Final Report for Period: 08/2008 - 07/2009 **Submitted on:** 08/31/2009

Principal Investigator: Embley, David W. **Award ID:** 0414644

Organization: Brigham Young University

Submitted By:

Embley, David - Principal Investigator

Title:

Collaborative Research: TANGO: Table Analysis for Semiautomatic Generation of Ontologies

Project Participants

Senior Personnel

Name: Embley, David

Worked for more than 160 Hours: Yes

Contribution to Project:

Actively participating in the project as PI.

Name: Tijerino, Yuri

Worked for more than 160 Hours: No

Contribution to Project:

Dr. Tijerino has left the Computer Science Department at Brigham Young University. He has taken a position at Kwansei Gakuin University in Japan. In our revised budget submitted just before the project started, we removing him as a Co-PI. Nevertheless, he still actively colloborates with us on the project, but his involvement is currently less than 160 hours per year.

Name: Lonsdale, Deryle

Worked for more than 160 Hours: Yes

Contribution to Project:

Actively participating in the project as Co-PI.

Post-doc

Graduate Student

Name: Tao, Cui

Worked for more than 160 Hours: Yes

Contribution to Project:

Developed software to do table interpretation using sibling pages. Based on this work, she is investigating the problem of automatically generating ontologies for user-selected components of interpreted tables. Cui has received support from this award, and it is anticipated that she will again receive support in the Fall. Currently, she is doing an internship at the Mayo Clinic in Rochester, Minnesota, where she is using some of her skills learned while working on the TANGO project.

Name: Lian, Zonghui

Worked for more than 160 Hours: Yes

Contribution to Project:

Developing software to support the integration of mini-ontologies into a growing ontology (part 3 of 3, as originally proposed for TANGO). Finished his MS thesis in March, 2008. Continues to work some with the project to help others enhance and integrate his code into the larger system. Zonghui is not currently receiving financial support from this award.

Name: Al-Kamha, Reema

Worked for more than 160 Hours: Yes

Contribution to Project:

Reema worked on conceptual XML, which contributes to the interface documents we need to exchange data between the subsystems of TANGO. Reema received support from this award. She graduated with her PhD in June, 2007.

Name: Lynn, Stephen

Worked for more than 160 Hours: Yes

Contribution to Project:

Developed software to generate mini-ontologies from interpreted tables (part 2 of 3, as originally proposed for TANGO). Finished his MS thesis in April, 2008. Continues to work on the project: As a special project for school credit, he is currently organizing management aspects of the project (which has grown to become a huge problem). Stephen is not currently receiving financial support from this award.

Name: Al-Muhammed, Muhammed

Worked for more than 160 Hours: Yes

Contribution to Project:

Developed an ontology-based query system. We can use the system to query information under a TANGO-generated extraction ontology. Muhammed was partially supported by this award. He graduated with his PhD in August, 2007. Until he returned to Damascus University in Syria in October, 2007, he continued to work on the project.

Name: Ding, Yihong

Worked for more than 160 Hours: Yes

Contribution to Project:

Developed an OSM-to-OWL converter. This converter lets us transform a TANGO-generated ontology in our proprietary ontology language to a standard ontology language. Developed an OSM-to-RDF converter. This converter lets us transform the data in a populated TANGO-generated ontology to RDF so that it can be queried with SPARQL. Yihong is also developing a 2-phase ontology extractor that will let us convert a TANGO-generated extraction ontology into a layout-based extractor. Yihong is being partially supported from this award.

Name: Woodbury, Charla

Worked for more than 160 Hours: Yes

Contribution to Project:

Is developing a genealogical application to be used in connection with the TANGO project. Charla is receiving support from this award.

Undergraduate Student

Name: Hathaway, Chris

Worked for more than 160 Hours: No

Contribution to Project:

Developed software to (manually) convert ordinary tables found on the web into mini-ontologies. Did not receive funding, but was working on a senior thesis. Graduated and left to pursue a PhD elsewhere.

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Name: Peters, Jeff

Worked for more than 160 Hours: Yes

Contribution to Project:

Debugs and enhances tools used in the project. Jeff has been receiving support from this award for about a year, and now in particular, from the REU associated with this award. It is anticipated that he will again receive support in the Fall, when he returns from his summer internship at Microsoft.

Years of schooling completed: Junior

Home Institution: Same as Research Site

Home Institution if Other:

Home Institution Highest Degree Granted(in fields supported by NSF): Doctoral Degree

Fiscal year(s) REU Participant supported: 2008

REU Funding: REU supplement

Name: Zitzelberger, Andrew

Worked for more than 160 Hours: Yes

Contribution to Project:

Andrew worked on a free-form back-end query engine for querying populated ontologies created by TANGO. Andrew has also been our main trouble-shooter and build master. Andrew has been supported by this award, and in particular, by the REU for this award. He worked on the project beginning in May, 2008.

Years of schooling completed: Junior

Home Institution: Same as Research Site

Home Institution if Other:

Home Institution Highest Degree Granted(in fields supported by NSF): Doctoral Degree

Fiscal year(s) REU Participant supported: 2008 2008

REU Funding: REU supplement

Name: Clawson, Robert

Worked for more than 160 Hours: Yes

Contribution to Project:

Robert worked in recoding our mapping discovery algorithms for integration into the TANGO project. Robert also worked on putting the diverse software projects together to allow TANGO to function as envisioned. Robert was supported by this award, and in particular, by the REU for this award. He worked on the project from May, 2008 through April, 2009.

Years of schooling completed: Sophomore

Home Institution: Same as Research Site

Home Institution if Other:

Home Institution Highest Degree Granted(in fields supported by NSF): Doctoral Degree

Fiscal year(s) REU Participant supported: 2008 2008

REU Funding: No Info

Name: Watts, Robby

Worked for more than 160 Hours: Yes

Contribution to Project:

Worked on data integration for merged ontologies. This is an added component to the TANGO project, as originally specified. It is needed so that merged ontologies also have associated data -- data we can query as we evaluate TANGO. Robby also revised the ontology editor to work with a new internal format. The ontology editor is needed as part of the MapMerge part of TANGO. Robby was supported by this award, and in particular, by the REU for this award. He worked on on the project May, 2008, through August, 2008.

