PROJECT DESCRIPTION

1 Introduction

The exponential increase in new knowledge that characterizes our modern age of information technology precludes depending solely on scholarly individual effort to keep up with new information. We must therefore develop new ways of "keeping up," and we must develop them quickly.

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 and filled-in forms 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). TANGO thus exploits tables and filled-in forms to generate a domain-specific ontology with minimal human intervention.

The implications of this challenge are at the same time theoretically intriguing and practically significant. For a domain of interest, can we establish a minimal kernel of intentional and extentional objects and relationships, constraints among them, and computations over them that can serve as a basis for recognizing and assimilating new knowledge? Given this kernel of knowledge, can we derive semantics from syntactic clues in the layout and content of metadata and data? Then, with the semantics in hand, can we automatically recognize the overlap with kernel and other previously obtained knowledge, and thus also recognize the differences and add these differences to the growing body of knowledge? Can we also 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 tasks—answer queries, extract knowledge from unstructured documents within the domain of knowledge, resolve semantic interoperability, and enable information exchange between disparate software agents working within the same domain?

We propose TANGO as an approach to investigate the resolution of these challenges. TANGO builds domain-specific ontologies. "An ontology is a formal explicit specification of a shared conceptualization," where conceptualization means "how people think about things" and specification means "the concepts and relationships of the abstract model are given explicit terms and definitions" [GL02]. Our work can be considered as semi-automated, applied "ontological engineering," which has as its goal "effective support of ontology development and use throughout its life cycle—design, evaluation, integration, sharing, and reuse" [GL02]. As an analogy for what we are proposing, consider that instead of humans collaborating to design an ontology [HJ02], we provide an approach in which *tables* "collaborate" to design an ontology. In a sense, this is the same because information is assembled from specific instances of tables created by humans.

We will design an information-gathering engine that expands from an embryonic kernel rather than one that grows from scratch. Relevant web pages or tables will be interpreted with the help of current application and tool ontologies (i.e. ontologies for tables, and ontological knowledge about a domain and about semantic integration within a domain). From each experience, new facts, relations, and interpretive techniques will be used to expand, correct, and consolidate the growing application ontology. Where necessary, human help may be invoked: one of our major goals is to find out how little human interaction is sufficient.

We will demonstrate the feasibility of automated knowledge gathering in the domain of geo-political facts and relations, where relevant empirical data is widely scattered but often presented in the form of lists, tables, and forms. The geo-political application ontology will be constructed using tool ontologies that encapsulate a growing understanding of coordinate systems, geo-political subdivisions, and conventions for reading tables. The chosen domain of geography spans many important human activities: natural resources, travel, culture, commerce, and industry.

Our research will make use of robust web wrapper generation systems for on-line data, and of OCR-based table analysis for page images. As a basis model for ontology construction, we intend to use a formally defined conceptual modeling language that has a direct translation into predicate calculus [EKW92]. This provides a theoretical foundation for formal property analysis. Another key element of our approach to ontology building is searching for direct and indirect schema-element matches [EJX01, XE02b] between populated database schemas (i.e. between a new document and ontologically organized, previously seen documents). We will also depend on (1) subject-specific lexicons and thesauri, (2) specialized data frames [Emb80, ECJ$^+$99] for commonly occurring fields (like latitude-longitude pairs or dates), (3) object-class congruency principles [CEW96], (4) formally consistent tools for manipulating meta-data [LEW00], (5) OCR [RNN99], (6) analysis tools and techniques [LN99b, LN02, TE02], and (7) ontology-maintenance tools developed by others, e.g. [SMMS02].

Our earlier experiments with ontology-based information extraction have been successful on relatively narrow domains: census records [TE02], automobile want ads and obituaries [ECJ$^+$99], and several other domains [DEG]. Earlier work on related issues, including isolated tables [LN99b], topographic maps [LNS$^+$00], satellite images [EN89, NME90], and geographic data processing [NW79, Nag00a], has also been successful. We believe that we are now ready to integrate what we have learned and build a system to tackle the much larger, but still bounded, domain of geographic information.

The principal investigators of our research team consist of a conceptual-modeling/database specialist, a linguist/ontologist, and a document engineer, all linked by previous research projects. We hope that the project will allow one student to complete doctoral research under each of the three professors, and that it will motivate three undergraduate students to extend their studies.

We proceed with our proposal for TANGO as follows. Section 2 sets forth our research objectives. Section 3 presents our ideas on growing ontologies by means of an extended example. Section 4 builds on the extended example in Section 3 to provide some applications for what we are proposing. Section 5 explains how we expect to evaluate our work and measure its effectiveness. Section 6 shows how our work builds on and is complementary to some of our own previous work and the work of others, and Section 7 describes our plan for accomplishing the proposed research.

2 Objectives

The goal of our research is to transcend the current limitations—of scale, variety and affordability—to organize scattered, heterogeneous data into useful collections.

We propose to integrate recent developments in conceptual modeling and in document understanding to build a knowledge-gathering engine that will operate with minimal human supervision. We expect to demonstrate empirically that a domain-specific ontology can be expanded at an increasing rate by matching already known groups of items and relations with groups of new items linked by only partially analogous relations. We hope to show the benefits of a unified and formal approach where not only the information extracted from each source, but also the analysis tools themselves, are realized as intermeshing ontologies.

To discover the effectiveness of our methods, we will conduct experiments on selected "greenhouse documents," which we plan to find or create to illustrate both potential opportunities and difficulties, as well as on real-world web pages and images of available paper documents. We will compare the relative merits of beginning with kernel ontologies of various sizes. We will test the growth in the quality, size, and scope of the knowledge base as a function of the number of accessed documents through periodic graded queries, interoperability resolution, and automatic sub-ontology elicitation for particular tasks such as data extraction, database application modeling, and agent communication. We will determine the amount and type of human intervention necessary by automated logging of all interactions.

Our research will identify and quantify what can be accomplished by combining the best available ideas and tools (1) in the geo-knowledge domain of high global interest, (2) in the growing field of ontology analysis and development, and (3) in a sphere of knowledge engineering where further invention is necessary.

3 Ontology Generation

Ontology generation makes use of auxiliary knowledge sources, including an ontology-based system for (1) table understanding, (2) data extraction, and (3) data integration. Based on completed research, we offer the following specifics.

- Our ontology-based table-understanding system allows us to take an arbitrary[1] table as input and produce attribute-value pairs as output [LN99b, LN00, ETL02, TE02].
- Our ontology-based data-extraction system allows us to take semistructured text as input, including in particular attribute-value pairs extracted from tables [ETL02], and produce as output a database corresponding to a given application ontology and populate it with the given semistructured data. (We have developed resilient web wrapper-generation systems that do not break when pages change[2] or new pages come on-line because the basis for the extraction is an ontology rather than a page grammar and its variations [ECJ$^+$99, LRNdST02, ETL02, DEG].)
- Our ontology-based integration system produces schema-element matches between populated database schemas: direct matches when schema elements in two schemas have the same meaning, and many indirect matches when schema elements have overlapping meanings or have different encodings [BE03, EJX01, XE02b]. The key ideas for matching, which we explore in this integration work, are (1) value characteristics, (2) expected values based on our data-extraction techniques, (3) attribute names and their synonyms, and (4) the structure of a schema.

