Annual Report for Period: 08/2007 - 07/2008 **Submitted on:** 05/30/2008

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 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: No

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: Crabtree, Jordan

Worked for more than 160 Hours: Yes

Contribution to Project:

Is developing a way to generate and populate ontologies for applications whose metadata can be organized hierarchically and whose data can be extracted from web pages. Jordan is being supported by this award, and in particular, now by the REU for this award. He initially started working on this project last fall semester as a special project for school credit.

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: REU supplement

Name: Zitzelberger, Andrew

Worked for more than 160 Hours: No

Contribution to Project:

Is working on a free-form back-end query engine for querying populated ontologies created by TANGO. Andrew is being supported by this award, and in particular, by the REU for this award. He began working on the project at the beginning of May, and has thus not yet contributed 160 hours to the project.

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: No Info

Name: Clawson, Robert

Worked for more than 160 Hours: No

Contribution to Project:

Is working in recoding our mapping discovery algorithms for integration into the TANGO project. Robert is being supported by this award, and in particular, by the REU for this award. He began working on the project at the beginning of May, and has thus not yet contributed 160 hours to the project.

Years of schooling completed: Freshman

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: Watts, Robby

Worked for more than 160 Hours: No

Contribution to Project:

Is working on data integration for merged ontologies. This is an added component to the TANGO project, as originally specified. It's needed so that merged ontologies also have associated data -- data we can query as we evaluate TANGO. Robby is being supported by this award, and in particular, by the REU for this award. He began working on the project at the beginning of May, and has thus not yet contributed 160 hours to the project.

Years of schooling completed: Freshman

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

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 is also providing a QBT (Query By Table) interface to the data in populated ontologies.

BYU's principal role in the TANGO
project is parts 2 and 3 of the project:
-- develop software to convert interpreted tables to mini-ontologies and
-- integrate mini-ontologies with a growing ontology.
BYU is also providing data merge (only schema merge was in the original proposal), and BYU is also providing a free-form query interface to the data in the populated ontologies.

BYU and RPI will conduct the final evaluation together.

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. Currently, Dr. Tijerino is planning to take a more active role in developing a form-based interface that will allow users to specify arbitrary, but simple ontologies and annotate web pages with respect to the declared ontology.

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, Provo, Utah. 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 û 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 û Long term collaboration on document layout analysis, symbolic interactive classification, interactive visual pattern recognition and VLSI test generation.

S. Veeramachaneni û 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 û 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 those tangentially related activities that are contributing to the success of the project. As a result of working on TANGO, we have discovered some additional activities we must do in order to properly evaluate the project, and we have seen some additional, highly related, 'low-hanging' fruit we should pick because of its intellectual merit and potential impact.

For each item in the outline, we present our progress. Overall, we are behind schedule. (NB: Nagy could not engage a graduate student until May 2006 because the project approval did not filter through RPI till October 2005, and he was on sabbatical leave that year.) It will likely take us until the end of the summer to integrate the software we have developed into the TANGO system we proposed. This integration is prerequisite to an overall system evaluation. Doing the evaluation properly will take much longer. (We expected this software integration project to be a major effort, but it is even more problematic than we anticipated.)

During the current reporting period, we worked on all activities except those we completed earlier (1a,b) and those not yet started (5b,d,e,f).

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. Sibling-page technique for table interpretation (completed)

   e. An interactive Wang Notation Tool for web table interpretation (Version 3 completed)

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. Tool to automatically create mini-ontologies from interpreted tables (completed)

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 (completed)

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

4. Development of auxiliary tools to aid in the project and investigation and tool development for additional 'low-hanging'

fruit.

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 QBT (Query-By-Table) tool to query the data of a populated TANGO-generated ontology (nearing completion)

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 (underway)

g. Several annotation tools based on generated ontologies (one done, one well underway, more underway)

h. Based on generated ontologies and automatic annotation, development of a WoK (Web of Knowledge) (underway)

5. Project evaluation

a. Evaluation of sibling-page tool (completed)

b. Evaluation of interactive Wang Notation Tool (one round of evaluation completed)

c. Evaluation of mini-ontology creation tool (completed)

d. Evaluation of auxiliary tools and 'low-hanging fruit' tools (some done, most underway)

e. Overall evaluation of TANGO (to do)

f. Overall evaluation of WoK (to do)

Findings:

We list findings for each of the activities in the activities report. Thus, each finding relates to expected results (as originally proposed) or relates to new goals (discovered along the way as worthwhile, low-hanging fruit to be picked). 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 except that 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.
MS thesis completed, conference paper submitted.

