AN ONTOLOGY FOR TABLE PROCESSING

David W. Embley, CS Department, Brigham Young University, Provo, UT
George Nagy, ECSE Department, Rensselaer Polytechnic Institute, Troy, NY

1. INTRODUCTION

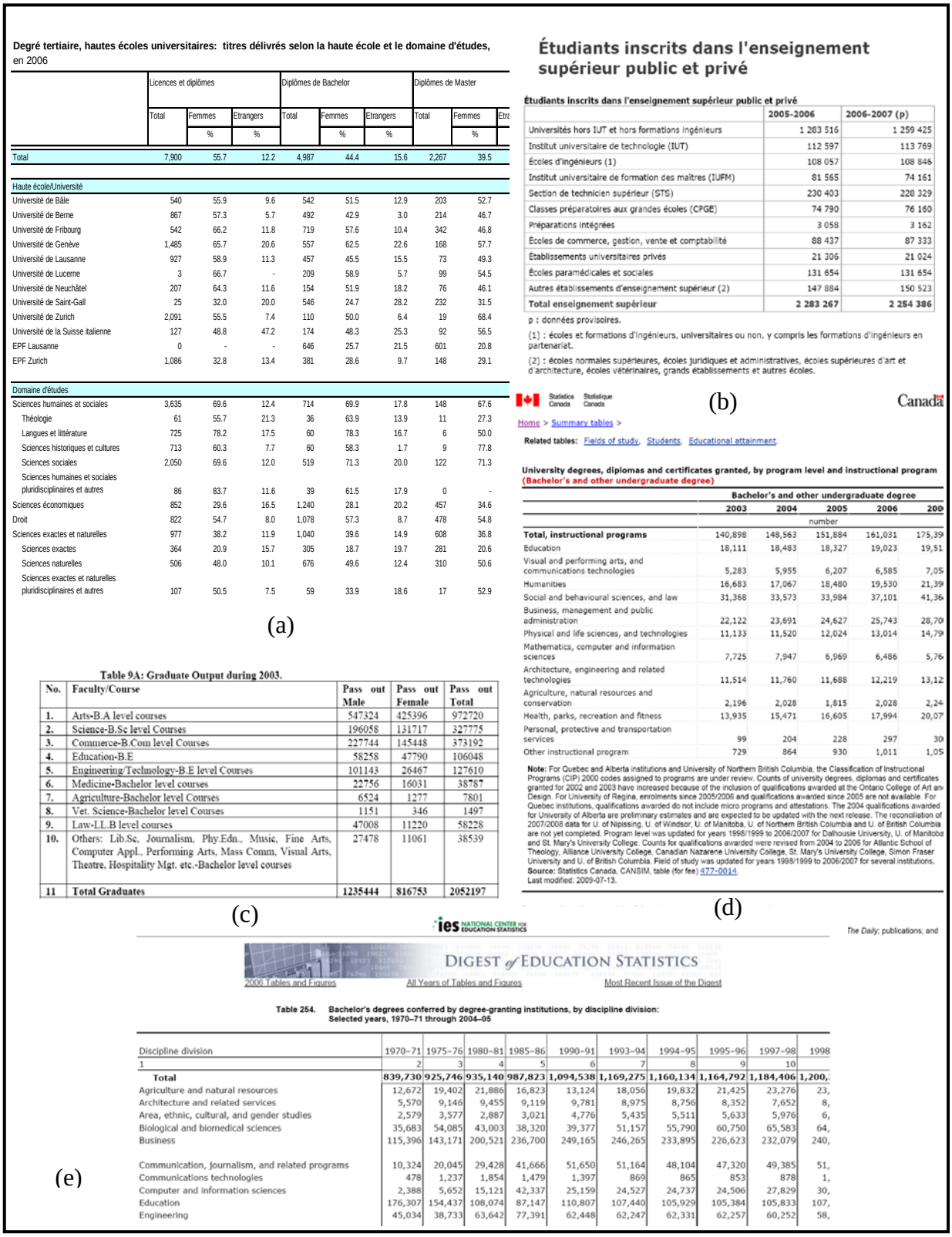
Consider the plight of a researcher, government official, or business person seeking to compile information about the number of university graduates in various fields in selected countries. With moderate effort a knowledge worker can find on the web the tables in Figure 1, interpret them, and compile them—perhaps into the simple Excel table of Figure 2.

The goal of this proposal is to aid knowledge workers like this in their quest to harvest information from heterogeneous collections of tables and to organize it into queriable knowledge repositories. A complete solution requires (1) geometric analysis of the underlying grid format of the source tables, (2) topological analysis of the relationship between the row and column header cells and the data-content cells, (3) interpretation and reconciliation of the semantics of both header and data cells, and (4) integration of the data into a readily accessible form. For the example in Figures 1 and 2 the third step can only be accomplished with advanced domain expertise of the educational infrastructure of each country. In general, understanding table data requires considerable knowledge external to the table. The fourth step requires semantic data integration—a task defying full automation. We intend to concentrate on the first two steps and tap the best techniques available for the last two.

The significance of developing capabilities for harvesting semi-structured data from web tables cannot be overestimated. Almost all nations post quantitative data such as the lengths of rivers or coast lines, heights of mountains, areas of lakes, population, age, ethnic origin, birth and death rates, immigration and emigration, education, employment, industrial production, commerce, and transportation. Canada Statistics (www.statcan.gc.ca), for example, has over 38 million series/vectors in over 2800 tables [STAa]. The Swiss site, from which Figure 1(a) was taken, currently has 50,033 tables [STAc]. Indiastat is even larger [STAb]. In the US more specialized sites are maintained by various government departments: Agriculture, Energy, Health, … . The CIA World Factbook and several international organizations like UNICEF and the Worldbank offer tables of worldwide data. These sites are consulted frequently by the general public and by decision makers.

Although tables remain the accepted method for displaying data for human access, table layout and structure has been undergoing rapid change since our first table studies twenty years ago [Kry90]. Layout used to be governed primarily by human visual acuity and by page paper size (with rules promulgated by the US Government Printing Office and the University of Chicago manuals of style). However, advances in digital technology for page layout, typesetting, spread sheets, and browsers (e.g., scrolling, zooming, dynamic tables) have had significant effect on best practices of table construction. In this two-year proposal, we focus on a large subset of web tables we call *grid tables*. We postpone for future research, work on tables not laid out on a (perhaps invisible) grid, nested tables, concatenated tables, and tables containing graphics.

