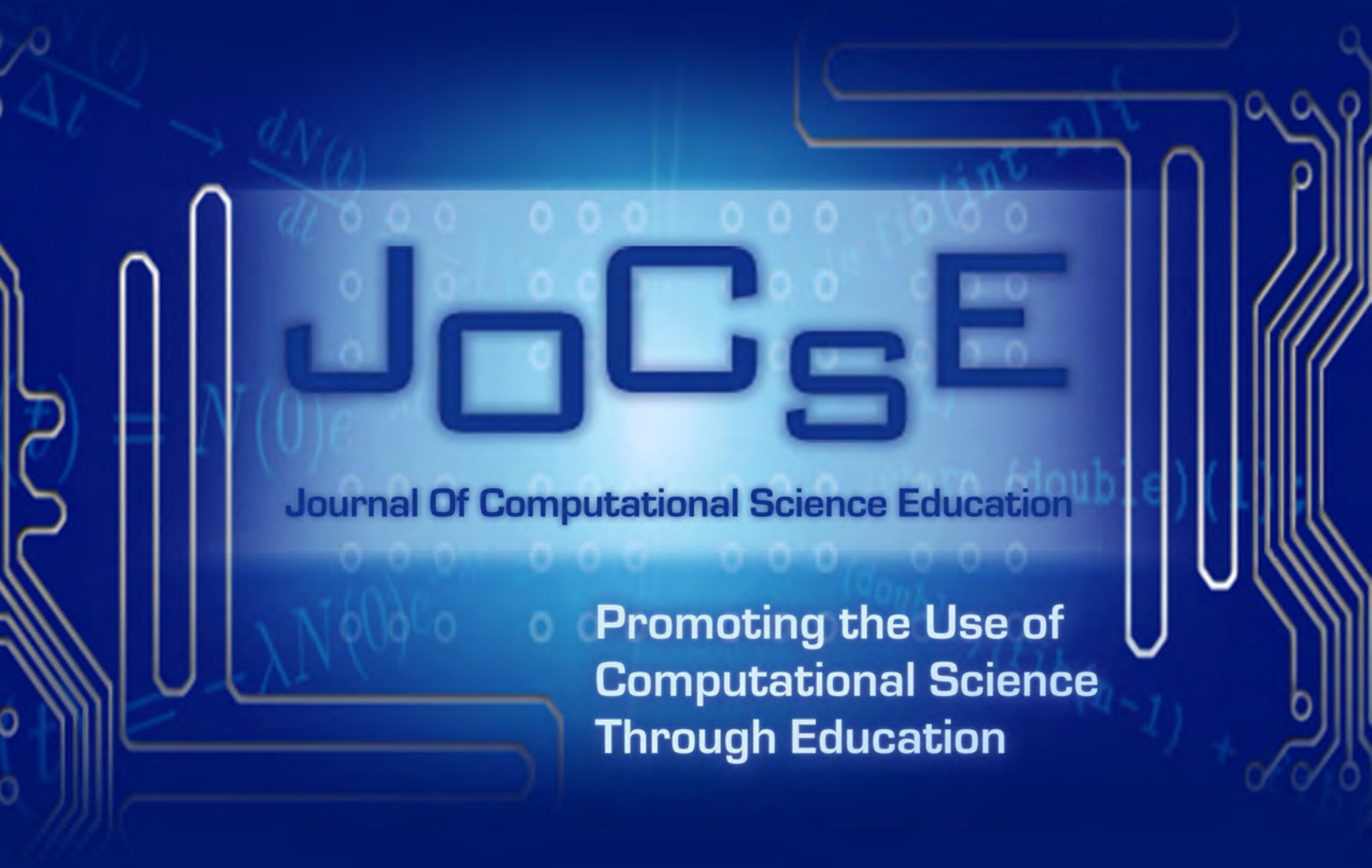
June 2012

Volume 3 Issue 1



JOCSE

Journal Of Computational Science Education

Promoting the Use of
Computational Science
Through Education

ISSN 2153-4136 (online)

Volume 3, Issue 1 **June 2012**

JOCSE

Journal Of Computational Science Education

The Journal Of Computational Science Education (JOCSE), ISSN 2153-4136, is published quarterly in online form, with more frequent releases if submission quantity warrants, and is a supported publication of the Shodor Education Foundation Incorporated. Materials accepted by JOCSE will be hosted on the JOCSE website, and will be catalogued by the Computational Science Education Reference Desk (CSERD) for inclusion in the National Science Digital Library (NSDL).

Subscription: JOCSE is a freely available online peer-reviewed publication which can be accessed at `http://jocse.org`.

Contents

Introduction to the CI-TEAM Issue

Steven I. Gordon
Editor
Ohio Supercomputer Center
Columbus, OH
sgordon@osc.edu

Forward

The Cyberinfrastructure Training, Education, Advancement, and Mentoring for Our 21st Century Workforce (CI-TEAM) program at the National Science Foundation supported projects that integrated science and engineering research with educational activities that integrated the use of information, communication, and computational technologies. The projects funded under this program encompassed a wide range of applications and activities from both formal and informal education programs. The articles in this issue of the Journal of Computational Science Education (JOCSE) provide descriptions of a cross-section of projects that should be of interest to the broader computational science community. A comprehensive list of recent awards can be found on the NSF website.

Given the wide variety of topics, it is impossible to put all of the projects into very distinct groups. However, there are several components of the projects that are common across large sub-groups. Many of the projects apply Internet and computer technologies toward some educational goal. This includes adaptation of web-based tools to support new communities or provide on-line access to educational materials. Some projects have created portals not only to information but also to simulations and models on a variety of topics with educational objectives. They have used the simulations to reach specific audiences and to stimulate interest in a broad range of science and engineering topics.

Other projects have chosen to study how particular groups of researchers currently use computational tools in their work. Some use this information to create collaboratories that allow researchers to more effectively share that information and potentially improve their research productivity. Still others assist researchers in finding the appropriate cyberinfrastructure tools that will most benefit their research endeavors.

Each of the projects also focused on a different target audience. Many focus on improving the outcomes of education in the STEM (science, technology, engineering, and mathematics) fields by engaging students in the learning process using computer modeling tools and games. A number focus on retraining a specific portion of the workforce, providing them with the skills they need to compete in the 21st century workforce.

The idea for this edition of JOCSE emerged from a meeting of the CI-TEAM principle investigators and their research teams sponsored by NSF and hosted at the University of Illinois in May 2011. At that time, the PIs presented the preliminary results of their projects and shared information on the nature of the topics, methods, and audiences associated with their projects. Many of the projects integrated computer modeling and related cyberinfrastructure tools in their implementation. This led the editors to offer to assemble this special issue of the journal. A call for papers was released late in 2011. A number of papers were received, peer reviewed, and revised to constitute this issue of JOCSE.