Years of schooling completed: Sophomore

Home Institution: Same as Research Site

Home Institution if Other:

Home Institution Highest Degree Granted(in fields supported by NSF): Doctoral Degree

Fiscal year(s) REU Participant supported: 2008

REU Funding: No Info

Name: McDaniel, Derek

Worked for more than 160 Hours: Yes

Contribution to Project:

Derek is working on putting the TANGO software we have developed together into a single working whole. Diverse pieces of software need to pass data to one another and work harmoniously together to enable TANGO to function as envisioned. Derek started working in May, 2009, and is being supported by REU funds associated with the grant.

Years of schooling completed: Sophomore

Home Institution: Same as Research Site

Home Institution if Other:

Home Institution Highest Degree Granted(in fields supported by NSF): Doctoral Degree
Fiscal year(s) REU Participant supported: 2009
REU Funding: REU supplement

Organizational Partners

Rensselaer Polytechnic Institute

Our grant is a joint, collaborative grant between BYU and RPI.
BYU Grant Number: 0414644 BYU PI: David W. Embley.
RPI Grant Number: 0414854 RPI PI: George Nagy

RPI's principal role in the TANGO project is Part 1 of the 3 major parts of the project: develop software to interpret tables. RPI developed WNT (Wang Notation Tool) and TAT (Table AbstractionA Tool). Both take HTML tables as input and produce a populated table ontology as output. Currently, we use TAT in our TANGO implementation. RPI also developed QBT (Query By Table), an interface to the data in populated ontologies.

BYU's principal role in the TANGO project is Parts 2 and 3:
-- develop software to convert interpreted tables to mini-ontologies and
-- integrate mini-ontologies with a growing ontology.
BYU developed MOGO (Mini-Ontology GeneratOr) for Part 2 and developed MapMerge for Part 3. BYU also developed OntologyWorkbench -- the framework for integrating the component parts of TANGO. The workbench includes additional tools developed during and useful in the TANGO project:
-- the OntologyEditor for developing formal conceptual models of the table ontology and related conceptualizations;
-- a converter to derive XML-Schema documents from conceptual-model diagrams;
-- a layout generator for deriving graphical layouts of generated conceptualizations of tables; and
-- a data editor for viewing and updating data stored in conceptual models.

Other Collaborators or Contacts

Yuri Tijerino -- Having left BYU, Professor Tijerino should be considered as a collaborator, rather than a Co-PI. He is currently at Kwansei Gakuin University in Japan. Dr. Tijerino acts as a consultant -- discussing the project with us whenever he visits us or we visit him -- usually about twice a year.

Daniel Lopresti -- Department of Computer Science and Engineering, Lehigh University, Bethlehem, Pennsylvania. Professor Lopresti provided some initial help as a consultant for the table interpretation phase of our project. He also developed an initial prototype tool for manual table interpretation. Several visits, some including students, between Lehigh and Rensselaer Polytechnic Institute. Joint NSF Cyber Trust grant
(http://perfect.cse.lehigh.edu/).

Stephen W. Liddle -- Department of Information Systems, Brigham Young
University. Dr. Liddle acts as the chief architect for the broader software systems we are developing at BYU. As such, he contributes generously to the software being developed for the TANGO project.

Mukkai Krishnamoorthy -- Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY. Professor Krishnamoorthy has been helping since February 2008 to find contradictory quantitative information on web pages and in Google Books. Earlier he has provided advice to the Rensselaer Polytechnic Institute graduate students on effective data structures for table representation.

C.L. Liu -- Institute of Automation, Chinese Academy of Science. Collaborate and co-publish on document conversion and search methods.

S. Seth -- University of Nebraska, Lincoln. Long term collaboration on document layout analysis, symbolic interactive classification, interactive visual pattern recognition and VLSI test generation.

S. Veeramachaneni -- Thomson-Reuters Research. Collaboration that resulted in three recent journal articles and a dozen other publications on style-constrained and surrogate classification. (Publications not listed below because they are not directly relevant to current grant. Our on-going discussions have, however, been most useful because of Dr. Veeramachaneni's experience with mining giga datasets.).

H. Fujisawa -- Since Dr. Fujisawa became Corporate Chief Scientist of Hitachi Corporation two years ago, our collaboration has taken mainly the form of extended email, review of each others publications in preparation, and face-to-face meetings once or twice per year. Nagy assisted Drs. Fujisawa and Liu in the preparation of the winning proposal to host the International Conference on Document Analysis Systems in Beijing in 2011.

Activities and Findings

Research and Education Activities:

The following outline lists the three main activities in the TANGO project, plus related activities contributing to the success of the project. For each item in the outline, we present our progress. We have completed most activities, but a few are still underway. We intend to complete these activities.

1. Development of a table-interpretation system

 a. Research survey on table processing paradigms (completed)

 b. Initial notes on table recognition (completed)

 c. An initial manual system for table interpretation (completed)

 d. A tool to do Table Interpretation with Sibling Pages (TISP) (completed)

 e. An interactive Wang Notation Tool (WNT) for web table interpretation (completed)

 f. A more advanced Table Abstraction Tool (TAT) for web table interpretation (completed)

 g. An automatic tool for table interpretation based on X-Y trees (underway)

2. Development of a system to convert interpreted tables to mini-ontologies

 a. Basic tool to manually create mini-ontologies from interpreted tables (completed)

 b. A Mini-Ontology GeneratOr (MOGO) to automatically create mini-ontologies from interpreted tables (completed)

 c. Enhancements to MOGO to accommodate more advanced features (underway)

3. Development of a system to integrate mini-ontologies with a growing ontology

 a. Basic mapping and merging tool (MapMerge) to manually integrate mini-ontologies with a growing ontology (completed)

 b. Plug-in modules for MapMerge to allow for automatic integration of mini-ontologies into a growing ontology (completed for 1-1 mapping modules)

c. More advanced plug-in modules for MapMerge to handle 1-n, n-1, and m-n mappings (underway)

