PROJECT DESCRIPTION

1 Introduction

The exponential increase in new knowledge that characterizes our modern age of information technology precludes depending solely on individual effort to keep up with new information. We must therefore develop new ways of "keeping up," and we must develop them quickly. The Semantic Web [BLHL01] offers a promise that we can "keep up" by allowing software agents to roam in cyberspace in our behalf, where they can gather information of interest and synergistically assist us in decision making and in negotiating for our wants and desires. This dream, however, relies on agents being able to find and manipulate useful information, which, in turn, relies on having an abundance of ontologically described repositories [DFvH03]. Hence, the fundamental enabling component for the Semantic Web is an ontological description of information, which provides for a shared understanding of a repository of information.

Unfortunately, creating ontological descriptions for information repositories is nontrivial. If we could automate the process, or at least make the process semi-automatic, we could significantly improve our chances of making the Semantic Web a reality. We thus propose a way to meet this challenge.

Motivated by our belief that inference about unknown objects and relations in a known context can be automated, we propose to develop an information-gathering engine to assimilate and organize knowledge. While understanding context in a natural-language setting is difficult, structured information such as tables[1] make it easier to interpret new items and relations. We organize the new knowledge we gain from "reading" tables as an ontology [Bun77] and thus we call our information-gathering engine *TANGO* (Table ANalysis for Generating Ontologies) [TELN03].

The implications of meeting this challenge of automatically generating ontologies are at the same time theoretically intriguing and practically significant. For a domain of interest and a set of tables within the domain, can we automatically establish intentional and extensional objects and relationships and constraints among them? Can we derive semantics from syntactic clues in the layout and content of metadata and data? Can we automatically recognize overlapping information and thus also recognize differences and add these differences to a growing body of knowledge? Can we recognize conflicts between new knowledge and previously obtained knowledge and then either resolve the conflicts or hold in abeyance alternative knowledge for later reconciliation? Finally, can we use the constructed and growing body of knowledge to support knowledge intensive Semantic-Web tasks such as answering queries, extracting knowledge, resolving semantic interoperability, and enabling information exchange between disparate software agents working within the same domain?

Because this challenge is intriguing and significant, others have also taken on this task [MS00, MGJ01, DF02a, DF02b, Ont03, Rub03, Gom03]. One area of agreement among all researchers is that this is an important problem, especially for Semantic-Web applications. OntoBuilder [Ont03] is the most advanced of these systems. OntoBuilder starts with a user-selected web page that contains a form. It analyzes the form (its fields and value options) and constructs an initial ontology. Once an ontology exists, the user refines the ontology and then suggests additional web sites with other forms. OntoBuilder attempts to interactively adapt the original ontology to cover concepts from

[1]Tables have a particular spatial layout of material [Wan96] that carries significant meaning [DeM80, FD92, Car00, CL00, Sow00]. [Lem98] describes tables as "organizational resources to enable meaningful relations to be recovered from bare thematic items in the absence of grammatical constructions," and argues that there is always "an implied grammar, and a recoverable textual sentence or paragraph for every table."

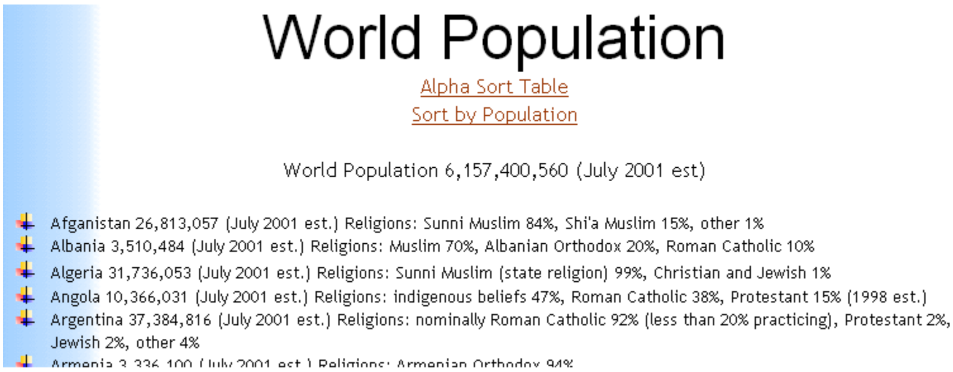

Figure 1: Partial Page of World Religious Populations [dlb03].

the new web pages. TANGO, our proposed approach, is similar to OntoBuilder except that we use tables rather than forms, and we choose different techniques for many low-level but essential details.

Our proposed work can be considered as semi-automated, applied "ontological engineering" [GL02]. As an analogy for what we are proposing, consider that instead of humans collaborating to design an ontology [HJ02], we enable tables to "collaborate" to design an ontology. In a sense, this is the same because TANGO assembles information from specific instances of human-created tables.

We plan to demonstrate the feasibility of automated knowledge gathering in the domain of geopolitical facts and relations, where relevant empirical data is widely scattered but often presented in the form of tables.[2] Using this domain, we illustrate the specifics of our ideas in Section 2, where we show that most semi-structured, factual data is table-equivalent, and in Section 3, where we show how we generate ontologies from sets of table-equivalent data. Section 4 explains how we expect to evaluate our work and measure its effectiveness. Section 5 presents our plan for accomplishing the proposed research, and Section 6 describes the expected significance for what we are proposing.[3]

The collective experience of the principal investigators positions them to succeed in the proposed research endeavor. The research team consists of a conceptual-modeling/database specialist, a linguist/ontologist, a document engineer, and a pragmatist with recent industrial experience with ontologies. The project will allow two PhD and two MS students to complete their graduate work, one under each of the professors, and should motivate four additional undergraduate students to extend their studies.

2 Table Normalization

Although many consider the idea of a table to be simple, a careful study (e.g., [LN00]) reveals that the question "What constitutes a table?" is indeed difficult to answer. As only two of thousands of examples, does the information in Figure 1 constitute a table? What about the information in Figure 2?

[2] The chosen domain of geography spans many important human activities: natural resources, travel, culture, commerce, and industry. It is also an application domain in which we have done some previous research, including topographic maps [LNS+00], satellite images [Nag84, Nag85, EN89, NME90], and geographic data processing [NW79, EN91, Nag00a].

[3] The many references to our own work in Sections 2, 3, and 6 show how we build on previous work and particularly how we build on a previous NSF grant (IIS-0083127).

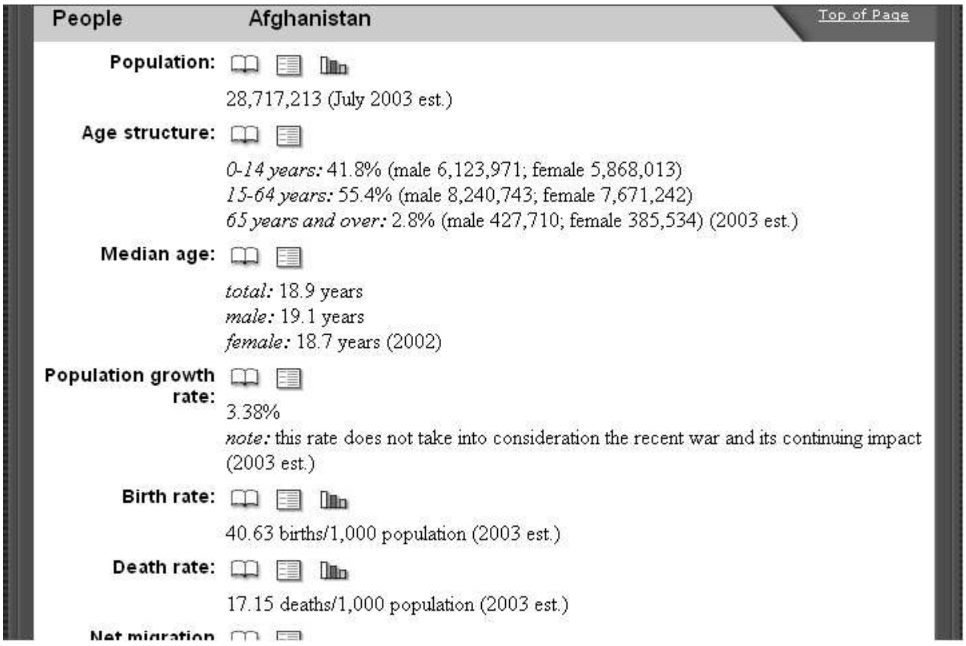

Figure 2: Partial Page from People in the 2003 CIA World Factbook [Wor03b].

We choose to define a table indirectly through information normalization. Working backwards, we first consider relations in a relational database to be tables in a normalized form. Using a standard, formal definition of a relational table [Mai83], we can define a normalized table as follows. A *schema* for a normalized table is a finite set $\{L_1, ..., L_n\}$ of label names or phrases, which are simply called *labels*. Corresponding to each label L_i, $1 \leq i \leq n$, is a set D_i, called the *domain* of L_i. Let $D = D_1 \cup ... \cup D_n$. A *normalized table* T with table schema S is a set of functions $T = \{t_1, ..., t_n\}$ from S to D with the restriction that for each function $t \in T$, $t(L_i) \in D_i$, $1 \leq i \leq n$.

As is common for relational databases, we often display tables in two dimensions. When we display a table two dimensionally, we fix the order of the labels in the schema for each function and factor these labels to the top as column headers. Each row in the table constitutes the domain values for the corresponding labels in the column headers. Thus, for example, we can display the normalized table $\{\{(A, 1), (B, 2), (C, 3)\}, \{(A, 4), (B, 5), (C, 6)\}\}$ as follows.

A	B	C
1	2	3
4	5	6

Displayed in this form, a normalized table is simply called a *table*. Whether the original information should be called a "table" may be debatable. To avoid the argument, whenever there may be doubt, we will refer to the information as *table-equivalent data*.