Tables appear to be simple objects, but in fact the rules governing their layout and composition are recondite. It is now widely accepted that table understanding is a high-level cognitive skill that is not easily programmed. The intellectual challenge of the proposal is the systematic analysis and formalization of geometric and topological table syntax and of intra- and inter-table



Degré tertiaire, hautes écoles universitaires: titres délivrés selon la haute école et le domaine d'études, en 2006

	Licences et diplômes			Diplômes de Bachelor			Diplômes de Master		
	Total	Femmes	Etrangers	Total	Femmes	Etrangers	Total	Femmes	Etra
		%	%		%	%		%	
Total	7,900	55.7	12.2	4,987	44.4	15.6	2,267	39.5	
Haute école/Université									
Université de Bâle	540	55.9	9.6	542	51.5	12.9	203	52.7	
Université de Berne	867	57.3	5.7	492	42.9	3.0	214	46.7	
Université de Fribourg	542	66.2	11.8	719	57.6	10.4	342	46.8	
Université de Genève	1,485	65.7	20.6	557	62.5	22.6	168	57.7	
Université de Lausanne	927	58.9	11.3	457	45.5	15.5	73	49.3	
Université de Lucerne	3	66.7	-	209	58.9	5.7	99	54.5	
Université de Neuchâtel	207	64.3	11.6	154	51.9	18.2	76	46.1	
Université de Saint-Gall	25	32.0	20.0	546	24.7	28.2	232	31.5	
Université de Zurich	2,091	55.5	7.4	110	50.0	6.4	19	68.4	
Université de la Suisse italienne	127	48.8	47.2	174	48.3	25.3	92	56.5	
EPF Lausanne	0	-	-	646	25.7	21.5	601	20.8	
EPF Zurich	1,086	32.8	13.4	381	28.6	9.7	148	29.1	
Domaine d'études									
Sciences humaines et sociales	3,635	69.6	12.4	714	69.9	17.8	148	67.6	
Théologie	61	55.7	21.3	36	63.9	13.9	11	27.3	
Langues et littérature	725	78.2	17.5	60	78.3	16.7	6	50.0	
Sciences historiques et cultures	713	60.3	7.7	60	58.3	1.7	9	77.8	
Sciences sociales	2,050	69.6	12.0	519	71.3	20.0	122	71.3	
Sciences humaines et sociales pluridisciplinaires et autres	86	83.7	11.6	39	61.5	17.9	0	-	
Sciences économiques	852	29.6	16.5	1,240	28.1	20.2	457	34.6	
Droit	822	54.7	8.0	1,078	57.3	8.7	478	54.8	
Sciences exactes et naturelles	977	38.2	11.9	1,040	39.6	14.9	608	36.8	
Sciences exactes	364	20.9	15.7	305	18.7	19.7	281	20.6	
Sciences naturelles	506	48.0	10.1	676	49.6	12.4	310	50.6	
Sciences exactes et naturelles pluridisciplinaires et autres	107	50.5	7.5	59	33.9	18.6	17	52.9	

(a)

Étudiants inscrits dans l'enseignement supérieur public et privé

Étudiants inscrits dans l'enseignement supérieur public et privé

	2005-2006	2006-2007 (p)
Universités hors IUT et hors formations ingénieurs	1 283 516	1 259 425
Institut universitaire de technologie (IUT)	112 597	113 769
Écoles d'ingénieurs (1)	108 057	108 846
Institut universitaire de formation des maîtres (IUFM)	81 565	74 161
Section de technicien supérieur (STS)	230 403	228 329
Classes préparatoires aux grandes écoles (CPGE)	74 790	76 160
Préparations intégrées	3 058	3 162
Écoles de commerce, gestion, vente et comptabilité	88 437	87 333
Établissements universitaires privés	21 306	21 024
Écoles paramédicales et sociales	131 654	131 654
Autres établissements d'enseignement supérieur (2)	147 884	150 523
Total enseignement supérieur	**2 283 267**	**2 254 386**

p : données provisoires.

(1) : écoles et formations d'ingénieurs, universitaires ou non, y compris les formations d'ingénieurs en partenariat.

(2) : écoles normales supérieures, écoles juridiques et administratives, écoles supérieures d'art et d'architecture, écoles vétérinaires, grands établissements et autres écoles.

(b)

Statistics Canada Statistique Canada

Canada

Home > Summary tables >

Related tables: Fields of study, Students, Educational attainment.

University degrees, diplomas and certificates granted, by program level and instructional program (Bachelor's and other undergraduate degree)

	Bachelor's and other undergraduate degree				
	2003	2004	2005	2006	200
	number				
Total, instructional programs	140,898	148,563	151,884	161,031	175,39
Education	18,111	18,483	18,327	19,023	19,51
Visual and performing arts, and communications technologies	5,283	5,955	6,207	6,585	7,05
Humanities	16,683	17,067	18,480	19,530	21,39
Social and behavioural sciences, and law	31,368	33,573	33,984	37,101	41,36
Business, management and public administration	22,122	23,691	24,627	25,743	28,70
Physical and life sciences, and technologies	11,133	11,520	12,024	13,014	14,79
Mathematics, computer and information sciences	7,725	7,947	6,969	6,486	5,76
Architecture, engineering and related technologies	11,514	11,760	11,688	12,219	13,12
Agriculture, natural resources and conservation	2,196	2,028	1,815	2,028	2,24
Health, parks, recreation and fitness	13,935	15,471	16,605	17,994	20,07
Personal, protective and transportation services	99	204	228	297	30
Other instructional program	729	864	930	1,011	1,05

Note: For Quebec and Alberta institutions and University of Northern British Columbia, the Classification of Instructional Programs (CIP) 2000 codes assigned to programs are under review. Counts of university degrees, diplomas and certificates granted for 2002 and 2003 have increased because of the inclusion of qualifications awarded at the Ontario College of Art and Design. For University of Regina, enrolments since 2005/2006 and qualifications awarded since 2005 are not available. For Quebec institutions, qualifications awarded do not include micro programs and attestations. The 2004 qualifications awarded for University of Alberta are preliminary estimates and are expected to be updated with the next release. The reconciliation of 2007/2008 data for U. of Nipissing, U. of Windsor, U. of Manitoba, U. of Northern British Columbia and U. of British Columbia are not yet completed. Program level was updated for years 1998/1999 to 2006/2007 for Dalhousie University, U. of Manitoba and St. Mary's University College. Counts for qualifications awarded were revised from 2004 to 2006 for Atlantic School of Theology, Alliance University College, Canadian Nazarene University College, St. Mary's University College, Simon Fraser University and U. of British Columbia. Field of study was updated for years 1998/1999 to 2006/2007 for several institutions.
Source: Statistics Canada, CANSIM, table (for fee) 477-0014.
Last modified: 2009-07-13.

(d)

Table 9A: Graduate Output during 2003.