4. Development of auxiliary tools.

a. An ontology-to-XML converter to generate XML-Schema interface specifications between table-interpretation work and mini-ontology generation (completed)

b. A natural-language query processor for querying information extracted with respect to populated, TANGO-generated extraction ontologies (completed)

c. A Query-By-Table tool (QBT) to query the data of a populated TANGO-generated ontology (completed)

d. A tool to convert TANGO-generated ontologies (in our proprietary OSM ontology language) to OWL ontologies (representative of standard ontology languages) (completed)

e. A tool to generate RDF data instances with respect to generated OWL ontologies (completed)

f. A data-integration tool, in addition to a model integration tool (manual version completed, automated version underway)

g. Several annotation tools based on generated ontologies (TISP completed, more underway)

h. Based on generated ontologies and automatic annotation, development of a WoK (Web of Knowledge) (initial prototype completed, additional tools underway)

5. Project evaluation

a. Evaluation of sibling-page tool (completed)

b. Evaluation of interactive Wang Notation Tool (completed)

c. Evaluation of Mini-Ontology GeneratOr (completed)

d. Evaluation of interactive Table Abstraction Tool (completed)

e. Evaluation of auxiliary tools (some done, some underway)

f. Overall evaluation of TANGO (underway)

Findings:

Findings (BYU and RPI):

We list findings for each of the activities in the activities report. Each finding relates to expected results (as originally proposed) or relates to new goals (discovered along the way as worth a detour). For each finding we give the problem addressed and the solution obtained; we also give the publication status of the finding. For those 'findings' we are currently working on, we comment briefly on our expectations. For those activities still to be done, we say very little, but we expect the outcomes to be positive.

1a. Research survey on table processing paradigms (published)

While everyone seems to know what a table is, a precise, analytical definition of 'tabularity' remains elusive. Our survey shows that recent research on table composition and table analysis has improved our understanding of the distinction between the logical and physical structures of tables, and has led to improved formalisms for modeling tables. Further, our survey argues that that progress on half-a-dozen specific research issues would open the door to several applications dependent on automated table understanding.

1b&c. Initial notes on table recognition & construction of the initial manual system for table interpretation (published)

The shift of interest to web tables in HTML and PDF files, coupled with the incorporation of table analysis and conversion routines in commercial desktop document processing software, are likely to turn table recognition into more of a systems than an algorithmic issue. We suggest that the appropriate target format for table analysis is a representation based on the abstract table introduced by X. Wang in 1996. We show that the Wang model is adequate for some useful tasks that prove elusive for less explicit representations, and outline our plans to develop a semi-automated table processing system to demonstrate this approach.

1d. Sibling-page technique for table interpretation (published)

The longstanding problem of automatic table interpretation still eludes us. We offer a solution for the common special case in which so-called sibling pages are available. Sibling pages, which are the pages commonly generated by underlying web databases, are compared to identify and connect non-varying components (category labels) and varying components (data values). We tested our solution using more than 2,000 tables in source pages from three different domains---car advertisements, molecular biology, and geopolitical information. Experimental results show that the system can successfully identify sibling tables, generate structure patterns, interpret different tables using the generated patterns, and automatically adjust the structure patterns as needed.

1e. An interactive Wang notation tool for web table interpretation (published)

The Wang Notation Tool (WNT) is a semiautomatic, interactive tool that converts tables to Wang notation -- a layout independent representation of tables where all relationships between cells are recorded without relying on the physical structure of tables. WNT requires minimal interaction to select categories from which it deduces relationships. However, if WNT is incorrect, user correction is available to generate correct Wang notation.

1f. A more advanced tool for web table abstraction (published)

Two hundred web tables from ten sites were imported into Excel. The tables were edited as needed, then converted into layout independent Wang Notation using the recently developed Table Abstraction Tool (TAT). The output generated by TAT consists of XML files to be used for constructing narrow-domain ontologies. On an average each table required 104 seconds for editing. Augmentations like aggregates, footnotes, table titles and notes were also extracted. Every user intervention was logged and audited. The logged interactions were analyzed to determine the relative influence of factors like table size, number of categories (Wang dimension), and various types of augmentations on the processing time. The analysis suggests which aspects of interactive table processing can be automated in the near term, and how much time such automation would save.

1g. A theoretical investigation of table interpretation based on X-Y trees (published)

The extraction of the relations of nested table headers to content cells is automated with a view to constructing narrow domain ontologies of semistructured web data. A taxonomy of tessellations for displaying tabular data is developed. X-Y tessellations that can be obtained by a

divide-and-conquer method are asymptotically only an infinitesimal fraction of all partitions of a rectangle into rectangles. Admissible tessellations are the even smaller subset of all partitions that correspond to the structures of published tables and that contain only rectangles produced by successive guillotine cuts. Many of these can be processed automatically. Their structures can be conveniently represented by X-Y trees, which facilitate relating hierarchical row and column headings to content cells. A formal grammar is proposed for characterizing the X-Y trees of layout-equivalent admissible tessellations. Algorithms are presented for transforming a tessellation into an X-Y tree and hence into multidimensional, layout-independent Category Trees (Wang abstract data types).

2a. Basic tool to manually create mini-ontologies from interpreted tables

Although we thought this would be a special tool in the TANGO project, it turned out to just be our ontology editor, which we had already developed, coupled with a window displaying the form to be turned into a mini-ontology. We still use this tool, but in a different way from our original expectations. Rather than interactively create a mini-ontology, we either create a mini-ontology with the ontology editor completely by hand, or we create a mini-ontology automatically and then either accept the generated mini-ontology or use the ontology editor to make adjustments.

2b. Tool to automatically create mini-ontologies from interpreted tables (published)

Enabling a system to automatically conceptualize and annotate a human-readable table is one way to create interesting semantic web content. But exactly 'how?' is not clear. With conceptualization and annotation in mind, we investigate a semantic-enrichment procedure as a way to turn syntactically observed table layout into semantically coherent ontological concepts, relationships, and constraints. Our semantic enrichment procedure shows how to make use of auxiliary world knowledge to construct rich ontological structures and to populate these ontological structures with instance data. The system uses auxiliary knowledge (1) to recognize concepts and which data values belong to which concepts, (2) to discover relationships among concepts and which data-value combinations represent relationship instances, and (3) to discover constraints over the concepts and relationships that the data values and data-value combinations should satisfy. Experimental evaluations indicate that the automatic conceptualization and annotation processes perform well, yielding F-measures of 90% for concept recognition, 77% for relationship discovery, and 90% for constraint discovery in web tables selected from the geopolitical domain.