When we normalize the table-equivalent data in Figure 1, we obtain the table in Figure 3.[4] To normalize the table-equivalent data in Figure 2 to obtain the table in Figure 4, we first recognize that the data is split across many web pages; each page has the same data but for a different country. Thus, each page is itself a function from the labels, which consist of the bold label phrases on the left composed with the sublabel phrases on the right, to domain values, which are non-label

[4]In this and other tables, "missing" values are null values, which we assume are elements of every domain.

Country	Population (July 2001 est.)	Religion					
		Albanian Orthodox	Muslim	Roman Catholic	Shi'a Muslim	Sunni Muslim	other
Afghanistan	26,813,057				15%	84%	1%
Albania	3,510,484	20%	70%	10%			
...							

Figure 3: Partial Normalized Table for World Religious Populations [dlb03].

Country	Population (July 2003 est.)	Median Age (2002)			Population Growth Rate (2003 est.)
		Total	Male	Female	
Afghanistan	28,717,213	18.9 years	19.1 years	18.7 years	3.38%*
Albania	3,582,205	26.5 years	24.8 years	28.1 years	1.03%
...					

* Note: this rate does not take into consideration the recent war and its continuing impact

Figure 4: Partial Normalized Table for People in the 2003 CIA World Factbook [Wor03b].

values on the right. In addition, there are explanatory comments, which we can standardize by adding them as footnotes.

So, how can we determine whether we have table-equivalent data, and how can we turn table-like information into normalized tables? Since we have defined a table indirectly and by construction, we only need to answer the second question. If we can turn semi-structured information into a normalized table, we can declare that the semi-structured information is table-equivalent data and that the normalized table is a table.

There is a spectrum of cases to be considered. On the one extreme, we may already have information presented as a normalized table. All relational database tables, for example, are normalized tables, and many tables on the web appear in or essentially in normalized form. Other web tables, however, pose problems such as tables displayed piecemeal, tables spanning multiple pages, tables with no <table> tag, folded tables, tables with factored rows, tables with linked subtables, table rows with additional linked row values, all of which we have dealt with in previous work [ETL02, ETL04]. Some tables, more difficult to interpret, include features such as tables nested within table rows, folded table rows, and tables with both column and row headings. Table-equivalent data that does not have a typical two-dimensional layout is more difficult, but we have experimented with techniques to interpret them. Using ideas developed in [ETL02, ETL04], for example, we can obtain a basic resolution of which text is label text and which text is value text from the World Factbook in Figure 2 by comparing the pages—the label text stays constant from page to page whereas the value text changes. We can recognize the nested table in Figure 1 by recognizing the lists of label-value pairs (<Religion Name>, <Percentage Value>) in each row. Tables with images, such as GIF images for labels or values, and tables in non-HTML documents, such as in PDF documents, present even more challenges. Not only do we need OCR [RNN99] and image layout analysis [Nag00b], but these documents also provide even more freedom in table layout (for surveys see [Han99, LN00, ZdBC04]). We have also experimented with these types of tables in previous work [KNNR95, Haa98, LN99b, LN00, Nag00b, HKL$^+$01, LN02, TE02].

Our general approach to table normalization will be to create an ontology for table understanding. We have already begun this process in previous work [Haa98, LN00]. In [Haa98] we created a conceptual model for image-based table understanding that allows us to catalog information about tables and table cells, including the presence of lines and their location and thickness, the presence and location of text, and the hierarchical, *XY-tree* representation [NS84, AT98] of the document. In [LN00] we described a document taxonomy, a schema for document and table image analysis; we characterized tables in terms of their jargon, representation, and dimensionality; and we discussed logical/physical dichotomies leading to multiple tabular views of the same information. Much more knowledge about forms and form layout needs to be added to create the ontological knowledge we need to recognize table-equivalent data and normalize tables.[5] In obtaining and assembling this knowledge about tables, we will rely not only on our own work, but also on the work of many others, including: (a) linguistic [HD97, Kie98, Han01] and geometric [PCA97, RS97, HKLW00, HKLW01] characteristics that distinguish tables from text; (b) title/label/caption/footnote characteristics [Wan96, HD97] (c) frame (box) and ruling properties (topology and line type) [GK95a, GK95b, HD95, WQS95, TBB96, Zuy97]; (d) horizontal and vertical segmentation rules (alignment and spacing) [PC97]; (e) typesetting and linguistic rules for cell similarity (color, typeface and size, case, normal/bold/italic, alpha/digit, indentation, punctuation, leaders, lexical and grammatical categories) [PC97]; and (f) markup tags [LN99a]. Further, we will rely on a large corpus of sample tables and table-equivalent data, which we intend to gather and organize for general use.[6]

Based on our experience, we are confident that we can interpret most tables, including image-based tables, and we are confident that we can interpret the most typical kinds of table-equivalent data. We do not, however, expect to achieve 100%, nor do we need to in order for our TANGO project to be successful.

3 Ontology Generation

Our table-analysis approach to ontology generation addresses the principled creation of ontologies based on the content of normalized tables. TANGO operates in four steps:

1. Recognize and normalize table information.
2. Construct mini-ontologies from normalized tables.
3. Discover inter-ontology mappings.
4. Merge mini-ontologies into a growing application ontology.

In support of these four steps TANGO relies on auxiliary information. This auxiliary information includes dictionaries and thesauri, natural language parsers, and data frames [Emb80], which are similar in intent to the base knowledge for ontologies proposed in [SMJ02]. Specifically, we use WordNet [Fel98] for auxiliary lexicon information and shallow parsing (e.g. [Abn91]) for natural language processing. We are creating our own data frame library. Each data frame[7] in the library is a snippet of knowledge that encapsulates the essential properties of common data items such as dates, currencies, numbers, percentages, weights, measures, and so forth. A data frame extends an abstract data type to include not only an internal data representation and applicable operations

[5]Producing this ontological body of knowledge is itself a contribution, which we wish to share with others.

[6]Gathering this corpus also constitutes a contribution. We intend, in particular, to focus on HTML tables and table-equivalent data found on the web so that we can augment, rather than duplicate, the work of others [PCH93, GJK99].

[7]The name "data frame" was coined because of the similarities to abstract data types [LZ74] and Minsky frames [Min75]. Minsky's theory of frames is a theory of rich symbolic structure where a frame represents a particular situation. Data frames represent data items instead of situations, but the information included and its purpose are quite similar.

Country	Location Description	Geographic Coordinates
Afghanistan	Southern Asia, north and west of Pakistan, east of Iran	33 00 N, 65 00 E
Albania	Southeastern Europe, bordering on the Adriatic Sea and Ionian Sea, between Greece and Serbia and Montenegro	41 00 N, 20 00 E
...		

Figure 5: Partial Normalized Table for Geography in the 2003 CIA World Factbook [Wor03b].

	Population
Asia	3,674,000,000
Africa	778,000,000
...	
New York City, New York	8,040,000
Los Angeles, California	3,700,000
...	
Mumbai, India	12,150,000
Buenos Aires, Argentina	11,960,000
...	
China	1,256,167,701*
India	1,017,645,163*
...	

*January 15, 2000

Figure 6: Partial Normalized Table for Largest Populations [Wor03a].

but also detailed representational and contextual information that allows a string that appears in a text document to be classified as belonging to the data frame. Thus, for example, a data frame for a longitude/latitude location on the earth's surface has regular expressions that recognize all forms of longitude and latitude values and regular expression recognizers for keywords such as "lon.", "lat.", "degrees north", "degrees east", and "position".[8]

Given this auxiliary information, we begin with the first step: recognize and normalize table information. We illustrated this step in the previous section except that we did not mention that we not only normalize the structure, as explained, but we also use data frames to normalize the values. Hence, for each common data item we have the values all in the same units, and we can display values with the same (or different) precision, as desired. For example, we use meters rather than feet or yards, and we can display population values in millions, if we wish.

We discuss and illustrate the remaining three steps in this section. For these examples, we assume that we have all the information from the partial tables in Figures 3 and 4, and from the partial normalized tables[9] in Figures 5, 6, 7, and 8.

[8]Creating this library of data frames is itself a contribution. To our knowledge no one has created a publicly available library of recognizers for lexical representations of common data items.

[9]These normalized tables are subparts of actual tables found on the web—subparts in the same sense that the table in Figure 4 is a subpart of the table in Figure 2. A reference for each original table from which we drew the information appears in the bibliography. We chose the subset presented here for purpose of illustration.

Place	Type	Elevation*	USGS Quad	Lat	Lon
Bonnie Lake	reservoir	unknown	Seivern	33 72 N	81 42 W
Bonnie Lake	lake	unknown	Mirror Lake	40 71 N	110 88 W
...					
New York	town/city	unknown	Jersey City	40 71 N	74 01 W
New York	town/city	149 meters	Leagueville	32 17 N	95 67 W
New York	mine	unknown	Heber City	40 62 N	111 49 W
...					

*Elevation values in this table are approximate, and often subject to a large degree of error. If in doubt, check the actual value on the map.

Figure 7: Partial Normalized Table for US Topographical Maps [Top02].

Pos	Language	Speakers	Where Spoken (Major)
1	Mandarin	885,000,000	China, Malaysia, Taiwan
2	Spanish	332,000,000	South America, Central America, Spain
3	English	322,000,000	USA, UK, Australia, Canada, New Zealand
...			

Figure 8: Partial Normalized Table for Most Spoken Languages [Mos03].

