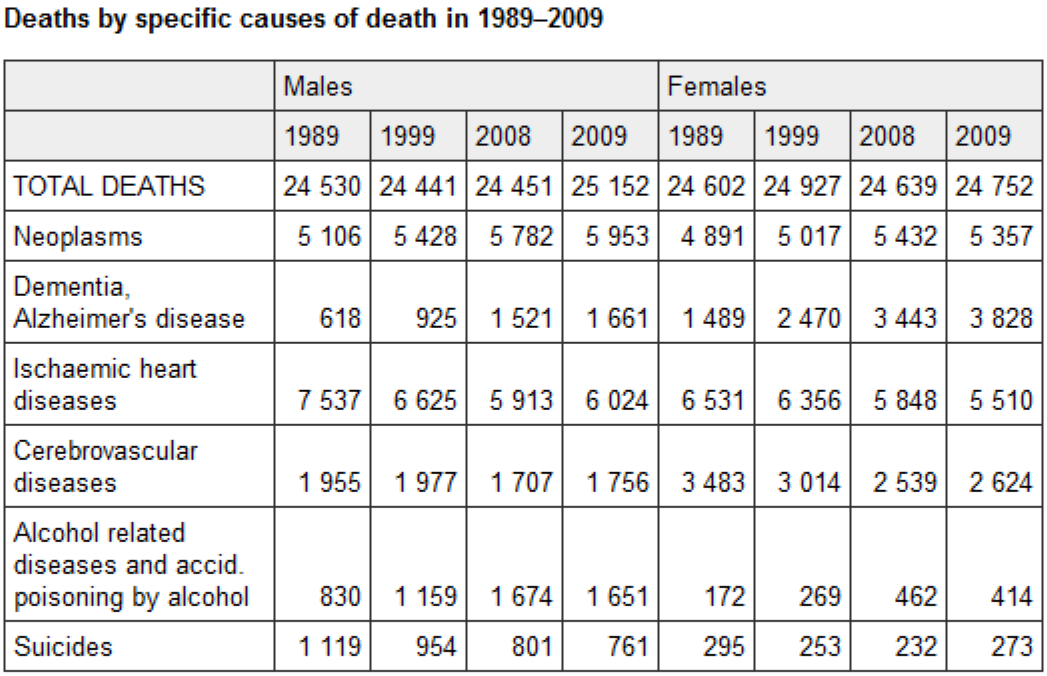
Source Statistics Austria
http://www.statistik.at/web_en/statistics/population/population_censuses/population_by_demographic_characteristics/028545.html, January 2011

Figure 5.1: Table form Statistics Austria

5.3.2 Statistics Finland

Figure 5.2 shows an example for a table from Statistics Finland [56]. The table was successfully tested for three components, without a need for integration or partitioning.

In this example, no measure unit is present, so the system will retrieve *Deaths in number* as the measure. With regards to the dimension label and the extractions, the title provides no clue about the label for the words *Male* and *Female,* which should be *Sex*. Consequently, the domain integration needs to be checked for what was already collected from other tables, so that the correct label can be indicated. The Time range 1989, 1999, 2008, and 2009 are indicated as time, based on the keyword pattern. The last dimension extracted from the row header will be labeled, *Causes of death*.



Deaths by specific causes of death in 1989–2009

	Males				Females			
	1989	1999	2008	2009	1989	1999	2008	2009
TOTAL DEATHS	24 530	24 441	24 451	25 152	24 602	24 927	24 639	24 752
Neoplasms	5 106	5 428	5 782	5 953	4 891	5 017	5 432	5 357
Dementia, Alzheimer's disease	618	925	1 521	1 661	1 489	2 470	3 443	3 828
Ischaemic heart diseases	7 537	6 625	5 913	6 024	6 531	6 356	5 848	5 510
Cerebrovascular diseases	1 955	1 977	1 707	1 756	3 483	3 014	2 539	2 624
Alcohol related diseases and accid. poisoning by alcohol	830	1 159	1 674	1 651	172	269	462	414
Suicides	1 119	954	801	761	295	253	232	273

Source Statistics Finland, http://www.stat.fi/til/ksyyt/2009/ksyyt_2009_2010-12-17_tie_001_en.html , January 2011