No.	Faculty/Course	Pass out Male	Pass out Female	Pass out Total
1.	Arts-B.A level courses	547324	425396	972720
2.	Science-B.Sc level Courses	196058	131717	327775
3.	Commerce-B.Com level Courses	227744	145448	373192
4.	Education-B.E	58258	47790	106048
5.	Engineering/Technology-B.E level Courses	101143	26467	127610
6.	Medicine-Bachelor level courses	22756	16031	38787
7.	Agriculture-Bachelor level Courses	6524	1277	7801
8.	Vet. Science-Bachelor level Courses	1151	346	1497
9.	Law-LL.B level courses	47008	11220	58228
10.	Others: Lib.Sc, Journalism, Phy.Edn., Music, Fine Arts, Computer Appl., Performing Arts, Mass Comm, Visual Arts, Theatre, Hospitality Mgt. etc.-Bachelor level courses	27478	11061	38539
11	**Total Graduates**	**1235444**	**816753**	**2052197**

(c)

ies NATIONAL CENTER FOR EDUCATION STATISTICS

The Daily; publications; and

DIGEST of EDUCATION STATISTICS

2008 Tables and Figures | All Years of Tables and Figures | Most Recent Issue of the Digest

Table 254. Bachelor's degrees conferred by degree-granting institutions, by discipline division: Selected years, 1970–71 through 2004–05

Discipline division	1970–71	1975–76	1980–81	1985–86	1990–91	1993–94	1994–95	1995–96	1997–98	1998
1	2	3	4	5	6	7	8	9	10	
Total	**839,730**	**925,746**	**935,140**	**987,823**	**1,094,538**	**1,169,275**	**1,160,134**	**1,164,792**	**1,184,406**	**1,200,**
Agriculture and natural resources	12,672	19,402	21,886	16,823	13,124	18,056	19,832	21,425	23,276	23,
Architecture and related services	5,570	9,146	9,455	9,119	9,781	8,975	8,756	8,352	7,652	8,
Area, ethnic, cultural, and gender studies	2,579	3,577	2,887	3,021	4,776	5,435	5,511	5,633	5,976	6,
Biological and biomedical sciences	35,683	54,085	43,003	38,320	39,377	51,157	55,790	60,750	65,583	64,
Business	115,396	143,171	200,521	236,700	249,165	246,265	233,895	226,623	232,079	240,
Communication, journalism, and related programs	10,324	20,045	29,428	41,666	51,650	51,164	48,104	47,320	49,385	51,
Communications technologies	478	1,237	1,854	1,479	1,397	869	865	853	878	1,
Computer and information sciences	2,388	5,652	15,121	42,337	25,159	24,527	24,737	24,506	27,829	30,
Education	176,307	154,437	108,074	87,147	110,807	107,440	105,929	105,384	105,833	107,
Engineering	45,034	38,733	63,642	77,391	62,448	62,247	62,331	62,257	60,252	58,

(e)

Fig. 1. Tables of data about university degrees in (a) Switzerland, (b) France, (c) India, (d) Canada, and (e) the US. Tables (a), (b) and (d) are from national statistical web sites, (c) from the Worldbank Education site, and (e) from International Center for Educational Statistics. Note the heterogeneity of these sites.

Degrees by discipline and year								
	Canada		France		India		Switzerland	
	2005	2006	2005	2006	2005	2006	2005	2006
Humanities								
Social Sciences								
Science	18,993	19,500			327,775		578	1040
Engineering	11,720	12,219	108,057	108,846	127,610		369	633
Business								

Fig 2. Target format for compiling data from tables in Figure 1 for further analysis. (We realize that the numbers here are incomplete and are beyond and inconsistent with the data in the tables in Figure1. This table is only an illustration of a plausible partial result.)

constraints and relations. In addition to refining our and others' techniques for transforming individual tables into a form suitable for combining information from several tables, we plan to formalize table-related information. We propose to develop a table *structure ontology* that codifies useful aspects of tables and a *table task ontology* for table processing methods, algorithms, and software. Both will be open and extendable. We will interact—and let other interested researchers interact—with the developing ontology through the project web site.

We present the details of our approach to meeting the challenge of largely automating table understanding as follows. Section 2 describes our approach to analyzing grid tables. Section 3 lays the foundations for integrating data from multiple tables. Section 4 proposes an experimental evaluation of both tasks, based on the extraction and consolidation of data from ten diverse websites. Section 5 outlines the scope of the proposed table ontology. Our research plan and the educational contributions integrated therein are presented in Section 6. Section 7 summarizes our research on related topics under prior NSF sponsorship. Finally, in Section 8 we point out the expected significance of our proposed research.

2. TABLE GEOMETRY AND TOPOLOGY

After a brief review of previous work, this section explains how we propose to transform web tables into layout-independent XML-based Augmented Wang Notation (AWN) for back-end semantic analysis.

Comprehensive reviews of two decades of research on table processing appear in [ZBC04, EHLN06]. Researchers first developed algorithms for specifying cell location based on rulings [LV92, Ito93] or, in the case of unruled [Han01] and ASCII tables [PC97, KD98], developed algorithms to determine typographic similarity of cell content and alignment [KK01, KHG05]. More recently researchers have addressed the information organizational aspects of tables, including associating content cells with header cells [Hur00, ETL05, ELN06, JNS+09]. We have devised methods to exploit the similarity of multiple tables from the same hidden-web source [TE09] and initiated analysis of augmentations that are not part of the primary grid, such as table titles, captions, units, footnotes, and aggregates [PJK+09]. A reported initiative for an end-to-end system divided the table-understanding task into table detection, segmentation, function analysis, structural analysis, and interpretation, but it was not implemented and did not define which tables could or could not be processed [SJT06]. None of the methods that address web tables (e.g., [PSC+07, GBH+07]) carries the analysis to the layout-independent multi-category level.

Our emphasis is on processing heterogeneous tables from diverse web sources and compiling their content for narrow-domain decision-support systems. The following flow diagram summarizes the steps detailed in the rest of this section. Section 3 shows how we begin with

AWN (Augmented Wang Notation) and move through several more steps to the ultimate goal of table understanding.

Web-table → Spreadsheet table → XY tree→ P-Notation → AWN

2.1 Table Geometry and XY trees

Figure 3 displays the canonical structure of a *grid table*. Its support is a rectangular grid. The table has four rectangular regions separated by a horizontal and a vertical ruling. Their point of intersection uniquely defines *stub-header (SH), column-header (CH), row-header (RH),* a*nd data-cell (DC) regions* of a grid table. Our challenges are to (1) analyze the possibly complex and hierarchical cellular structure of the header regions and (2) detect and recognize the augmentations and aggregations in the data cell regions.

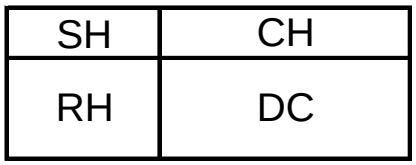


Fig. 3 A grid table, with the stub, column, row-header, and data-cell regions.

In addition to having the four identified regions, a *grid table* is one of several particular patterns of discrete rectilinear tessellations, or rectangular tilings. The tilings partition an isothetic rectangle into rectangles defined on an $m \times n$ lattice, which allow for a unique representation of its geometry by the locations and types of all its *junction points* at which two orthogonal cell boundaries meet or cross. Some tilings, called *XY-tessellations,* can be obtained by a divide-and-conquer method based on successive horizontal and vertical guillotine cuts. The number of tilings, $N_{all(m)} \equiv N_{all(m,m)}$, increases exponentially with the size of the grid. A quick count reveals that even a 4×4 grid has 70,878 different partitions. XY tilings represent the miniscule but indispensable, fraction of all tilings that are likely to be encountered as tables. Klarner and Magliveras proved that the number $N_{xy}(m)$ of XY-tessellations decreases quickly relative to the size of the grid [KM88]. Although $N_{xy}(4) = 68{,}480$, which does not differ in order of magnitude from 70,878, $\lim_{m \to \infty} N_{XY}(m) / N_{all}(m) = 0$.