3a. Basic tool to manually create mini-ontologies from interpreted tables (published MS thesis)

This work addresses the problem of tool support for semi-automatic ontology mapping and merging. We have built a tool that will take a mini-ontology and a growing ontology as input and make it possible to produce manually, semi-automatically, or automatically an extended growing ontology as output. Characteristics of this tool include: (1) a graphical, interactive user interface with features that will allow users to map and merge ontologies, and (2) a framework supporting pluggable, semi-automatic, and automatic mapping and merging algorithms.

3b&c. Plug-in modules to allow for automatic integration of mini-ontologies into a growing ontology (several papers published)

In progress. Expected findings: We expect that the mapping and merging algorithms we have developed in an earlier NSF-sponsored project (0083127) will allow us to successfully merge mini-ontologies into a growing ontology.

4a. An ontology-to-XML converter to generate XML-Schema interface specifications between table-interpretation work and mini-ontology generation (several papers published)

As part of a larger project to create a conceptual-modeling language for XML, which we call C-XML, and to map to and from a C-XML model instance and an XML-Schema instance, we have implemented a tool to generate an XML-Schema instance from a C-XML model instance. We use this tool in the TANGO project to generate our interface between the RPI part of the project and the BYU part of the project. Specifically, we are able to model what we want in C-XML and let the system generate the interface for us.

4b. A natural-language query processor for querying information extracted with respect to populated, TANGO-generated extraction ontologies (several papers published)

As part of a larger project to develop ontology-based web services, we have developed a server for free-form requests. In the TANGO project, we use this free-form-request server as a convenient way to query the results obtained after constructing populated ontologies based on input

tables. We expect to use this tool as part of our evaluation of TANGO.

4c. A QBT (Query-By-Table) tool to query the data of a populated TANGO-generated ontology (published)

Querying any information system requires the knowledge of some formal language, making it inaccessible to computer-na´ve potential users. We propose a new intuitive querying mechanism where the query is a (well-formed) table. We extract the underlying logical structure of the table to retrieve values from a database. Query tables are interpreted to perform simple SELECT & JOIN operations. We demonstrate that query tables with different layouts but with the same underlying logical structure yield correct answers. This approach can be extended to form complicated conditional queries and queries involving aggregates.

4d&e. A tool to convert TANGO-generated ontologies (in our proprietary OSM ontology language) to OWL ontologies (representative of standard ontology languages) and a tool to generate RDF data instances with respect to generated OWL ontologies (part of this published, but it is mostly just code)

As part of initial work on a WoK (Web of Knowledge), we have developed tools to convert TANGO-generated populated ontologies to OWL and to RDF. Thus, we are able to generate, query, and store 'knowledge' on the web (as opposed to merely having pages on the web that contain knowledge). (Fully developing the envisioned WoK would be an ideal follow-on project to TANGO.)

4f&g&h. A tool to generate RDF data instances with respect to generated OWL ontologies & several annotation tools based on generated ontologies and automatic annotation, development of a WoK (Web of Knowledge) (several papers published)

Expectations: a reasonable tool will be produced that will allow a user to manually, semi-automatically, and automatically clean and merge data. Originally, our overall goal for TANGO was simply to generate ontologies, not necessarily populated ontologies. We discovered, along the way, however, that we really need populated ontologies in order to evaluate TANGO. Even more interesting, we realized that populated ontologies are the fundamental components of a WoK (Web of Knowledge) and that the TANGO project provides a sensible way to create a WoK. We are continuing work on the WoK project.

5. Project evaluation

We have reported the results of completed evaluations of individual components above. The rest are on our 'to do' list. We have analyzed many tables from a large representative site, Canada Statistics, and noted the obstacles encountered in the way of wholly automated interpretation. Completed survey of this large site to enumerate the number of 'facts' and dimensions, and estimated the number of tables on the site (2000). We expect positive results, but until we do the evaluations, we won't know for sure what the results will be.

Training and Development:

Project participants have developed research and teaching skills in
the following ways.

1. Research group meetings:
* discussion of research papers
* presentations of student work
* coordination of prototype development

2. Face-to-face visits with collaborators at partner institutions:
* RPI student presentations of their work to BYU PI
* BYU student presentations of their work to RPI PI
* Presentations by BYU PI to RPI students and RPI PI to BYU students
* Visit by RPI student P. Jha to BYU

3. One-on-one student presentations to DocLab visitors
(S. Seth (UNL), S. Veeramachaneni (Thomson-Reuters, W. Barrett (BYU), D. Stork (Ricoh), H. Baird, D. Lopresti, X. Huang (all at Lehigh), G. Tan (Boston U.), R. Hoch (Mannheim), M. Mukherjee (IBM), E. Olivetti (IRST Trento), C.L. Liu (CASIA, Beijing), S. Rice (U. Mississipi), H. Fujisawa (Hitachi).

4. Assignments and projects related to TANGO were introduced in Pattern Recognition and in Digital Picture Processing, both graduate courses taught by Nagy. Similarly, TANGO-related material was introduced in a database course taught by Embley.

5. Graduate student training
RPI:
* Three students, A. Joshi, S. Andra, and X. Zhang, completed PhDs in 2006-2007 with Nagy on related projects. (Drs. Joshi and Andra now work in commercial data mining, Dr. Zhang is at NLM/NIH.)

* B. Yamadala completed MS on Evaluation of Classifiers in December 2006.
* P. Jha completed MS thesis on Wang Notation Tool in May 2008 (US cit, F).
* A. Miller completed MS in Comm'ns in December 2008, co-supervised by Nagy (US cit, F).
* J. Kline completed MS in CS in July 2009, co-supervised by Nagy (US cit, F).
* R. Padmanabhan completed MS thesis on Table Abstraction TOOL (TAT) in May 2009 and is currently enrolled in the RPI ECSE PhD program.
* R. Jandhyala is working on MS thesis on table analysis (expected May 2010).
* W. Silversmith, URA, graduated in May 2009, and is now working on tables in DocLab. He will enter graduate school next year (US cit).
* M. Muthathlau, URA, graduated May in 2009 and started in RPI MS program (US cit).
* Four other MS students currently work in DocLab on other document analysis projects.
* During the project period, Nagy served on 11 PhD committees (three abroad).

