Annual Report for Period: 08/2006 - 07/2007 **Submitted on:** 07/23/2007

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

Organization: Brigham Young University

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 and wrote a paper about 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 is receiving support from this award.

Name: Lian, Zonghui

Worked for more than 160 Hours: Yes

Contribution to Project:

Is developing software to support the integration of mini-ontologies into a growing ontology. Has received support from this award; is not currently receiving support -- working part-time elsewhere.

Name: Al-Kamha, Reema

Worked for more than 160 Hours: Yes

Contribution to Project:

Reema is working on conceptual XML. Her work contributes to the interface documents we need to exchange data between the subsystems of TANGO. Reema has received support from this award. She has graduated with her PhD.

Name: Lynn, Stephen

Worked for more than 160 Hours: Yes

Contribution to Project:

Is developing software to generate mini-ontologies from interpreted tables. Is not currently receiving support -- working part-time elsewhere.

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 the grant.

Name: ding, Yihong

Worked for more than 160 Hours: Yes

Contribution to Project:

Developed an OSM-to-OWL converter. The converter lets us transform a TANGO-generated ontology in our proprietary ontology language to a standard ontology language. Yihong is also developing a 2-phase ontology extractor that will let us convert a TANGO-generated extraction ontology into a layout-based extractor. He is also working on converting extracted pages to semantic-web pages and on a system to locate and obtain information from these semantic web pages. Yihong was partially supported by the grant.

Undergraduate Student

Name: Hathaway, Chris

Worked for more than 160 Hours: Yes

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.

Name: Peters, Jeff

Worked for more than 160 Hours: Yes

Contribution to Project:

Debugs and enhances tools used in the project. Is receiving support from this award.

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Rensselaer Polytechnic Institute

Our grant is a joint, collaborative grant between BYU and RPI.

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.

Daniel Lopresti -- Department of Computer Science and Engineering, Lehigh University, Bethlehem, Pennsylvania.

Stephen W. Liddle -- Department of Information Systems, Brigham Young University, Provo, Utah.

Activities and Findings

Research and Education Activities:

The following outline lists the three main activities in the TANGO project, plus those tangentially related activities that are contributing to the success of the project.

1. Development of a table-interpretation system

 a. Research survey on table processing paradigms

 b. Initial notes on table recognition

 c. An initial manual system for table interpretation

 d. Sibling-page technique for table interpretation

 e. An interactive Wang notation tool for web table interpretation

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

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

 b. Tool to automatically create mini-ontologies from interpreted tables

 c. Form-like specification of ontologies

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

 a. Basic tool to manually integrate mini-ontologies with a growing ontology

 b. Plug-in modules to allow for automatic integration of mini-ontologies into a growing ontology

4. Development of auxiliary tools to aid in the project

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

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

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

Findings:

We list findings and status information for each of the project activities. Each finding is in the form of an abstract that gives the problem addressed the solution obtained.

1a. Research survey on table processing paradigms

Tables are a ubiquitous form of communication. While everyone seems to know what a table is, a precise, analytical definition of ``tabularity'' remains elusive because some bureaucratic forms, multicolumn text layouts, and schematic drawings share many characteristics of tables. There are significant differences between typeset tables, electronic files designed for display of tables, and tables in symbolic form intended for information retrieval. Most past research has addressed the extraction of low-level geometric information from raster images of tables scanned from printed documents, although there is growing interest in the processing of tables in electronic form as well.

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. This review, which is structured in terms of generalized paradigms for table processing, indicates that progress on half-a-dozen specific research issues would open the door to using existing paper and electronic tables for database update, tabular browsing, structured information retrieval through graphical and audio interfaces, multimedia table editing, and platform-independent display.

1b. Initial notes on table recognition &
1c. Construction of the initial manual system for table interpretation

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 illustrate the transition by some actual examples of web table conversion. We then suggest that the appropriate target format for table analysis, whether performed by conventional customized programs or by off-the-shelf software, 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. Screen-snapshots of a prototype tool to allow table mark-up in the style of Wang are also presented.

1d. Sibling-page technique for table interpretation

The longstanding problem of automatic table interpretation still illudes us. Its solution would not only be an aid to table processing applications such as large volume table conversion and information extraction, but would also be an aid in solving related problems such as ontology learning and semi-structured data management. In this paper, 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

In progress. Expected findings: Users can generate table interpretations faster with the Wang notation tool than by hand. (This work is being done at RPI. See the RPI report for more details.)

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

In progress. Status: We have built and are continuing to improve a basic ontology editor. The ontology editor allows a user to create an ontology by hand. We have also built an initial tool to display an interpreted table side-by-side with an ontology-editor window so that a user can conveniently create a mini-ontology for the table.

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

In progress. A thesis proposal to do this conversion is well underway.

2c. Form-like specification of ontologies

In conjunction with 1e, we have begun to develop a way to do form-like specifications of ontologies. We have implemented a preliminary version for a prototype system in the biological domain with the following results.

Biologists usually focus on only a small, individualized, sub-domain of the huge domain of biology. With respect to their sub-domain, they often need data collected from various different web resources. In this research, we provide a tool with which biologists can generate a sub-domain-size, user-specific ontology that can extract data from web resources. The central idea is to let a user provide a seed, which consists of a single data instance embedded within the concepts of interest. Given a seed, the system can generate an extraction ontology, match information with the user's view based on the seed, and collect information from online repositories. Our initial experimentations indicate that our prototype system can successfully match source data with an ontology seed and gather information from different sources with respect to user-specific, personalized views.