Figure 5.2: Table from statistics Finland

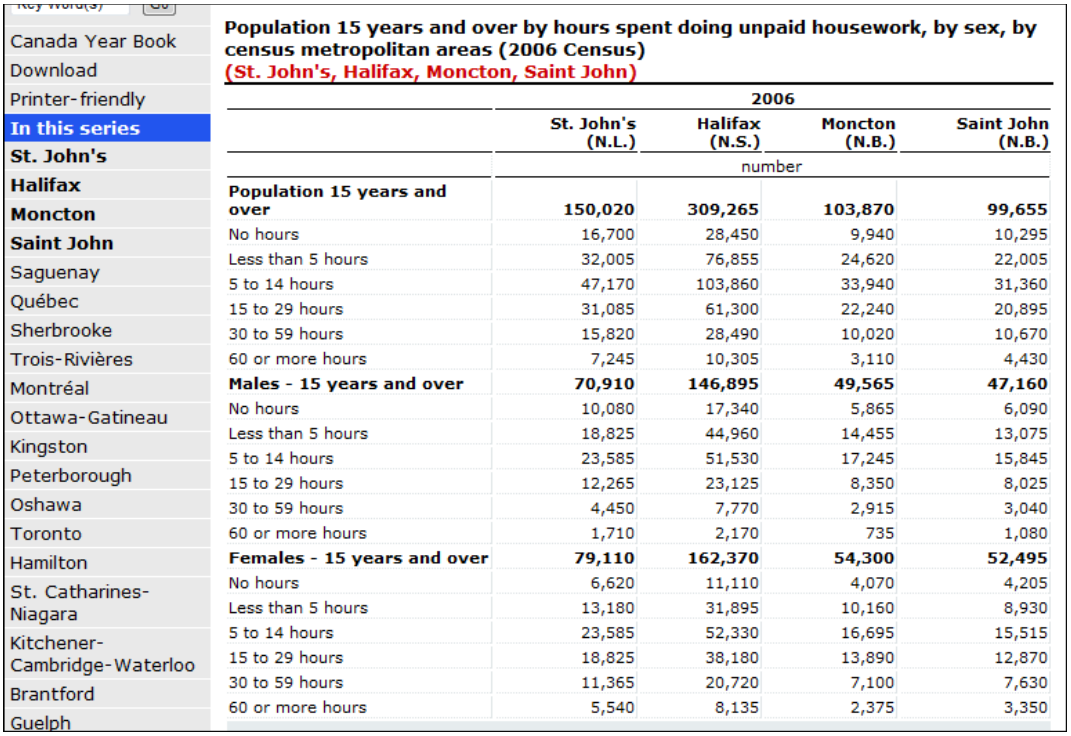
5.4 Analysis of Failed Cases

Our experiments show that a good number of components had successful results. Nevertheless, some false results were seen, which are discussed in the following sections.

5.4.1 Deriving the Dimensions and Assigning Dimension Labels

Most of our experiments showed that a suitable dimension label was assigned; however, for some dimensions, the system failed to distinguish a proper label; for example:

- Figure 5.3 includes a dimension that should have two dimension labels – *age* and *sex*; however, in our method, we assume that only one dimension label is used for each dimension.
- Figure 5.4 includes some dimension members that change over the years. For example, Y.T. could stand for both Yukon Territory and Yukon. Our abbreviation rule can determine that Y.T. stands for Yukon Territory; however, it cannot distinguish that Y.T. stands for Yukon, which appears in the special links. This is actually solved later by applying the domain integration.
- Figure 5.5 shows an example where a more suitable dimension label can be found within the table title, instead of from the dimension header. Based on our algorithm, the dimensions under the header, '*persons in family household*' are assigned this as their dimension label. Still, the table title contains a much better dimension label: '*living arrangements'*.
- Figure 5.6 illustrates that some dimension members are not well represented in WordNet, in terms of retrieving their ancestors and finding a match within the table title. For example, ancestors retrieved by WordNet do not allow for the matching of full-time or part-time with '*registration status'*.

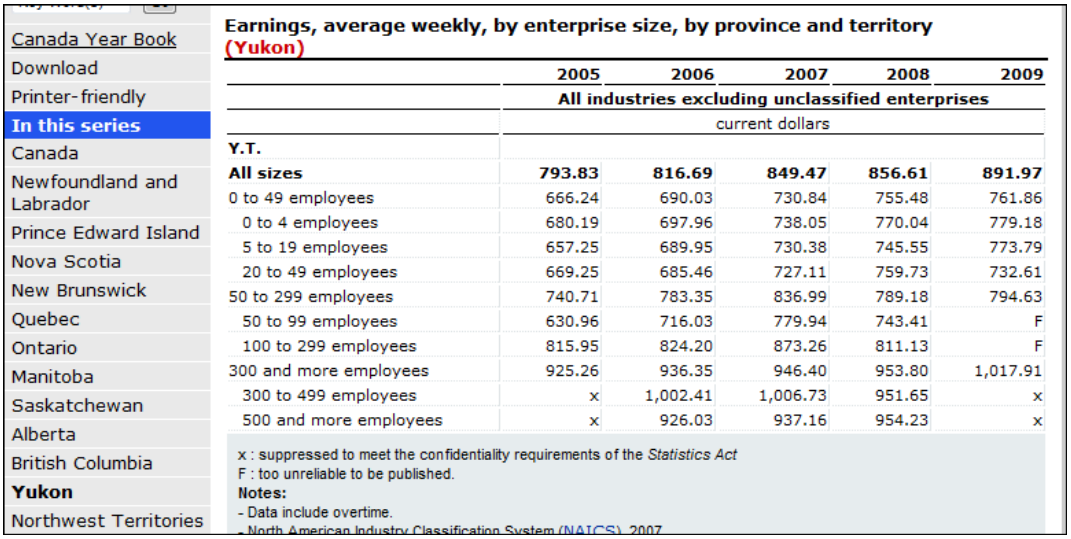


Population 15 years and over by hours spent doing unpaid housework, by sex, by census metropolitan areas (2006 Census)
(St. John's, Halifax, Moncton, Saint John)

	2006			
	St. John's (N.L.)	Halifax (N.S.)	Moncton (N.B.)	Saint John (N.B.)
	number			
Population 15 years and over	**150,020**	**309,265**	**103,870**	**99,655**
No hours	16,700	28,450	9,940	10,295
Less than 5 hours	32,005	76,855	24,620	22,005
5 to 14 hours	47,170	103,860	33,940	31,360
15 to 29 hours	31,085	61,300	22,240	20,895
30 to 59 hours	15,820	28,490	10,020	10,670
60 or more hours	7,245	10,305	3,110	4,430
Males - 15 years and over	**70,910**	**146,895**	**49,565**	**47,160**
No hours	10,080	17,340	5,865	6,090
Less than 5 hours	18,825	44,960	14,455	13,075
5 to 14 hours	23,585	51,530	17,245	15,845
15 to 29 hours	12,265	23,125	8,350	8,025
30 to 59 hours	4,450	7,770	2,915	3,040
60 or more hours	1,710	2,170	735	1,080
Females - 15 years and over	**79,110**	**162,370**	**54,300**	**52,495**
No hours	6,620	11,110	4,070	4,205
Less than 5 hours	13,180	31,895	10,160	8,930
5 to 14 hours	23,585	52,330	16,695	15,515
15 to 29 hours	18,825	38,180	13,890	12,870
30 to 59 hours	11,365	20,720	7,100	7,630
60 or more hours	5,540	8,135	2,375	3,350