Four students (Murphy, Mutathlau, Clifford, Silversmith) participated in REU.

BYU:
* Three students, Reema Al-Kamha, Muhammed Al-Muhammed, and Cui Tao completed PhDs. (Drs. Al-Kamha and Al-Muhammed now have professorial positions at the University of Damascus. Dr. Tao has a research position with the Mayo Clinic.)
* Two students, Zonghui Lian and Stephen Lynn, completed MS degrees.
* Three additional PhD students and two additional MS student are continuing their work on TANGO-related projects.
* Four undergraduates participated in the REU in the last year of the project.

6. Conducting research and data analysis and preparing publications: see publications section of this report for a list of publications and authors.

Outreach Activities:

1. Colloquium talks:
* 'Concepts, Ontologies and Project TANGO' by Deryle Lonsdale, at
  BYU, October 2005.
* 'Semantic Understanding: An Approach Based on Information
  Extraction Ontologies' at the Technical University of Vienna,,
  by David W. Embley, October 2005.
* 'WoK: A Web of Knowledge' by David W. Embley at
  Kwansei Gakuin University, Kobe-Sanda, Japan, January 2008
  Earth Environmental Research Labs, Kyoto, Japan, January 2008
  University of Innsbruck, Austria, April 2008
  Dagstuhl, Germany, April 2008
* 'Interactive Visual Pattern Recognition,' George Nagy, at
  CASIA, Beijing, October 2005
  Tsinghua University, Beijing, October 2005
  Queens University, Kingston, Canada, December 2005
  University of Mississippi, Oxford, March 2006
  UniversitÚ de MontrÚal, June 2006
  University of Salerno, Italy, June 2006
  Google, Sunnyvale, January 2007

2. Other presentations (related to TANGO):

Student presentations at international conferences and workshops:
* at the 5th International Semantic Web Conference, 'Toward Making Online Biological Data Machine Understandable' (poster) by Cui Tao.
* at the Biotechnology and Bioinformatics Symposium, 'HTML Table Interpretation by Sibling Page Comparison in the Molecular Biology Domain' by Cui Tao, October 2006.
* at the ICDE PhD Workshop, 'Using Data-Extraction Ontologies to Foster Automating Semantic Annotation' by Yihong Ding, April 2006.
* at the Fourth International Workshop on Semantic Web for Services and Processes, 'Bringing Web Principles to Services: Ontology-Based Web Services' by Muhammed Al-Muhammed (with Yuri Tijerino), July 2008.
* at MKM 2009, 'From Tesselations to Table Interpretation' by Ramana C. Jandhyala, July 2009

Student presentations at the annual BYU College of Physical and Mathematical Sciences Spring Research Conference:
* Table Structure Understanding by Sibling Page Comparison' by Cui Tao, March 2006.
* Semi-Automatic Generation of Mini-Ontologies from Canonicalized Relational Tables' by Chris Hathaway, March 2006.
* A Tool to Support Onotlogy Creation Based on Incremental Mini-Ontology Merging by Zonghui Lian, March 2006.
* Ontology Generation Based on a User-Specified Ontology Seed by Cui Tao, March 2007.
* A Tool to support Ontology Creation Based on Incremental Mini-Ontology Merging' by Zonghui Lian, March 2007.
* Semi-Automatic Semantic Annotation for Hidden-Web Tables, March 2008.
* Semi-Supervised, Knowledge-Based Information Extraction for the Semantic Web by Thomas Packer, March 2009.

Weekly student presentations during Summer 2009 to Rensselaer Center for Open SOftware.

PhD dissertation defense by Reema Al-Kamha: Conceptual XML for Systems Analysis, June 2007.

MS thesis defense by Zonghui Lian: A Tool to Support Ontology Creation Based on Incremental Mini-Ontology Merging, March 2008.

MS thesis defense by Stephen Lynn: Automating Mini-ontology Generation from Canonical Tables, April 2008.

PhD dissertation defense by Cui Tao: Ontology Generation, Information Harvesting and Semantic Annotation for Machine-Generated Web Pages, December 2008.

MS thesis proposal: A Tool to Support Ontology Creation Based on Incremental Mini-Ontology Merging by Zonghui Lian, October 2006.

TANGO-related term paper: Table Extraction Using MaxEnt by Zonghui Lian, December 2006.

PhD dissertation proposal: Toward Making Online Biological Data machine Understandable by Cui Tao, June 2005.

Opening talk by G. Nagy at Digital Libraries Conference in Lyon, France, 2006.

Workshop presentation by G. Nagy, Notes on Contemporary Table Recognition, Wellington, NZ 2006.

Workshop presentation by G. Nagy, Sequence Clustering, U. Conn, 2007.

Invited talk by D.W. Embley, Foundational Data Modeling and Schema Transformations, Klagenfurt, Austria, April 2008.

Workshop presentation by D.W. Embley, Seed-based Generation of Personalized Bio-Ontologies for Information Extraction, Auckland, New Zealand, November 2008.

Conference presentation by D.W. Embley, Table Interpretation by Sibling Page Comparison, Auckland, New Zealand, November 2008.

Conference presentation by G. Nagy, Camera Based Ballot Reader, at Document Recognition and Retrieval in San Jose, CA, January 2009. Conference followed as usual by a dinner of ~20 DocLab alumni and guests.

Conference presentation by D.W. Embley, Semantically Conceptualizing and Annotating Tables, Bangkok, Thailand, February 2009.

Conference presentations by G. Nagy, Interactive conversion of Large Web Tables, La Rochelle, France, and on other document image analysis topics at CBDAR and ICDAR in Barcelona, Spain in July 2009.

Sessions chaired G. Nagy at GREC and ICDAR in Barcelona, Spain, July 2009.