Our ontology-generation procedure has three steps, the first of which we do only once:

1. We build a kernel application ontology, which should be *small* (having only a few concepts), *central* (containing the most important concepts for the application), and *example-rich* (containing typical sample data, descriptions of common data values such as dates and times, and typical operations over this data).

[1] Since fully general table-understanding can be extremely difficult, even for humans [HKL$^+$01], "arbitrary" means "most common table formats." We intend, however, to expand our ontology for table understanding and contribute to this research, as well.

[2] [KK02] reports that on the average pages change twice per year.

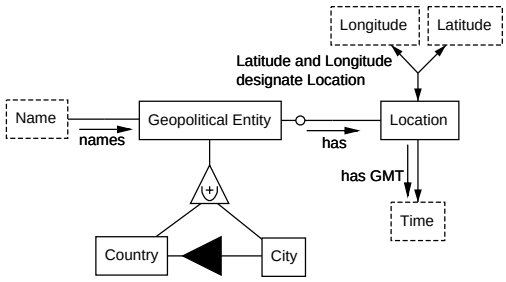


Figure 1: Initial Geopolitical Ontology.

2. For any given table, we create a mini-ontology based on our understanding of the table. This yields a schema of object and relationship sets, values for the object sets as attribute-value pairs, and tuples for the relationship sets each representing a relationship among attribute-value pairs.

3. We integrate each new mini-ontology with the ontology we are building. Integration may raise several issues: (a) there may be alternative ways we can integrate the mini-ontology into the evolving global ontology, (b) constraints may be inconsistent, (c) adjustments to the evolving ontology may be necessary, and (d) we may need human intervention. To resolve these issues, we can use congruency principles [CEW96] and principles of ontology construction [Bun77, Gua98, WSW99, WG01, EW01]; and when we need human intervention we can use Issue/Default/Suggestion (IDS) statements as in [BE03] as well as tools for cleaning ontologies, e.g. [WG01, GW02].

We explain these three steps by examples. Figure 1 shows a graphical representation of a proposed kernel application ontology for geopolitical entities. We briefly explain the notation and the knowledge associated with the notation.[3] In the notation a box represents an object set—dashed if printable (e.g. *Longitude* in Figure 1) and not dashed if not printable (e.g. *Geopolitical Entity*). 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 named *Latitude and Longitude designate Location*). Names of binary relationship sets have a labeled reading-direction arrow, which along with the names of connected object sets form its name (e.g. *Location has GMT Time*). Optional or mandatory participation constraints 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 ontology in Figure 1 declares that, geopolitical entities must have specified names, but need not have specified locations. 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. Open triangles denote generalization/specialization hierarchies (ISA hierarchies, subset constraints, or

[3]The particular notation 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], ontological table understanding [ETL02], ontological data extraction [DEG], and ontological data integration [EJX01, XE02b].

inclusion dependencies), so that both *Country* and *City*, for example, are subsets of *Geopolitical Entity*. We can constrain ISA hierarchies by partition ($\uplus$), union ($\cup$), or mutual exclusion ($+$) among specializations or by intersection ($\cap$) among generalizations. Thus, for example, the ontology in Figure 1 declares that countries and cities are all the (currently) known geopolitical entities, and that countries and cities are mutually exclusive.[4] Filled in triangles denote part/whole, part-of, or aggregation hierarchies (e.g., a city is part of a country).

Each object set in an application ontology has an associated a data frame.[5] We provide seed values for our initial, kernel application ontology. For example, we initialize a lexicon with a few entries for *Country* such as *United States, Germany, Hungary, Japan, Brazil*, and another lexicon with a few entries for *City* such as *New York, Philadelphia, Los Angeles, Chicago, Salt Lake City, Berlin, Frankfurt, Budapest, Tokyo, Yokohama, Sao Paulo*. We also provide regular expressions for infinite-value sets. For *Time*, for example, we let $([1-9]|10|11|12):[0-5]\setminus d(\setminus s*(a|p)\setminus .?\setminus s*m\setminus .?)?$, which denotes strings such as *2:00 pm* and *11:49 a.m.*, be part of the recognizer.[6] Finally, we add appropriate procedural knowledge that may be useful. Examples include distances between locations based on latitude and longitude, the duration between two times, or the number of time zones between two geopolitical entities.

Having exemplified Step 1, production of a kernel ontology, we now give three examples to illustrate Steps 2 and 3. Besides illustrating these steps, we also illustrate the types of input tables we intend to consider in our research. Note (1) that the examples range from full tables directly available on the web to partial tables hidden behind forms on the web, (2) that they range from electronic tables to scanned table images, and (3) that their diversity ranges from simple tables to semistructured tables with auxiliary information.

- www.gazetteer.de/home.htm on 17 September 2002 (Figure 2). Given this table, we create the mini-ontology in Figure 3(a) and then integrate this ontology into the ontology we are constructing (initially the ontology in Figure 1). The result is the ontology in Figure 3(b). This is the heart of our research, and there are a host of problems to resolve. Briefly, we reach the ontology in Figure 3(b) by reasoning as follows.

  Understand Table: Table "understanding" means to associate the attributes with values and obtain atomic attribute-value pairs. This is straightforward for the table in Figure 2.[7]

  Discover Constraints: (1) By looking at the data, we can obtain the functional dependencies (FDs) with reasonable, but not absolute, certainty. Since *Agglomeration* is a key—and particularly, a left-most-column key—we have *Agglomeration* $\rightarrow$ *Population*, *Continent*, *Country*. We have overwhelming evidence that *Population* $\rightarrow$ *Agglomeration*, *Continent*, *Country*.[8] We also have overwhelming evidence that *Country* $\rightarrow$ *Continent*, plus overwhelming counter-evidence that *Continent* $\not\rightarrow$ *Population*, *Country*, *Agglomeration*, and that *Country* $\not\rightarrow$ *Population*,

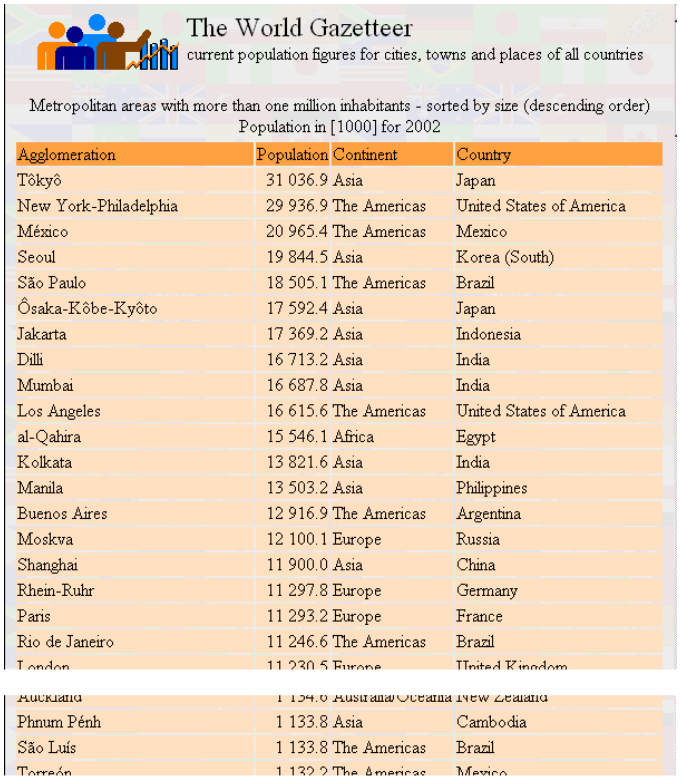
[4] Note that for city-states like Singapore, one object represents the City and another object represents the Country—both can have the same name.