The papers presented herein represent a significant cross-section of the CI-TEAM projects, spanning several topic areas and objectives congruent with those of JOCSE and the computational science education community. We expect you will find the articles provide excellent examples of innovative approaches to cyberinfrastructure-based education programs.

We would like to acknowledge the partial support for the compilation of this issue from the University of Illinois through a subcontract on the NSF grant that supported the CI-TEAM principle investigator meeting.

Application of the Occupational Analysis of Computational Thinking-Enabled STEM Professionals as a Program Assessment Tool

Joyce Malyn-Smith
Education Development Center
55 Chapel Street
Newton MA 02458-1060
jmalynsmith@edc.org

Irene Lee
Santa Fe Institute
1399 Hyde Park Road
Santa Fe, NM 87505
lee@santafe.edu

ABSTRACT

This paper describes the application of findings from the National Science Foundation's project on Computational Thinking (CT) in America's Workplace to program assessment. It presents the process used to define the primary job functions and work tasks of CT-Enabled STEM professionals in today's scientific enterprise. Authors describe three programs developing CT skills among learners in secondary and post secondary programs and how the resulting occupational analysis was used to review these programs. The article presents ways this analysis can be used as a framework to guide the development of STEM learning outcomes and activities, and sets of directions for future work.

1. INTRODUCTION

Over the past several years, thought leaders within the computer science and education communities have defined computational thinking within their own communities of practice and discussed the importance of computational thinking as a key ingredient in technology-enabled discovery and innovation. [1, 2, 3, 9, 10, 22, 23] This national conversation has provoked questions about what computational thinking looks like in practice among scientists, engineers and technologists in various industry sectors and how core computational thinking skills might be nurtured in students as they prepare for STEM careers.

Clearly these conversations have made a significant contribution to the field by developing momentum for dialogue within the STEM education community and focusing attention on the importance of defining CT for purposes of developing programs and curricula. Thus far, however, the conversations have represented the perspectives of university-based thought leaders and others somewhat distanced from scientific, technical, and industry workplaces. And while they have succeeded in creating theoretical constructs for CT, these conversations represent only one side of the education-to-employment continuum and thus provide only an approximation of how CT is integrated into daily work activity and shapes problem solving in scientific work settings. There is an urgent need to build on and strengthen this good work by expanding the conversation to include STEM workers in a variety of settings who can ground these ongoing efforts to define CT in real work activities. Without clear, authentic examples and artifacts illustrating what CT looks like "in action" at work, educators will have difficulty building programs that lead to successful use of CT in STEM careers. Articulating authentic examples of CT in action requires the conversation to focus on expert computational thinkers who currently work in STEM fields. Expert workers can contribute accurate examples with the specificity and authenticity educators will need to integrate CT into K–20 learning.

The NSF-funded "Computational Thinking in America's Workplaces" project (NSF award #OCI 1057672 9/1/10 – 8/31/12) advances understanding of computational thinking by exploring CT as a foundational skill for STEM workers and developing a Profile that describes the ways scientists and other STEM professionals engage in CT as they carry out routine job tasks and solve problems associated with their work. Additionally, this project generated language and examples that promote new ways of talking about computational thinking and clarified the definition of CT by contextualizing it within the work of scientists and engineers. Products of the research include an occupational definition and profile of the Computational Thinking Enabled STEM Professional, and concrete examples of what CT looks like "in action" in America's scientific and engineering workplaces.

2. METHODOLOGY AND JUSTIFICATION

Building on the successes of ATE's IT Across Careers (ITAC) Project and a legacy of experience in developing national skill standards [4, 6, 7, 8, 15, 18, 20, 21], the project developed the "Learning Occupation" of a CT-enabled STEM worker then identified and validated with expert CT workers the "computational thinking" skills/competencies that are used by scientists and engineers in STEM careers. The process employed was one that had been used successfully to develop national skill standards for emerging industries (Biosciences) [11] and industries undergoing substantive changes in professional and technical job responsibilities (Human Services). [21] The Learning Occupation was used in the Bioscience and Human

Services Skill Standards Projects to represent an outcome goal for education and training designed for workers who will be able to perform a broad variety of related work tasks suitable to a large cluster of occupations. A similar situation exists in STEM, making this an appropriate application of the concept. The work process involved four distinct steps: building a team, defining a learning occupation, developing a profile of the CT-enabled STEM worker and creating examples of CT in action.

Building a team: EDC (Education Development Center, Inc.) assembled a project team that included 3 experienced, highly qualified skill standards developers; and a technical committee consisting of 4 computer scientists involved in national discussions on CT representing major universities, research institutions, and industry (University of Washington, Massachusetts Institute of Technology (MIT), Williams College, Santa Fe Institute, and Raytheon Corporation). This technical committee's role was to ensure that the project team's products would connect to the interests of national thought leaders in the field of computational thinking. A panel of 11 expert CT workers representing a range of STEM careers, occupational levels, and work settings was recruited to participate in the rigorous occupational analysis. Expert panelists included research scientists, theoretical physicists, software engineers, mathematicians, applied scientists, engineers and security specialists drawn primarily from National Laboratories.

Defining a learning occupation: The following learning occupation was developed by the technical committee and revised by the expert panel as a result of the analysis workshop:

"A computational thinking enabled STEM professional engages in a creative process to solve problems, design products, automate systems, or improve understanding by defining, modeling, qualifying and refining systems, processes or mechanisms generally through the use of computers. Computational thinking often occurs in collaboration with others."

Developing a profile of the CT-enabled STEM worker: Once the learning occupation was defined and agreed upon, the expert panel developed a profile of the CT-enabled STEM worker. The Learning Occupation proposed by the project's technical committee became the subject of a modified DACUM analysis. DACUM (Developing A CUrriculuM) [14] is an internationally-known methodology used by expert practitioners in an occupational field to identify the major areas of work and the constituent tasks that define successful job performance. The DACUM method has been used internationally for more than half a century to identify core workforce competencies. This process rests upon three basic principles:

- Expert workers can describe and define their jobs more accurately than anyone else.
- An effective way to define a job is to precisely describe the tasks that expert workers perform.
- All tasks, in order to be performed correctly, demand certain knowledge, skills, resources, and behaviors.

3. DACUM ANALYSIS

Traditional DACUM analyses invite expert practitioners representing a single occupation. The "modified" DACUM approach used successfully by EDC engaged expert workers *from a range of related occupations* who share a common core of work tasks, knowledge, and skills. The first task undertaken by this panel of experts was to discuss and refine the proposed Learning Occupation so that it captured the essence and commonalities of their own work. The ensuing guided dialogue provided descriptions of concrete, observable activities for which the panelists use CT and that met the definition of the Learning Occupation. From the set of CT activities described, the panel identified 11 large functional groupings or "job functions".