Source: Statistics Canada: http://www40.statcan.gc.ca/l01/cst01/famil126a-eng.htm

Figure 5.3: Age and sex appears in the same dimension member



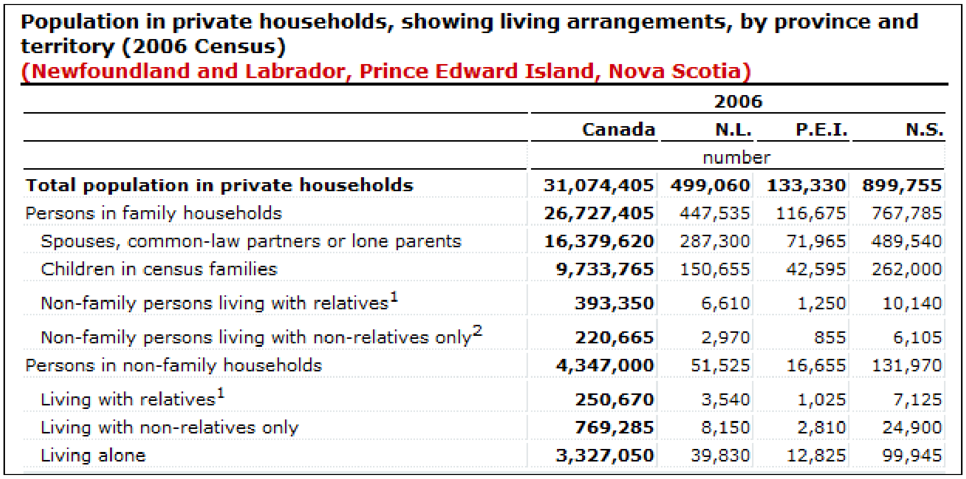
Earnings, average weekly, by enterprise size, by province and territory
(Yukon)

	2005	2006	2007	2008	2009
	All industries excluding unclassified enterprises				
	current dollars				
Y.T.					
All sizes	**793.83**	**816.69**	**849.47**	**856.61**	**891.97**
0 to 49 employees	666.24	690.03	730.84	755.48	761.86
0 to 4 employees	680.19	697.96	738.05	770.04	779.18
5 to 19 employees	657.25	689.95	730.38	745.55	773.79
20 to 49 employees	669.25	685.46	727.11	759.73	732.61
50 to 299 employees	740.71	783.35	836.99	789.18	794.63
50 to 99 employees	630.96	716.03	779.94	743.41	F
100 to 299 employees	815.95	824.20	873.26	811.13	F
300 and more employees	925.26	936.35	946.40	953.80	1,017.91
300 to 499 employees	x	1,002.41	1,006.73	951.65	x
500 and more employees	x	926.03	937.16	954.23	x

x : suppressed to meet the confidentiality requirements of the *Statistics Act*
F : too unreliable to be published.
Notes:
- Data include overtime.
- North American Industry Classification System (NAICS) 2007

Source: Statistics Canada: http://www40.statcan.gc.ca/l01/cst01/labr83l-eng.htm

Figure 5.4: *Yukon* should be written as *Yukon Territory*

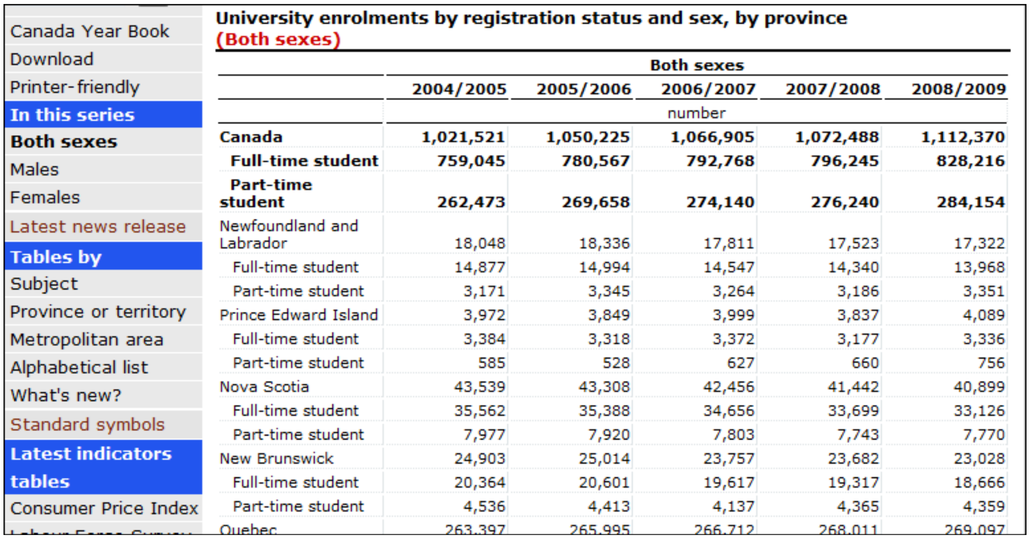


Population in private households, showing living arrangements, by province and territory (2006 Census)
(Newfoundland and Labrador, Prince Edward Island, Nova Scotia)