[5] Using regular expressions and lexicons, a data frame for a concept C recognizes self-describing constant values of C and keywords that signal the presence C objects or C values. Data frames also include transformations between internal and external representations and computational knowledge as multi-sorted algebras over the concepts within the knowledge domain. See [Emb80].

[6] We use Pearl-like syntax in our regular expressions.

[7] It takes considerable knowledge to recognize that the populations are in thousands and that they are for 2002. In [ETL02] we show how to extract header and footer information, but only if anticipated; thus, we do not assume that this knowledge comes from the table in Figure 2. We do, however, keep all knowledge sources so that we can refer back to them as we continue to update the ontology.

[8] We do not always seek for 100%. As Figure 2 shows *Phnum Pénh* and *São Luis* happen to have the same population value. We often consider near 100% as sufficient evidence that a constraint should hold. Indeed, for this particular case, we will want to reject this FD. To solve such problems, we intend to use additional reasoning and new information sources to confirm or contradict "known" information.



The World Gazetteer

current population figures for cities, towns and places of all countries

Metropolitan areas with more than one million inhabitants - sorted by size (descending order)
Population in [1000] for 2002

Agglomeration	Population	Continent	Country
Tôkyô	31 036.9	Asia	Japan
New York-Philadelphia	29 936.9	The Americas	United States of America
México	20 965.4	The Americas	Mexico
Seoul	19 844.5	Asia	Korea (South)
São Paulo	18 505.1	The Americas	Brazil
Ôsaka-Kôbe-Kyôto	17 592.4	Asia	Japan
Jakarta	17 369.2	Asia	Indonesia
Dilli	16 713.2	Asia	India
Mumbai	16 687.8	Asia	India
Los Angeles	16 615.6	The Americas	United States of America
al-Qahira	15 546.1	Africa	Egypt
Kolkata	13 821.6	Asia	India
Manila	13 503.2	Asia	Philippines
Buenos Aires	12 916.9	The Americas	Argentina
Moskva	12 100.1	Europe	Russia
Shanghai	11 900.0	Asia	China
Rhein-Ruhr	11 297.8	Europe	Germany
Paris	11 293.2	Europe	France
Rio de Janeiro	11 246.6	The Americas	Brazil
London	11 230.5	Europe	United Kingdom
Auckland	1 134.6	Australia/Oceania	New Zealand
Phnum Pénh	1 133.8	Asia	Cambodia
São Luís	1 133.8	The Americas	Brazil
Torreón	1 132.2	The Americas	Mexico

Figure 2: Partial City Population Table.

Agglomeration. Figure 3(a) shows the mini-ontology for the table after having determined the FDs and after having removed those that are redundant. (2) The data shows (nearly 100%[9]) that the relations over (*Continent*, *Country*) and (*Country*, *Agglomeration*), are irreflexive, asymmetric, and transitive.

Match: (1) *Country* matches *Country*. (2) We parse the strings under *Agglomeration* and, using techniques in [EJX01], discover that they are cities. Moreover, using techniques in [ETL02], we discover that some are city groups when we recognize, for example, both *New York* and *Philadelphia* in *New York-Philadelphia*. This leads us to believe that *Agglomeration* is a group of one or more hyphen-separated cities.[10] (3) The value characteristics of *Agglomeration*, *City*, *Continent*, and *Country* all correspond to the expected characteristics for *Name* of *Geopolitical Entity*. *Population*, however, does not.

Merge: (1) Based on ISA for *Country* and *City* in Figure 1, plus importantly that the names satisfy the name constraints for *Name* of *Geopolitical Entity*, we are led to believe that *Continent* and *Agglomeration* should be added as specializations of *Geopolitical Entity*. (2) Since the FDs are consistent with the typical 1-*n* relationships of aggregation, the names satisfy the name constraints for *Name* of *Geopolitical Entity*, and the relations are irreflexive, asymmetric, and transitive, we are led to believe that *City isPartOf Agglomeration isPartOf Country isPartOf Continent*. (3) We do not include *Population* since it satisfies neither the name constraints nor the 1-*n* constraints. (4) Because of the *isPartOf* constraints and the relationship of both *Agglomeration* and *City* with *Population*, we are led to the conclusion that *Population* should be an attribute of all the specializations under *Geopolitical Entity*. We thus relate *Population* directly to *Geopolitical Entity*. Its functional constraints, however, are in question. We observe that the

[9]Exception examples: The city *Singapore* is in the country *Singapore*, and *Istanbul* is in both Asia and Europe.

[10]There are some really interesting contradictions: *al-Qahira* is only one city and *Rhein-Ruhr* describes several dozen cities in Germany extending from *Köln* to *Dortmund*.

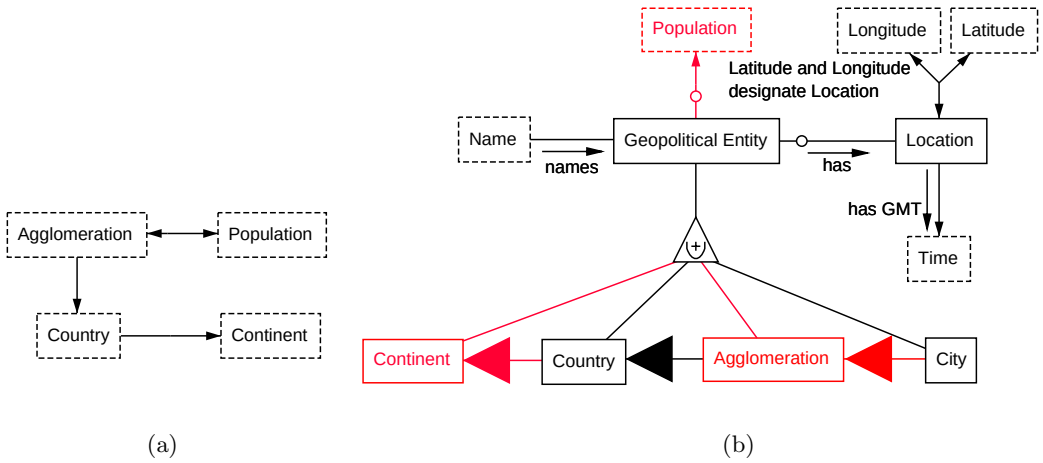


Figure 3: Mini-Ontology Constructed from the Table in Figure 2 (left), and Updated Ontology after Integrating Figure 2 into Figure 1 (right). (The red elements are new.)

counter-evidence *Continent* $\not\rightarrow$ *Population* and *Country* $\not\rightarrow$ *Population* suggests that *Geopolitical Entity* $\not\rightarrow$ *Population*, and we observe that we have sometimes split *Agglomeration* into cities with no population value and have a few counter examples for *Population* $\rightarrow$ *Agglomeration*, *Continent*, *Country*. These observations raise too many questions, so we let the user resolve the problems (the resolutions should be synergistic, based on ontological principles and tool support [WG01, GW02, BE03]). We assume that this resolution yields the optional FD from *Geopolitical Entity* to *Population* in Figure 3(b).