Eight job functions (A-H) were organized into four categories as follows: "A computational thinking enabled STEM professional….":

Defines:
A. Identifies problem.
B. Specifies constraints.
Models:
C. Designs the model/system.
D. Builds the model.
E. Develops experimental design.
Qualifies:
F. Verifies the model.
Refines:
G. Optimizes the model and user-interface.
H. Facilitates knowledge/discovery.

The panelists identified 68 activities/tasks in which the STEM professionals described in the learning occupation use computational thinking. Each of the 68 tasks was grouped under the job function category to which it best corresponded (see Table 1).

Three cross cutting job functions were also identified: Engages in a creative process, Collaborates, and Documents. In addition, the panelists developed lists of the Knowledge, Abilities (skills), Desirable Behaviors of CT enabled STEM professionals as well as selected Tools and Techniques used as they are engaged in the activities listed (see profile of a Computational Thinking-enabled STEM Professional).

Creating examples of CT in action: Although the profile identified the work tasks in which STEM professionals engage when they are thinking computationally, concrete examples that described what computational thinking "looks like in action" would be needed to build a strong dialog bridge between computer scientists who understood CT and non-computer scientist educators who were struggling to understand CT and connect it to learning objectives in their classes. To build this bridge project staff worked with the expert panel to draft twenty-nine examples of routine tasks and problems the expert panelists solved using CT. Two examples follow:

A computational scientist verifies the numerical convergence of a solid mechanics finite-element model, by refining the mesh associated with a mechanical assembly, for the purpose of assessing the correct implementation of the mathematical equations. [Qualifies: Verifies the model]

A nuclear engineer validates a coupled thermo-mechanical computer model, by comparing the model predictions with existing thermal stress experimental data, to assess the performance of a nuclear fuel element for the purpose of extending the operational lifetime of the fuel in the reactor. [Qualifies: Validates the model]

Although the examples of "CT in Action" emerging from this expert group were clarifying for non-computer scientists and

STEM education professionals, the technical depth and complexity of the tasks described limited the use of these examples within the K-12 curricula. Additional examples more relevant to the experiences of K-12 students, and simpler tools/strategies would be needed to help learners and their guides/teachers recognize and nurture computational thinking in the K-12. The Job Functions and Tasks identified in the DACUM analysis serve as a simple tool to bridge that gap.

4. APPLICATION OF THE CT DACUM AS A PROGRAM ASSESSMENT TOOL

The Occupational Profile (DACUM) of a computational thinking enabled STEM professional was used to evaluate the core computational thinking skills nurtured in three programs serving students ranging from the middle school and graduate school levels. The three programs, one at the middle school level, one primarily at the high school level, and one at the university level, were selected because they actively engage students in computational thinking through computational modeling and simulation. The focus on computational modeling and simulation programs was intentional as “the underlying idea in computational thinking is developing models and simulations of problems that one is trying to study and solve.” [13]

The analysis was conducted by reviewing each program’s curricular materials, lesson plans, student assessments and rubrics, pedagogy (as reflected in teacher professional development materials), and students’ work products. Subsequently the authors interviewed individuals responsible for implementing each program’s curriculum.

Santa Fe Institute’s Project GUTS (Growing Up Thinking Scientifically) is an afterschool program at the middle school level that engages students in computational thinking through modeling and simulation in StarLogo TNG. In Project GUTS students actively engage in computational thinking as they design and implement models of local relevance and then use the models to run simulations. Students use the process of abstraction to narrow the problem down to something that could be implemented on a computer using StarLogo TNG, an agent based modeling tool. Students design and create models as test beds to answer questions about real-world concerns. For example, as part of the Project GUTS unit on Epidemiology, a group of students wanted to investigate whether a disease would spread throughout their school population given the layout of the school, the number of students, the movement of the students, the virulence of the disease, and the number of students initially infected. Mapping this question and scenario onto an agent based model, agents were used as abstractions or simplified representations of students and the number of agents matched the number of students in their school. Agents were given movement behaviors that were abstractions of moving from classroom to classroom, and decisions were made about which features of the school were important to take into consideration before a 3-D virtual model of the school building was created. For instance, students decided that recreating the number and location of passages and doors at the school was important. Additionally students modeled the characteristics of the contagion being spread: how often contact between students spread the disease from one to the other, and how many students were initially infected. To make the model a test-bed capable of running experiments, it was equipped with interface sliders to control individual variables. One interface element controlled the number of initially infected agents and another controlled the virulence of the contagious element.

A three-stage progression is used within Project GUTS to first engage and prepare youth in CT. This progression, called Use-Modify-Create [10], describes a pattern of engagement that was seen to support and deepen youth’s acquisition of CT in the authors’ NSF projects. It is based on the premise that scaffolding increasingly deeper interactions will promote the acquisition and development of CT. In the *use* stage, students run experiments using pre-existing simulations. Over time they begin to *modify* the model with increasing levels of sophistication. For example, a student may initially want to change the color of a character or some other purely visual attribute. Later the student may want to change the character’s behavior in a way that entails developing new pieces of code. Modification of this kind necessitates an understanding of at least a subset of the abstraction and automation contained within a model. Through a series of modifications and iterative refinements, new skills and understandings are developed as what was once someone else’s creation becomes one’s own. As youth gain skills and confidence, they can be encouraged to develop ideas for new computational projects of their own design that address issues of their choosing.

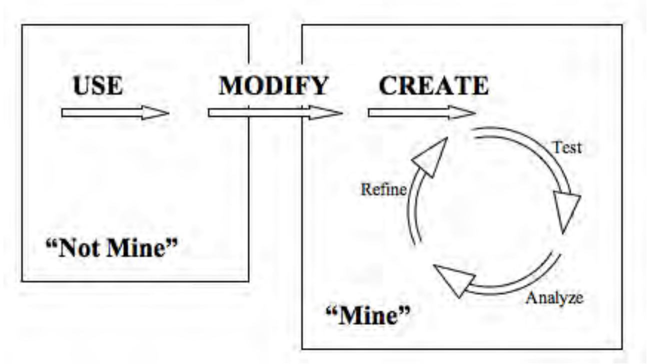


Diagram 1: Use-Modify-Create Learning Progression

Analyzing the CT learning within Project GUTS using the CT DACUM: Project GUTS engages middle school students in a creative process and encourages collaboration using pair-programming techniques, but only rudimentary coverage of the scope of tasks of a CT-enabled STEM professional are addressed. Of the 8 job functions delineated in the CT DACUM, the three primarily addressed during the course of participation in Project GUTS are: A) Identifies problem, C) Designs the model and D) Builds the model. (See Table 1.) Constraints are rarely addressed and students develop limited experimental designs. Rarely do they verify or refine a model. Note that the test-analyze-refine cycle in the diagram 1 above refers to iterative refinement in code development not model verification and validation.