	2006			
	Canada	N.L.	P.E.I.	N.S.
	number			
Total population in private households	**31,074,405**	**499,060**	**133,330**	**899,755**
Persons in family households	**26,727,405**	447,535	116,675	767,785
Spouses, common-law partners or lone parents	**16,379,620**	287,300	71,965	489,540
Children in census families	**9,733,765**	150,655	42,595	262,000
Non-family persons living with relatives[1]	**393,350**	6,610	1,250	10,140
Non-family persons living with non-relatives only[2]	**220,665**	2,970	855	6,105
Persons in non-family households	**4,347,000**	51,525	16,655	131,970
Living with relatives[1]	**250,670**	3,540	1,025	7,125
Living with non-relatives only	**769,285**	8,150	2,810	24,900
Living alone	**3,327,050**	39,830	12,825	99,945

Source: Statistics Canada : http://www40.statcan.gc.ca/l01/cst01/famil52a-eng.htm

Figure 5.5: Wrong dimension label assignment to show living arrangements



University enrolments by registration status and sex, by province
(Both sexes)

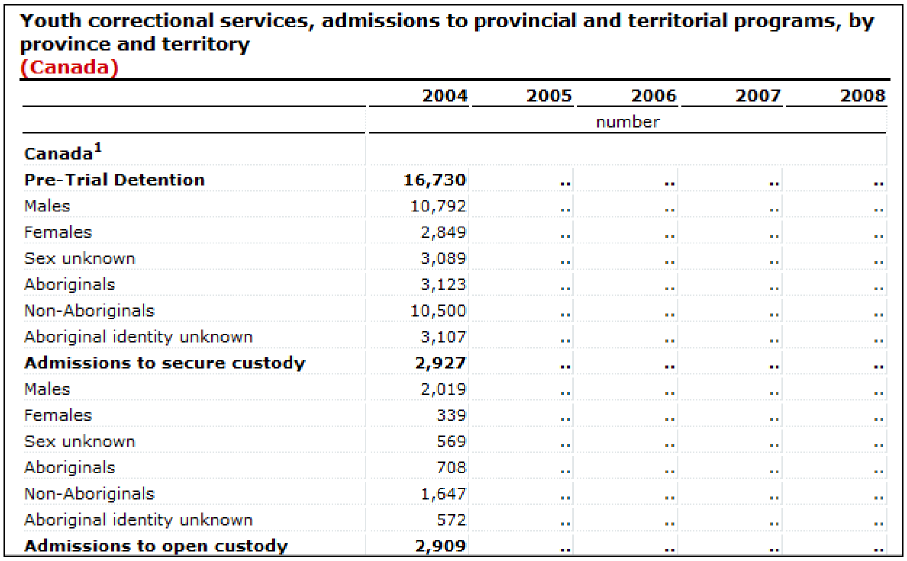
	Both sexes				
	2004/2005	2005/2006	2006/2007	2007/2008	2008/2009
	number				
Canada	**1,021,521**	**1,050,225**	**1,066,905**	**1,072,488**	**1,112,370**
Full-time student	**759,045**	**780,567**	**792,768**	**796,245**	**828,216**
Part-time student	**262,473**	**269,658**	**274,140**	**276,240**	**284,154**
Newfoundland and Labrador	18,048	18,336	17,811	17,523	17,322
Full-time student	14,877	14,994	14,547	14,340	13,968
Part-time student	3,171	3,345	3,264	3,186	3,351
Prince Edward Island	3,972	3,849	3,999	3,837	4,089
Full-time student	3,384	3,318	3,372	3,177	3,336
Part-time student	585	528	627	660	756
Nova Scotia	43,539	43,308	42,456	41,442	40,899
Full-time student	35,562	35,388	34,656	33,699	33,126
Part-time student	7,977	7,920	7,803	7,743	7,770
New Brunswick	24,903	25,014	23,757	23,682	23,028
Full-time student	20,364	20,601	19,617	19,317	18,666
Part-time student	4,536	4,413	4,137	4,365	4,359
Quebec	263,397	265,995	266,712	268,011	269,097

Source: Statistics Canada: http://www40.statcan.gc.ca/l01/cst01/educ53a-eng.htm

Figure 5.6: Registration status does not match full-time with part-time

5.4.2 Measure Dimension

In a small number of cases, the first part of the table title does not contain the measure, though in our heuristics, we assume it comes first. The table title is usually concise and does not contain extra information; for example, Figure 5.7, which contains *admissions to provincial and territorial programs* as the second part of the title.



Youth correctional services, admissions to provincial and territorial programs, by province and territory
(Canada)