- www.topozone.com/findresults.asp?place=Bonnie+Lake&statefips=0&... on 8 April 2002 (Figure 4). This site uses a form: we entered "Bonnie Lake" to obtain the upper table in Figure 4. We can in addition use the site's form to look for places other than Bonnie Lake, such as New York as the lower table in Figure 4 shows.[11]. We reason, using the same *Unterstand-Discover-Match-Merge* steps as before. The result is in Figure 5.

  Understand Table: This is straightforward, except that we have at our disposal a huge table behind the form, made up of many small tables, one for each *Place* in the hidden database.

  Discover Constraints: (1) First we have to observe that *Place* is not a key; yet it also contains the name we entered in our search. We conclude that the places are all different. Hence, we give each row a tuple identifier, which makes it a member of a nonlexical object set, which we call *Place*. In addition, we make a lexical object set *Place Name*, which contains the lexical name in the table under *Place*—*Bonnie Lake* in upper table and *New York* in the lower table in Figure 4. (2) We obtain the FDs by looking at the data and the optionals from the *unknown* values in the table. Figure 5(a) shows these constraints. (3) Next we observe that *Type* includes *City*, which we already have in our growing ontology (Figure 3(b)). With some more investigation into other tables using cities we know about such as *New York* and *Philadelphia*, we eventually conclude that *Type* values, like cities, are each a specialization of *Place*.[12]

  Match: (1) *Longitude* and *Latitude* in Figure 5(a) match with *Longitude* and *Latitude* in our growing ontology in Figure 3(b). (2) The newly created lexical object set, *Place Name*, matches *Name*

[11] We can also automatically retrieve other parts of the table using techniques for crawling the hidden web [RGM01, LYE01, LDEY02]

[12] To keep the illustrative ontology small, we show only the types from Figures 4—there are many more.

TopoZone.com TopoZone.com click here to find out more

Get a Map
Place name search
Decimal degrees
Deg/min/sec
UTM coords
Download Maps
TopoFactory Login
About TopoFactory
Specifications
TopoFactory Store
How to...
Put topo maps on your Web site
Get digital data
Link to us
What's New?
Get the TopoTimes
Press releases
New @ TopoZone
Awards
Help
Map tips
Topo map symbols
FAQ
Support
Privacy policy
About us

Here's what we found

Here are all the places we know about that match your search. Select the one you'd like and **click on its name** to go to the map.

Place	County	State	Type	Elevation	USGS Quad	Lat	Lon
Bonnie Lake	Matanuska-Susit	Alaska	lake	unknown	Anchorage D-4	61.814°N	148.303°W
Bonnie Lake	Fresno	California	lake	9531 feet	Kaiser Peak	37.298°N	119.194°W
Bonnie Lake	Mono	California	lake	unknown	Tower Peak	38.189°N	119.576°W
Bonnie Lake	Jackson	Iowa	lake	unknown	Green Island	42.193°N	90.365°W
Bonnie Lake	Crow Wing	Minnesota	lake	1204 feet	Pelican Lake	46.553°N	94.126°W
Bonnie Lake	Lake	Minnesota	lake	1396 feet	Kekekabic Lake	48.086°N	91.217°W
Bonnie Lake	Klamath	Oregon	lake	6186 feet	Willamette Pass	43.549°N	122.106°W
Bonnie Lake	Aiken	South Carolina	reservoir	unknown	Seivern	33.723°N	81.420°W
Bonnie Lake	Duchesne	Utah	lake	unknown	Mirror Lake	40.711°N	110.876°W
Bonnie Lake	King	Washington	lake	unknown	Big Snow Mountain	47.566°N	121.271°W
Bonnie Lake	Spokane	Washington	lake	1790 feet	Chapman Lake	47.273°N	117.568°W

MAPCARD Map Your Property! CLICK HERE! Aerial Photos Road Maps Back Next

Place	County	State	Type	Elevation*	USGS Quad	Lat	Lon
New York	Santa Rosa	Florida	town/city	231 feet	Chumuckla	30.838°N	87.201°W
New York	Wayne	Iowa	town/city	1657 feet	Corydon	40.852°N	93.260°W
New York	Ballard	Kentucky	town/city	460 feet	Blandville	36.989°N	88.953°W
New York	Caldwell	Missouri	town/city	800 feet	Hamilton East	39.685°N	93.927°W
New York	Shelby	Missouri	area	unknown	Unknown	unknown	unknown
New York	Cibola	New Mexico	town/city	6033 feet	Cubero	35.059°N	107.527°W
New York	New York	New York	town/city	unknown	Jersey City	40.714°N	74.006°W
New York	Henderson	Texas	town/city	488 feet	Leagueville	32.168°N	95.669°W
New York	Summit	Utah	mine	unknown	Heber City	40.615°N	111.489°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 4: Table of Bonnie Lakes (above) and New Yorks (below).

of *Geopolitical Entity*. (3) *town/city* matches *City*.[13].

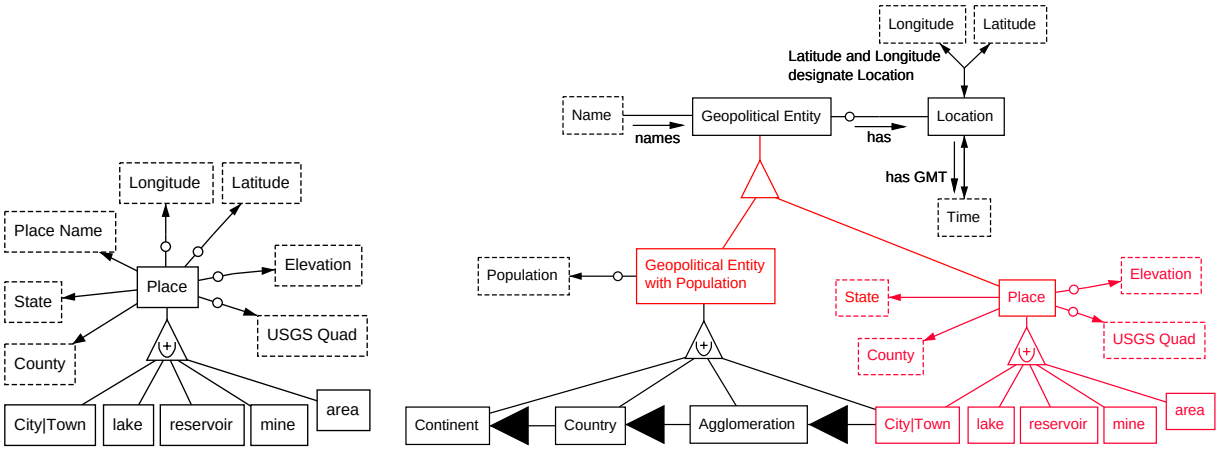
Merge: (1) Given that *City* is a specialization of *Geopolitical Entity*, and that each *Place* has a name and location (*Longitude* and *Latitude*), we conclude that *Place* is either a *Geopolitical Entity* or a specialization of a *Geopolitical Entity*. Further, since its specializations do not include *Continent*, *Country*, or *Agglomeration*, we rule out *Place* as being equivalent to *Geopolitical Entity* and conclude that *Place* must be a specialization. (2) Since we have no evidence about populations for places, by congruency [CEW96], there must be a missing specialization object set of *Geopolitical Entity*, which we call *Geopolitical Entity with Population*. (3) We note that *City|Town* is in both *Geopolitical Entity with Population* and *Place*. Thus, we cannot have mutual exclusion between the two object sets and thus also no partition. We could have a union constraint, but as mentioned there are many, many more types; thus, we do not place a union constraint in the diamond under *Geopolitical Entity*.[14]