3.1 Construction of Mini-Ontologies

Figure 9 gives a graphical representation of each of the mini-ontologies for our six sample normalized tables in Figures 3 - 8. In the notation[10] boxes represent *object sets*—dashed if displayable (e.g. *Population* in Figure 9(b) and *Longitude* in Figure 9(e)) and not dashed if not displayable because their objects are represented by object identifiers (e.g. *Geopolitical Entity* in Figure 9(d)). With each object set we can associate a data frame to give it a rich description of its value set. We represent actual objects by labeled dots (e.g. *July 2001* in Figure 9(a)). Lines connecting object sets are *relationship sets*; these lines may be hyper-lines (hyper-edges in hyper-graphs) when they have more than two connections to object sets (e.g. the relationship set among the attributes *Country*, *Religion*, and *Percent* in Figure 9(a)). Optional or mandatory *participation constraints* respectively specify whether objects in a connected relationship may or must participate in a relationship set (an "o" on a connecting relationship-set line designates *optional* while the absence of an "o" designates *mandatory*). Thus, for example, the mini-ontology in Figure 9(e) declares that a *Place* must have a *Name* and may, but need not have an *Elevation*. Arrowheads on lines specify *functional constraints*—for n-ary relationship sets, $n > 2$, acute versus obtuse angles disambiguate situations where tuples of two or more tails or heads form the domain or co-domain in the function. Thus, according to Figure 9(e), a *Place* has a single *USGS Quad*, and *Geographic Coordinates* and the pair *Longitude* and *Latitude* have a one to one correspondence. Open triangles denote *generalization/specialization hierarchies* (ISA hierarchies, subset constraints, or inclusion dependencies), so that in Figure 9(c) *Continent*, *Country*, and *City* are all specializations of *Geopolitical Entity* and thus are each themselves geopolitical entities. We can constrain ISA hierarchies by partition ($\uplus$), union ($\cup$), or mutual exclusion ($+$) among specializations or by intersection ($\cap$) among gener-

[10] The particular notation we use to represent ontologies is not significant, but the concepts it represents are significant. We choose it because (1) it is fully formal in terms of first-order predicate calculus [EKW92], (2) it covers the typical ontological properties of interest—ISA hierarchies, part/whole hierarchies, relationships, and concepts including lexical appearance, representation, and computational manipulation, and (3) it has specialized tools for ontology creation and manipulation [Hew00, LEW00], ontological table understanding [ETL02, ETL04], ontological data extraction [DEG, ECJ+99], and ontological data integration [EJX01, XE03b].

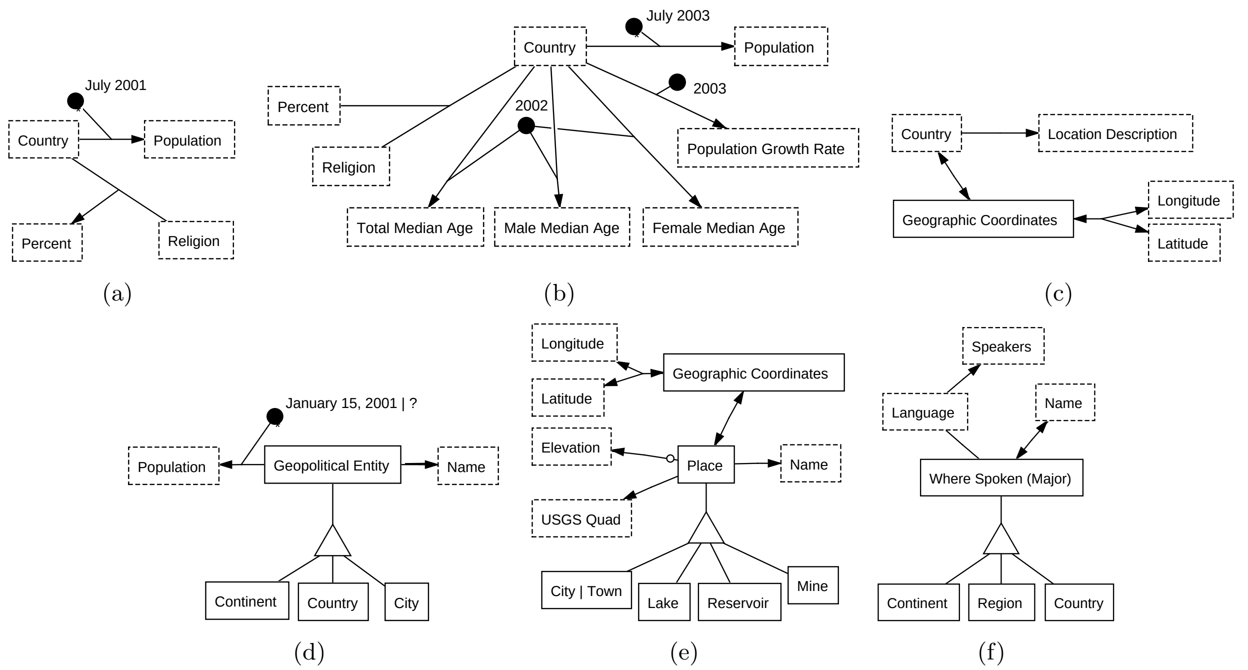

Figure 9: Mini-Ontologies Constructed from the Tables in Figures 3 - 8.

alizations. Filled-in triangles denote part/whole, part-of, or *aggregation hierarchies*. (We have no examples of aggregations in our mini-ontologies.)

To construct mini-ontologies from tables, we must discover what concepts (object sets) are involved and how they are related (relationship sets). We must also determine the constraints that hold over the relationship sets (functional, mandatory/optional participation, aggregations) and among the object sets (generalization/specialization). We do so by mining the table values for constraints such as functional dependencies and inclusion dependencies [SM79, KMRS92, MLPT03], by observing mandatory and optional patterns in the data; by using lexicons to find hypernyms/hyponyms and kind-of relationships among terms; and by using data frames to recognize values in labels, tables with multiple concept values in a column, and tables with columns whose values should be split into two or more concepts.

As an example, we obtain the mini-ontology in Figure 9(a) from the table in Figure 3 as follows. *Country* is a key and appears in a leftmost column, strongly suggesting that it should be the tail side of functional dependencies. *Population* depends on *Country* but also depends on *July 2001*. Knowledge from the data frame library recognizes that the values in the *Religion* columns are *Percent* values. The religions, which could either be object sets or values, are values since there are many (our current threshold is five). Given that religions are values, we therefore have a ternary relationship among *Country*, *Religion*, and *Percent*. Based on constraint mining, we can determine that *Country* and *Religion* together functionally determine *Percent*. Similarly, we obtain the mini-ontology in Figure 9(b) from the table in Figure 4. This time, however, the *Median Age* subcategories should be object sets rather than values because there are fewer than five.

Although creation of the remaining five mini-ontologies is also similar, there are several interesting observations we can make.

(1) For Figure 9(c), our data frame library can help us recognize the *Longitude* and *Latitude* values and place them pairwise in a one-to-one correspondence with *Geographic Coordinates*. Further, since both *Country* and *Geographic Coordinates* are keys, they are in a one-to-one correspondence.

(2) For Figure 9(d), WordNet not only knows about continents, countries, and cities, it also knows specific continents and some specific countries and cities. WordNet can therefore help us realize that the unnamed column in Figure 6 contains three categories, and it can give us *Object* as a common hypernym for the name of the generalization. Further, recognition that *Object* is a common hypernym for thousands of terms would prompt an IDS (*Issue/Default/Suggestion*) statement [BE03] raising the *Issue* that the term *Object* is likely to be far too general, stating that the *Default* is to do nothing, and making a *Suggestion* that the user choose a more meaningful name. We assume that the user follows the suggestion and chooses *Geopolitical Entity* as the name.

(3) For Figure 9(e), natural language processing can help us recognize that the column whose label is *Type* contains concepts that should become object sets. Since each *Place* is one of these objects, each of which has a *Name*, we make *Place* a generalization of these objects and then factor out *Name* from each object and associate it with *Place*. Our data frame library lets us recognize that *Lat* and *Lon* are *Latitude* and *Longitude* and that together they are *Geographic Coordinates*. Evidence from Figure 7 indicates that the *Geographic Coordinates* functionally determines *Place* and also that *Place* is unique. Further, some of the *Elevation* values are *unknown*, which lets us conclude that the *Elevation* can be optional.

(4) For Figure 9(f), we can recognize and disregard the rank (*Pos*) numbers in Figure 8. Further, for Figure 9(f), natural language processing and WordNet can find continents, countries, and regions as concepts that are all specializations of *Where Spoken*. Further, they can tell us that *Major* is an adjective, not another object or concept. Constraint mining leads to an understanding that the relationship from *Language* to *Speakers* is functional, that the relationship between *Language* and *Where Spoken* is many-to-many, and that the relationship between *Where Spoken* and the *Name* of each *Continent*, *Region*, and *Country* is one-to-one.

3.2 Discovery of Inter-Ontology Mappings

Our approach to discovering inter-ontology mappings is multi-faceted [EJX01, EJX02], which means that we use all evidence at our disposal to determine how to match concepts. In using this evidence we look not only for direct matches as is common in most schema matching techniques [BLN86, MZ98, BCV99, LC00, PTU00, DDH01, EJX01, MBR01] but also for indirect matches [BE00, MHH00, BE03, XE03b, XE03c]. Thus, for example, we are able to split or join columns to match the single *Geographic Coordinates* column in Figure 5 with the pair of columns, *Lat* and *Lon*, in Figure 7, and we are able to divide the values in the *Place* column in Figure 7 into several different object sets. For TANGO, we intend to continue with our multi-faceted approach to schema mapping. We discuss the techniques we plan to use in the following paragraphs.

Label Matching. We have successfully experimented with machine-learned decision trees over WordNet features such as synonyms,[11] word senses, hypernyms/hyponyms from WordNet [EJX01]. In [Cha03] we have also successfully experimented with modified soundex matching [HD80], Levenshtein edit-distance [Lev65], and longest common subsequences. These modified measures are particularly useful when name matching is obscured by shortened mnemonic names, abbreviations, and acronyms, which are sometimes found in table headers.