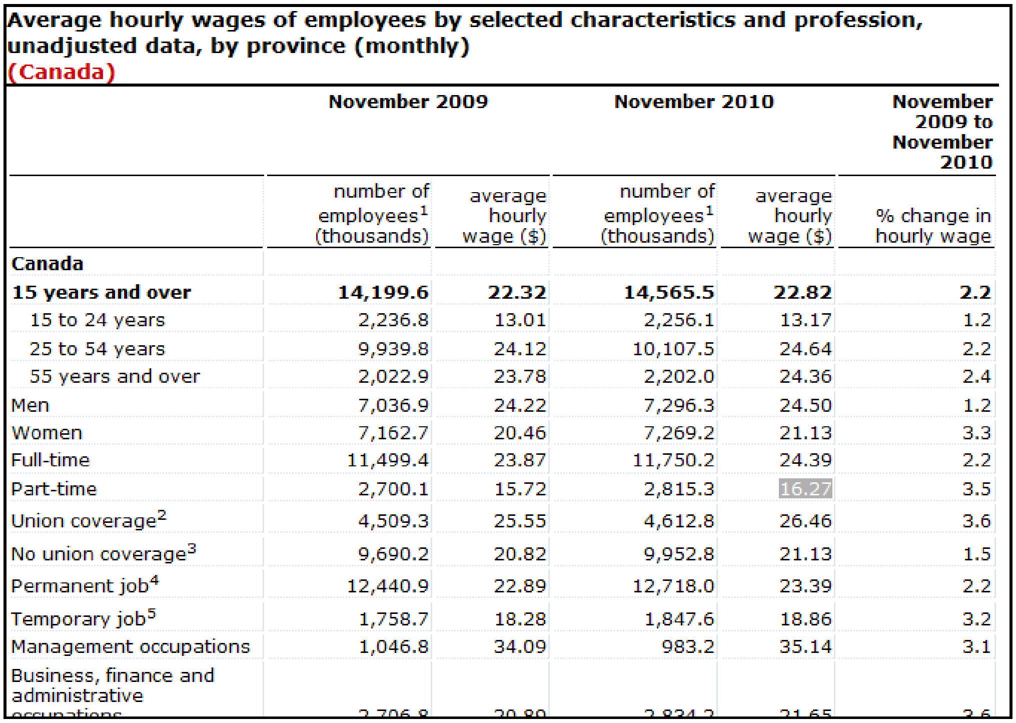
	2004	2005	2006	2007	2008
	number				
Canada[1]					
Pre-Trial Detention	**16,730**	..	..	..	..
Males	10,792	..	..	..	..
Females	2,849	..	..	..	..
Sex unknown	3,089	..	..	..	..
Aboriginals	3,123	..	..	..	..
Non-Aboriginals	10,500	..	..	..	..
Aboriginal identity unknown	3,107	..	..	..	..
Admissions to secure custody	**2,927**	..	..	..	..
Males	2,019	..	..	..	..
Females	339	..	..	..	..
Sex unknown	569	..	..	..	..
Aboriginals	708	..	..	..	..
Non-Aboriginals	1,647	..	..	..	..
Aboriginal identity unknown	572	..	..	..	..
Admissions to open custody	**2,909**	..	..	..	..

Source: Statistics Canada: http://www40.statcan.gc.ca/l01/cst01/legal42a-eng.htm

Figure 5.7: The measure represented in the second part of the title

5.4.3 Table Partitioning

Table partitioning can be problematic if the visual clues outlining the header are not clear (see Figure 5.8). This may lead to double counting in some instances, for the same visual representation. The problem can be solved by using the summation rule and testing a group of members within the same visual representation and determining if each permutation is equal to the other permutation. If they are equal, the table can be split, though this is not considered in our system and is dealt with in future work.



Average hourly wages of employees by selected characteristics and profession, unadjusted data, by province (monthly)
(Canada)

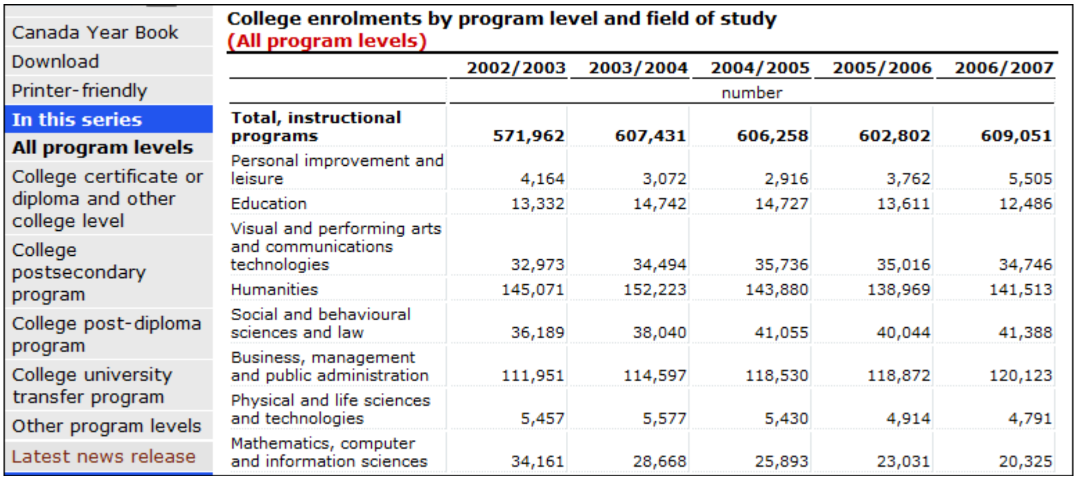
	November 2009		November 2010		November 2009 to November 2010
	number of employees[1] (thousands)	average hourly wage ($)	number of employees[1] (thousands)	average hourly wage ($)	% change in hourly wage
Canada					
15 years and over	**14,199.6**	**22.32**	**14,565.5**	**22.82**	**2.2**
15 to 24 years	2,236.8	13.01	2,256.1	13.17	1.2
25 to 54 years	9,939.8	24.12	10,107.5	24.64	2.2
55 years and over	2,022.9	23.78	2,202.0	24.36	2.4
Men	7,036.9	24.22	7,296.3	24.50	1.2
Women	7,162.7	20.46	7,269.2	21.13	3.3
Full-time	11,499.4	23.87	11,750.2	24.39	2.2
Part-time	2,700.1	15.72	2,815.3	16.27	3.5
Union coverage[2]	4,509.3	25.55	4,612.8	26.46	3.6
No union coverage[3]	9,690.2	20.82	9,952.8	21.13	1.5
Permanent job[4]	12,440.9	22.89	12,718.0	23.39	2.2
Temporary job[5]	1,758.7	18.28	1,847.6	18.86	3.2
Management occupations	1,046.8	34.09	983.2	35.14	3.1
Business, finance and administrative occupations	2,706.8	20.80	2,834.2	21.65	3.6