- www.nara.gov/cgi-bin/starfinder/6881/images.txt on 18 July 2002 (Figure 6).

  Here we use an image of a paper table rather than an HTML table. This bolsters our claim that not

[13]The bar notation in OSM lets us specify synonyms (aliases); thus the object set named *City|Town* contains objects that we can call either *City* or *Town*

[14]As our ontologies grow, diagrams eventually become too large to display. They are useful for display of small portions of a large ontology, but our actual representation is textual and equivalent in every respect with the OSM diagramatic notation [LEW00].



(a) (b)

Figure 5: Mini-Ontology Constructed from the Tables in Figure 4 (left), and after processing the Tables (right). (The red elements are new.)

only can we use HTML tables ranging from very high level (continents) to very low level (small bodies of water), but we can also range across all kinds of media—HTML tables, spreadsheets, electronic human-constructed forms (tabbed by hand), and images of paper tables where we need OCR to read the data—and all kinds of structured and semistructured information (complete tables, semistructured tables, filled-in forms).

Understand Table: Table understanding is not straightforward for the document in Figure 6. Although there are no attribute headings to guide us, we are able to use the extraction techniques we have developed to recognize the two dates,[15] the military-style times, and the latitudes and longitudes. We, of course, can also recognize the strings. In addition, we can also use the geometric layout information to observe the following nesting patterns.

```
USS Franklin
    (Date
        String1
            (Time String2)*)*
```

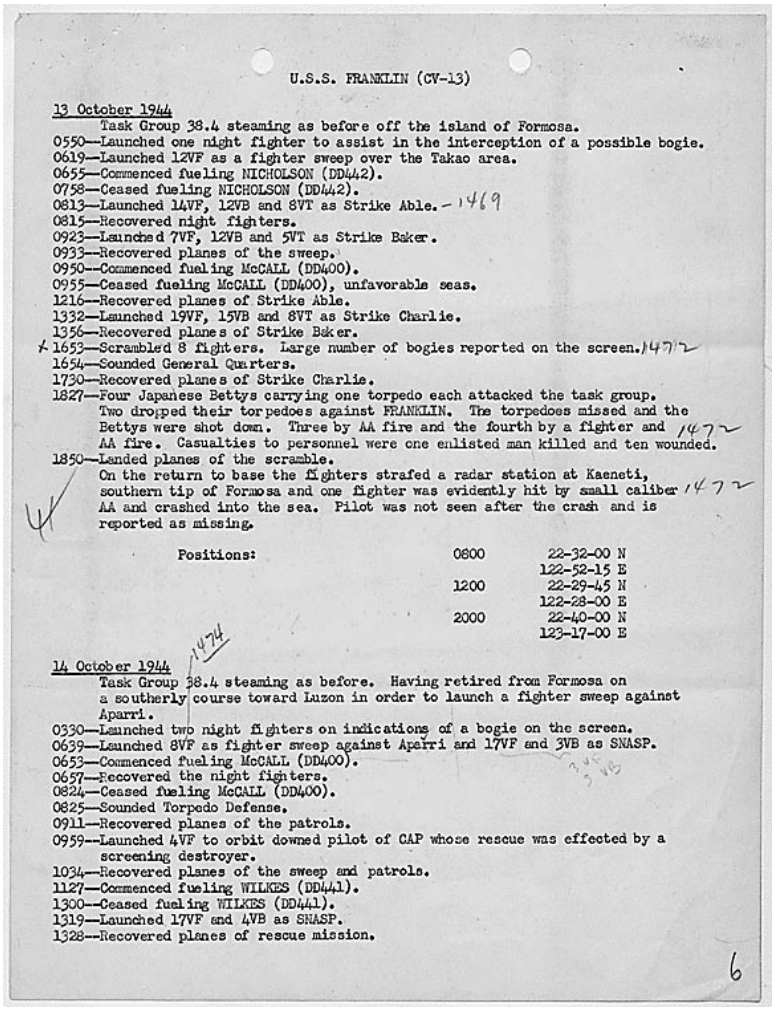
```
USS Franklin
    (Date
        String1
            (Time
                (Latitude
                Longitude)*)*)*
```

We then unnest to obtain tables whose headers and first two rows are as follows.

	Date	String1	Time	String2
USS Franklin	13 October 1944	"Task Group ..."	0550	"Launched one ..."
USS Franklin	13 October 1944	"Task Group ..."	0619	"Launched 12VF ..."

	Date	String1	Time	Latitude	Longitude
USS Franklin	13 October 1944	"Task Group ..."	0800	22-32-00 N	122-52-15 E
USS Franklin	13 October 1944	"Task Group ..."	1200	22-29-45 N	122-28-00

[15] As part of our kernel ontology, we assume the existence of common data items such as dates, currencies, distances in various units, and so forth.



U.S.S. FRANKLIN (CV-13)

13 October 1944

Task Group 38.4 steaming as before off the island of Formosa.
0550--Launched one night fighter to assist in the interception of a possible bogie.
0619--Launched 12VF as a fighter sweep over the Takao area.
0655--Commenced fueling NICHOLSON (DD442).
0758--Ceased fueling NICHOLSON (DD442).
0813--Launched 14VF, 12VB and 8VT as Strike Able.
0815--Recovered night fighters.
0923--Launched 7VF, 12VB and 5VT as Strike Baker.
0933--Recovered planes of the sweep.
0950--Commenced fueling McCALL (DD400).
0955--Ceased fueling McCALL (DD400), unfavorable seas.
1216--Recovered planes of Strike Able.
1332--Launched 19VF, 15VB and 8VT as Strike Charlie.
1356--Recovered planes of Strike Baker.
1653--Scrambled 8 fighters. Large number of bogies reported on the screen.
1654--Sounded General Quarters.
1730--Recovered planes of Strike Charlie.
1827--Four Japanese Bettys carrying one torpedo each attacked the task group. Two dropped their torpedoes against FRANKLIN. The torpedoes missed and the Bettys were shot down. Three by AA fire and the fourth by a fighter and AA fire. Casualties to personnel were one enlisted man killed and ten wounded.
1850--Landed planes of the scramble.
On the return to base the fighters strafed a radar station at Kaeneti, southern tip of Formosa and one fighter was evidently hit by small caliber AA and crashed into the sea. Pilot was not seen after the crash and is reported as missing.