Value Similarity. We [EJX01] and others (e.g. [LC94]) have successfully used machine-learned rules to match object sets based on value characteristics such as alphanumeric features including length, alpha/numeric ratio, and space/nonspace ratio and numeric features such as mean and

[11] Surprisingly, neither direct word match nor synonym match mattered in our machine-learned decision-tree rule. Instead, the number of common hypernym roots and the distances to common hypernyms dominated the rule. Of course, identical words and synonyms have common hypernym roots at a minimal distance from the words, which mitigates our surprise.

variance. We intend to also consider Gaussian value matching [SSZ98] and regression matching [HL86], which should, for example, allow us to match imprecise but highly correlated value sets such as population values and import/export estimates.

Expected Values. Using constant value recognizers in data frames, we have shown that finding and matching expected values in value sets provides significant leverage in schema matching [ETL02, XE03a, ETL04]. Being able to recognize values such as latitudes, longitudes, distances, dates, times, and percent values can help us match object sets. Data frame recognizers can also help us tell when table labels might be values or when table values might be labels, decompose or compose value strings for matching, and help us determine whether value sets are unions or subsets of other value sets [ETL02, ETL04].

Constraints. In [BE03] we studied constraints in the context of schema matching. These include keys in tables (as well as nonkeys), functional relationships, one-to-one correspondences, subset/superset relationships, optional and mandatory constraints in connection with unknown and null values. Others have derived constraints from typed hierarchies [NAM97, NAM98] and recurrent subpatterns [WL97]. Although we can capitalize on some of these constraints, and indeed others have via data mining [DP95, dSMH01], we have also discovered that the many points of view and the many different objectives often prompt the need for IDS interaction [BE03].

Structure. We [EJX01, XE02, EJX02, XE03a] and others [CAFP98, CDSS98, MZ98, Coh99, MHH00, DDH01, MBR01, SH01, MGMR02] have developed matching algorithms based on structural context. We have been able to use proximity, node importance as measured by in/out-degree, and neighbor similarity to help match object sets.

3.3 Ontology Merge

Once we have discovered mappings between mini-ontologies or between a mini-ontology and the ontology we are building, we can begin the merge process. Sometimes the match is such that we can directly fuse two ontologies by simply keeping all the nodes and edges of both and merging nodes and edges that directly correspond [LNE89, SG89]. Often, however, merging induces conflicts that must be resolved [NG82, SP94, GSSC95].

We use three basic approaches to conflict resolution: (1) automatic adjustment based on constraint satisfaction, (2) synergistic adjustment based on Issue/Default/Suggestion (IDS) statements, and (3) multiple adjustments leading to multiple ontological views with mappings between them. All three of these approaches rely on being able to determine plausible merges. Then, for automatic adjustments, we can take the best among the plausible merges; for synergistic adjustments, we can raise the important issues and make suggestions, letting a user make the final decisions; and for multiple adjustments, can keep all plausible merges, later eliminating those discarded in synergistic evaluations and those that no longer stand up to new evidence gathered as the process continues.

To determine plausible merges based on discovered mappings, we consider constraint violations and congruency principles. Constraint violations include functional/non-functional mismatches, optional/mandatory participation, displayable/non-displayable object sets, and subset/superset constraints. Congruency principles [CEW96, Emb98, Gua98b] attempt to ensure that all objects in an object set have the same properties; the objects in an object set are *congruent* when this principle holds and are otherwise *incongruent*. Other similar principles of formal ontology construction also apply [Gua98a, Gua99, WSW99, Gua00, GW00, EW01, WG01], as well as related work on merging ontologies (e.g. [MFRW00]) and comparing and aligning ontologies (e.g. [BB01]). We illustrate these ideas by merging the mini-ontologies in Figure 9.

We look initially for mini-ontologies that exhibit as large of an overlap as possible (as measured by the number of inter-ontology mappings); thereafter we select mini-ontologies with the largest

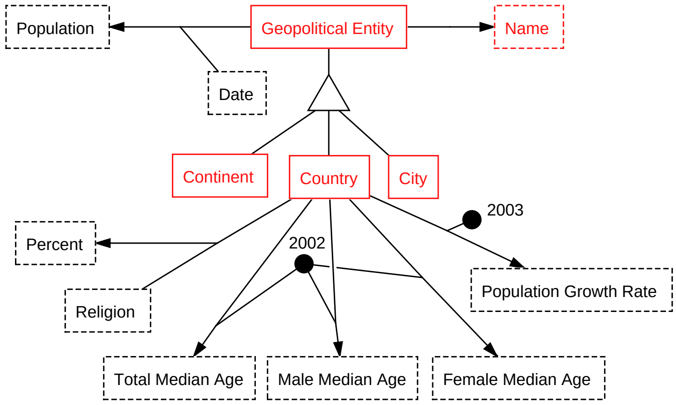

Figure 10: Growing Ontology after Merging the Mini-Ontologies in Figures 9(a), 9(b), and 9(c). (Red object sets are those added in the latest merge—2nd Merge.)

overlap with our growing ontology. In our example we begin by merging the mini-ontologies in Figures 9(a) and 9(b).

1st Merge *Country* matches *Country* and *Population* matches *Population*. Both *July 2001* and *July 2003* are date components associated with *Population*, and we merge them as *Date*.

2nd Merge Building on the 1st Merge, we add the mini-ontology in Figure 9(d) and obtain the emerging ontology in Figure 10. Here, we must reconcile the displayable/non-displayable *Country* object sets, but this is straightforward based on the inherited *Name* property in Figure 9(c). According to congruency principles, we also let *Population* be an inherited property and thus omit it from the *Country* specialization.

3rd Merge Continuing, we merge the mini-ontology in Figure 9(f) with the growing ontology in Figure 10. Here, the data in the object sets *Geopolitical Entity* and *Where Spoken* largely overlap, but it is not 100% clear whether one set should be a subset of the other, whether they are overlapping siblings in an ISA hierarchy, or whether they should be the same set. An IDS statement is therefore appropriate, and we assume the issue is resolved by declaring that the sets are the same and should be called *Geopolitical Entity*.

4th Merge Continuing, we next add the mini-ontology in Figure 9(c). Here, the constraints on the *Location Description* in Figure 9(c) declare that the relationship is mandatory for both *Country* and *Location Description* and functional from *Country* to *Location Description*. Because of the lack of location descriptions for most countries in our growing collection, however, we have enough evidence to override the mandatory declaration and make the relationship for *Country* optional. Later, when we see more location descriptions for countries, which will most certainly not be the same as the ones we already have, we will also override the functional declaration (but for now we leave it functional).

5th Merge Continuing, we next add the mini-ontology in Figure 9(e) and obtain the growing ontology in Figure 11. Here, TANGO must recognize that *Geopolitical Entity* is a subset of *Place*. Other adjustments, including inheriting *Name* only from *Place* and making the existence of *USGS Quad* optional for *Place*, come readily.

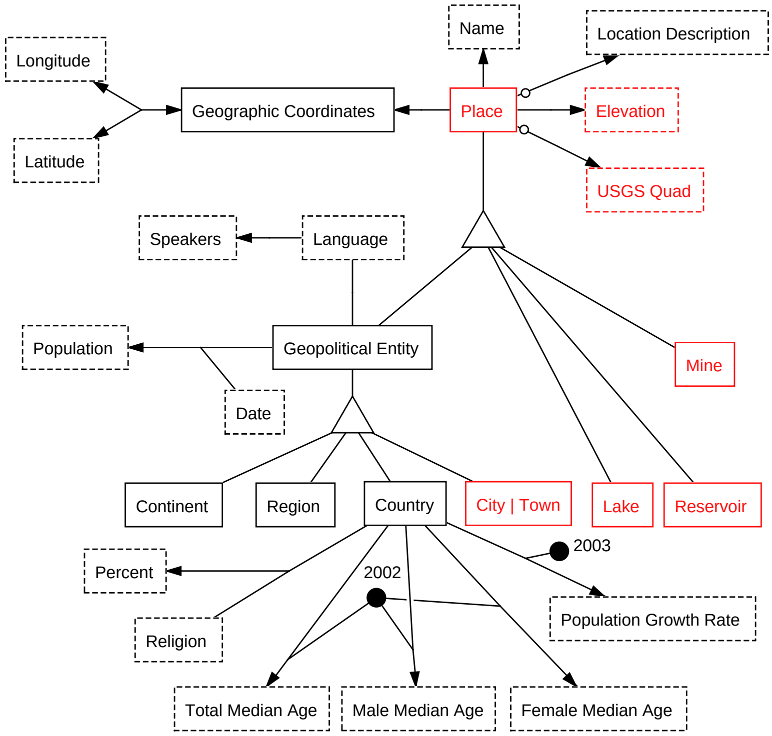

Figure 11: Growing Ontology after Merging all Mini-Ontologies. (Red object sets are those added in the last merge—5th Merge.)

4 Experimental Evaluation

The basic measure we intend to investigate is cost reduction, where the cost is an application dependent convex combination of user-times T_1, T_2, and T_3. Time T_1 measures ontology construction; T_2 measures time required to retrieve desired information using the ontology; and T_3 measures extra time required to retrieve information from the original source material because the ontology does not contain the necessary information. We will compare, on identical information utilization tasks, the cost of five scenarios (one vacuous) of creating ontologies:

1. *Null*: No ontology will be built; all data remains, as it is, in source documents.
2. *Human-Built*: TANGO will be run entirely by a user.
3. *Synergistic*: TANGO will be run synergistically under the guidance of IDS statements.
4. *Automatic+User*: The user will make corrections to the TANGO-generated ontology.
5. *Automatic*: TANGO will generate the ontology without help.

Our experimental design is standard and similar to prior experiments we have conducted [Emb78, EN81b]. Evaluating the interactive and automated components of TANGO requires specifying the following items for each experiment:

- *Ontology*: an organized body of information constructed under one of the scenarios.
- *Subjects*: skilled information users unconnected with the project.

- *Source Documents*: one of two document datasets described below.
- *Application*: a set of queries and corresponding unique answers based on a dataset.