The Wang Notation Tool (WNT) is a semiautomatic, interactive tool that converts tables to Wang notation û 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, the operator can intervene to generate correct Wang notation. Initial evaluation of Version 3 was conducted on 12 subjects and 16 large tables from Canada Statistics.

In progress. Expected findings: Users can generate table interpretations faster with the Wang notation tool than by hand.

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 (MS thesis completed and published, paper submitted)

Ontology creation is a daunting task: manual creation is tedious and time consuming, and automatic creation is disappointingly inaccurate. We have developed a tool to automate the generation of ontologies from ordinary web tables. The process is akin to reverse-engineering relational tables to conceptual models, but must account for a much greater variety of table layout patterns. Experimental evaluations show that the automatic ontology acquisition process can 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 (MS thesis completed and published)

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. 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. Conference paper submitted. 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.

In progress. Expectations: In the TANGO project, we also expect to use this QBT tool as a convenient way to query the results obtained and for our evaluation of TANGO.

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 the larger project to develop 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)

In progress. 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 one fundamental way to create a WoK.

5. Project evaluation

We have reported the results of completed evaluations of individual projects 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 total number of tables (~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.)
* P. Jha completed an MS on Wang Notation Tool in May 2008.
* B. Yamadala completed MS on Evaluation of Classifiers in December 2006.
* R. Padmanabhan is preparing MS thesis on Query by Table, expected December 2008.
* Two other MS students currently work in DocLab on unrelated projects.
* During the project period, Nagy served on 9 PhD committees on unrelated projects.

BYU:
* Two students, Reema Al-Kamha and Muhammed Al-Muhammed, completed PhDs.
* Two students, Zonghui Lian and Stephen Lynn, completed MS degrees.
* Two additional PhD and one additional MS student are continuing their work on TANGO and are expected to graduate within the next year.

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 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.

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.

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.

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

Workshop presentation by G. Nagy, Document Analysis Systems, Wellington, NZ 2006.

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

3.International Outreach
ò 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.
ò In Summer 2007, Emanuele Olivetti, a full-time researcher and PhD student, spent two months at RPI DocLab to familiarize himself with our projects.
ò 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.
ò In June 2008, Nagy is presenting three lectures and supervising exercises at the Summer School on Digital image Processing offered by the Indian Institute of Science, Bangalore, as part of its Centennial Celebration. He will also give a lecture on table analysis at First Indian Corporation Private Limited (FIC), Bangalore, a leading web transaction processor.
ò 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.
ò Embley presented a colloquium talk at the General Earth Environmental Research Laboratories, Research Institute for Humanity and Nature located in Kyoto, Japan.

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. Another student, about to graduate from High School, will work with Nagy this summer on the URP grant and probably continue in the Fall.

5. Professional societies:
* Embley, Lonsdale and Nagy participated extensively in refereeing proposed books and manuscripts for technical journals
* We reviewed of 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.
* 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. , vol. , (2008). Submitted,

Books or Other One-time Publications

M. Al-Muhammed, D.W. Embley, and S.W. Liddle, "Conceptual model based semantic web services", (2005). Proceedings, Published
Bibliography: Proceedings of the Twenty Fourth International Conference on Conceptual Modeling (ER'05), Klagenfurt, Austria, 288-303

D.W. Embley, D. Lopresti, and G. Nagy, "Notes on Contemporary Table Recognition", (2006). Proceedings, Published
Bibliography: Document Analysis Systems VII, 7th International Workshop, Procs. DAS 2006, H. Bunke and A. L. Spitz, Eds., vol. 3872, LNCS, pp. 164-175, Springer, Nelson, New Zealand

G. Nagy and D. Lopresti, "Interactive Document Processing and Digital Libraries", (2006). Proceedings, Published
Bibliography: Procs. 2nd IEEE International Conference on Document Image Analysis for Libraries, Lyon, France, April 27-28, pp. 1-9, IEEE Computer Society Press

M.J. Al-Muhammed and D.W. Embley, "Resolving underconstrained and overconstrained systems of conjunctive constraints for service requests", (2006). Book, Proceedings
Bibliography: Proceedings of the 18th International Conference on Advanced Information Systems Engineering (CAiSE'06, LNCS 4001), Luxembourg City, Luxembourg, 5-9 June, 223-238