All grid tables are XY-tessellations, but not all XY-tessellations are grid tables. Figure 4 shows several XY-tessellations—(a), (b), (d), and (e)—and one non-XY-tessellation—(c). In the VLSI literature non-XY-tessellations are known as *nonslicing structures* [KO90]; they seldom, if ever, occur in real table layouts. Figure 4a shows a concatenated table with two different column headers for the two *concatenated* tables, and Figure 4b shows a *nested* table. In the near-term (and for this proposal), we limit our objective for these tables to recognizing them for possible routing to a human interpreter. Figure 4 also shows two non-tabular XY-tessellations, (d) and (e).

Although not all XY-tessellations are grid tables, XY-tessellations provide a way to recognize table topology via table geometry. A key is to recognize that XY-tessellations, like polar graphs[1], abstract away the geometry of rectangular tilings but preserve the adjacency relationships between the tiles. It is known that horizontal and vertical polar graphs (which are duals of each other) can be drawn for any rectangular tessellation. For a *slicing structure* (an XY-tessellation), polar graphs are series-parallel. Thus, XY tessellations provide a *structural* representation of the

[1] Polar graphs can be traced to a 1940 paper on the dissection of rectangles into squares [BSST40].

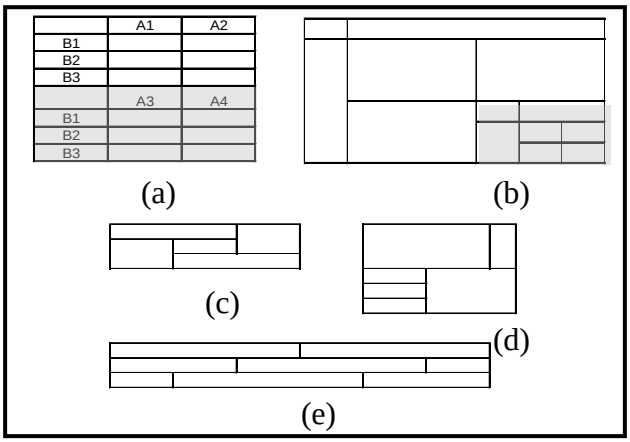


Fig. 4. Rectilinear Tessellations rejected as grid tables.

rectangles obtained by horizontal and vertical cuts at alternating levels of a tree, whose resulting partitions can be represented by *XY-trees*[2].

XY-trees for grid tables provide the bridge we seek to transform web tables into AWN. Since web tables can be readily imported into Excel, we analyze their Excel incarnation instead of attempting to parse the original HTML or PDF files (some sites, like the source of Figure 1a, expressly provide for Excel downloads). In collaboration with Prof. S. Seth at UNL and Prof. M. Krishnamoorthy at RPI, we have developed an algorithm based on recursive horizontal and vertical subdivision to transform an Excel table into a linear lexical representation of its XY-tree, called *P-Notation*. P-Notation is the input to the transformation to Wang notation (Section 2.2), which we then further carry into AWN—Augmented Wang Notation (Section 2.3).

2.2 Table Topology and Wang Notation

XY-trees represent only the physical layout of a table, which can be modified to suit page size, column width, or display characteristics. The first step in *understanding* a table is to analyze its category structure, which is independent of its presentation aspects. Interpretation requires understanding the relationship between *headings* and *content cells*. In 1996 Wang proposed a new representation for this purpose [Wan96]. It models headings as category trees (*labeled domains*) whose Cartesian product provides the paths to every data content cell, which Wang calls *delta cells*. The number of category trees is the *dimensionality* of the table. Figure 5 displays the P-Notation and the category trees for a simple three-category table. Its *size* is the product of the number of rows and columns of delta cells.

Any *well formed table* (WFT) can be represented in Wang Notation. The necessary condition is that any combination of paths, one through each category tree, must specify a unique delta cell [JN08]. Equivalently, the cardinality of the Cartesian product of the unique paths through the category trees must be equal to the number of delta cells. Figure 6 shows a WFT with four Wang categories. WFTs are seldom encountered in practice, but most tables can be readily transformed to a WFT without loss of content. Often only a category root is missing: for example, in Figure 2,

[2] We originally proposed XY trees for page layout analysis [NS84, KNSV93]. They have been periodically rediscovered and are also known by other names like *puzzle tree* or *treemap* [Samet06]. They transform a 2-D structure into two interlaced 1-D structures.

Vertical-cut-first P-Notation:

[{ [C D][C1 {D1 D2 }][C2 {D1 D2 }]} {A [{A1 [A11A12]}A2][d11 d12 d13][d21 d22 d23][d31 d32 d33][d41 d42 d43]}].

Category notation:

(A,{(A1,{(A11,Φ),(A12,Φ)}),(A2,Φ)})
(C, {(C1,Φ),(C2,Φ)})
(D, {(D1,Φ),(D2,Φ)})

		A		
		A1		A2
C	D	A11	A12	
C1	D1	d11	d12	d13
	D2	d21	d22	d23
C2	D1	d31	d32	d33
	D2	d41	d42	d43

Delta notation:

δ({A.A1.A11,C.C1,D.D1}) = d11
δ({A.A1.A12,C.C1,D.D1}) = d12
...

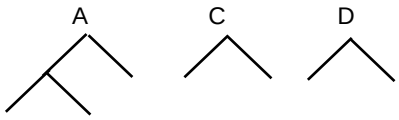


Fig. 5. P-Notation and Wang Notation for a simple 3-dimensional table. P-Notation is a breadth-first traversal of the XY tree. Category-notation is a breadth-first traversal of each category tree (here A, C and D). Delta notation is a (comma-separated) list of the paths to every delta cell.

			A													
			A1				A2				A3					
			B				B				B					
			B1		B2		B3	B1		B2		B3	B1		B2		B3
			B11	B12	B21	B22		B11	B12	B21	B22		B11	B12	B21	B22	
C	D																
C1	D1	D11															
		D12															
	D2	D21															
		D22															
C2	D1	D11															
		D12															
	D2	D21															
		D22															

Fig. 6. A well-formed table. This table has four Wang categories. Its size (the number of delta cells) is the Cartesian product of the number of category paths (3 x 5 x 2 x 4 = 120).

the root-category headings *Year*, *Country*, and *Discipline* in this three-dimensional table were omitted because they are obvious to the reader. Since Wang Notation requires rooted category trees, our programs would automatically add *virtual headings*. The creation of virtual headings does not require "understanding" the categories: arbitrary unique labels are acceptable.

We can already convert some types of WFTs into Wang Notation [PJK+09], but must generalize our algorithms to at least the following common cases: (1) Headers of WFTs vary in the absence or presence and location of their category roots. (2) Top-level headings are often above both row and column headers, which destroys their otherwise symmetric structure (e.g., category roots C and D in Figure 6). (3) Stub headers may be empty, or contain either or both category roots, or indicate the content of the delta cells. (4) Units may give rise to additional rows of spanning cells (e.g., "number" in Figure 1(d)). Generally, the variability encountered in the layout of the row and column headers of WFTs approaches that of word-order in grammatically correct sentences, with the additional challenge of two physical dimensions.