Source: Statistics Canada: http://www40.statcan.gc.ca/l01/cst01/labr69a-eng.htm

Figure 5.8: Double counting (non-linear hierarchy)

5.4.4 Table Integrator

The goal of table integrator is to integrate multiple tables that carry the same dimensions and measure, but appear in different web pages. Thus, the dimension members will be integrated and represented within the same dimension and the same cube. For example, in Figure 5.9, the dimension members that should be integrated are under *Program level*, which is not included in the table. In this case, since we are unable to identify the dimension, we cannot carry out the integration.


Canada Year Book
Download
Printer-friendly
In this series
All program levels
College certificate or diploma and other college level
College postsecondary program
College post-diploma program
College university transfer program
Other program levels
Latest news release

College enrolments by program level and field of study
(All program levels)

	2002/2003	2003/2004	2004/2005	2005/2006	2006/2007
			number		
Total, instructional programs	**571,962**	**607,431**	**606,258**	**602,802**	**609,051**
Personal improvement and leisure	4,164	3,072	2,916	3,762	5,505
Education	13,332	14,742	14,727	13,611	12,486
Visual and performing arts and communications technologies	32,973	34,494	35,736	35,016	34,746
Humanities	145,071	152,223	143,880	138,969	141,513
Social and behavioural sciences and law	36,189	38,040	41,055	40,044	41,388
Business, management and public administration	111,951	114,597	118,530	118,872	120,123
Physical and life sciences and technologies	5,457	5,577	5,430	4,914	4,791
Mathematics, computer and information sciences	34,161	28,668	25,893	23,031	20,325

Source: Statistics Canada: http://www40.statcan.gc.ca/l01/cst01/educ60a-eng.htm

Figure 5.9: The dimension *program level* is not part of the table

5.4.5 Domain Integrator

As illustrated earlier, the goal for domain integration is to collect similar dimensions and combine them into the same group. This depends on the data in the collection. Therefore, the more dimensions being tested, the more accurate the results, since we rely on members overlapping between the dimensions. One of the failures in the testing was due to the combining of domains for US states. On the Statistics Canada website, two tables list the top 15 states; however, the two dimensions share a small number of members (i.e., states). Consequently, the two dimensions do not pass the threshold for the merging of the two dimensions.

6. Conclusion

A vast amount of useful information is available online, especially from the well maintained websites of government agencies. The statistical data is often published online in the form of multidimensional tables, for the purpose of answering user questions and providing information, and may also be valuable for ongoing research studies and applications. In this thesis, an information and metadata extractor was designed and developed for use with multidimensional tables.

6.1 Summary

Extensive research has been conducted on the extraction of information from online sources. While many of the studies tend to extract large volumes of information from relational tables, the metadata extracted from multidimensional tables is of an inferior quality. This thesis contributes to the knowledge in several areas:

- By presenting the problem of extracting from multidimensional tables, we show how certain applications can benefit. We also present an overview of related work.
- We created a method to understand the components of multidimensional tables, and to identify the dimensions representing the cube. The components can be extracted by knowing the visual clues that are common to the design of multidimensional tables. In addition, we determined whether a dimension is a single-level or a multi-level dimension, as a form of Generalization-Specialization. This was accomplished with a linguistic approach to check if the higher member in the header hierarchy can act as the generalized word for the

lower group of members. Furthermore, we can also confirm this relationship by using the summation method to check whether or not the higher member numeric value is a summation for the lower member group.

- We assigned labels for the dimensions, taking advantage of the table title appearance, and the multi-level representation in the row or column headers. Assigning dimension labels is especially challenging, because the members of the dimensions can be phrases, and typically are not common knowledge (i.e., they depend on the purpose of the table). We can assign a label for the dimension members by applying our techniques, using WordNet ontology, keyword patterns, and the results of domain integration.
- We developed strategies for table partitioning and table generation in the context of table canonicalization. Where most studies tend to transform one table to one component, which is either a relational table or a cube, in this research, we show that one multi-dimensional table can be divided into more than one cube, to guarantee summarizability and the integration of multiple multi-dimensional tables into one cube, for the full classification of a dimension. Use of the summation rule is crucial for the decision to partition tables. Alternatively, checking the dimensions and their members in different tables is crucial for the decision to integrate tables.

6.2 Future Directions

From our results for extracting information from multidimensional tables, several possible enhancements may be explored in future research:

The summation rule, discussed in Section 4.6.1, can help in determining the aggregation member of the dimension and the decision to partition tables. The summation total is retrieved for all members with the same visual clues. In some rare cases, the aggregated members could have the same visual clues as those of the other members. Therefore, the summation for permutations along the members with the same visual clues needs to be calculated, and the results compared with all other members of that group. In addition, in this research, we only applied the summation function, and in some cases, the average was used; however, the kind of equation being used by a given multidimensional table must be ascertained, and the results must be checked using that equation.

