REQUEST FOR REU SUPPLEMENT

The TANGO project requires executing a number of tasks that require sound judgment and understanding of complex project objectives, but no advanced preparation in mathematics or computer science. Assigning these tasks to qualified undergraduates, even sophomores, would allow the PIs and the graduate students to concentrate on the further development information extraction algorithms and of the schemas required for the representation and combination of ontologies in the form of conceptual models.

Rensselaer attracts excellent undergraduates, with most in the top 10% of their high school graduating class. President Shirley Jackson has instituted effective measures for diversity. Over the last five years the number of underrepresented minorities has more than doubled, and so has the number of women in the engineering programs. We therefore have excellent opportunities to recruit students in these categories. Nagy has a strong record of engaging women in research at the doctoral, masters, and undergraduate levels, and has mentored several minority undergraduates. Many of his undergraduate project participants chose graduate school over remunerative engineering jobs. His most recent research assistant (on another information-technology oriented project) is a history major with nine years' experience as a journalist.

RPI strongly encourages undergraduate research participation, but the current maximum of $600 per semester (awarded only against matching external funds) is insufficient for financially insecure students. We will, of course, apply for the RPI supplement to the REU supplement.

The REU students will be full participants in our research, join our project meetings, make written and oral reports, and have weekly or biweekly conferences with the PI. They will also be assigned space in DocLab with the graduate students, and invited to the DocLab hikes. We also make it a practice to introduce all our students to our many visitors from the US and abroad, and to have them give a short presentation or demo. The tasks we propose for the two students are challenging and instructive. They will increase the students' computer skills as well as their geopolitical domain knowledge.

Recruitment

Students will be recruited through the following Rensselaer organizations:

Black Students Alliance http://bsa.union.rpi.edu
National Society of Black Engineers (NSBE) http://www.nsbe.rpi.edu
SHPE (Society of Hispanic Professional Engineers) http://shpe.union.rpi.edu
SWE (Society of Women Engineers) http://swe.union.rpi.edu

The two positions will also be posted on the website of Undergraduate Research Program (URP) of the Rensselaer Office of Undergraduate Education
http://www.rpi.edu/academics/undergraduate/urp.html ,
and on the websites of selected academic departments.

Proposed assignment of Student #1:
Data Collection, Organization, and Transformation

The task proposed for Student #1 is data collection. This consists of the following steps:

1. Location of suitable web pages that contain one or more tables with geopolitical data. At least some of the tables must contain partially overlapping data, but designated by different headers. Sources of such tables are primarily US and Canadian government sites.
2. Extraction of the tables using the Java tools already developed for this purpose at BYU and Rensselaer Polytechnic Institute, and addition of a unique identifier, table title, table caption, and any annotation pertaining to footnotes, units, aggregates. Portions of the web pages that contain HTML table tags that do not correspond to tables must be rejected. (Table tags are often used for formatting text and illustrations.). At this point all of the necessary data is translated into ASCII files.
3. Conversion of the ASCII table files into XML tagged Wang notation using the WNT program developed at RPI. The tool is interactive and requires sufficient familiarity with table structures to determine categories and subcategories as defined in the Wang abstract data type. This is a challenge for many tables (e.g., the Periodic Table or a corporate Annual Report) designed for readers with specialized interests.
4. Validation of each XML schema, and correction of any errors. This will expose the student to the important notions of validation and verification.
5. Organization of the original web pages, of the HTML subsets, and of the XML-tagged table files for transmission to BYU, where they will be converted into mini-ontologies. The mini-ontologies are eventually merged into growing-ontologies. This step requires close liaison with BYU researchers.

Performance of this task will give the student insight into web infrastructure, the interfaces between JAVA, Matlab and C++, experimental design, and the communications requirements of a large, distributed research project.

Proposed assignment of Student #2:
Performance Evaluation

The second student's assignment will be to help with the evaluation of the proposed approach to information consolidation. In simple terms, we intend to compare information retrieval with and without TANGO. An information consumer can find answers to queries by consulting a large prespecified set of web pages that contain the desired facts. Alternatively, he or she can address queries directly to the TANGO ontology. Both textual queries and queries-by-table (i.e., filling out blank tables) will be emtertained. The student's responsibilities will include the following:

1. Tracking the time spent in harvesting experimental tables, including the rejection of inappropriate web pages that the automated filter accepts. This information must be aggregated over tables and subjects.
2. Timing, through the logging routines now being developed, the specification of Wang categories and subcategories, and the correction of WNT errors.
3. Comparing table interpretation (i.e., conversion to Wang Notation) with WNT against table interpretation without WNT.
4. Reporting and analyzing the types of errors encountered for subsequent improvement of the algorithms and of the Matlab graphic user interface.
5. Collecting timing data from the logging routines for the back-end, i.e., answering queries with and without TANGO. Aggregating results in Excel across subjects and tables.

Performance of this task will give the student insight into the evaluation of human-machine systems, and of the care required to produce replicable results. The student will also become familiar with elementary statistical concepts like the role sample mean and sample variance in hypothesis testing, and develop an appreciation for the importance of sample size for results at a given level of significance. (Our last undergraduate with a similar assignment is now a doctoral student at MIT.)

Schedule

We will remain flexible enough to let the undergraduates participate, depending on their abilities and interests, in other aspects of TANGO. We must, however, give priority at all times during the academic year to the students' course work, assignments, and examinations. We plan to engage the students for 10 hours per week during the semester, and 20 hours per week during two summer months. If we can afford it, we will arrange a week's visit for each student to our partner institution, BYU (such a visit by one of our graduate assistants has proved immensely valuable.)

To the extent that the arrangements prove mutually satisfactory, the REU students will be expected to remain with the project until its conclusion.