Because P-Notation is a string-like sentence, it is natural to attempt to analyze it by means of a grammar. We have made a start on this with a simple grammar that produces Wang Notation for categories trees with arbitrary breadth and depth [JNS+09]. Consider Figure 7 as an example. Since only the structure matters for P-Notation, we replace all the textual labels in the table by the

Employment Status					
Unemployed			Employed		
Education					
High School or Less	College		High School or Less	College	
	BS/BA	Graduate Degree		BS/BA	Graduate Degree

Fig. 7. Sample table column heading for grammar $\mathfrak{G}1$.

generic symbol c. Then by XY-cuts (horizontal first), we obtain the following XY-tree "sentence" S_{XY} for this tessellation:

S_{XY} = {c[c c]c[c{c[c c]}c{c[c c]}]}.

In the notation, the first c corresponds to "Employment Status", the second to "Unemployed", and so forth. We alternate brackets and braces for ease of reading, but they are equivalent. We also carry this alternation of brackets into our grammar G1:

S := A
A := {B}
B := c[X]B | c[X]
X := cX | AX | A | c

where the non-terminals in G1 serve the following functions.

S is the start symbol (eventually to generate all admissible strings for tables).
A is the non-terminal that generates all admissible strings for column headers.
B generates one or more instances of categories in the form "c[X]".
Each c becomes a root category and X generates its subcategory tree.
X generates strings of length ≥ 1, with arbitrary occurrences of c and A.

Grammar G1 can *shift-reduce* parse fully parenthesized input for column headers of tables with arbitrary dimensions and any number of levels in each dimension. It is a simple matter to add a mirror-image grammar to parse the row headings and delta cells. From the parse tree we can obtain the Wang Notation.

We are confident that in the course of the proposed research we can expand G1 to G2 (and perhaps to G3, G4, …) that will cope with most of the variations encountered in practice. The fraction of tables that the expanded grammars G2 … accepts will be estimated as part of the Evaluation (Section 5).

2.3 Augmentations and Aggregates

A*ugmentations* appearing in a table do not depend on the header-to-content cell mappings. An augmentation may apply to the entire table (e.g., *Table Title, Table Caption, Notes*), to one or more rows or columns (e.g., *Unit, Footnote*), or to a single cell of the table (e.g., *Footnote, Note*). We distinguish between *Footnote citations* and *Footnote references*.

In contrast to augmentations, *aggregates* are data appearing in delta cells, rather than supplementary explanations of such data. All of the tables in Figure 1 display a *total*. Typical geographic aggregates include totals, averages, and medians under region designators. An aggregate can function both as a category root and as a category leaf cell. For instance, in Canadian census data, the row header Canada can denote both the root node of the Province subcategories, and yet have value cells of its own.

In a recently conducted experiment on 193 tables from ten web sites, we found that 87 tables contained aggregates (up to 43 in a single table), and 73 tables contained footnotes (up to 214 in one table) [PJK+09]. Multilinear estimation of interactive processing time, based on the number of aggregates, footnotes, size, and dimensionality, yielded a source-specific correlation coefficient of 0.75 between predicted and observed processing time (ρ_{global} = 0.67).

Tables with aggregates took much more time to mark and classify than tables without them. Our current method of selecting and annotating aggregate cells and footnotes is cumbersome. Tagging

these cells is human intensive, time consuming and error prone. Fortunately automating the identification and annotation of aggregates and footnotes appears quite feasible. Spanning cells containing *units* should also be relatively easy to detect automatically.

Analyzing the logical structure of a table and recording the augmentations and aggregations is necessary but by no means sufficient for understanding it or for combining its contents with the contents of other tables. Both require context and knowledge that extend beyond the table under consideration. There is ample evidence that automating table understanding, or even merely verifying claims to this effect, is difficult [LN00, HKL+01, NL02]. We shall retain the XML format for AWN that we developed with prior NSF support. This format, although disagreeably verbose, serves as the bridge between the analysis of individual tables and semantic analysis leading to table understanding and integration. This aspect of the project is described in the next section. RPI and BYU have complementary expertise in front-end table processing and semantic analysis, so we expect research on each phase to inform and advance the other.

3. SEMANTIC ENRICHMENT AND INTEGRATION

As stated in the Introduction, we intend to facilitate the compilation and interpretation of data from heterogeneous sources such as the table sources in Figure 1. To reach this goal, we must do much more than just index the delta cells using Wang category trees and label the augmentations. For Table (c) in Figure 1, for example, after establishing the relationship of Wang trees to delta cells as Figure 8a shows and the augmentations as Figure 8b, shows, we also need the semantically enriched conceptualizations exemplified by the conceptual-model hypergraph of Figure 9.

The semantic conceptualizations we seek to derive (Figure 9) comprise linguistically grounded named objects (e.g., *Bachelor's_and_other_undergraduate_degree* and *Canada*), named object sets (e.g., *Year*, *instructional_program*, and *number*), optionally named relationship sets (e.g., the 5-ary hyperedge), and constraints (e.g., the functional dependency *Bachelor's_and_other_ undergraduate_degree*, *Year*, *Country*, *instructional_program* → *Number*, and the summation constraint). We call the derivation of semantic conceptualizations from augmented AWN table interpretations *semantic enrichment*.

(Bachelor's and other undergraduate degree.2003, ⊥.Education) → 18,111
(Bachelor's and other undergraduate degree.2003, ⊥.Visual and performing arts, and communications technologies) → 5,283
…

(a)

Title("University degrees, diplomas and certificates granted, by program level and instructional program")
Note("(Bachelor's and other undergraduate degree)")_applies_to_Title("University degrees, diplomas and certificates granted, by program level and instructional program") Note("Canada")
Unit("number")_applies_to_Data_Cell($Data\_Cell_{11}$)
Unit("number")_applies_to_Data_Cell($Data\_Cell_{12}$)
…

(b)

Fig. 8. Augmented Wang-Interpretation of Table (c) in Figure 1.

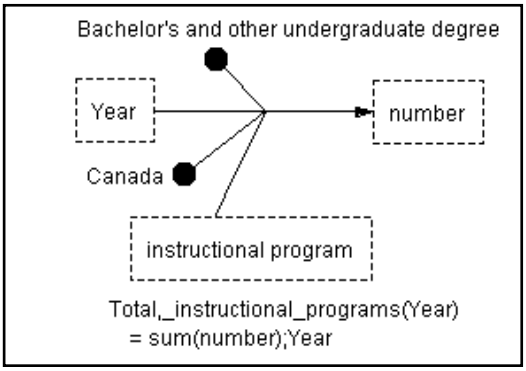


Fig. 9. Conceptualization of Table (c) in Figure 1.