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, (2006). Beijing, China, 3-7 September, 400-414

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

M. Al-Muhammed and D.W. Embley, "Ontology-Based Constraint Recognition for Free-Form Service Requests", (2007). Proceedings, Published
Bibliography: Proceedings of the 23rd International Conference on Data Engineering (ICDE 2007), Istanbul, Turkey, 16--20 April, 366-375

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), (2007). Paris, France, 27--29 June, 131-142

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

Cui Tao and David W. Embley, "Seed-based Generation of Personalized Bio-Ontologies for Information Extraction", (2007). Proceedings, Published
Bibliography: Proceedings of the First International Conference on Conceptual Modelling for Life Sciences Applications (CLMSA'07), Auckland, New Zealand, 5--9 November, 74--84

Cui Tao and David W. Embley, "Automatic Hidden-Web Table Interpretation by Sibling Page Comparison", (2007). Proceedings, Published

Bibliography: Proceedings of the 26th International Conference on Conceptual Modeling (ER'07). Auckland, New Zealand, 5--9 November, 566?581

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

D. Lopresti, G. Nagy, S. Seth, and X. Zhang, "Multi-character Field Recognition for Arabic and Chinese Handwriting", (2008). Book Chapter, Published
Bibliography: Arabic & Chinese Handwriting Recognition (D. Doermann, S. Jaeger, editors), Springer LNCS # 4768, pp. 218-230, April

G. Nagy, "Digitizing, coding, annotating, disseminating, and preserving documents", (2008). Proceedings, Accepted
Bibliography: Procs. IWRIDL-2006 workshop on Digital Libraries, Kolkota, India, ACM 1-59593-608-4, April

P. Jha and G. Nagy, "Wang Notation Tool: Layout Independent Representation of Tables", (2008). Proceedings, Submitted
Bibliography: International Conference on Pattern Recognition, December

R. Padmanabhan and G. Nagy, "Query by Table", (2008). Proceedings, Submitted
Bibliography: International Conference on Pattern Recognition, December

Stephen Lynn and David W. Embley, "Automatic Generation of Ontologies from Canonicalized Web Tables", (2008). Proceedings, submitted; to be improved and submitted again
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

Sharing Information:

In progress. 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

Sharing Information:

In progress. Will be posted on the TANGO website after thorough verification.

Product Type:

Software (or netware)

Product Description:

Mapping and Merging tool: allows users to integrate ontologies

Sharing Information:

In progress. Will be posted on the TANGO website after thorough verification.

Product Type:

Software (or netware)

Product Description:

Automated Mapping and Merging tool

Sharing Information:

In progress. Will be posted on the TANGO website after thorough verification.

Product Type:

Software (or netware)

Product Description:

TANGO: end-to-end ontology generator

Sharing Information:

In progress. Will be posted on the TANGO website after thorough verification.

Product Type:

Software (or netware)

Product Description:

WoK: Web of Knowledge

Sharing Information:

In progress. Will be posted on the TANGO website after thorough verification.

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 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 16 large tables and 12 subjects.

We have analyzed 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.

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 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 that allows users to integrate ontologies. The tool also supports plug-ins for automatic mapping and merging algorithms.

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, six students working on the project have graduated: two PhD and two MS students at BYU, and one MS and one BS student at RPI.

Students and recent graduates are coauthors of 18 published papers, 1 accepted paper, and 4 submitted papers.

4 students in the TANGO research groups are female. two have graduated, one is a current PhD student, and one is an MS student. None of the other students working on the project is in an under-represented or minority group |
(Nagy has another US Female research assistant on an NSF Cybertrust project.)

Since the award was granted, in forums with the public invited, students have given 4 major presentations (colloquium talks or conference/workshop presentations) and have given 6 minor presentations (local workshops).

Since the award was granted, four students working on the project have graduated -- two PhD students and two MS students.

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

Since the award was granted, in forums with the public invited, students have given 4 major presentations (colloquium talks or conference/workshop presentations) and have given 6 minor presentations (local workshops).

Three students in our TANGO research group are female. One graduated, one is a current PhD student, and one is a current MS student. None of the other students working on the project is in an under-represented or minority group.

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 on Web of Knowledge, specifically retrieving facts by combining information from several tables meant for human readers.

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.

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

Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Contributions: To Any Other Disciplines

Contributions: To Any Beyond Science and Engineering