Positions:

	0800	22-32-00 N 122-52-15 E
	1200	22-29-45 N 122-28-00 E
	2000	22-40-00 N 123-17-00 E

14 October 1944

Task Group 38.4 steaming as before. Having retired from Formosa on a southerly course toward Luzon in order to launch a fighter sweep against Aparri.
0330--Launched two night fighters on indications of a bogie on the screen.
0639--Launched 8VF as fighter sweep against Aparri and 17VF and 3VB as SNASP.
0653--Commenced fueling McCALL (DD400).
0657--Recovered the night fighters.
0824--Ceased fueling McCALL (DD400).
0825--Sounded Torpedo Defense.
0911--Recovered planes of the patrols.
0959--Launched 4VF to orbit downed pilot of CAP whose rescue was effected by a screening destroyer.
1034--Recovered planes of the sweep and patrols.
1127--Commenced fueling WILKES (DD441).
1300--Ceased fueling WILKES (DD441).
1319--Launched 17VF and 4VB as SNASP.
1328--Recovered planes of rescue mission.

Figure 6: Deck Log of the U.S.S. Franklin.

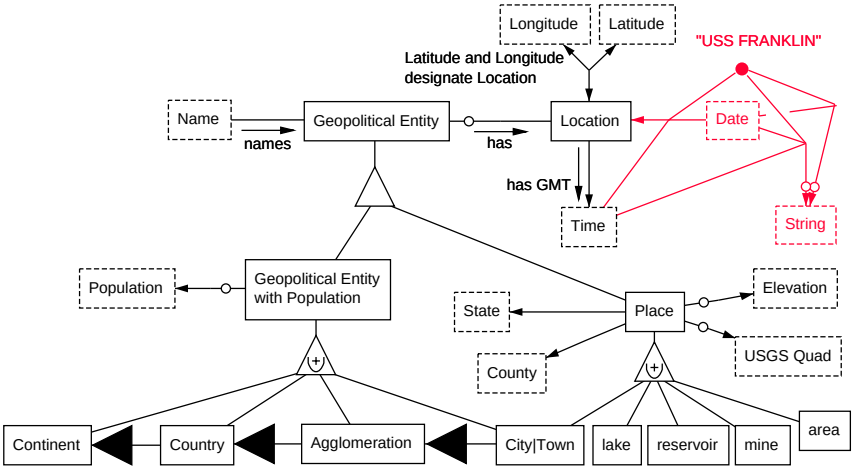


Figure 7: Ontology after Processing the U.S.S. Franklin Document. (The red elements are new.)

Discover Constraints: Next we observe the functional dependencies within the nestings: *USS Franklin, Date → String1*, *USS Franklin, Date, Time → String2*, and *USS Franklin, Date, Time → Latitude, Longitude*.

Match: We directly match *Latitude*, *Longitude*, and *Time* and observe that *Latitude* and *Longitude* are in a one-to-one correspondence with *Location*. We add *Date* and *String* as object sets from the common-knowledge part of our data-frame library.

Merge: Merge is straightforward: we simply add the discovered FDs as edges, with *Location* standing for a *Latitude-Longitude* pair, and thus, we arrive at Figure 7.

These examples only illustrate the kind of thing we want to do. TANGO should be capable of taking any readable tables in the geopolitical domain, understand them, discover constraints in them, match them with the growing ontology, and merge them such that the knowledge contained in them expands the growing ontology. Although we plan to illustrate our work only in the geopolitical domain, we intend to create TANGO so that, given a reasonable kernel ontology in any domain, it can grow ontologies for that domain.

4 Expected Significance

The focus of our TANGO project is semi-automated ontology creation, a worthy goal in and of itself. Having constructed an ontology of the type we are proposing, however, also puts us in a position to resolve many interesting and challenging problems. Examples[16] follow:

Multiple-Source Query Processing: We can use the ontology as an integrated global schema against which we can pose queries over multiple sources [ETL02, XE02a, Che02]. Examples: Did the USS Franklin dock at a city with a population of more than one million inhabitants? What towns are within 30 miles of Bonnie Lake in Duchesne County, Utah? (Hopefully none because we're looking for a wilderness experience.)

Extraction Ontologies: We can use the ontology as a guide for constructing wrappers to extract geopolitical information from as-yet-unseen, semistructured or even unstructured web pages [ECJ+99].

Extraction-Ontology Generation: As [LRNdST02] points out, our methodology [ECJ+99] creates resilient wrappers—wrappers that do not "break" or need to be rewritten or regenerated when the wrapper encounters a changed page or a newly developed page in the same application domain. Resiliency depends on approaching the problem ontologically. Manual creation costs of ontology-based wrappers, however, are high (although the costs are mitigated by amortizing over resiliency-enabled reuse). In an effort to reduce the cost of creating extraction ontologies, we have experimented with the possibility of generating them automatically given a general global ontology and a general data-frame library. Our implementation using the Mikrokosmos ontology [Mik, BN97] shows that this is possible, but that it works even better when the ontology is richer in relationship structure and more tightly integrated with the data-frame library (like the ontologies we intend to build in this proposed project) [Din02, LDEM02].

Data Integration: Automating data integration tends to work best when when rich auxiliary knowledge sources provide a basis for analyzing sources from multiple points of view, including dictionaries of synonyms and hypernyms, value characteristics, expected values, and structure [EJX01]. Indeed, we can achieve over 90% precision and recall both for direct as well as many indirect matches between data sources [XE02b]. We intend to endow TANGO ontologies with the characteristics it needs to assist in data integration.

[16] 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.

Semantic Web Creation 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 semistructured, data-rich documents on the web as they now stand into semantic-web documents. In [Cha02] we proposed to show how to bridge the gap between the current web and the semantic web by semiautomatically converting Resource Description Framework Schemas (RDFS's) [BG02] and DAML-OIL ontologies [HM00] into data-extraction ontologies [ECJ+99]. Extracted data will then be converted to RDFS, making it accessible to semantic-web agents and, in addition, will superimpose 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-created ontologies will work even better for this application.

Agent Interoperability: We have begun a project in which we wish to experiment with scalable ontology-based matching for agent communication [AM02]. 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.

Document Image Analysis: The proposed techniques can eliminate some common shortcomings of current table-reading and forms-processing software [LN99b].