For the measure dimension, the measure-related dimension must be found; it can be retrieved by knowing the relationship with the first part of the table title. Since the table titles are human-generated, other rules must also be determined, especially for different kinds of multidimensional tables. Therefore, we can utilize the previous assigned measures to improve the measure identification.

In this research, table integration was accomplished by merging tables with the same dimensions, though one of the dimensions is not a complete classification for the members. The initial list of tables used to test the integration was found from the special links in the same webpage. Furthermore, a dimension member needs to be displayed in both special links and in the table header. Thus, this integration could be improved by finding if a relation exists between a set of special links and the table title, without considering the appearance of the dimension member in the table header. As a future area of research, we intend to perform the table integration testing without a reliance on the special links. In addition, the domain integration needs to be improved, so that the best possible title can be assigned to the results.

In Chapter 1, the kinds of applications that could benefit from this research were mentioned; specifically, question answering, faceted search, and ontology generation. Further work is necessary to clarify how these applications would be improved. The multidimensional extractor also, should be able to extract tables in PDF documents or other format that are able to be converted to the HTML format.

7. Bibliography

[1] X. Wei,B. Croft and D. Pinto, "Question answering performance on table data," in *Proceedings of national conference on Digital government research* , 2004.

[2] W. Dakka and P.G. Ipeirotis, "Automatic Extraction of Useful Facet Hierarchies," in *ICDE 2008*, 2008.

[3] D.W. Embley, M. Hurst, D. Lopresti, and G. Nagy, "Table-Processing Paradigms: A Research Survey," *International Journal of Document Analysis*, vol. 8, no. 2, pp. 66-86, 2006.

[4] D.E. Appelt and D. Israel, "Introduction to Information Extraction technology," in *IJCAI-99, Tutorial*, 1999.

[5] H. Elmeleegy, J. Madhavan, and A. Halevy, "Harvesting Relational Tables from Lists on the Web," in *VLDB 2009*, Lyon, France, August, 2009.

[6] Wikipedia. (2010, November) [Online]. http://en.wikipedia.org/wiki/Semi-structured_data

[7] M.J. Cafarella, "Extracting and Querying a Comprehensive Web Database," in *CIDR 2009*, Monterey, California, 2009.

[8] E. Ahmed and H.M. Jamil, "Post Processing Wrapper Generated Tables for Labeling Anonymous Datasets," in *WIDM '09*, Hong Kong, 2009.

[9] L.A. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, "Automatic Annotation of Data Extracted from Large Web Sites," in *International Workshop on the Web and Databases*, San Diego, California, 2003.

[10] D. Hu and X. Meng, "Automatic Data Extraction from Data-rich Web Pages," in *DASFAA 2005*, Beijing, China, 2005, pp. 828-839.

[11] W. Su, J. Wang, and F. Lochovsky, "ODE: Ontology-Assisted Data Extraction," *ACM TODS*, vol. 34, no. 2, June 2009.

[12] J. Gray, A. Bosworth, A. Layman, and H. Prahesh, "Data Cube: A relational Aggregation Operator Generalizing Group-By, Cross-Tabs, and Sub-Totals," in *ICDE '96*, New Orleans, 1996.

[13] D.W. Embley, C. Tao, and S.W. Liddle, "Automating the Extraction of Data from HTML Tables with Unknown Structure," *Data & Knowledge Engineering*, vol. 54, no. 1, pp. 3-28, July 2005.

[14] D. Pinto et al., "Qusam: A System for Question Answering Using Semi-structured Data.," in *Joint Conference on Digital Libraries*, 2002, pp. 46-55.

[15] H. Wang et al., "Semantic search on internet tabular information extraction," in *International Conference on Information and Knowledge Management*, 2000, pp. 243-249.

[16] M. Tanaka and T. Ishida, "Ontology Extraction from Tables on the Web," in *International Symposium on Applications on Internet*, 2006, pp. 284-290.

[17] W. Gatterbauer, P. Bohunsky, and M. Herzog, "Towards Domain-Independent Information Extraction from Web Tables," in *WWW 2007*, Banff, Alberta, Canada,

2007.

[18] M.J. Cafarella, A. Halevy, Y. Zhang, D.Z. Wang, and E. Wu, "Webtables: Exploring the Power of Tables on the Web," in *VLDB, 2008*, Auckland, New Zealand, 2008.

[19] S. Agrawal, S.Chaudhuri and G. Das , "DBXplorer: A System for Keyword-Based Search over Relational Databases," in *18th International Conference on Data Engineering (ICDE'02)*, San Jose, California, 2002.

[20] K.P. Yee, K. Swearingen, K. Li, and M. Hearst, "Faceted Metadata for Image Search and Browsing," in *CHI 2003*, Ft. Lauderdale, Fl. , 2003.

[21] S. T. Dumais , Z. Gutt J. Teevan, "Challenges for Supporting Faceted Search in Large, Heterogeneous Corpora like the Web," in *the Second Workshop on Human-Computer Interaction and Information Retrieval (HCIR '08)*, Redmond, WA, October 2008.

[22] Statistics Denmark. (2010) Number of persons and course participants by area, educational area, highest education previously completed, age, national origin, sex and time. [Online]. http://www.statistikbanken.dk/statbank5a/SelectVarVal/Define.asp?Maintable=VEU21&PLanguage=1

[23] P. Wu, Y. Sismanis and B. Reinwald , "Towards keyword-driven analytical processing," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data* , Beijing, China, 2007.

[24] F.M. Suchanek, G. Kasneci, and G. Weikum, "YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia," in *WWW 2007*, Banff, Alberta, Canada, 2007.

[25] A. Pivk, "Automatic Ontology Generation from Web Tabular Structures," *AI Communications*, vol. 19, pp. 83-85, 2006.