3.International Outreach

Nagy was a guest of the Institute of Automation, Chinese Academy of Science (CASIA), Beijing, in October 2006. Among the topics of organized discussions with professors and graduate students at CASIA was the prospective convergence of digital libraries and the semantic web. He also gave a lecture and met document researchers at Tsinghua University.

In Summer 2007, Emanuele Olivetti, a full-time researcher and PhD student, spent two months at RPI DocLab to familiarize himself with our projects.

Nagy visited the Istituto per la Ricerca Scientifica et Technologica (IRST) in Trento on several occasions to as part of a long-standing collaboration on document technologies. In March 2008 he participated in the doctoral defense of Emanuele Olivetti.

In June 2008, Nagy presented three lectures and supervised exercises at the Summer School on Digital image Processing offered by the Indian Institute of Science, Bangalore, as part of its Centennial Celebration. He also gave a lecture on table analysis at First Indian Corporation Private Limited (FIC), Bangalore, a leading web transaction processor.

Nagy has responded by personalized email to about 200 foreign student applicants. He carries on email correspondence with half-a-dozen foreign students on their specific research, including a doctoral dissertation on table analysis in Australia.

Embley presented a colloquium talk at the General Earth Environmental Research Laboratories, Research Institute for Humanity and Nature located in Kyoto, Japan.

Embley participated as a panelist at ER08, Barcelona, Spain, October 2008.

Nagy participated on the doctoral jury of Joel Gardes in Grenoble, France, July 2009.

Students Jandhyala and Siversmith made presentations to DocLab visitor Dr. Harsha Veeramachaneni of Thomson-Reuters in August 2009.

Yves Saint-Pierre, Manager of University Relations for Canada Statistics, visited DocLab and gave Nagy, Krishnamoorthy and three TANGO students a two-hour on-line demo in August 2009.

4. High School

Nagy has judged projects presented by students Questar III Rensselaer Education Center where students take both introductory RPI courses and regular high-school courses on Campus. He also reviews blogs posted by Questar students on their studies in US history and economics. One student who did a fine arts oriented project with him in Grade 12 took two of his courses later and just graduated from RPI. Another student has been working with Nagy since he graduated from high school in summer 2008 and will continue in the Fall 2009.

5. Professional societies:

Embley, Lonsdale and Nagy participated extensively in refereeing proposed books and manuscripts for technical journals.

We reviewed several proposals submitted to US and foreign government agencies.

Nagy participates in the Steering Committee to review applications for hosting the next DIAL (Digital Image Analysis for Libraries) workshop. He also reviewed an application for hosting ICDAR 2011.

Embley actively participates as a steering committee member for the international conferences on conceptual modeling.

Journal Publications

Yuri A. Tijerino, David W. Embley, Deryle W. Lonsdale, and George Nagy, "Towards Ontology Generation from Tables", World Wide Web: Internet and Web Information Systems, p. 261, vol. 8, (2005). Published,

David W. Embley, Matthew Hurst, Daniel Lopresti, and George Nagy, "Table Processing Paradigms: A Research Survey", International Journal on Document Analysis and Recognition, p. 66, vol. 8, (2006). Published,

L. Xu and D.W. Embley, "A composite approach to automating direct and indirect schema mappings", Information Systems, p. 697, vol. 31, (2006). Published,

Deryle W. Lonsdale, David W. Embley, Yihong Ding, Li Xu, and Martin Hepp, "Reusing Ontologies and Language Components for Ontology Generation", Data & Knowledge Engineering, p. , vol. , (2008). Accepted,

Cui Tao and David W. Embley, "Automatic Hidden-Web Table Interpretation, Conceptualization, and Semantic Annotation", Data & Knowledge Engineering, p. 683, vol. 68, (2008). Published,

Books or Other One-time Publications

D. Goldin, R. Mardales, G. Nagy, "In search of meaning for time series subsequence clustering: Matching algorithms based on a new distance measure", (2006). Proceedings, Published
Bibliography: pp. 347-356, Procs. Conference on Information and Knowledge Management (CIKM06), Arlington, VA, November

Reema Al-Kamha, "Conceptual XML for Systems Analysis", (2007). Thesis, Published
Bibliography: Department of Computer Science, Brigham Young University

Muhammed Al-Muhammed, David W. Embley, Stephen W. Liddle, and Yuri Tijerino, "Bringing Web Principles to Services: Ontology-Based Web Services", (2007). Proceedings, Published
Bibliography: Proceedings of the Fourth International Workshop on Semantic Web for Services and Processes, (2007). Salt Lake City, Utah, 9 July, 73?80

Muhammed J. Al-Muhammed, "Ontology Aware Software Service Agents: Meeting Ordinary User Needs on the Semantic Web", (2007). Thesis, Published
Bibliography: Brigham Young University, July

Zonghui Lian, "A Tool to Support Ontology Creation Based on Incremental Mini-Ontology Merging", (2008). Thesis, Published
Bibliography: Brigham Young University, March

Stephen Lynn, "Automating Mini-Ontology Generation from Canonical Tables", (2008). Thesis, Published
Bibliography: Brigham Young University, April

G. Nagy and S. Veeramachaneni, "Adaptive and Interactive Approaches to Document Analysis", (2008). Book Chapter, Published
Collection: Machine Learning in Document Analysis and Recognition
Bibliography: (S. Marinai, H. Fujisawa, editors), Springer, Studies in Computational Intelligence,
Vol. 90, ISBN 978-3-540-76279-9, pp. 221-257, April

G. Nagy, "Digitizing, coding, annotating, disseminating, and preserving documents", (2006). Proceedings, Published
Bibliography: Procs. IWRIDL-2006 workshop on Digital Libraries, Kolkota, India, ACM 1-59593-608-4, April

Stephen Lynn and David W. Embley, "Automatic Generation of Ontologies from Canonicalized Web Tables", (2008). Technical Report, submitted; to be improved and submitted again
Bibliography: Technical Report, Brigham Young University

D.W. Embley, S.W. Liddle, D.W. Lonsdale, A. Stewart, and C. Tao, "KBB: A Knowledge-Bundle Builder for Research Studies", (2009). Proceedings, Accepted
Bibliography: Proceedings of the 2nd International Workshop on Active Conceptual Modeling of Learning, Gramado, Brazil, November, 2009.