5 Evaluation

Initially, we will test TANGO on a set of 75–100 carefully selected documents. With the limited tools available at the beginning, we cannot expect to be able to demonstrate self-organization on any but the simplest data. Once TANGO begins to perform with some reliability, we will collect a set of 100 "greenhouse" documents of graded difficulty, and a set of 100 documents subject only to the constraint that they contain semi-formatted geographic data. Using these 200 documents, we will elaborate the "reasoning" ability of the integration tools, and expand our data frames and keyword lexicons. We will test the system on both the controlled and the uncontrolled data, initializing it with both the kernel ontologies and subsets of the ontologies that incorporate the data from the initial document data set. During the experiments, will monitor the dependence of the expansion of the ontologies on the order of presentation of the test documents, as well as the amount of human intervention. Since experimentation on the same data set leads to statistically unreliable conclusions, when the system is deemed ready, we will "freeze" it, collect another 100 documents of graded difficulty and another 100 "free-style" documents, and conduct a bona fide test.

We will implement TANGO so that it can be run across the full spectrum of human intervention—from fully automatic, where it will do its best even when encountering ambiguous and contradictory information, to fully user driven, where it will do nothing more than build ontologies as directed by its users. Between these extremes, we will allow for synergistic Issue/Default/Suggestion (IDS) usage [BE03], where TANGO will do all it can to resolve difficulties, but will point out issues it encounters, state what its default action will be, and suggest possible alternatives a user may choose instead. We will also instrument TANGO with a monitoring system that will log both system and user actions.

The basic measure we intend to investigate is cost reduction. Can TANGO reduce the cost of creating a geopolitical ontology based on a kernel ontology and the information provided in the given set of tables? TANGO can reduce the cost if, either on its own or with the help of a user, it can create an ontology faster than a human can create the ontology. Since ontology creation is a complex task and one, in which, even different human experts may produce different results given the same information, we must provide a way to determine whether an ontology is satisfactory.

	Time		
	$TestSet_1$	$TestSet_2$	...
Human Built	...	...	
Synergistic	...	...	
Automatic+User	...	...	

Table 1: Ontology Build Times.

We say that an ontology is *satisfactory* with respect to a given set of documents D if it (1) is complete—contains an object set for every identifiable object contained within a table of D, (2) is consistent—the predicate-calculus formulas generated from the ontology's structure [EKW92] are not contradictory, and (3) has a valid interpretation—the conjunction of the closed predicate-calculus formulas over the collected data evaluates to *true*. Note that to be satisfactory, the view over the data need not be the same, although the views should be equivalent in the sense that they are complete, consistent, and have valid interpretations.

Table 1 shows the framework for the result data. *Human Built* means that TANGO will be run fully driven by a user. *Synergistic* means that TANGO will be run interactively under the guidance of IDS statements. *Automatic+User* means that TANGO will be run fully on its own; then if it does not produce a satisfactory ontology, a user will step in and complete the ontology in a fully user-driven mode of operation. Our research will be successful if we can speed up the ontology-building process, and will be highly successful if we can significantly ($p \leq .05$) speed up the process on a wide-ranging set of documents and web pages.

6 Related Efforts

Two areas within the purview of philosophy and linguistics have a direct bearing on the issues at hand: epistemology and semiotics. A third area, document image analysis and table understanding, completes the triangle of background knowledge to achieve the results proposed in our project. [Sow00] recently speculated on possible relationships among these areas in discussing ontologies, semiotic primitives, and metadata, particularly with respect to the formatting and presentation of text. We agree, and follow up with further details in the rest of this section.

6.1 Epistemology

Epistemology includes the study of knowledge and its foundations, organization, and formalization [Aud98]. One significant focus of attention in this field is the motivation for, the specification of, and the creation of ontologies. In some respects user specification of an ontologically based target schema is an epistemological effort, though our work focuses on the empirical applications, leaving detailed philosophical investigations for others.

The construction of ontologies [Gua98] has systematized the effort of specifying principled relationships between source and target schemas or templates [Gua99, Gua00]. Ontology building can be partially automated to some extent, by leveraging such techniques as hyperlink traversal [KRRT01] or traditional text mining [MS00]. As ever more ontologies emerge from disparate research efforts, the need for merging [MFRW00] and comparing (or aligning) [BB01] ontologies has been apparent. Despite these technological developments, the hand-crafting of ontological knowledge sources by domain experts (who are nonetheless not typically knowledge representation experts) is still widely practiced. Ontology development tools therefore fill a crucial role (and likely

will for some time) in helping humans manage and organize these (often very subtle) conceptual models [GW00, WG01, Kim02]. Such tools use standard representations from such disparate fields of endeavor as software engineering [EW01] and philosophy [GW02].

Ontology building through merging is similar to schema integration [BLN86]. Early work on schema integration [NG82] questions the possibility of efficiently and accurately integrating schemas using only structure and constraints. Low-level techniques such as computing attribute equivalence [LNE89] or value equivalence [SG89] help, but are not enough. Consequently, work turned to integrating schemas via mappings of conceptual models rather than brute-force attribute matching, and the use of conceptual models was deemed preferable [SL90]. One approach converts the various schemas to a common standardized conceptual representation [GSSC95], whereas another allows users to write assertions describing correspondences that serve to resolve structural conflicts [SP94]. More recently, research efforts have proposed several techniques and have built a few tools to automate the attribute matching problem [CAFP98, CDSS98, Coh99, DDH01, MBR01, MHH00, MZ98, SH01]. Included among these is our own work [EJX01, XE02b], which places us in a position to enter the much larger arena of automated ontology building.

Constraint discovery is also important to our work. [DP95] and [dSMH01] suggest the use of data mining to help with constraint discovery. Other work has focused on discovering and extracting schemas from semistructured data, including the discovery and positing of inter-conceptual relationships such as typed hierarchies [NAM97b, NAM97a] or recurrent subpatterns [WL97].

6.2 Semiotics

Semiotics is the study of signs and symbol systems and the meaning behind them [Eco79, Cha01]; by extension, this can include the metaphorical description of anything nonlinguistic as being language-based. Spatial layout of material in texts (e.g. diagrams, graphs, and tables) and hypertext (e.g. online search and information extraction results), even if not primarily linguistic in form, may have significant semiotic value [DeM80, FD92, Car00, CL00]. [Lem98] argues that for "visual semiotics" tables are "organizational resources to enable meaningful relations to be recovered from bare thematic items in the absence of grammatical constructions," and further argues that there is always "an implied grammar, and a recoverable textual sentence or paragraph for every table." Indeed, lists and tables are the prevalent form of presenting and communicating structured data in books, technical papers, and web pages, and their construction, representation, and understanding have been thoroughly explored [Wan96].

Electronic Tables. Various approaches have been implemented for low-level recognition of raw ASCII tables in electronic text. [HKLW00, HKLW01] uses hierarchical clustering to identify columns, as well as spatial and lexical criteria to categorize headers. TINTIN [PC97] locates structural clues (primarily aligned whitespace) indicating the presence of a table in text; it has been used to analyze a homogeneous collection of over 6000 ASCII tables from the *Wall Street Journal*. Other systems leverage the conventions used for specifying tables via SGML and HTML [LN99a, ETL02].