The Supercomputing Challenge is a year-long program for middle and high school students culminating in a student competition. Middle and high school students are introduced to computational thinking and computational modeling at a Kickoff Conference held annually each fall. Students primarily use StarLogo TNG, NetLogo, and Java as the basis for their computational models and simulation. Teachers are prepared to sponsor and mentor student teams through the joint Supercomputing Challenge / Project GUTS Summer Teacher Institute. “Challenge” teams, working in small groups, develop computational modeling projects of their own choosing. They use a framework, called the "Computational Science Cycle" to guide them through the process of designing, implementing and analyzing a computational model. "This design-based approach has been effective in engaging learners in exploring computational ideas" [16, 17, 19]. Students are guided through the stages in the process as follows:

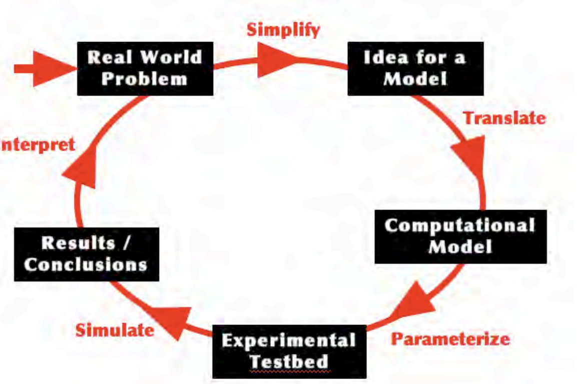


Diagram 2: The Computational Science Cycle

Stage 1: Select a real-world problem to study.
Discuss what makes a problem suitable for studying using computational methods. Describe the simplifications made in models through abstraction. Specify the measurable aspects of the problem and the questions that will be answered through modeling and simulation.

Stage 2: Simplify the scope of the model using abstraction. Specify the aspects of the problem that are important to include in the model and narrow the scope of the problem to one that can be modeled given the software and computing resources available.

Stage 3: Translate the idea for a model into a computational model. Decompose the problem. Abstract real-world objects into computational analogs. Abstract the physical behavior of the objects. Define interactions between variables, objects or elements. Choose appropriate representations. Use existing code and technology. Writes algorithms and programs. Debug and troubleshoot.

Stage 4: Parameterize the model. Discuss relevant variables and parameter and experiment design. Discuss what constitutes proof when using data output from models.

Stage 5: Simulate and collect data. Use the computational model as a test bed for running experiments. In some cases this involves writing another program that runs the model repeatedly over a set of input values; called a parameter sweep.

Stage 6: Analyze / Interpret: Discuss the limitations of the computer model, the assumptions were made, and what the model tell us, if anything, about the real world. Introduction to how models are verified and validated. Demonstrate the exploratory uses of models when no theory exists. [Verification is the demonstration that the model is logically correct and follows from the physical and mathematical laws used. Validation is the demonstration that the model correctly predicts the phenomena modeled.]

Repeat. The Computational Science Cycle is an iterative process. In evaluating the model one might find verification errors (e.g., bugs in code) or validation errors (e.g. when comparing model behavior to real-world data there are difference that suggest that the wrong assumptions or simplifications were made). In either case, the whole computational cycle repeats. It is an iterative refinement process.

Analyzing the CT learning within the Supercomputing Challenge: The Supercomputing Challenge engages middle and high school students in all three cross-cutting job functions; student teams engage in a creative process, collaborate in project work, and document all phases of their project. The Supercomputing Challenge requires student teams to submit project descriptions (abstracts), interim and final reports. In comparison to the middle school Project GUTS, a larger subset of the tasks of a CT-enabled STEM professional is addressed. Of the 8 job functions delineated in the CT DACUM, the four primarily addressed during the course of participation in the Supercomputing Challenge are: A) Identifies problem, C) Designs the model, D) Builds the model, and E) Develops experimental design. (See Table 1.) Top rated student projects address constraints and attempt at verification and validation of models however, the majority of projects do not reach this level of sophistication.

The NM EPSCoR (Experimental Program to Stimulate Competitive Research) program at New Mexico Tech offers opportunities for undergraduate and graduate students to participate on research projects in Computational Science, High Performance Computing, and Cyber-Infrastructure. Three themes, problem solving, collaboration and communication, are emphasized as keys to student success at NMT. In project-based courses, students conduct research projects as a major component of their coursework. They read and analyze project requirements, brainstorm project ideas then select and propose a project. Within each project, students working in project teams develop a solution to a real-world problem and communicate their results. Specific tasks include:

- Define and analyze requirements and specifications
- Determine feasibility and scope of problem
- Determine relevant assumptions, limitations, relationships among data
- Prototype problem to verify requirements
- Select modeling methodology, language / modeling environment, visualization
- Build model leading to hardware/software solution
- Develop experimental design (parameters & value ranges)
- Facilitate knowledge discovery (via model & vis.)
- Communicate results / accomplishment

Analyzing the CT learning within university courses at NM Tech: Undergraduate students enrolled in project-based CS courses at NM tech engage in all three cross-cutting job functions; they engage in a creative process, collaborate, and document and present their project. In comparison to the Supercomputing Challenge, a larger subset of the tasks of a CT-enabled STEM professional is addressed at the university level. Of the 8 job functions delineated in the CT DACUM, the five primarily addressed are: A) Identifies problem, B) Determines / specifies constraints, C) Designs the model, D) Builds the model, and E) Develops experimental design. (See Table 1.)

Within Senior Design projects, students are required to deliver a complete product to a customer. Students encounter tasks associated with the job function "Determines/Specifies Constraints". Senior design classes require that students work with clients/stakeholders to specify the requirements of the solution and resources, attend to stakeholder/customer communications and satisfaction, conduct a needs analysis and resolve conflicting requirements. Unlike earlier university project work, model verification and iterative refinement are necessary to improve the solution until the client/stakeholder is satisfied. Graduate research is expected to advance the state of the art in their discipline and communicate these advances to the field. This is in alignment with the expectations of a modern university.