One way to convert discovered category trees, delta cells, and augmentations into semantic conceptualizations is via an application-specific conceptual-modeling tool with a human in the loop. Allowing human intervention ensures that any desired result can be obtained and that any results derived automatically can be corrected if necessary. We strive, however, to automate semantic enrichment by resorting to semantic resources such as WordNet [Fel98],linguistically grounded extraction ontologies, ontology snippets, and value recognizers [ECJ+99, BCHS09]. In [LE09] we show how to match nodes in dimension trees, data in data cells, and textual components of augmentations with semantics in community-established semantic resources. For example, we recognize the leaf nodes in the first dimension tree in Figure 8a as a year concept and let *Year* be an object set in Figure 9. We recognize "Canada" as the name of a country object, and we pick up the root node of the first category tree as the object *Bachelor's and other undergraduate degree*. A side benefit of our research will be the development and deployment of a library of highly tuned value recognizers and ontology snippets extending those we have heretofore created [ECJ+99].

Although semantic enrichment goes a long ways toward the goal compiling data from a heterogeneous collection of tables into a unified user view, an additional integration step is necessary to complete the task. Semantic data integration is said to be *AI-complete*, which is a euphemism for "believed to be impossible to solve completely." However, we are in a good position to build "best-effort," "pay-as-you-go" integration systems because

> (1) We have considerable experience with data integration [EJX01, EJX02, BE03, XE03, EXD04, XE06], and our approach of using ontology snippets in schema-mapping lends itself particularly well to table-data integration [ETL05, TE09].
>
> (2) We have implemented a tool, *MapMerge* (Figure 10), that synergistically supports automatic, semi-automatic, and manual integration [Lia08]. The human in the loop can resolve subtle problems such as determining declared equivalences among international academic degrees and non-uniform designations for disciplines in Figure 1.
>
> (3) The overall task of integrating information from a collection of heterogeneous tables is broken down into manageable subdivisions: (a) interpret table, (b) semantically enrich table, (c) map related table conceptualizations, (d) allow human interaction but relegate as much as possible to the system.

As stated earlier, we cannot hope to resolve all semantic-enrichment and map/merge problems. What we propose here is only to automate what we can by exploiting some of the best methods that we and others have discovered for semantic enrichment and integration.

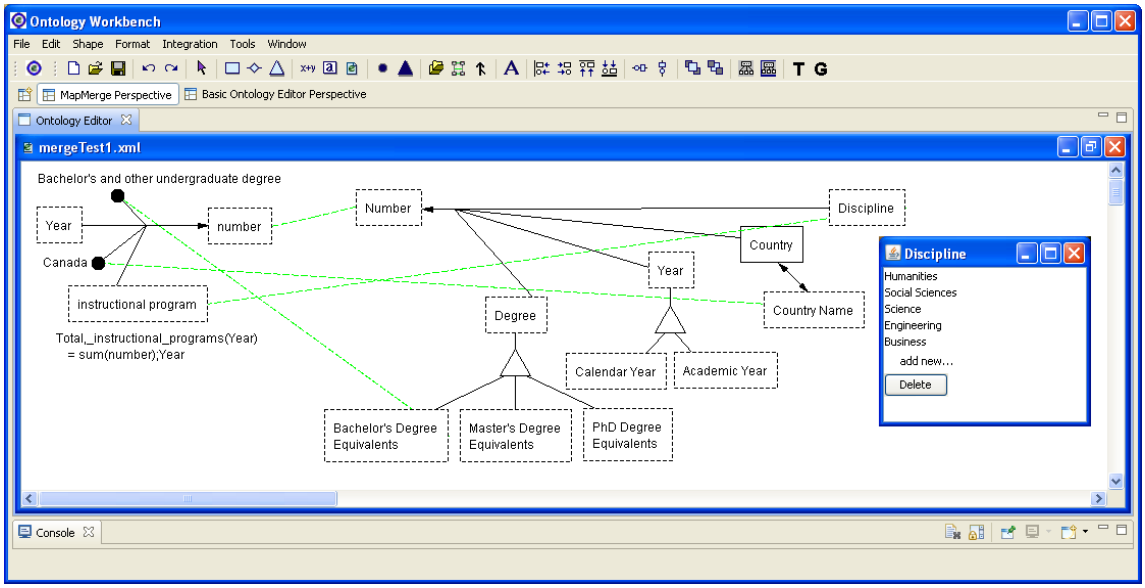


Fig. 10. *MapMerge* Example Showing the Mapping of the Conceptualization in Figure 9 into a Growing Conceptualization in a Form Displayable as the Table in Figure 2.

4. EXPERIMENTAL EVALUATION

The evaluation plan is simple, but will require careful preparation and considerable effort and time. It is a major component of this proposal because it will help to assess not only new ideas but also already available table processing methods.

E1. Collect 2000 tables from ten large institutional research sites in roughly equal proportions. Save locally all the tables in both original and Excel format.

E2. Convert interactively as many of the 2,000 tables as possible to Augmented Wang Notation using our developed table-interpretation software. Time stamp and log any and all user interventions. Save all rejected tables in their original format for further analysis.

E3. Select one of the institutions (probably Canada Statistics) as the primary source and construct an empty database (OUR_DB) with the same schema as the DBMS from which the data posted on the web was generated (SOURCE_DB). Also obtain read-only SQL-equivalent access to the data on SOURCE_DB.

E4 Populate OUR_DB with the data extracted from tables from the primary source using our developed semantic-enrichment and our integration software.

E5. Compare the SOURCE_DB data with OUR_DB data. Analyze shortcomings and refine the software.

E6. Apply the modified software, without further testing on data from the other sites.

E7. Determine how much of the tabular data from the secondary sources can be accessed via OUR_DB.

E8. Analyze the cost (in terms of human time) of all the steps and report, by source and by table characteristics, the data successfully transferred to OUR_DB.

The objective of the evaluation is to measure the reduction in *human time and effort* by means of interactive table processing and the *accuracy and completeness* of the retained information.

Notes

(a) Only the evaluation scheme, rather the proposed paradigm, depends on the existence of a database from which tables are generated. The actual operation will harvest and organize web table data independently of any underlying DB.

(b) As a matter of statistical integrity, all of the analysis methods and the software for comparing data must be developed and frozen before the experiment is performed.

(c) The tables to be used will be selected blindly. We will draw tables with a pseudo-random scheme until 2000 potentially useable tables are selected. We will tag and store, but not attempt to convert, nested and concatenated tables and other table-like pages that do not conform to our definition of grid table.

(d) Much of the table selection and processing will be performed by undergraduates. We must decide how many different "operators" would produce the most representative results, and what prior training operators should have.

(d) The number of tables to be processed is based on analysis of the results of our prior 200-table experiment, and on balancing the expected source-and-feature specific mean-to-variance ratio against the time required to perform the experiment.

(f) In the course of our prior project the Director of Educational Outreach from Canada Statistics (Ottawa) visited us to discussed possible follow-up projects. We are not bound to Canada Statistics as the primary site, but we do need a collaborative organization.