D.W. Embley, S.W. Liddle, and C. Tao, "Conceptual Modeling for a Web of Knowledge", (2009). Invited Papers from a Dagstuhl Seminar, Submitted
Editor(s): R. Kaschek, L. Delcambre, and H. Mayr
Bibliography: Lecture Notes in Computer Science

C. Tao, D.W. Embley, and S.W. Liddle, "FOCIH: Form-based Ontology Creation and Information Harvesting", (2009). Conference Proceedings, Accepted
Bibliography: The 28th International Conference on Conceptual Modeling (ER'09), Gramado, Brazil, 9--12 November

S. Lynn and D.W. Embley, "Semantically Conceptualizing and Annotating Tables", (2009). Conference Proceedings, Published
Bibliography: Proceedings of the Third Asian Semantic Web Conference (ASWC 2008), Bangkok, Thailand, 2-5 February, 345--359.

C. Tao, "Ontology Generation, Information Harvesting and Semantic Annotation for Machine-Generated Web Pages", (2008). Thesis, Published
Bibliography: Brigham Young University

R. Padmanabhan, R. C. Jandhyala, M. Krishnamoorthy, G. Nagy, S. Seth, W. Silversmith, "Interactive Conversion of Large Web Tables", (2009). Conference Proceedings, Published
Bibliography: Proceedings of Eighth International Workshop on Graphics Recognition, GREC 2009, Published by City University of La Rochelle, La Rochelle, France, July 22-23

Ramana C. Jandhyala, Mukkai Krishnamoorthy, George Nagy, Raghav Padmanabhan, Shared Seth, and William Silversmith, "From Tessellations to Table Interpretation", (2009). Conference Proceedings, Published
Bibliography: Proceedings of the 8th International Conference on Mathematical Knowledge Management, MKM 2009, Grand Bend, Ontario (CANADA), 10-12 July 2009, Springer-Verlag Lecture Notes in Art

D.W. Embley and A. Zitzelberger, "Theoretical Foundations for Enabling a Web of Knowledge", (2009). Technical Report, Submitted
Bibliography: Technical Report, Brigham Young University

Web/Internet Site

URL(s):
tango.byu.edu
Description:
This is the homepage of the NSF TANGO project. The site includes a list of participants, papers and theses published, tools available for download, presentations, and proposals. The site includes a reference to NSF as well as the grant numbers.

Other Specific Products

Product Type:
Software (or netware)
Product Description:
Wang Notation Tool, JAVA/MATLAB software for interactive HTML table entry.
Sharing Information:
Available on demand. Will be posted on the TANGO website after thorough verification.

Product Type:
Software (or netware)
Product Description:
Query by Table, Excel/MATLAB software for querying a database (in progress)
Sharing Information:
Will be posted on the TANGO web site after thorough verification.

Product Type:
Software (or netware)
Product Description:
MOGO: generates mini-ontologies, given canonicalized tables as input (in progress)
Sharing Information:
Will be posted on the TANGO website after thorough verification.

Product Type:
Software (or netware)
Product Description:
MapMerge: semi-automatically integrates ontologies (in progress)
Sharing Information:
Will be posted on the TANGO website after thorough verification.

Product Type:
Software (or netware)
Product Description:
TANGO: end-to-end ontology generator (in progress)
Sharing Information:
Will be posted on the TANGO website after thorough verification.

Product Type:
Software (or netware)
Product Description:
WoK: Web of Knowledge (in progress)
Sharing Information:
Will be posted on the TANGO website after thorough verification.

Product Type:
Software (or netware)
Product Description:
Table Abstraction Tool (TAT) embedded Excel VBA software for interactive table interpretation
Sharing Information:
Posted on the TANGO website

Contributions

Contributions within Discipline:

We have written surveys of the current state-of-the art of table processing paradigms, on table processing methods, and on the evaluation of automated table and graphics processing methods. We served on conference committees and on external doctoral committees on related topics.

We have completed the third version of an interactive software tool for converting HTML tables to layout-independent Wang Notation and tagged XML format suitable for the extraction of mini-ontologies, and tested it on 200 large Web tables. The Table Analysis Tool (TAT) is embedded in Excel and is freely available to other researchers. In addition to basic table interpretation, TAT includes facilities for capturing footnotes, units, aggregates, titles, captions and notes.

We have reported aspects of tables designed for human understanding to determine what aspects are difficult to implement algorithmically. We have incorporated some formal rules and some heuristics related to Well-Formed Tables and Virtual Headings into Query-by-Table and TAT.

We have shown that table interpretation by sibling page comparison is possible. Further, we have tested our solution using more than 2,000 source tables from three different domains obtaining near 100% accuracy.

We have built a tool called MOGO (Mini-Ontology GeneratOr) to generate mini-ontologies from canonicalized table input. Preliminary tests show that the tool performs reasonably well: 90% accuracy for concept recognition, 77% accuracy for relationship discovery, and 90% for constraint discovery for a set of web tables selected from the geopolitical domain.

We have built a tool called MapMerge that allows users to integrate ontologies. The tool also supports plug-ins for automatic mapping and merging algorithms. We have added several plug-ins so that the integration can be run automatically or semi-automatically, allowing for user interaction to resolve issues the automatic version cannot resolve by itself.

In addition to the basic tools (TAT, MOGO, and MapMerge) We have built several significant auxiliary software tools in connection with the TANGO project: (1) OntologyEditor: the basic ontology editor that gives TANGO users the opportunity to modify mini-ontologies generated from tables, ontologies to be merged, and the growing TANGO ontology; (2) C-XML-to-XML-Schema Converter: to generate XML-Schema instances from C-XML conceptual model instances, as an aid to converting output from our TANGO table interpreter to our TANGO mini-ontology generator; (3) OntologyWorkbench: an environment in which TANGO processes can be embedded to run end-to-end; (4) OSM-to-OWL converter: to render ontologies in OWL and RDF so that they can be queried; (5) AskOntos: a free-form query processor to process queries against the data in a growing TANGO ontology; (6) QBT (Query By Table): to query the data by a table interface.