Instrumentation. The measures proposed above are all based on time. We will therefore implement a monitoring system that will log user actions at the keystroke level and computer actions at the subroutine level. (We consider the computer time insignificant as long as the response time is less than the 0.5 seconds considered acceptable for complex interactive tasks [EN81a].) We will transfer the results of the log to preformatted Excel sheets as we did for experiments with CAVIAR [NZ02, ZN04a, ZN04b, ZN04c], for both individual and aggregate performance analysis. We exclude collection of source documents from the evaluation, since it must be done regardless of how, and whether, the ontology is built. Therefore T_1 for Scenario 1 is zero. We must, however, measure the time expended to answer the queries, both in consulting the ontology (T_2) and in accessing any source documents when the ontology is either non-existent or proves deficient (T_3).

Subjects. We will recruit subjects from the professional staff of the BYU and RPI libraries. They will therefore be expert information retrieval specialist. Although most will be familiar with the concept of ontologies, they will not hesitate to consult the source documents directly when expedient. The subjects will not be rewarded, but will report their freeform suggestions that we will solicit by email immediately after their experimental session. We do not expect difficulty in obtaining the necessary permissions for experimentation on human subjects from both the BYU and RPI Review Boards, because none of the investigators are in a position of authority over library personnel.

Test Data. The experimenters will naturally test all aspects of the evolving system on gradually increasing sets of documents. Since experimentation on the same data set leads to statistically unreliable conclusions, when the system is deemed ready, we will "freeze" it, and conduct formal "arms-length" evaluations on two databases. One will consist a set of 100 new "greenhouse" documents of limited difficulty, and the other of a set of 100 documents collected subject only to the constraint that they contain table-equivalent data for geographic information.

Ontology Construction. Subjects who will construct the ontologies will be different from those who attempt to use it. Because ontology creation is a complex task in which even human experts may produce different results given the same information, we will test three subjects for each of the five scenarios for building a TANGO ontology for each of the two databases (24 experiments). We will use a PowerPoint presentation to instruct external ontology builders, but an experimenter will remain present to answer any questions about TANGO commands. We consider the learning time to be considered "gratis," since it does not have to be repeated for new ontology constructions.

Quality of Ontologies. We will also test three subjects for each of the five query-answer tasks (15 experiments). The subjects will be directed to find the answers using either TANGO or the original source data, or any combination of the two. A high ratio of T_2 to T_3 implies that a subject spends little, if any, time accessing material beyond the material already accessible in the ontology and thus will indicate that the ontology being tested is satisfactory. Each subject will answer the same queries. They will not be penalized for errors, because we expect that inadequacies in an ontology will be reflected only by an increase in T_3, and that the quality of responses will remain high with all methods. Both the ontology construction and the Q/A tasks will be sized so that they can be performed in a single experimental session of an hour or two.

Criteria. Our research will be successful if we can speed up the ontology-building process without compromising the quality of the product, and it will be highly successful if we can significantly ($p < 0.05$) speed it up on a wide-ranging set of documents and web pages.

A larger number of subjects and source documents would of course be advantageous. We may be able to increase the number if we find qualified students to conduct the experiments. We will make both TANGO and our databases available to interested parties through the Internet as soon as we have reasonably glitch-free versions. We expect that exposure will generate ideas for further improvement.

5 Research Plan

The principal investigators have collaborated (in pairs) for years (and in one pair for decades); therefore no special provision is needed to facilitate communication between them. We will simply continue to exchange email, telephone calls and visits as required.

The students will, however, be new to the project and require appropriate mentoring.[12] In addition to weekly meetings with them, as we have with all of our students, it will be beneficial for each student to spend a summer at the "other" university. To maximize the students' exposure to each other, the BYU graduate students will spend the first summer in Troy, and the RPI students, including the RPI undergraduate student, will spend the second summer in Provo. During the third year, at least one graduate student from each university will have the opportunity to participate in at least one conference germane to our topic.

Year 1. The major task will be the construction of the the infrastructure for the ontology generation system at BYU and the basic table ontology at RPI. Also, under our direction the RPI undergraduate student will implement a monitoring system to log both system and user actions. By the end of the first year each graduate student will present a plausible thesis topic within the scope of the research.

Year 2. Based on our first year experience, we will conduct repeated experiments on the same data and improve the system by gradually eliminating weak points. Also, in an effort to show the usefulness and applicability of TANGO-constructed ontologies, three BYU undergraduate students will undertake some of the projects described in Section 6. These undertakings will continue during the third year of the project.

Year 3. We will conduct the evaluation experiments on the new data during the first half of the year. The last half of the year will be devoted to disseminating the results at appropriate conferences and to preparing them for publication in archival technical journals. Our ontology for table understanding, plus our fully developed infrastructure (including our data frame library, ontology editors, IDS interaction system, and the ontology mapping and merging components), plus our corpus of tables, plus our experimental results and all the raw web pages used in the tests will be made available to other researchers through our web sites.

6 Expected Significance

The intellectual merit and broader impacts of the proposed work have the potential to make a significant difference in universal access to dispersed knowledge on the web.

6.1 Intellectual Merit

The TANGO project addresses fundamental issues in information systems: *data* (isolated attribute value pairs), *information* (data in a conceptual framework), and *knowledge* (information with a degree of certainty or community agreement).[13] We directly address each of these three issues—

[12] The principal investigators hope to maintain their successful recent record of attracting women (BYU: 4, RPI: 3), minorities (RPI: 1), and citizens of developing countries (BYU: 2, RPI: 10) to their research projects.

[13] These definitions are a variation of those offered in [Mea92].

data with data frames that include fine-grained recognizers to locate and classify text strings, information with conceptual modeling of table-equivalent data, and knowledge with community agreement based on merging overlapping source repositories.

Further, having constructed data, information, and knowledge as an ontology of the type proposed in our TANGO project puts us in a position to resolve many interesting and challenging problems. Examples[14] include: (a) robust information extraction from semi-structured web pages [ECJ+99], as opposed to brittle information extraction (e.g. [HGMC+97, KWD97, HD98, Mus99, BLP01, LRNS02]) requiring wrapper maintenance [LM00] or generation/regeneration [CMM01] for new or changed pages [LRNdST02]; (b) extraction ontology generation [LDEM02, Din03]; (c) high-precision classification of semi-structured web pages [RL94, ENX01, KN03]; (d) data integration, which tends to work best when rich auxiliary knowledge sources provide a basis for analyzing sources from multiple points of view, especially when considering both direct and indirect schema matching [EJX01, XE03b, XE03c]; (e) multiple-source query processing [XE02, Xu03] which has advantages over other approaches (e.g. Global-as-View and Local-as-View approaches [CGMH+94, LRO96, GKD97, CLL01]); and (f) document image analysis for which the proposed techniques can eliminate some common shortcomings of current table understanding software [LN99b].

6.2 Broader Impact

Semantics is a grand challenge for the current generation of computer technology. It is the key for unlocking the door, for example, to personal agents that can roam the Semantic Web and carry out sophisticated tasks for their masters, to information exchange and negotiation in e-business, and to automated, large-scale, in-silico experiments in e-science. We do not claim that the work proposed here will resolve this challenge, but we do claim that it addresses issues related to this grand challenge and that its successful realization would help us move a step closer to a resolution. As specific research in this direction, we offer the following.

Semantic-Web Construction and Superimposed-Information Generation. As the Semantic Web becomes more popular, a question of increasing importance will be how to convert some of the interesting unstructured and semi-structured, data-rich documents on the web as they now stand into Semantic-Web documents. In [Cha03] we proposed a way to bridge the gap between the current web and the Semantic Web by semi-automatically converting Resource Description Framework Schemas (RDFS's) [BG02] and DAML-OIL ontologies [HM00] into data extraction ontologies [ECJ+99]. The prototype system we built [DEG] does this conversion, extracts data, and then converts it to RDFS, making it accessible to Semantic-Web agents. In addition, the prototype system superimposes the meta-data of the extracted information over the document for direct access to data in context, as suggested in [MD99]. We believe that the TANGO-constructed ontologies will work even better for this application.

Agent Interoperability. We are experimenting with and have built an initial prototype system that allows "on the fly" communication [Usc02] among heterogeneous software agents [AM02, AME03]. Rather than relying on a specified shared ontology, a common communication language, and a specified message format to achieve interoperability, we intend to use an independent global ontology to encode and decode messages exchanged among agents. TANGO can help us create the independent knowledge we need for an application of interest.

[14] As the references in these examples indicate, the basis for the resolution of these problems is our current work, which is supported by the National Science Foundation under grant No. IIS-0083127.

References

[Abn91] S.P. Abney. Parsing by chunks. In R.C. Berwick, S.P. Abney, and C. Tenny, editors, *Principle-Based Parsing: Computation and Psycholinguistics*, pages 257–278. Kluwer Academic Publishers, Boston, Massachusetts, 1991.

[AM02] M. Al-Muhammed. Dynamic matchmaking between messages and services in multi-agent systems. Technical report, Brigham Young University, Provo, Utah, 2002. (thesis proposal, currently at www.deg.byu.edu/proposals/index.html).

[AME03] M. Al-Muhammed and D.W. Embley. Dynamic matchmaking between messages and services in multi-agent information systems. In *Proceedings of the Workshop on Agent-Oriented Information Systems (AOIS@ER03)*, pages 244–246, Chicago, Illinois, October 2003.

[AT98] A. Abu-Tarif. Table processing and table understanding. Master's thesis, Rensselaer Polytechnic Institute, May 1998.

[BB01] A. Burgun and O. Bodenreider. Comparing terms, concepts, and semantic classes in WordNet and the Unified Medical Language System. In *WordNet and Other Lexical Resources: Applications, Extensions, and Customizations; An NAACL-01 (North American Association for Computational Linguistics) Workshop*, pages 77–82, Pittsburgh, Pennsylvania, June 2001.

[BCV99] S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Record*, 28(1):54–59, March 1999.

[BE00] J. Biskup and D.W. Embley. Mediated information gain. In *International Database Engineering and Applications Symposium (IDEAS2000)*, pages 360–370, Yokohama, Japan, September 2000.