5. TABLE ONTOLOGY

Although some may disagree, most see ontology as the formalized conceptualization of a domain of interest [Gru93, Jep09]. Acceptable purposes of ontology are to enlighten a common-interest based community, foster communication, speak with a common voice, and even enable interoperability among interested parties, human or otherwise (e.g., UMLS [UML], GO [GO]). Ontology must be based on some sort of community agreement. Ultimately, however, ontology can and should increase intellectual productivity, enhance learning, disseminate knowledge, and aid decision makers.

In the course of years of research, we have developed a detailed conceptualization of the world of tables [TEL+05, ELN06]. We have interacted with many researchers with similar interests, written several surveys [LN99, LN00, EHLN06], and processed tables from a variety of sources. As part of the contribution of this proposal, we intend to formalize our comprehensive conceptualization of tables as a table ontology—both a structure ontology defining what tables are and a process ontology describing various table-processing tasks ranging from table detection to table understanding. We shall seek community agreement on the conceptualization among the table processing community and make access to the ontology practical and beneficial to knowledge workers in other communities.

5.1 Table Structure Ontology

Although finding the most appropriate language for table ontology is part of the proposed research, OSM-O (Object-oriented Systems Modeling [EKW92] for Ontologies [EZ10]) has

some (but potentially not all) of the required properties. It provides a graphical representation of a decidable fragment of predicate calculus with reasonable complexity bounds that includes object and relationship sets for concepts and relationships among concepts, generalization/specialization (is-a) hierarchies, aggregation (part-of) hierarchies, and constraints for relationship-set cardinalities and object set disjointness and completeness. To enable reasoning, we can add safe, positive horn-clause rules without jeopardizing decidability and complexity bounds [Ros05]. Here we show how OSM-O can be used to formalize table properties that could also be captured by other ontology languages.

Formally, *OSM-O* is a triple (*O*, *R*, *C*):

- *O* is a set of object sets; each is a one-place predicate (like *City*(*x*) or *CityName*(*x*)); and each predicate has a *lexical* designation (human-readable objects like "Troy" for *CityName*) or a *non-lexical* designation (OID-identified objects like the actual city of Troy for *City*).
- *R* is a set of *n*-ary relationship sets ($n \geq 2$); each is an *n*-place predicate (like *Country*("Canada")*_has_estimated_Population*(30000000)*_in_Year*(1999))
- *C* is a set of constraints: referential integrity, min:max participation, hyponym/hypernym (e.g., Bachelor's Degree *is-a* University Degree) (including potentially disjointness or completeness or both), and meronym/holonym (e.g., Wasatch Mountain Range *is-part-of* Rocky Mountains). Each constraint is a closed well-formed formula (e.g., for referential integrity of an n-ary relationship set *R* connecting object sets (S_1, …, S_n): $\forall x_1 \ldots \forall x_n \left(R(x_1, \ldots, x_n) \Rightarrow S_1(x_1) \wedge \ldots \wedge S_n(x_n) \right)$.

Interpretations of OSM-O model instances in which all closed-formula constraints hold are *valid interpretations* (*models*, although we avoid this term because we are in a conceptual-modeling context where model instances refer to conceptual diagrams). We can now ontologically describe the conceptualization of some thing *t* as any valid interpretation for a conceptual-model instance of *t*. In particular, we can ontologically describe tables as any valid interpretation of an OSM-O model instance that has a particular configuration of objects, relationships, and constraints.

Figure 11 shows an example of an OSM-O model instance that describes a certain class of tables. Each solid box is a non-lexical object set and each dashed box is a lexical object set. Each line connecting object sets is a relationship set. An arrowhead on a relationship-set line denotes a functional relationship and an o denotes optional participation. Open triangles denote hypernym/hyponym *is-a* constraints with the generalization connected to the apex of the triangle and the opposite edge connected to the specializations. Completeness or union constraints (∪) or disjointness constraints (+) or both (⊎) may apply. Any valid interpretation for the OSM-O model instance in Figure 11 must have *n Category Root Nodes*, which, with a formal reading of the model instance in Figure 11, turn out to be the root nodes of the category trees in the Wang Notation discussed in Section 2.2. Further, any valid interpretation must also have the proper number of Data Cells, appropriately aligned with the leaf nodes of the category trees. Optionally, a Table can have a Title and a Caption and may have Augmentations associated with Titles, Captions, Category Nodes, and Data Cells. Additional constraints must be added to restrict this model to grid tables. Our table ontology will formally describe grid tables as well as other classes of tables.

5.2 Table Task Ontology

For the main part of the task table ontology, we take as input a source conceptualization and try to map it into a valid interpretation in the structural ontology. The task ontology is a framework for transformations from a source to the table structure ontology.

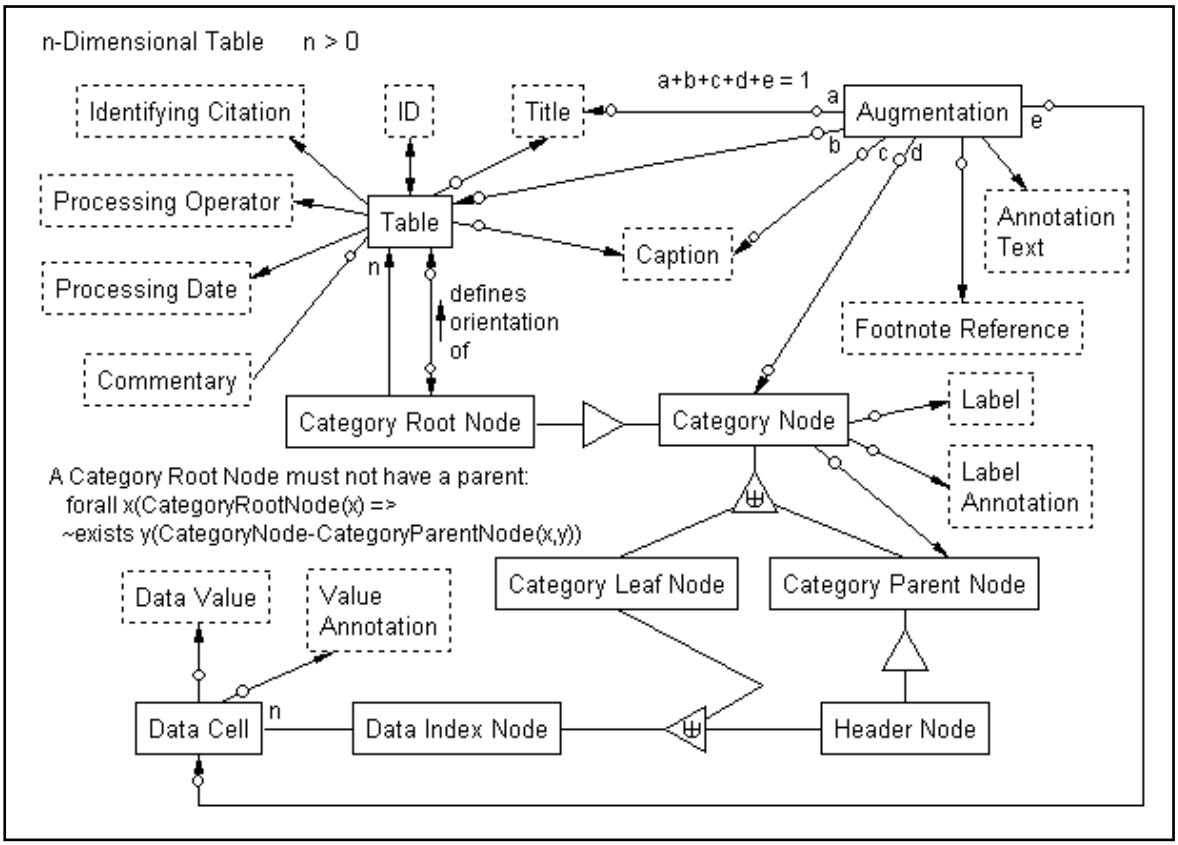


Fig. 11. Essentials for the Table Structure Ontology.

