SUMMARY

At the heart of today's information-explosion problems are issues involving semantics, mutual understanding, concept matching, and interoperability. Ontologies and the Semantic Web are offered as a potential solution, but the assumption that ontology creation is a human-intensive activity with little room for automation impedes progress. Contrary to this assumption, we believe that for many domains, nearly automatic ontology generation is possible.

We base this belief on our collective experience in several areas relevant to this work: ontology building, conceptual modeling, information extraction, natural language processing, computational linguistics, machine learning, image processing, table understanding, document image analysis, OCR, and geographic information systems. Although these fields are each distinct and interesting as separate fields of research, it is our unique approach to combining them in new and groundbreaking ways that leads to a potential solution for automatic ontology generation.

In our research we address ontology generation by developing an approach called TANGO (Table ANalysis for Generating Ontologies). We apply ideas we have developed, particularly in conceptual-model-based extraction and table recognition, in new and innovative ways to: (i) understand a table's structure and its conceptual content; (ii) discover the constraints that hold between concepts extracted from the table; (iii) match the recognized concepts with ones recognized in other tables; (iv) merge the resulting structures to create a domain ontology; and (v) adjust the created domain ontology so that it is a clean, complete, accurate, and redundancy-free conceptualization of the source tables.

For purposes of illustration and testing, we have chosen geographic information as the application domain—one that has a large and important field of knowledge where data often appears in a tabular format. Concrete results we seek in our TANGO project include: (i) an ontology of geographic relationships produced from various tables; (ii) a tool ontology describing tabular formats, arrangements, styles, and content types; (iii) a publicly available demonstration system deployed on the web, which allows users to build ontologies in their own domains of interest, and (iv) a publicly available collection (or corpus) of a wide range of sample tables in various formats that can serve for future testing and development.

Intellectual Merit. As part of the solution to the grand challenge for automatically understanding semantics for free interoperation of heterogeneous software components, we offer the prospect of building a tool for generating ontologies as intermediate components for resolving heterogeneity. In particular, we combine original techniques from our collective research backgrounds to show that a system can be built that can extract and organize not only data but metadata as well, and can automatically create an ontological description of a body of knowledge.

Broader Impact. We consider our TANGO project not only as a means to advance discovery and understanding, but also to promote training and cross-fertilization between departments and universities. Our research team is housed in three departments (Computer Science, Linguistics, and Electrical, Computer, and Systems Engineering) at two institutions (BYU and RPI). We also promote participation of underrepresented groups and demographic diversity. Our research groups have and do include female students and students from the US, the Middle East, and the Far East. More directly for the research itself, we are building infrastructure by making corpora and demonstration systems available and by building tools (ontology generators) for enhancing infrastructure. Achieving the goals of the TANGO project are of benefit the larger community because they enable technology for harnessing the information explosion and they are helping to make the Semantic Web a reality.