3a. Basic tool to manually create mini-ontologies from interpreted tables &
3b. Plug-in modules to allow for automatic integration of mini-ontologies into a growing ontology

In progress. Expected findings based on a defended thesis proposal: This thesis will address the problem of tool support for semi-automatic ontology mapping and merging. Solving this problem can contribute to ontology creation and evolution by relieving users from tedious and time-consuming work. As proposed in the NSF-supported TANGO project, this work will show that a tool can be built 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.

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

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.

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.

4c. A tool to convert TANGO-generated ontologies (in our proprietary OSM ontology language) to OWL ontologies (representative of standard ontology languages).

We have implemented a tool to convert OSM ontologies to OWL ontologies. This allows us to present the results of ontology construction in a standard ontology language rather than in our proprietary ontology language. It also demonstrates the relative ease with which we can render the results of ontology creation in the TANGO project in any ontology language.

Training and Development:

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

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

2. Face-to-face visits with collaborators at RPI:
   * student presentations of their work to RPI co-PI
   * Co-PI presentations in our research group
   * visit by RPI student to BYU

3. Student presentations at international conferences and workshops:
   * 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).
   * 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.

4. Student presentations at the annual BYU College of Physical and Mathematical Sciences Spring Research Conference:
   * '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.
   * '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.

5. Colloquium talks:
   * at BYU, 'Concepts, Ontologies and Project TANGO' by Deryle Lonsdale, October 2005.
   * at the Technical University of Vienna, 'Semantic Understanding: An Appproach Based on Information Extraction Ontologies' by David W. Embley, October 2005.

6. Other presentations:
   * PhD dissertation defense: 'Conceptual XML for Systems Analysis' by Reema Al-Kamha, June 2007.
   * 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.

7. Doing research and data analysis and writing papers: see publications section of this report for a list of publications and authors.

Outreach Activities:

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,

Books or Other One-time Publications

David W. Embley, Daniel Lopresti, and George Nagy, "Notes on Contemporary Table Recognition", (2006). Proceedings, Published
Bibliography: Proceedings of the Seventh International Association for Pattern Recognition Workshop on Document Analysis Systems

Cui Tao and David W. Embley, "Automatic Hidden-Web Table Interpretation by Sibling Page Comparison", (2007). Proceedings, Accepted
Bibliography: Proceedings of the 26th International Conference on Conceptual Modeling (ER'07)

Cui Tao and David W. Embley, "Seed-based Generation of Personalized Bio-Ontologies for Information Extraction", (2007). Proceedings, Accepted
Bibliography: Proceedings of the First International Conference on Conceptual Modelling for Life Sciences Applications (CLMSA'07)

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

Yihong Ding, Deryle Lonsdale, David W. Embley, Martin Hepp, and Li Xu, "Generating Ontologies via Language Components and Ontology Reuse", (2007). Proceedings, Published
Bibliography: Proceedings of the 12th International Conference on Applications of Natural language to Information Systems (NLDB'07)

Yihong Ding, David W. Embley, and Stephen W. Liddle, "Automatic Creation and Simplified Querying of Semantic Web Content: An Approach Based on Information-Extraction Ontologies", (2006). Proceedings, Published
Bibliography: Proceedings of the 1st Asian Semantic Web Conference

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

Web/Internet Site

URL(s):
tango.byu.edu
Description:

This is the homepage of the NSF project.

Other Specific Products

Product Type:

Software (or netware)

Product Description:

* Ontology Editor: allows users to manually create ontologies.

* C-XML-to-XML-Schema Converter: generates an XML-Schema model instance from a C-XML model instance, created manually in the Ontology Editor.

* Sibling-Table-Based Table Interpretation: given a set of sibling pages, finds sibling tables and discovers label-value pairs within the tables.

Sharing Information:

Available on demand. Will be posted on the TANGO website after thorough verification.

Contributions

Contributions within Discipline:

1. We have written a survey of the current state-of-the art of table processing paradigms.

2. 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---car ads, molecular biology, and geopolitical information. Experimental results show the system (with near 100% accuracy) can identify sibling tables, generate structure patterns, interpret different tables using the generated patterns, and automatically adjust the patterns as needed while processing tables from a web site.

3. Some significant auxiliary software for the TANGO project has been completed: (1) OntologyEditor: the basic ontology editor, (2) SerFER: a server for free-form requests, (3) C-XML-to-XML-Schema converter: to generate XML-Schema instances from C-XML conceptual model instances, (4) OSM-to-OWL converter: to render ontologies for the OntologyEditor in OWL.

Contributions to Other Disciplines:

Contributions to Human Resource Development:

* Since the award was granted, one student working on the project has graduated.

* Since the award was granted, students have participated as coauthors of several published papers. In addition, students are coauthors of several papers currently in press or under review.

- Students are coauthors of 2 published papers.
- Students are coauthors of 2 accepted papers.
- Students are coauthors of 1 submitted paper.

* Since the award was granted, in forums with the public invited, students have given 4 major presentations (colloquium talks or

workshop presentations) and have given 5 minor presentations (college-sponsored workshops).

* Two students in our TANGO research group are female. Both are seeking a PhD, and one just graduated. None of the other students working on the project is in an under-represented or minority group.

Contributions to Resources for Research and Education:

BYU has provided $12,520 of the $15,000 it promised as cost sharing for the project. We have used this money to purchase new computers and associated equipment.

Contributions Beyond Science and Engineering:

Special Requirements

Special reporting requirements:

There are no changes to our basic plan for the coming year.

Change in Objectives or Scope: None

Unobligated funds: $ 0.00

Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Activities and Findings: Any Outreach Activities

Contributions: To Any Other Disciplines

Contributions: To Any Beyond Science and Engineering