[26] A Simitsis, A. Baid, Y. Sismanis, and B. Reinwald, "Multidimensional Content eXploration," in *VLDB*, Auckland, New Zealand, 2008.

[27] CHP Leung, "Understanding, interpreting and querying web statistical tables," M.S. thesis, 2005.

[28] A. Shoshani, "OLAP and Statistical Databases," in *Fifth International Conference on Information and Knowledge Management*, Rockville, Maryland, 1996.

[29] W. Luk and P. Leung, "Extraction of Semantics From Web Statistical Tables," in *IEEE/WIC/ACM International Workshop on Semantic Web Mining and Reasoning*, Beijing, China, 2004.

[30] Y. Yang and W. Luk, "A Framework for Web Table Mining," in *ACM WIDM*, Washington, DC, 2002.

[31] M.S. Amin and H. Jamil, "Ontology Guided Autonomous Label Assignment in Wrapper Induced Tables with Missing Column Names," in *IEEE International Conference on Information Reuse & Integration*, Las Vegas, NV, 2009.

[32] J. Wang and F.H. Lochovosky, "Data Extraction and Label Assignment for Web Databases," in *WWW 2003*, Budapest, Hungary, 2003.

[33] K. Lerman, L. Getoor, S. Minton, and C. Knoblock, "Using the Structure of Web Sites for Automatic Segmentation of Tables," in *SIGMOD 2004*, Paris, France, 2004.

[34] C. Tao and D.W. Embley, "Automatic Hidden-Web Table Interpretation, Conceptualization, and Semantic Annotation," *Data &Knowledge Engineering*, vol. 68, no. 7, pp. 683-703, July 2009.

[35] M.J. Cafarella, A. Halevy, D.Z. Wang, E. Wu and Y. Zhang, "Uncovering the relational web," in *Proceedings of the 11th International Workshop on Web and Databases* , Vancouver, Canada, 2008.

[36] Thomson Reuters. OpenCalais. [Online]. www.opencalais.com/

[37] Inxight. ThingFinder. [Online]. http://inxightfedsys.com/products/sdks/tf/

[38] E. Rahm and P.A. Bernstein, "A Survey of Approaches to Automatic Schema Matching," *VLDB Journal*, vol. 10, no. 4, pp. 334-350, 2001.

[39] A. Doan and A.Y. Halevy, "Semantic Integration Research in the Database Community: A Brief Survey," *AI Magazine*, vol. 26, no. 1, 2005.

[40] A. Doan, P. Domingos and A. Y. Halevy, "Reconciling schemas of disparate data sources: a machine-learning approach," in *Proceedings of the 2001 ACM SIGMOD international conference on Management of data* , Santa Barbara, CA, USA, 2001, pp. 509--520.

[41] D. W. Embley, D. Jackman and L. Xu, "Multifaceted exploitation of metadata for attribute match discovery in information integration," in *Proceedings of the International Workshop on Information Integration on the Web (WIIW'01)*, Rio de Janeiro, Brazil, 2001, pp. 110--117.

[42] S. Lynn and D.W. Embley, "Semantically Conceptualizing and Annotating Tables," in *ASWC '08 Proceedings of the 3rd Asian Semantic Web Conference*, 2008.

[43] O. Lassila and R.R. Swick, "Resource Description Framework (RDF) Model and Syntax Specification," W3C Recommendation, 1999.

[44] Y.A. Tijerino, D.W. Embley, D.W. Lonsdale, and Y. Ding, "Towards Ontology Generation from Tables," *World Wide Web: Internet and Web Information Systems*, vol. 8, pp. 261-285, 2005.

[45] T. Priebe and G. Pernul, "Ontology-based Integration of OLAP and Information Retrieval," in *DEXA 2003 Workshop on Web Semantics*, Prague, Czech, 2003.

[46] T. Niemi and M. Niinimaki, "Ontologies and Summarizability in OLAP," in *ACM SAC'10*, Sierre, Switzerland, 2010.

[47] C. Humphrey, "Tips for Reading a Statistical Table," in *Winter Institute on Statistical Literacy for Librarians (WISLL)*, 2010.

[48] Statistics Finland. (2010, September) Statistics Finland - Online Statistics Course - How to read and use statistics - Ways of displaying statistics - Table. [Online]. http://www.stat.fi/tup/verkkokoulu/data/tlkt/03/01/index_en.html

[49] K. Tsoukalas, "Extracting and tagging tabular data on the web ," Thesis (M.Sc.) , 2009.

[50] C. Fellbaum, "WordNet: An Electronic Lexical Database," *Cambridge, MA: MIT Press.*, 1998.

[51] T. Simpson and T. Dao. (2010, Jan) Code Project: WordNet-based semantic similarity measurement. [Online]. http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx?msg=2776502

[52] T. Simpson and T. Dao. (2010, Jan) WordNet-based semantic similarity measurement, Source code. [Online]. http://wordnetdotnet.googlecode.com/svn/trunk/Projects/Thanh/

[53] HJ. Lenz and A. Shoshani, "Summarizability in OLAP and Statistical Data Bases," in

Ninth International Conference on Scientific and Statistical Database Management, Olympia, WA , USA , 1997 , pp. 132 - 143.

[54] Statistics Canada. [Online]. http://www.statcan.gc.ca/start-debut-eng.html

[55] Statistics Austria. [Online]. http://www.statistik.at/

[56] Statistics Finland. [Online]. http://www.stat.fi/index_en.html

[57] UN Data. [Online]. http://data.un.org/