[BE03] J. Biskup and D.W. Embley. Extracting information from heterogeneous information sources using ontologically specified target views. *Information Systems*, 28(3):169–212, 2003.

[BG02] D. Brickley and R. Guha. RDF vocabulary description language 1.0: RDF schema. Technical report, World Wide Web Consortium, 2002. (www.w3.org/TR/rdf-schema).

[BLHL01] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 36(25), May 2001.

[BLN86] C. Batini, M. Lenzerini, and S.B. Navathe. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys*, 18(4):323–364, December 1986.

[BLP01] D. Buttler, L. Liu, and Calton Pu. A fully automated object extraction system for the world wide web. In *Proceedings of the 21st International Conference on Distributed Computing Systems (ICDC'01)*, Mesa, Arizona, April 2001.

[Bun77] M.A. Bunge. *Treatise on Basic Philosophy: Vol. 3: Ontology I: The Furniture of the World*. Reidel, Boston, 1977.

[CAFP98] S. Castano, V. De Antonellis, M.G. Fugini, and B Pernici. Conceptual schema analysis: Techniques and applications. *ACM Transactions on Database Systems*, 23(3):286–333, September 1998.

[Car00] M.E. Carmack. Technical information types: A peircean analysis. Master's thesis, Linguistics Department, Brigham Young University, 2000.

[CDSS98] S. Cluet, C. Delobel, J. Siméon, and K. Smaga. Your mediators need data conversion! In *Proceedings of 1998 ACM SIGMOD International Conference on Management of Data*, pages 177–188, Seattle, Washington, June 1998.

[CEW96] S.W. Clyde, D.W. Embley, and S.N. Woodfield. Improving the quality of systems and domain analysis through object class congruency. In *Proceedings of the International IEEE Symposium on Engineering of Computer Based Systems (ECBS'96)*, pages 44–51, Friedrichshafen, Germany, March 1996.

[CGMH+94] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J Widom. The TSIMMIS project: Integration of heterogeneous information sources. In *IPSJ Conference*, pages 7–18, Tokyo, Japan, October 1994.

[Cha03] T. Chartrand. Ontology-based extraction of RDF data from the world wide web. Master's thesis, Brigham Young University, Provo, Utah, March 2003.

[CL00] M. Carmack and D. Lonsdale. Information structure and hypertext search results. In *Information Doors: Workshop on where Information Search and Hypertext Link*, Proceedings of the ACM Hypertext and Digital Librairies Conference, pages 5–10, San Antonio, Texas, May 2000.

[CLL01] D. Calvanese, D. Lembo, and M. Lenerini. Survey on methods for query rewriting and query answering using views. Technical report, University of Rome, Roma, Italy, April 2001.

[CMM01] V. Crescenzi, G. Mecca, and P. Merialdo. Roadrunner: Towards automatic data extraction from large web sites. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB'01)*, pages 109–118, Rome, Italy, September 2001.

[Coh99] W.W. Cohen. Some practical observations on integration of web information. In *Proceedings of the ACM SIGMOD Workshop on The Web and Databases (WebDB'99)*, pages 55–60, Philadelphia, Pennsylvania, June 1999.

[DDH01] A. Doan, P. Domingos, and A. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data (SIGMOD 2001)*, pages 509–520, Santa Barbara, California, May 2001.

[DEG] Homepage for BYU data extraction research group. URL: http://osm7.cs.byu.edu/deg/index.html.

[DeM80] M. DeMey. The relevance of the cognitive paradigm for information science. In O. Harbo, editor, *Theory and Application of Information Research*. Mansell, London, 1980.

[DF02a] Y. Ding and S. Foo. Ontology research and development. part 1: A review of ontology generation. *Journal of Information Science*, 28(2):123–136, February 2002.

[DF02b] Y. Ding and S. Foo. Ontology research and development. part 2: A review of ontology mapping and evolving. *Journal of Information Science*, 28(5):375–388, October 2002.

[DFvH03] J. Davies, Dieter Fensel, and F. van Harmelen, editors. *Towards the Semantic Web: Ontology-Driven Knowledge Management*. John Wiley & Sons, LTD, Hoboken, New Jersey, 2003.

[Din03] Y. Ding. Semiautomatic generation of relilient data-extraction ontologies. Master's thesis, Brigham Young University, Provo, Utah, June 2003.

[dlb03] dlbeck.com, December 2003. www.dlbeck.com/population.htm.

[DP95] S.K. Dao and B. Perry. Applying a data miner to heterogeneous schema integration. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pages 93–101, 1995.

[dSMH01] R. dos Santos Mello and C.A. Heuser. A rule-based conversion of a DTD to a conceptual schema. In *Proceedings of the 20th International Conference on Conceptual Modeling (ER2001)*, pages 134–148, Yokohama, Japan, November 2001.

[ECJ+99] D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.-K. Ng, and R.D. Smith. Conceptual-model-based data extraction from multiple-record web pages. *Data & Knowledge Engineering*, 31(3):227–251, November 1999.

[EJX01] 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)*, pages 110–117, Rio de Janeiro, Brazil, April 2001.

[EJX02] D.W. Embley, D. Jackman, and L. Xu. Attribute match discovery in information integration: Exploiting multiple facets of metadata. *Journal of the Brazilian Computing Society*, 8(2):32–43, November 2002.

[EKW92] D.W. Embley, B.D. Kurtz, and S.N. Woodfield. *Object-oriented Systems Analysis: A Model-Driven Approach*. Prentice Hall, Englewood Cliffs, New Jersey, 1992.

[Emb78] D.W. Embley. Empirical and formal language design applied to a unified control construct for interactive computing. *International Journal of Man-Machine Studies*, 10(2):197–216, March 1978.

[Emb80] D.W. Embley. Programming with data frames for everyday data items. In *Proceedings of the 1980 National Computer Conference*, pages 301–305, Anaheim, California, May 1980.

[Emb98] D.W. Embley. *Object Database Development: Concepts and Principles*. Addison-Wesley, Reading, Massachusetts, 1998.

[EN81a] D.W. Embley and G. Nagy. Behavioral aspects of text editors. *ACM Computing Surveys*, 13(1):33–70, March 1981.

[EN81b] D.W. Embley and G. Nagy. Empirical and formal methods for the study of computer editors. In M.J. Coombs and J.L. Alty, editors, *Computing Skills and the User Interface*, chapter 14, pages 456–496. Academic Press, London, England, 1981.

[EN89] D.W. Embley and G. Nagy. On the integration of lexical and spatial data in a unified high-level model. In *Proceedings of the International Symposium on Database Systems for Advanced Applications*, pages 329–336, Seoul, Korea, April 1989.

[EN91] D.W. Embley and G. Nagy. A multi-layered approach to query processing in geographic information systems. In *Proceedings of the International Workshop on Database Management Systems for Geographical Applications*, pages 214–238, Capri, Italy, May 1991.

[ENX01] D.W. Embley, Y.-K. Ng, and L. Xu. Recognizing ontology-applicable multiple-record web documents. In *Proceedings of the 20th International Conference on Conceptual Modeling (ER2001)*, pages 555–570, Yokohama, Japan, November 2001.

[ETL02] D.W. Embley, C. Tao, and S.W. Liddle. Automatically extracting ontologically specified data from HTML tables with unknown structure. In *Proceedings of the 21st International Conference on Conceptual Modeling (ER'02)*, pages 322–327, Tampere, Finland, October 2002.

[ETL04] D.W. Embley, C. Tao, and S.W. Liddle. Automating the extraction of data from tables with unknown structure. *Data & Knowledge Engineering*, 2004. (to appear).

[EW01] J. Evermann and Y. Wand. Towards ontologically based semantics for UML constructs. In *Proceedings of the 20th International Conference on Conceptual Modeling (ER2001)*, pages 513–526, Yokohama, Japan, November 2001.

[FD92] P.W. Foltz and T. Dumais. Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM*, 35(12):51–60, 1992.

[Fel98] C. Fellbaum. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, Massachussets, 1998.

[GJK99] M. Garris, S. Janet, and W. Klein. Federal register document image database. In *Proceedings of Document Recognition and Retrieval VI (IS&T/SPIE Electronic Imaging '99)*, volume 3651, pages 97–108, San Jose, California, 1999.

[GK95a] E.A. Green and M.S. Krishnamoorthy. Model-based analysis of printed tables. In *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR'95)*, pages 214–217, Montréal, Canada, August 1995.

[GK95b] E.A. Green and M.S. Krishnamoorthy. Recognition of tables using table grammars. In *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval*, pages 261–277, Las Vegas, Nevada, 1995.

[GKD97] M.R. Genesereth, A.M. Keller, and O.M. Duschka. Infomaster: An information integration system. In *Proceedings of 1997 ACM SIGMOD International Conference on Management of Data*, pages 539–542, Tucson, Arizona, May 1997.

[GL02] M. Gruninger and J. Lee. Ontology applications and design. *Communications of the ACM*, 45(2):39–41, February 2002.

[Gom03] J.M. Gomez. Ontobuilder: Generating ontologies for the semantic web, December 2003. www.nextwebgeneration.org/presentations/ontobuilder.pdf.

[GSSC95] M. Garcia-Solaco, F. Slator, and M. Castellanos. A structure based schema integration methodology. In *Proceedings of the 11th International Conference on Data Engineering (ICDE'95)*, pages 505–512, Taipei, Taiwan, 1995.

[Gua98a] N. Guarino. Formal ontologies and information systems. In N. Guarino, editor, *Proceedings of the First International Conference on Formal Ontology in Information Systems (FOIS98)*, pages 3–15, Trento, Italy, June 1998.

[Gua98b] N. Guarino. Some ontological principles for designing upper level lexical resources. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, May 1998.

[Gua99] N. Guarino. The role of identity conditions in ontology design. *Lecture Notes in Computer Science*, 1661:221–234, 1999.