5. FINDINGS

The DACUM occupational analysis provided a useful framework with which to evaluate the breadth and sequencing of CT instruction within computational modeling and simulation programs. Clearly students engaged in high quality, technical STEM learning are developing foundational computational thinking skills in the K-12 experience. Specifically, it was uncovered that the breadth and depth of tasks addressed corresponded with increases of grade level of the participants. Within each of the eight job functions (A. Identifies problem, B. Specifies constraints, C. Designs the model/system, D. Builds the model, E. Develops experimental design, F. Verifies the model, G. Optimizes the model and user-interface, and H. Facilitates knowledge/discovery), certain tasks were found to be appropriate for introduction at different grade levels. For example, within the "Identifies problem" job function, a middle school student in Project GUTS will identify the scope of the problem, select relevant aspects of the problem and define assumptions and limitations whereas a high school student participating in the Supercomputing Challenge additionally would be expected to identify data sources, risks of failure, and existing tools and solutions, research existing knowledge, determine if the problem has already been solved, and argue the need for a computational approach.

Across major job function categories, several entire categories were not addressed at all at the middle school level. Qualification and refinement of models based on verification and validation results is not addressed at the middle school level due to the advanced nature of these tasks. On the other hand, Documentation, a cross-cutting job function in the CT DACUM, is not emphasized in Project GUTS but would be appropriate and beneficial for middle school students to practice. At the high school level, all of the major job function categories are addressed to some extent though iterative optimization of models is rarely achieved due to constraints on students' time.

The three programs analyzed were found to prepare student for future computational thinking endeavors in STEM fields by providing student with experiences that directly mimic the work of CT-enabled STEM professionals. Several factors common to the workplace setting and the educational programs were intrinsic to offering these experiences: project-based investigations addressing real-world problems conducted by students working in teams.

6. LIMITATIONS

This study presents preliminary findings in a larger effort to define the primary work functions and tasks of computational thinking enabled STEM professional and technical workers. These would include applied and research scientists, engineers, technicians, technologists and mathematicians employed in our national STEM enterprise. Authors identify the following limitations to this study.

The sample size of computationally enabled STEM professionals was small. Although the size of the sample is consistent with literature on DACUM analyses, ongoing national validation of the importance and frequency in which CT-enabled STEM professional engage in these work functions and tasks will strengthen this work.

The sample of computationally enabled STEM professionals did not include persons describing themselves as engineers and/or technicians. As a result a second analysis is underway to align this analysis with the work functions and tasks of computationally enabled product engineers. The resulting data will be aligned and/or integrated to provide an analysis welcoming to scientists and engineers alike.

Program assessment did not take into account multiple levels of implementation of work tasks. The analysis employed a binary categorization; either the program under consideration included and promoted an activity or task or it did not. More detailed analysis would be useful.

7. IMPLICATIONS

The project's occupational profile of a CT-enabled worker and the examples of CT in action provide a common framework and an authentic structure against which CT thought leaders can test their concepts and assumptions about what CT activities, skills and knowledge STEM professionals use on the job. At the same time, these materials are designed to inform the thinking of educators. The occupational analysis provides a framework that can inform the development and sequencing of CT instruction for both academic and technical programs. Alternatively, the occupational profile provides a framework for the evaluation of programs that offer CT education. Occupational profiles have been used in the past to evaluate curricula and programs to ensure that all of the content needed to meet workplace demands/expectations were included in courses designed to lead to a specific career. [14]

In addition to providing a more inviting common language for national dialog on CT, the examples of CT are useful in helping the non-computer scientist understand the ways CT is used to perform routine tasks and solve problems in the STEM workplace. The listing of tasks can help educators understand the ways CT is applied in STEM work and analyze the evolving CT skills of their students performing routine scientific experiments/assays and solving problems in STEM classrooms and laboratories. Furthermore, the grouping of tasks into job functions, the listings of skills, knowledge and abilities as well as industry trends can help learners understand more about what it takes to succeed in America's STEM workplaces focused on discovery and innovation.

As an assessment tool, the CT DACUM provides a framework for conducting a gap analysis to assess the degree to which programs or curricula are addressing topics identified by the scientific community as important to the work of computational thinking enabled STEM professionals. Community stakeholders can use the tasks to determine how aligned STEM education programs are to workplace needs and expectations for STEM professionals work performance. Educators can determine whether specific courses or course sequences are adequately preparing students for future courses and professional endeavors. Students can analyze what skills and experiences they lack and which programs may fill those gaps. Examples of CT in Action statements can be used to guide assessment of students' CT skills.

The findings from this analysis may benefit students, educators, representatives from industry, and researchers by clarifying which computational thinking educational experiences link to workforce needs. Program managers are provided with a tool with which they can evaluate their own programs relative to preparing students for the computational thinking-enabled STEM workforce. Educators may deepen their understanding of computational thinking, its place in curricula, and its role in preparing the next generation of computational scientists.

Table 1: Job functions and tasks for the Computational Thinking enabled STEM professional annotated with three educational programs that nurture their development.

JOB FUNCTIONS and TASKS of the Computational-Thinking enabled STEM professional

Educational Programs
PG = Project GUTS (Growing Up Thinking Scientifically) middle school students using StarLogo TNG for modeling and simulation
SC = Supercomputing Challenge year long program for middle and high school students culminating in a student competition.
NMT = New Mexico Tech EPSCOR (Experimental Program to Stimulate Competitive Research) undergraduate and graduate students participate in research projects in Computational Science, High Performance Computing and Cyber-Infrastructure.

Defines			
Identifies			
A1. Identifies the scope of the problem.	PG	SC	NMT
A2. Selects relevant aspects of the problem.	PG	SC	NMT
A5. Defines assumptions and limitations.	PG	SC	NMT
A3. Identifies data sources.		SC	NMT
A4. Identifies risks of failure.		SC	NMT
A6. Identify existing tools and solutions.		SC	NMT
A7. Researches existing knowledge.		SC	NMT
A8. Determine if problem is already solved.		SC	NMT
A9. Identifies the need for a computational approach.		SC	NMT
Determines/Specifies			
B5. Specifies requirements of the solution.		SC	NMT
B6. Specifies resource requirements.		SC	NMT
B1. Determines if stakeholder has articulated the correct problem.			NMT
B2. Identifies stakeholder.			NMT
B3. Conducts needs analysis.			NMT
B4. Resolves conflicting requirements.			NMT
Models			
Designs the model			
C1. Proposes solution(s) / outcome(s) related to the problem.	PG	SC	NMT
C9. Decomposes problem / objects / processes / data.	PG	SC	NMT
C10. Abstracts the real world scenario /object into an analog.	PG	SC	NMT
C11. Abstracts physical behavior of the problem.	PG	SC	NMT
C12. Selects salient features to be included in the model.	PG	SC	NMT
C13. Designs the user interface.	PG	SC	NMT
C2. Identify why proposed solution is better than existing solutions.		SC	NMT
C3. Strategizes computational approach.		SC	NMT
C4. Identifies what modeling technique/ approach to employ.		SC	NMT
C5. Defines relationships among data (1:1, isomorphism).		SC	NMT
C6. Reverse Engineers processes and/or products.			NMT
C7. Applies systematic techniques to isolate cause & effect.			NMT
C8. Selects common properties from examples of the model / scenario / process.			NMT
Builds the model			
D1. Defines variables.	PG	SC	NMT
D2. Defines interactions among variables, objects or elements.	PG	SC	NMT
D3. Chooses an appropriate representation (e.g. data structures).	PG	SC	NMT
D4. Uses applicable existing code / technology.	PG	SC	NMT