We can formally define a *source-to-table transformation* as a 5-tuple (R, S, T, Σ, Π), where R is a set of resources, S is the source conceptualization, T is a target table structure ontology for an S-to-T transformation, Σ is a set of declarative source-to-target transformation statements, and Π is a set of procedural source-to-target transformation statements. The *target conceptualization* is either the table structure ontology of Figure 11, or some alternative table conceptualization to be investigated. A specified *set of resources* could be WordNet and a library of value recognizers. *Source conceptualization* refers here to grid tables. *Declarative* and *procedural source-to-target transformation* statements can be written in any formal language based, for instance, on layout tessellations and grammars like G1 (Section 2).

The source-to-table transformations must preserve information and constraints given, discovered in or inferred from source documents. Let S be a predicate calculus theory with a valid interpretation, and let T be a populated OSM-O model instance constructed from S by a transformation t. Transformation t *preserves information* if there exists a procedure to compute S from T. Let C_S be the closed, well formed formulas of S, and let C_T be the closed, well formed formulas of T. Transformation t *preserves constraints* if $C_T \Rightarrow C_S$.

In previous work we have investigated information and constraint preserving transformations from OSM model instances to reduced OSM model instances [Emb98], relational databases to OSM [EX97], and XML-Schema documents to C-XML (an XML conceptual-modeling language) [AK07]. We now intend to do the same for grid tables and for semantic enrichment. This is, in essence, a reverse-engineering problem with some added twists: (1) the syntax for

specifying initial predicates and constraints has many more degrees of freedom and (2) we may access external semantic resources to help establish predicates and constraints.

For the evaluation we are specifically interested in mapping grid tables into relational tables through a series of information- and constraint-preserving transformations. If we can transform source grid tables to semantically-enriched OSM model instances, we can generate relational tables. Our work in [Emb98] describes this step in the process. Since we have already implemented the transformation, we need only integrate it into the proposed prototype system.

6. RESEARCH PLAN

Based on our previous research, results, and expertise, RPI will have primary responsibility for the front end procedures (web-tables to AWN), while BYU will concentrate on back-end analysis and interpretation of AWN tables. We will work together on the formulation of the ontology, experimental evaluation, website, progress reports, and publications.

We started working together at UNL (the University of Nebraska—Lincoln) in 1976 and by 1981 had co-authored several articles including an *ACM Computing Surveys* review of human-computer interaction in text editors. Since leaving UNL we have exchanged one or two visits per year. During our visits we present overviews of all our research and listen to prepared presentations by each others' students. We have also arranged student inter-university visits and plan to continue to do so. The proposed project is too small to require an elaborate management plan, but the necessary progression of activities is laid out in Table 1.

Table 1. Schedule (En refers to one of the steps of the Evaluation Plan of Section 4)

Month	BYU	RPI
1-3	E3 Engage graduate students. Examine alternative formulations and languages for table ontology.	E3 Engage graduate students. Construct examples of acceptable P-notation and AWN for grid tables of increasing complexity.
4-6	Collect and retarget semantic-enhancement and data-integration tools for populating DBMS with contents of AWN tables.	Integrate grid table-to-AWN software. Code matching of P-notation of XY-trees of new tables to P-notation of already processed tables.
7-9	Enhance semantic-enhancement and data-integration tools.	Generalize table grammar. Apply analysis algorithms to several dozen tables to determine weaknesses.
10-12	Initiate structural ontology.	Develop and document statistical analysis of experimental results in E7 and E8. Provide information for structural ontology.
13-15	E4 Complete structural ontology. Formulate task ontology.	E1, E2 Propose components for task ontology. Develop recognition of selected augmentations.
16-18	E5 Complete task ontology.	Test transformed P-notation of grid tables with new header formats to formats acceptable by the generalized grammar.
19-21	E6 Slack. Run experimental evaluation.	If necessary, fix P-transformation and grammar Slack Run experimental evaluation.
22-24	E7, E8 Prepare final report and publications	E7, E8 Prepare final report and publications.

Each institution will have one or two graduate students working on the project at any one time (which must necessarily include students partially supported by teaching assistantships), and 3 or 4 undergraduates (10 hr/wk Academic-Year and 40 hrs/wk Summer). In the past a weekly individual meeting with each student and a weekly group meeting has worked out well.

7. PRIOR NSF-SPONSORED RESEARCH

Our previous collaborative project "TANGO" leveraged the strengths of our research teams at Brigham Young University (IIS-0414644, PI: David Embley) and Rensselaer Polytechnic Institute (IIS 0414854, PI: George Nagy) during 2005-2008. TANGO is a framework for organizing domain-specific factual data appearing in independently generated web pages. Algorithms and software were developed for extracting and interpreting individual lists and tables and integrating them with the contents of other tables that present partially overlapping information. The TANGO framework automated some of the laborious data entry tasks for a domain when information exists in lists and tables that describe the domain, as demonstrated in a test on 200 web tables from large web sites. The tools developed in TANGO are the starting point for the proposed research. This cross-disciplinary and cross-university endeavor introduced 10 graduate students (including 3 women) and 4 undergraduate students to cutting-edge research. Developed tools (including source programs), technical reports, theses, and over 20 journal, book chapter and conference publications can be found on the TANGO website (http://www.tango.byu.edu).

8. SIGNIFICANCE

The need for *actionable* data is widely recognized [BLHL01, Buc06, HNP09]. The research proposed here will yield principled methods to:

(1) Process large collections of computer-constructed *web tables* from multiple, heterogeneous institutional sites into queriable knowledge repositories;

(2) Characterize table structure by *XY-trees* which transform the display-oriented 2-D format into interlaced 1-D lists for content analysis;

(3) Extract *annotations* and *aggregations* which, as Wang has already noted in 1996, are an essential component of most tables [Wan96];

(4) Enable table-level semantic *analysis* by determining the relationship of headers to content cells rather than only the geometric cell structure;

(5) Devise *table interpretation* procedures based on partial semantic analysis of cell contents to consolidate information from different tables (from either the same or different sources);

(6) Unite useful known facts, methods and algorithms discovered over two decades of research in systematic, accessible and extendable *table ontology*.