[Gua00] N. Guarino. A formal ontology of properties. In *The ECAI-00 Workshop on Applications of Ontologies and Problem Solving Methods*, pages 12.1–12.8, 2000.

[GW00] N. Guarino and C. Welty. Ontological analysis of taxonomic relationships. In A.H.F. Laender, S.W. Liddle, and V.C. Storey, editors, *Proceedings of the 19th International Conference on Conceptual Modeling (ER2000)*, Lecture Notes on Computer Science (LNCS 1920), pages 210–224, Salt Lake City, Utah, October 2000.

[Haa98] T.B. Haas. The development of a prototype knowledge-based table-processing system. Master's thesis, Brigham Young University, Provo, Utah, April 1998.

[Han99] J.C. Handley. Document recognition. In E.R. Dougherty, editor, *Electronic Imaging Technology*, pages 289–316, 1999.

[Han01] J.C. Handley. Table analysis for multi-line cell identification. In *Document Recognition and Retrieval VIII*, volume 4307 of *Proceedings of SPIE*, pages 34–43, 2001.

[HD80] P.A. Hall and G.R. Dowling. Approximate string matching. *ACM Computing Surveys*, 12(4):381–402, 1980.

[HD95] O. Hori and D.S. Doermann. Robust table-form structure analysis based on box-driven reasoning. In *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR'95)*, pages 218–221, Montréal, Canada, August 1995.

[HD97] M. Hurst and S. Douglas. Layout and language: Preliminary experiments in assigning logical structure to table cells. In *Proceedings of the Fifth Applied Natural Language Processing Conference*, pages 217–220, Washington, DC, 1997.

[HD98] C-N. Hsu and M-T. Dung. Generating finite-state transducers for semi-structured data extraction from the web. *Information Systems*, 23(8):521–538, December 1998.

[Hew00] K.A. Hewett. An integrated ontology development environment for data extraction. Master's thesis, Brigham Young University, Provo, Utah, April 2000.

[HGMC+97] J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo. Extracting semistructured information from the web. In *Proceedings of the Workshop on Management of Semistructured Data*, Tucson, Arizona, May 1997.

[HJ02] C.W. Holsapple and K.D. Joshi. A collaborative approach to ontology design. *Communications of the ACM*, 45(2):42–47, February 2002.

[HKL+01] J. Hu, R. Kashi, D. Lopresti, G. Nagy, and G. Wilfong. Why table ground-truthing is hard. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, pages 129–133, Seattle, Washington, September 2001.

[HKLW00] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. A system for understanding and reformulating tables. In *Proceedings of the 4th IAPR International Workshop on Document Analysis Systems (ICDAR'01)*, Rio de Janeiro, Brazil, December 2000.

[HKLW01] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. Table structure recognition and its evaluation. In P.B. Kantor, D.P. Lopresti, and J. Zhou, editors, *Proceedings of Document Recognition and Retrieval VIII*, volume SPIE-4307, pages 44–55, San Jose, California, January 2001.

[HL86] D.W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, Inc., New York, New York, 1986.

[HM00] J. Hendler and D. McGuinness. The DARPA agent markup language. *IEEE Intelligent Systems*, 15(4):72–73, November-December 2000.

[Kie98] T.G. Kieninger. Table structure recognition based on robust block segmentation. In *Proceedings of Document Recognition V (IS&T/SPIE Electronic Imaging'98)*, volume 3305, pages 22–32, San Jose, California, January 1998.

[KMRS92] M. Kantola, H. Mannila, K.-J. Räihä, and H. Siirtola. Discovering functional and inclusion dependencies in relational databases. *International Journal of Intelligent Systems*, 7:591–607, 1992.

[KN03] L.W. Kwong and Y.-K. Ng. Performing binary-categorization on multiple-record web documents using information retrieval models and application ontologies. *World Wide Web: Internet and Web Information Systems*, 6(3):281–303, September 2003.

[KNNR95] J. Kanai, G. Nagy, T. Nartker, and S. Rice. Automated evaluation of OCR zoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):86–90, 1995.

[KWD97] N. Kushmerick, D.S. Weld, and R. Doorenbos. Wrapper induction for information extraction. In *Proceedings of the 1997 International Joint Conference on Artificial Intelligence*, pages 729–735, 1997.

[LC94] W.-S. Li and C. Clifton. Semantic integration in heterogeneous databases using neural networks. In *Proceedings of the 20th Very Large Data Base Conference*, Santiago, Chile, 1994.

[LC00] W. Li and C. Clifton. SEMINT: A tool for identifying attribute correspondences in heterogeneous databases using neural networks. *Data & Knowledge Engineering*, 33(1):49–84, 2000.

[LDEM02] D.W. Lonsdale, Y. Ding, D.W. Embley, and A. Melby. Peppering knowledge sources with SALT: Boosting conceptual content for ontology generation. In *Proceedings of the AAAI Workshop: Semantic Web Meets Language Resources*, pages 30–36, Edmonton, Alberta, Canada, July 2002.

[Lem98] J. Lemke. Multiplying meaning: Visual and verbal semiotics in scientific text. In J.R. Martin and Robert Veel, editors, *Reading Science: Critical and Functional Perspectives on Discourses of Science*, pages 87–113. Routledge, 1998.

[Lev65] V.I. Levenshtein. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1:8–17, 1965.

[LEW00] S.W. Liddle, D.W. Embley, and S.N. Woodfield. An active, object-oriented, model-equivalent programming language. In M.P. Papazoglou, S. Spaccapietra, and Z. Tari, editors, *Advances in Object-Oriented Data Modeling*, pages 333–361. MIT Press, Cambridge, Massachusetts, 2000.

[LM00] K. Lerman and S. Minton. Learning the common structure of data. In *Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-2000)*, Menlo Park, California, July 2000.

[LN99a] S. Lim and Y. Ng. An automated approach for retrieving heirarchical data from HTML tables. In *Proceedings of the Eighth International Conference on Information and Knowledge Management (CIKM'99)*, pages 466–474, Kansas City, Missouri, November 1999.

[LN99b] D. Lopresti and G. Nagy. Automated table processing: An (opinionated) survey. In *Proceedings of the Third IAPR Workshop on Graphics Recognition*, pages 109–134, Jaipur, India, September 1999.

[LN00] D. Lopresti and G. Nagy. A tabular survey of table processing. In A.K. Chhabra and D. Dori, editors, *Graphics Recognition—Recent Advances*, Lecture Notes in Computer Science, LNCS 1941, pages 93–120. Springer Verlag, 2000.

[LN02] D. Lopresti and G. Nagy. Issues in ground-truthing graphic documents. In D. Blostein and Y-B. Kwon, editors, *Graphics Recognition—Algorithms and Applications*, Lecture Notes in Computer Science, LNCS 2390, pages 46–66. Springer Verlag, 2002. (selected papers from the Fourth International Workshop on Graphics Recognition, GREC 2001).

[LNE89] J. Larson, S. Navathe, and R. Elmasri. A theory of attribute equivalence in databases with application to schema integration. *IEEE Transactions on Software Engineering*, 15(4), 1989.

[LNS+00] L. Li, G. Nagy, A. Samal, S. Seth, and Y. Xu. Integrated text and line-art extraction from a topographic map. *International Journal of Document Analysis and Recognition*, 2(4):177–185, June 2000.

[LRNdST02] A.H.F. Laender, B.A. Ribeiro-Neto, A.S. da Silva, and J.S. Teixeira. A brief survey of web data extraction tools. *SIGMOD Record*, 31(2):84–93, June 2002.

[LRNS02] A.H.F. Laender, B. Ribeiro-Neto, and A.S. Da Silva. DEByE—data extraction by example. *Data and Knowledge Engineering*, 40(2):121–154, 2002.

[LRO96] A.Y. Levy, A. Rajaraman, and J.J. Ordille. Querying heterogeneous information sources using source descriptions. In *Proceedings of the Twenty-second International Conference on Very Large Data Bases*, Mumbai (Bombay), India, 1996.

[LZ74] B.H. Liskov and S.N. Zilles. Programming with abstract data types. *Proceedings of the ACM Symposium on Very High Level Languages, SIGPLAN Notices*, 9(4):50–59, April 1974.

[Mai83] D. Maier. *The Theory of Relational Databases*. Computer Science Press, Inc., Rockville, Maryland, 1983.

[MBR01] J. Madhavan, P.A. Bernstein, and E. Rahm. Generic schema matching with Cupid. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB'01)*, pages 49–58, Rome, Italy, September 2001.

[MD99] D. Maier and L. Delcambre. Superimposed information for the Internet. In S. Cluet and T. Milo, editors, *Proceedings of the ACM SIGMOD Workshop on the Web and Databases (WebDB'99)*, Philadelphia, Pennsylvania, June 1999.

[Mea92] C.T. Meadow. *Text Information Retrieval Systems*. Academic Press, San Diego, California, 1992.

[MFRW00] D.L. McGuinness, R. Fikes, J. Rice, and S Wilde. An environment for merging and testing large ontologies. In *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning*, pages 483–493, Breckenridge, Colorado, April 2000.

[MGJ01] G. Modica, A. Gal, and H. Jamil. The use of machine-generated ontologies in dynamic information seeking. In *Proceedings of the 9th International Conference on Cooperative Information Systems (CoopIS 2001)*, pages 433–448, Trento, Italy, September 2001.

[MGMR02] S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity flooding: A versatile graph matching algorithm and it application to schema matching. In *Proceedings of the 18th International Conference on Data Engineering (ICDE 2002)*, pages 117–128, San Jose, California, 2002.

[MHH00] R. Miller, L. Haas, and M.A. Hernandez. Schema mapping as query discovery. In *Proceedings of the 26th International Conference on Very Large Databases (VLDB'00)*, pages 77–88, Cairo, Egypt, September 2000.

[Min75] M. Minsky. A framework for representing knowledge. In P.H. Winston, editor, *The Psychology of Computer Vision*, pages 211–277. McGraw-Hill, 1975.