Hardcopy Tables. The conversion of scanned hardcopy tables to a searchable or editable digital form requires combining OCR with diagram image processing techniques. It is a major focus of the Document Image Analysis (DIA) community, as surveyed in [Han99, LN99b, LN00, Nag00b]. Unmarked tables are usually located by structural clues [PCA97, HKLW00, HKLW01], sometimes coupled with linguistic clues [Han01]. Columns, rows and cells in scanned tables can be identified by recognizing delimiters (rules or white spaces) [GK95a, GK95b], by "box-driven reasoning" [HD95], by graph analysis for fully-boxed tables [TBB96], by X-Y trees [AT98], by structure-description trees (which also help group tables by format, structure, and content) [WQS95], or by profile

analysis [Zuy97]. A detailed analysis of multi-line cell identification is in [Han01]. Ontological relations were used in [TE02]. Both model-driven and data-driven methods have been applied to the analysis of uncoded tables. Formalisms include modular interactive agents [RS97], cohesion domain templates, [HD97], and bottom-up word-aggregation [Kie98]. Linguistic clues are exploited in [Hur01]. [PCA97] used a PERL script to recover the reading order from data, vertical/horizontal indices, title, and footnotes extracted from a heterogeneous corpus of tables. We documented the difficulties of evaluating information extraction from tables in [KNNR95], [HKL+01], and [LN02].

Table-Analysis Ontology. From the cited publications and from many others, and from our own research experience (four graduate students have already completed theses on table analysis at BYU(2) and RPI(2)), we conclude that the key elements of a table ontology for extracting the content and relationships of cells in a principled manner are: (a) item linguistic and geometric characteristics that distinguish tables from text; (b) title/label/caption/footnote characteristics; (c) frame (box) and ruling properties (topology and line type); (d) item horizontal and vertical segmentation rules (alignment and spacing); (e) item typesetting and linguistic rules for cell similarity (color, typeface and size, case, normal/bold/italic, alpha/digit, indentation, punctuation, leaders, lexical and grammatical categories); and (f) indicators of intra/extra document references (superscript, asterisk, dagger).

7 Research Plan

The principal investigators have collaborated (in pairs) 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 and as permitted by other commitments.

The students will, however, be new to the project and require appropriate mentoring.[17] 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 each graduate student 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 kernel application ontology and the ontology integration 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, two BYU undergraduate students will undertake some of the projects described in Section 4. 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. The results and all of the raw web pages used in the test will be made available to other researchers through our web sites.

[17] The principle investigators hope to maintain their successful recent record of attracting women (BYU: 4, RPI: 3) and minorities (RPI: 1) to their research programs.

References

[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).

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

[Aud98] R. Audi. *Epistemology: A Contemporary Introduction to the Theory of Knowledge*, volume 2 of *Contemporary Introductions to Philosophy*. Routledge, 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, PA, June 2001.

[BE03] J. Biskup and D.W. Embley. Extracting information from heterogeneous information sources using ontologically specified target views. *Information Systems*, 28(1), 2003. (to appear).

[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).

[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.

[BN97] S. Beale and S. Nirenburg. Breaking down the barriers: The mikrokosmos generator. In *Proceedings of the 4th Natural Language Processing Pacific Rim Symposium (NLPRS'97)*, pages 141–148, Phuket, Thailand, 1997.

[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.

[Cha01] D. Chandler. *Semiotics: The Basics*. Routledge, 2001.

[Cha02] T. Chartrand. Ontology-based extraction of RDF data from the world wide web. Technical report, Brigham Young University, Provo, Utah, 2002. (thesis proposal, currently at www.deg.byu.edu/proposals/index.html).

[Che02] X. Chen. Query rewriting for extracting data behind html forms. Technical report, Brigham Young University, Provo, Utah, 2002. (thesis proposal, currently at www.deg.byu.edu/proposals/index.html).

[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.

[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.

[Din02] Y. Ding. Semiautomatic generation of data-extraction ontologies. Technical report, Brigham Young University, Provo, Utah, 2002. (thesis proposal, currently at www.deg.byu.edu/proposals/index.html).

[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.

[Eco79] U. Eco. *A Theory of Semiotics*. Indiana University Press, 1979.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

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

[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.

[Gua98] 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.

[GW02] N. Guarino and C. Welty. Evaluating ontological decisions with OntoClean. *Communications of the ACM*, 45(2):61–65, February 2002.

[Han99] J.C. Handley. Chapter 8: 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.

[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.

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

[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*, 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.

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

[Hur01] M. Hurst. Layout and language: Exploring text block discovery in tables using linguistic resources. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR'01)*, pages 523–527, Seattle, Washington, September 2001.

[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.

[Kim02] H.M. Kim. Predicting how ontologies for the semantic Web will evolve. *Communications of the ACM*, 45(2):48–54, February 2002.

[KK02] C. Knoblock and S. Kambhampati. Information integration on the web. In *Tutorial MA1: 8th National Conference on Artificial Intelligence*, Edmonton, Alberta, July 2002.

[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.

[KRRT01] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. On semi-automated web taxonomy construction. In *Proceedings of the Fourth International Workshop on the Web and Databases (WebDB 2001)*, Santa Barbara, California, May 2001.

[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.

[LDEY02] S.W. Liddle, D.T. Scott D.W. Embley, and S.H. Yau. Extracting data behind web forms. In *Proceedings of the Joint Workshop on Conceptual Modeling Approaches for E-business: A Web Service Perspective (eCOMO 2002)*, pages 38–49, Tampere, Finland, October 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.

[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.

[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.

[LYE01] S.W. Liddle, S.H. Yau, and D.W. Embley. On the automatic extraction of data from the hidden web. In *Proceedings of the International Workshop on Data Semantics in Web Information Systems (DASWIS-2001)*, pages 106–119, Yokohama, Japan, November 2001.

[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.

[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*, Breckenridge, Colorado, 2000.

[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.

[Mik] Mikrokosmos ontology web site. crl.nmsu.edu/Research/Projects/mikro/htmls/ontology-htmls/onto.index.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.

[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.

[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.

[NAM97a] S. Nestorov, S. Abiteboul, and R. Motwani. *Extracting Schema from Semistructured Data*, 1997.

[NAM97b] S. Nestorov, S. Abiteboul, and R. Motwani. Inferring structure in semistructured data. *SIGMOD Record*, 26(4), December 1997.

[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.

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

[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.

[RGM01] S. Raghavan and H. Garcia-Molina. Crawling the hidden web. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB'01)*, Rome, Italy, September 2001.

[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.

[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.

[SL90] A.P. Sheth and J.A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22(3):183–236, September 1990.

[SMMS02] L. Stojanovic, A. Maedche, B. Motik, and N. Stojanovic. User-driven ontology evolution management. In *Proceedings of the 13th European Conference on Knowledge Engineering and Management (EKAW-2002)*, Sigüenza, Spain, October 2002. Springer Verlag.

[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.

[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.

[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.

[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.

[XE02a] L. Xu and D.W. Embley. Combining the best of global-as-view and local-as-view for data integration. 2002. (submitted for publication, manuscript currently at www.deg.byu.edu/papers/index.html).

[XE02b] L. Xu and D.W. Embley. Discovering direct and indirect matches for schema elements. 2002. (submitted for publication, manuscript currently at www.deg.byu.edu/papers/index.html).

[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.