		D5. Leverages existing solutions, algorithms.	PG	SC	NMT
		D6. Writes programs.	PG	SC	NMT
		D9. Debugs / Troubleshoots.	PG	SC	NMT
		D7. Modularizes model.		SC	NMT
		D11. Builds the User interface.		SC	NMT
		D8. Identifies sources of error.			NMT
		D10. Conducts fuzz testing (permutation testing).			NMT
	Develops experimental design				
		E1. Defines parameter space.	PG	SC	NMT
		E2. Defines initial conditions under which the model operates.	PG	SC	NMT
		E4. Executes model (tests limits / sweeps parameter space) to calculate results.	PG	SC	NMT
		E5. Tests the user interface.		SC	NMT
		E3. Develops testing equipment.			NMT
Qualifies					
	Verifies the model				
		F1. Verifies the model.		SC	NMT
		F2. Generates potential solutions / possibilities.		SC	NMT
		F3. Compares the behavior of the model to a known solution (or analytic solutions).		SC	NMT
		F4. Compares model with manufactured solutions.			NMT
		F5. Tests interface.			NMT
		F6. Validates the model.			NMT
		F7. Assesses the degree to which solution meets specifications / intended results.			NMT
		F8. Analyzes the sensitivity of the solution with respect to model parameters.			NMT
Refines					
	Optimizes the user interface and model				
		G1. Improve input / Interface.		SC	NMT
		G2. Output / design visual representation of data.		SC	NMT
		G3. Optimize model.		SC	NMT
		G4. Use iterative refinement to focus on the problem.		SC	NMT
		G5. Propose strategies to improve solution.			NMT
		G6. Identifies risks (e.g. sub-optimal solution).			NMT
		G7. Executes improved strategies.			NMT
		G8. Selects improved solution.			NMT
	Facilitates knowledge / discovery				
		H2. Observes phenomena to determine relationships (emergent behavior).		SC	NMT
		H3. Explains observed phenomena.		SC	NMT
		H6. Assesses the degree to which the solution produces new findings / knowledge.		SC	NMT
		H7. Analyzes experimental data.		SC	NMT
		H1. Generates new hypotheses that feedback to experimental design.			NMT
		H4. Discovers new relationships.			NMT
		H5. Refines experimental design.			NMT

8. CONCLUSIONS AND FUTURE WORK

The information contained in the DACUM analysis provides valuable keys to the success of the next generation of STEM innovators preparing to compete in a highly technical, global workforce. The tasks help to demystify the ways computational thinking is used in the STEM workplace by providing concrete descriptions of routine and problem solving tasks normally performed by computational thinking-enabled STEM professionals. Tying CT to concrete tasks helps educators deepen their understanding of CT and its STEM applications. This deeper understanding increases educators' ability to recognize and observe computational thinking of students in their classes. It also helps educators identify how foundational CT skills/knowledge are connected to their own curricula and addressed in K-12 both in and out of school. As we can see from the examples described herein, and from our casual observations of today's tech savvy generation, America's youth are developing valuable and marketable computational thinking skills at an early age.

The recognition of the breadth and depth of computational thinking of today's youth can result in educators purposefully cultivating computational thinking among learners of all ages both as a way to deepen the learning of difficult concepts and to prepare youth for STEM careers. As youth progress along a STEM career path the language contained in the tasks can help students articulate what they know and are able to do as they prepare for college and job interviews.

This work research also raises the following questions:

- What does the CT skills trajectory based on this work look like? What will it take to develop these skills progressions?
- Would computational thinking professionals in all career fields employ CT in similar ways? Are there core computational thinking skills that all American's should master to prepare for success in a competitive workforce focused on discovery and innovation?
- If in today's highly technical STEM workplaces, modeling and simulation are key strategies to discovery, innovation and problem solving, what role should modeling and simulation play in the education of our K-12 youth?
- If computational thinking is central to discovery and innovation in a technology rich society, how will CT be taught in K-12? Who will teach it? How will it be assessed in the K-12 system and reported?
- What does computational thinking look like in America's other workplaces?

  The profile of the Computational Thinking Enabled STEM Professional resulting from the DACUM analysis adds useful language to the ongoing national dialog on computational thinking, a framework that can contribute to the evolution of CT education at all levels.

Typically, the next steps in this process would be to work with representatives of scientific industries and the expert panel to develop rubrics for each of the 11 job functions (3 cross-cutting and 8 CT categories) that articulate current employer expectations for "proficiency" in computational thinking. Project staff would organize additionally gathered "in action" examples along with the language generated throughout the DACUM process, into levels that concretize and contextualize CT from "novice" to "above proficiency". The rubrics would be used by curriculum developers to sequence instruction, by educators as a tool to "observe" and record computational thinking in action in their classes, and by learners to self-evaluate their progress towards workplace proficiency. The rubrics could also be used by employers to guide the ongoing professional development of scientists and engineering as they move from junior to senior levels, and novice to expert in computational thinking in STEM workplaces.

9. Acknowledgements

The authors wish to thank Joe Ippolito, Dr. Lorie Liebrock, Dr. Nadine Miner, and David Kratzer for their contributions to the Supercomputing 2011 panel on linking CT educational programs in K-20 with workforce needs. We also with to thank the middle and high school teachers who have served as Project GUTS club leaders and Supercomputing Challenge team sponsors for their many contributions to support computational thinking in K-12 education.

10. References

[1] Allan, W., Coulter, B., Denner, J., Erickson, J., Lee, I., Malyn-Smith, J., Martin, F. 2010. *Computational Thinking for Youth.* A white paper of the ITEST Learning Resource Center Working Group on Computational Thinking, Unpublished manuscript, Education Development Center, Inc.

[2] Committee for the Workshops on Computational Thinking, 2010. *Report of a Workshop on the Scope and Nature of Computational Thinking*, National Research Council, Washington, D.C., National Academies Press.

[3] Cuny, J.E., Snyder, L., Wing, J.M., 2010. *Computational Thinking: A Definition*. Unpublished manuscript

[4] Dahms, A.S., Leff, J.A. Industry Expectations for Entry-Level Technical Workers, *Biochemistry and Molecular Biology Education*, 30, 4 (2002) 260-264

[5] Denner, J., Bean, S., Martinez, J., Girl Game Company: Engaging Latina Girls in Information Technology, *Afterschool Matters*, 8 (2009) 26-35.