[MLPT03] F. De Marchi, S. Lopes, J.-M. Petit, and F. Toumani. Analysis of existing databases and the logical level: the DBA companion project. *SIGMOD Record*, 32(1):47–52, March 2003.

[Mos03] The 30 most spoken languages of the world, December 2003. www.krysstal.com/spoken.html.

[MS00] A. Maedche and S. Staab. Mining ontologies from text. In *Proceedings of the 12th International Conference of Knowledge Engineering and Knowledge Management (EKAW 2000)*, pages 189–202, Juan-les-Pins, France, October 2000.

[Mus99] I. Muslea. Extraction patterns for information extraction tasks: A survey. In *Proceedings of the American Association for Artificial Intelligence*, 1999.

[MZ98] T. Milo and S. Zohar. Using schema matching to simplify heterogeneous data translation. In *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB-98)*, pages 122–133, August 1998.

[Nag84] G. Nagy. A critical analysis of remote sensing technology. In S. Levialdi, editor, *Digital Image Analysis*, pages 88–193. Pitman Publishing, London, England, 1984.

[Nag85] G. Nagy. Image database. In *Image and Vision Computing*, volume 3, pages 111–117. Butterworth, 1985.

[Nag00a] G. Nagy. Geometry and geographic information systems. In C. Gorini, editor, *Geometry at Work*, Notes Number 53, pages 88–104. The Mathematical Association of America, 2000.

[Nag00b] G. Nagy. Twenty years of document image analysis in PAMI. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):38–62, January 2000.

[NAM97] S. Nestorov, S. Abiteboul, and R. Motwani. Inferring structure in semistructured data. *SIGMOD Record*, 26(4):39–43, December 1997.

[NAM98] S. Nestorov, S. Abiteboul, and R. Motwani. Extracting schema from semistructured data. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data (SIGMOD'98)*, pages 295–306, Seattle, Washington, June 1998.

[NG82] S. Navathe and S.G. Gadgil. A methodology for data schema integration in the entity-relationship model. In *Proceedings of the 8th International Conference on Very Large Databases*, pages 142–164, Mexico City, Mexico, September 1982.

[NME90] G. Nagy, M. Mukherjee, and D.W. Embley. Making do with finite numerical precision in spatial data structures. In *Fourth International Symposium on Spatial Data Handling*, pages 55–65, Zürich, Switzerland, July 1990.

[NS84] G. Nagy and S. Seth. Hierarchical representation of optically scanned documemts. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 347–349, 1984.

[NW79] G. Nagy and S. Wagle. Geographic data processing. *ACM Computing Surveys*, 11(2):139–181, June 1979.

[NZ02] G. Nagy and J. Zou. Interactive visual pattern recognition. In *Proceedings of Sixteenth International Conference on Pattern Recognition*, volume III, pages 478–481, Québec, Canada, August 2002.

[Ont03] Ontobuilder, December 2003. iew3.technion.ac.il/OntoBuilder/.

[PC97] P. Pyreddy and W.B. Croft. TINTIN: A system for retrieval in text tables. In *Proceedings of the 2nd ACM International Conference on Digital Libraries*, pages 193–200, Philadelphia, Pennsylvania, July 1997.

[PCA97] C. Peterman, C.H. Chang, and H. Alam. A system for table understanding. In *Proceedings of the Symposium on Document Image Understanding Technology (SDIUT'97)*, pages 55–62, Annapolis, Maryland, April/May 1997.

[PCH93] I.T. Phillips, S. Chen, and R.M. Haralick. CD-ROM document database standard. In *Proceedings of the Second International Conference on Document Analysis and Recognition (ICDAR'93)*, pages 478–483, Tsukuba, Japan, October 1993.

[PTU00] L. Palopoli, G. Teracina, and D. Ursino. The system DIKE: Towards the semi-automatic synthesis of cooperative information systems and data warehouses. In *Proceedings of ADBIS-DASFAA 2000*, pages 108–117, 2000.

[RL94] E. Riloff and W. Lehnert. Information extraction as a basis for high-precision text classification. *ACM Transactions on Information Systems*, 12(3):296–333, 1994.

[RNN99] S.V. Rice, G. Nagy, and T.A. Nartker. *Optical Character Recognition: An Illustrated Guide to the Frontier*. Kluwer Academic Publishers, Boston, Massachusetts, 1999.

[RS97] D. Rus and D. Subramanian. Customizing information capture and access. *ACM Transactions on Information Systems*, 15(1):67–101, 1997.

[Rub03] L. Rubén. NGW seminar, December 2003. www.nextwebgeneration.org/nextwebgen03/1.

[SG89] A. Sheth and S.K. Gala. Attribute relationships: An impediment in automating schema integration. In *Proceedings of the NSF Workshop on Heterogeneous Database Systems*, Evanston, Illinois, December 1989.

[SH01] M. Stonebraker and J.M. Hellerstein. Content integration for e-business. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data (SIGMOD 2001)*, pages 552–560, Santa Barbara, California, May 2001.

[SM79] A.M. Silva and M.A. Melkanoff. A method for helping discover the dependencies of a relation. In H. Gallaire, J.-M. Nicolas, and J. Minker, editors, *Advances in Data Base Theory*, pages 115–133. Plenum Press, New York, New York, 1979.

[SMJ02] P. Spyns, R. Meersman, and M. Jarrar. Data modeling versus ontology engineering. *SIGMOD Record*, 31(4):12–17, December 2002.

[Sow00] J.F. Sowa. Conceptual structures: Logical, linguistic, and computational issues. In B. Ganter and G. W. Mineau, editors, *Proceedings of the 8th International Conference on Conceptual Structures (ICCS)*, volume 1867 of *Lecture Notes in AI*, pages 55–81, Darmstadt, Germany, August 2000. Springer-Verlag.

[SP94] S. Spaccapietra and C. Parent. View integration: A step forward in solving structural conflicts. *IEEE Transactions on Knowledge and Data Engineering*, 6(2):258–274, April 1994.

[SSZ98] G. Schechtman, T. Schlumprecht, and J. Zinn. On the guassian measure of the intersection of symmetric convex sets. *Annals of Probability*, 26:346–357, 1998.

[TBB96] E. Turolla, Y. Belaid, and A. Belaid. Form item extraction based on line searching. In R. Kasturi and K. Tombre, editors, *Graphics Recognition—Methods and Applications*, volume 1072 of *Lecture Notes in Computer Science*, pages 69–79, Berlin, Germany, 1996. Springer-Verlag.

[TE02] K. Tubbs and D.W. Embley. Recognizing records from the extracted cells of microfilm tables. In *Proceedings of the Symposium on Document Engineering (DocEng'02)*, pages 149–156, McLean, Virginia, November 2002.

[TELN03] Y.A. Tijerino, D.W. Embley, D.W. Lonsdale, and G. Nagy. Ontology generation from tables. In *Proceedings of the 4th International Conference on Web Information Systems Engineering*, Rome, Italy, December 2003. 242–249.

[Top02] TopoZone, October 2002. www.topozone.com.

[Usc02] M. Uschold. Creating semantically communication on the world wide web, May 2002. Keynote address at the *Proceedings of the Semantic Web Workshop at the 11th International WWW Conference.*

[Wan96] X. Wang. *Tabular Abstraction, Editing, and Formatting.* PhD thesis, University of Waterloo, 1996.

[WG01] C.A. Welty and N. Guarino. Supporting ontological analysis of taxonomic relationships. *Data & Knowledge Engineering*, 39(1):51–74, 2001.

[WL97] K. Wang and H. Liu. Schema discovery for semistructured data. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 271–274, Newport Beach, California, August 1997.

[Wor03a] Worldatlas.com, December 2003. www.worldatlas.com/geoquiz/thelist.htm.

[Wor03b] The world factbook—2003, December 2003. www.cia.gov/cia/publications/factbook.

[WQS95] T. Watanabe, Q.L. Quo, and N. Sugie. Layout recognition of multi-kinds of table-form documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(4):432–445, 1995.

[WSW99] Y. Wand, V.C. Storey, and R. Weber. An ontological analysis of the relationship construct in conceptual modeling. *ACM Transactions on Database Systems*, 24(4):494–528, December 1999.

[XE02] L. Xu and D.W. Embley. Combining the best of global-as-view and local-as-view for data integration. Technical report, Brigham Young University, Provo, Utah, 2002. http://www.deg.byu.edu.

[XE03a] L. Xu and D.W. Embley. Automating schema mapping for data integration. Technical report, Brigham Young University, Provo, Utah, 2003. http://www.deg.byu.edu.

[XE03b] L. Xu and D.W. Embley. Discovering direct and indirect matches for schema elements. In *Proceedings of the 8th International Conference on Database Systems for Advanced Applications (DASFAA 2003)*, pages 39–46, Kyoto, Japan, March 2003.

[XE03c] L. Xu and D.W. Embley. Using domain ontologies to discover direct and indirect matches for schema elements. In *Proceedings of the Workshop on Semantic Integration (WSI'03)*, pages 105–110, Sanibel Island, Florida, October 2003.

[Xu03] L. Xu. *Source Discovery and Schema Mapping for Data Integration.* PhD thesis, Brigham Young University, August 2003.

[ZdBC04] R. Zanibbi, d. Blostein, and J.R. Cordy. Recognizing tables in documents. *International Journal of Document Analysis and Recognition*, 2004. (to appear).

[ZN04a] J. Zou and G. Nagy. Computer assisted interactive recognition: Formal description and evaluation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2004. (submitted).

[ZN04b] J. Zou and G. Nagy. Evaluation of model-based interactive flower recognition. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2004. (submitted).

[ZN04c] J. Zou and G. Nagy. A procedure for model-based interactive object segmentation. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2004. (submitted).

[Zuy97] K. Zuyev. Table image segmentation. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'97)*, pages 705–708, Ulm, Germany, August 1997.