Contributions to Other Disciplines:

Contributions to Human Resource Development:

Since the award was granted, ten students working on the project have graduated: three PhD and two MS students at BYU, and two MS and three BS students at RPI.

Students and recent graduates are coauthors of 22 published papers, 2 accepted paper, and 1 submitted paper.

Since the award was granted, students have given five major presentations in forums with the public invited (colloquium talks or conference/workshop presentations) and have given nine minor presentations (local workshops).

Four students in the TANGO research group have been female; three have graduated and one is still an MS student. None of the other students working on the project is in an under-represented or minority group. (In 2008-2009 two other US females obtained MS degrees while research assistants on other NSF projects directed by Nagy.)

Contributions to Resources for Research and Education:

We have undertaken and partially completed a census of a very large web collection of structured data, Canada Statistics. When completed, this will be a useful resource for research in a Web of Knowledge, specifically retrieving facts by combining information from several tables meant for human readers. We created a complete XML interpretation of 200 web tables that is freely available as ground truth for other research groups.

BYU has provided more than the $15,000 it promised as cost sharing for the project. We have used this money to purchase new computers and associated equipment.

The RPI Center for Open Software provided $8000 for undergraduate research assistants.

Contributions Beyond Science and Engineering:

Conference Proceedings

Al-Muhammed, M;Embley, DW;Liddle, SW, Conceptual model based semantic web services, "OCT 24-28, 2005", CONCEPTUAL MODELING - ER 2005, 3716: 288-303 2005

Tao, C;Embley, DW, Automatic hidden-web table interpretation, conceptualization, and semantic annotation, "NOV 05-09, 2007", DATA & KNOWLEDGE ENGINEERING, 68 (7): 683-703 Sp. Iss. SI JUL 2009

Embley, DW;Liddle, SW;Lonsdale, D;Nagy, G;Tijerino, Y;Clawson, R;Crabtree, J;Ding, YH;Jha, P;Lian, ZH;Lynn, S;Padmanabhan, RK;Peters, J;Tao, C;Watts, R;Woodbury, C;Zitzelberger, A, A Conceptual-Model-Based Computational Alembic for a Web of Knowledge, "OCT 20-23, 2008", CONCEPTUAL MODELING - ER 2008, PROCEEDINGS, 5231: 532-533 2008

Al-Kamha, R;Embley, DW;Liddle, SW, Foundational data modeling and schema transformations for XML data engineering, "APR 22-25, 2008", INFORMATION SYSTEMS AND E-BUSINESS TECHNOLOGIES, 5: 25-36 2008

Tao, C;Embley, DW, Seed-based generation of personalized bio-ontologies for information extraction, "NOV 05-09, 2007", ADVANCES IN CONCEPTUAL MODELING - FOUNDATIONS AND APPLICATIONS, 4802: 74-84 2007

Tao, C;Embley, DW, Automatic hidden-web table interpretation by sibling page comparison, "NOV 05-09, 2007", CONCEPTUAL MODELING - ER 2007, PROCEEDINGS, 4801: 566-581 2007

Al-Muhammed, MJ;Embley, DW, Ontology-based constraint recognition for free-form service requests, "APR 11-15, 2007", 2007 IEEE 23RD INTERNATIONAL CONFERENCE ON DATA ENGINEERING, VOLS 1-3, : 341-350 2007

Embley, DW;Hurst, M;Lopresti, D;Nagy, G, Table-processing paradigms: a research survey, "AUG 02, 2003", INTERNATIONAL JOURNAL ON DOCUMENT ANALYSIS AND RECOGNITION, 8 (2-3): 66-86 JUN 2006

Ding, YH;Embley, DW;Liddle, SW, Automatic creation and simplified querying of semantic web content: An approach based on information-extract ion ontologies, "SEP 03-07, 2006", SEMANTIC WEB - ASWC 2006, PROCEEDINGS, 4185: 400-414 2006

Embley, DW;Lopresti, D;Nagy, G, Notes on contemporary table recognition, "FEB 13-15, 2006", DOCUMENT ANALYSIS SYSTEMS VII, PROCEEDINGS, 3872: 164-175 2006

Tijerino, YA;Embley, DW;Lonsdale, DW;Ding, YH;Nagy, G, Towards ontology generation from tables, "DEC 10-12, 2003", WORLD WIDE WEB-INTERNET AND WEB INFORMATION SYSTEMS, 8 (3): 261-285 SEP 2005

Tijerino, YA;Lonsdale, DW;Embley, DW;Nagy, G, Ontology generation from tables, "DEC 10-12, 2003", FOURTH INTERNATIONAL CONFERENCE ON WEB INFORMATION SYSTEMS ENGINEERING, PROCEEDINGS, : 242-249 2003

Al-Muhammed, MJ;Embley, DW, Resolving underconstrained and overconstrained systems of conjunctive constraints for service requests, "JUN 05-09, 2006", ADVANCED INFORMATION SYSTEMS ENGINEERING, PROCEEDINGS, 4001: 223-238 2006

Lopresti, D;Nagy, G;Seth, S;Zhang, XL, Multi-character field recognition for Arabic and Chinese handwriting, "SEP 27-28, 2006", ARABIC AND CHINESE HANDWRITING RECOGNITION, 4768: 218-230 2008

Nagy, G;Lopresti, D, Interactive document processing and digital libraries, "APR 27-28, 2006", Second International Conference on Document Image Analysis for Libraries, Proceedings, : 2-9 2006

Ding, YH;Lonsdale, D;Embley, DW;Hepp, M;Xu, L, Generating ontologies via language components and ontology reuse, "JUN 27-29, 2007", Natural Language Processing and Information Systems, Proceedings, 4592: 131-142 2007

Jha, P;Nagy, G, Wang Notation Tool: Layout Independent Representation of Tables, "DEC 08-11, 2008", 19TH INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, VOLS 1-6, : 1861-1864 2008

Padmanabhan, R;Nagy, G, QUERY BY TABLE, "DEC 08-11, 2008", 19TH INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION, VOLS 1-6, : 1949-1952 2008

Categories for which nothing is reported:

Contributions: To Any Other Disciplines

Contributions: To Any Beyond Science and Engineering