[6] Hofstader, R., Chapman, K. 1997. *Foundations for Excellence in the Chemical Process Industries. Voluntary Industry Standards for Chemical Process Industries Technical Workers* (ERIC Document Reproduction Service No. ED405480), American Chemical Society, Washington, D.C.

[7] *Integrating IT Skills in Law, Public Safety, Corrections and Security Career Programs*, second ed. 2008. Education Development Center, Inc., Newton, MA (Accessed 1/11/12 from IT Across Careers Web site, http://itac.edc.org/)

[8] Ippolito, J., Latcovich, M., Malyn-Smith J. 2008. *In Fulfillment of their Mission: The Duties and Tasks of a Roman Catholic Priest*, NCEA Publication, New York, NY.

[9] Isbell, C., Stein, L.A., Cutler, R., Forbes, J., Fraser, L , Impagliazzo, J. et al. Re(defining) Computing Curricula by Re(defining) Computing, *ACM SIGCSE Bulletin*, 41, 4 (2009) 195-207.

[10] Lee, I., Martin, F., Denner, J., Coulter, B., Allan, W., Erickson, J., Malyn-Smith, J., and Werner, L. (2011). Computational Thinking for Youth in Practice, ACM Inroads Vol. 2 No. 1.

[11] Leff, J. 1995. *Gateway to the Future: Skill Standards for the Bioscience Industry*. Education Development Center, Inc., Newton, MA.

[12] Leff, J., Malyn-Smith, J., Hiles, E. 1999. *Making Skill Standards Work: Highlights from the Field*. Education Development Center, Inc., Newton, MA.

[13] Moursund, D. 2009. *Computational Thinking*, IAE-pedia.org. Available online at *http://iae pedia.org/Computational_Thinking*. Accesed August 8, 2010.

[14] Norton, R.E. 1997. *DACUM Handbook*, second ed., Publications, Center on Education and Training for Employment, Columbus, OH.

[15] *Profile of a Technology-Enabled Investigator (Duties and Tasks)*, 2009. Education Development Center, Inc., Newton, MA.

[16] Resnick, M. 2002. Rethinking Learning in the Digital Age. In *The Global Information Technology Report: Readiness for the Networked World*, G. Kirkman, Ed. Oxford University Press.

[17] Resnick, M. 2006. Computer as Paintbrush: Technology, Play, and the Creative Society. In *Play = Learning: How play Motivates and Enhances Children's Cognitive and Social-Emotional Growth*, Singer, D., Golikoff, R., Hirsh-Pasek, K. (eds.). Oxford University Press.

[18] *Rubrics to Access Basic IT User Skills*, 2006. Education Development Center, Inc. Newton, MA, 2006. (Accessed 1/11/12 from IT Across Careers Web site: http://itac.edc.org/)

[19] Rusk, N., Resnick, M., Berg, R., Pezalla-Granlund, M. New Pathways into Robotics: Strategies for Broadening Participation, *Journal of Science Education and Technology*, 17, 1 (2008) 59-69.

[20] Taylor, M., Bradley, V., Silver, J., Leff, J., Malyn-Smith, J. 1996. *Using the Community Support Skill Standards: A Guidebook for Human Service Educators & Trainers*, Human Services Research Institute, Cambridge, MA.

[21] Taylor, M., Bradley, V., Warren, Jr., R. (Eds.) 1996. *The Community Support Skill Standards: Tools for Managing Change and Achieving Outcomes. Skill Standards for Direct Service Workers in the Human Services* (ERIC Document Reproduction Service No. ED400646), Human Services Research Institute, Cambridge, MA.

[22] Wing, J., Computational Thinking, *Communications of the ACM*, 49, 3 (2006) 33-35.

[23] Wing, J.M. 2009. Computational Thinking, Presentation at *The SIGCT Forum of the ISTE Annual National Educational Computing Conference*, Washington, D.C

Building a Project Methodology to Provide Authentic and Appropriate Experiences in Computational Science for Middle and High School Students

Patricia Jacobs
Shodor
807 East Main Street, Suite 7-100
Durham, North Carolina 27701
pjacobs@shodor.org

Jennifer Houchins
Shodor
807 East Main Street, Suite 7-100
Durham, North Carolina 27701
jhouchins@shodor.org

ABSTRACT

Shodor [4], a national resource for computational science education, has successfully developed a model for middle and high school students to gain authentic and appropriate experiences in computational science. As we prepare students for the 21st century workforce, three of the most important skills for advancing modern mathematics and science are quantitative reasoning, computational thinking, and multi-scale modeling. Shodor's Computing MATTERS: Pathways to Cyberinfrastructure program [1], funded in part by the National Science Foundation Cyberinfrastructure Training, Education, Advancement, and Mentoring (CI-TEAM) program, provides opportunities for middle and high school students to explore all three of these areas. One of the wide range of programs offered through Computing MATTERS is the SUCCEED Apprenticeship Program [6].

The overall goal of the SUCCEED Apprenticeship Program [6] is to provide students with authentic and appropriate experiences in the use of technologies, techniques and tools of Information Technology (IT) with a particular focus on computational science and to produce evidence that students become proficient in these IT technologies, techniques and skills. The program combines appropriate structure (classroom-style training and project-based work experience) with meaningful work content, giving students a wide variety of technical and communication skills. The program uses innovative approaches to get students excited about computational science and enables students to grow from excitement to expertise in science, technology, engineering, and mathematics (STEM). Since its beginning in 2005, the SUCCEED Apprenticeship Program [6] has proven to be a successful model for enabling middle and high school students of both genders and of ethnically and economically diverse backgrounds to gain proficiency in STEM while learning, experiencing, and using information technologies.

Keywords

Computational Science, Mathematics, Science, Technology, Engineering, Modeling, Interactive

1. INTRODUCTION

Computing MATTERS: Pathways to Cyberinfrastructure [1], funded in part by the National Science Foundation Cyberinfrastructure Team (CI-TEAM), is an initiative of Shodor that provides a coherent continuum of other-than-schooltime activities from upper elementary grades through college for students to encounter the excitement of discovery, the power of inquiry, and the joy of learning enabled by cyberinfrastructure technologies. Computing MATTERS (Mentoring Academic Transitions Through Experiences in Research and Service) provides explorations that have content-rich context for students to learn computational science, technology, programming, and collaboration. Shodor believes that computing "matters" because quantitative reasoning, computational thinking, and multi-scale modeling are the intellectual heart of 21st century science and therefore are the essential skills needed for the future. Computing MATTERS combines the best of Shodor's efforts from workshops, apprenticeships, and internships and have proven to attract many individuals from groups otherwise underrepresented in STEM (science, technology, engineering and mathematics).

One component of Computing MATTERS is the SUCCEED Apprenticeship Program [6] which provides formal training in computational science and its practical applications within science and the world. In the Apprenticeship Program, middle and high school students are given the opportunity to gain experience in 21st century workforce skills through a coordinated set of apprentice level workshops and projects. Since the program began in 2005, the SUCCEED Apprenticeship Program has provided over 184 students - apprentices - with authentic and appropriate experiences in the use of the technologies, techniques, and tools with a particular focus on computational science and its associated areas. During their participation in the program, apprentices study, learn, and demonstrate knowledge of a wide variety of skills ranging from basic numerical methods, scientific programming, model design, validation and verification to research methods incorporating computational science.

2. PROGRAM OVERVIEW

The SUCCEED Apprenticeship Program [6] is only one of a wide range of programs provided by Shodor that helps middle and high school students encounter the excitement of discovery, the power of inquiry, and the joy of learning enabled by advanced technologies. The SUCCEED Apprenticeship Program [6] builds on Shodor's Stimulating Understanding of Computational science through Collaboration, Exploration, Experimentation, and Discovery (SUCCEED) program [3] which provides workshops to introduce middle and high school students to the technologies, techniques, and tools of computational science. Once students have shown interest and excitement for math and science by actively engaging in SUCCEED workshops, they have the opportunity to participate in the SUCCEED Apprenticeship Program [6]. In the program, upper middle and high school students work with Shodor staff and other scientists in a learning or "apprentice" mode to use computational science to conduct scientific research, create mathematical models of scientific phenomenon, and use those models to perform a variety of scientific and mathematical explorations.

The overall goal of the SUCCEED Apprenticeship Program [6] is to provide activities, support mechanisms, and mentoring to move students from an excitement for computational science and Information Technology (IT) to becoming an expert in one or more areas of computational science and associated IT components.Throughout the program, apprentices participate in authentic and appropriate experiences in the use of computational science and advanced technologies and techniques to study scientific events within the context of STEM and engage in hands-on activities and projects to demonstrate evidence that they have become proficient in these skills. In addition to the computational and technical skills, the program also enables apprentices to acquire a set of problem solving, collaboration and communication skills identified as valuable for 21st century workforce.

The program methodology involves workshops, projects and long-term, mentor- supported opportunities for upper middle and high school students in Durham, Raleigh and Research Triangle Park in North Carolina. Significant work has proceeded during the six years of the program to develop both structure and curriculum as well as to evaluate the project methodology of providing opportunities for upper middle and high school students for measuring the effectiveness of authentic experiences in computational science. Shodor has incorporated the "power of inquiry" based on the 5 E's (engage, explore, explain, elaboration, evaluate) [9] learning cycle in the SUCCEED Apprenticeship Program.

2.1 5E's Phases

2.1.1 Engagement

The classes and projects in the SUCCEED Apprenticeship Program [6] are designed to capture the student's attention, stimulate their thinking, and build upon prior knowledge. Apprentices learn and develop various skills such as computer modeling and simulation, computer programming, chemistry, web design, graphics, database design, and engineering. In addition to taking classes, apprentices have the opportunity to work on local, regional, and nationally funded projects.

2.1.2 Exploration

Throughout their participation in the program, apprentices are given hands-on assignments and projects that require them to think, plan, investigate and organize collected information. This structure provides numerous opportunities to explore and develop skills which culminates in the development of expertise in one or more areas of computational science.

2.1.3 Explanation

Apprentices are involved in analysis of their assignments and projects. They work with instructors and their mentors to ensure their understanding of concepts and processes is demonstrated through their assignments and projects.

2.1.4 Elaboration

Apprentices attend classes to learn and increase their knowledge of computational skills and complete assignments and projects to demonstrate and build on a learned skill set. Apprentices have numerous opportunities to demonstrate competence and confidence in the use of technology, critical and analytical thinking skills, and communication and leadership skills. Apprentices work with their mentor to rework assignments until they have successfully completed each one.

2.1.5 Evaluation

Throughout the program, the development of the apprentices are evaluated on participation, completion of assignments and projects, journals, surveys, and feedback from their mentor as well as staff. Through the combination of appropriate structure and meaningful work content, the SUCCEED Apprenticeship Program [6] provides outstanding opportunities for students while providing the project staff with the mechanism by which to measure the effectiveness of the program in providing authentic experiences for students to learn computational science and IT.

3. PROGRAM PARTICIPANTS

By its third year, the SUCCEED Apprenticeship Program had already surpassed its goal of reaching 100 students, called *apprentices*, in three years. Overall, a total of 184 middle and high school students have participated in the program. The program consists of students in Durham and the surrounding Research Triangle Park area with a particular emphasis on economically disadvantaged, but highly motivated students. Six years after its beginning, the SUCCEED Apprenticeship Program [6] continues to attract underrepresented groups to STEM fields through its dynamic, hands-on learning experiences and explorations. The program has not focused on a single gender or ethnic group but has successfully recruited both male and female students from diverse racial and ethnic groups and the structure of the program promotes interaction among all groups. Thus, the program does not mirror the atmosphere in many schools where students often cluster together with others of the same ethnic identity and miss the opportunity to work with students from other groups. One indicator of the need for this program is the finding that so many students came from schools where computer science is not available. In 2009, less than 50% of the students reported that their schools offered a course in computer science.

Apprentices are recruited from Shodor SUCCEED [3] workshops, local school-based programs and through Shodor's outreach programs. To participate in the program, students must complete an online application, obtain a teacher recommendation, and undergo an interview with Shodor staff. Students are interviewed and evaluated in the following areas: 1) their interest in the program, 2) commitment to the program, 3) teamwork, 4) communication: oral and written, and 6) leadership skills. The program is not limited to talented or high achieving students; the only requirement for admission is a demonstrated interest in STEM. As a result, this leads to a broad range of student outcomes; not all students reach the same level of skill or knowledge at the same pace. However, all students who remain in the program reach a basic level of achievement in a predetermined set of computational science and IT skills and many students go far beyond expectations.

Program Participant Overview:

- Rising **8th - 12th** graders
- Students are interviewed and admitted **based on their interests in STEM**
- Students and parents sign a contract **committing their support** for their student's participation in the project
- **24 apprentices** currently enrolled for 2011-2012
- Apprentices are recruited first from **Shodor SUCCEED Scholars Program Workshop** [3] and then from other **Shodor summer workshops and outreach programs**

The Apprenticeship Program [6] started in Fall 2005 with an enrollment of 17 students. By the Fall 2007, the excitement about the program had grown and we had approximately 60 applicants, twice as many applicants as we had space available for in the 2007-2009 cohort. In Summer 2011, The SUCCEED Apprenticeship Program [6] had reached a total of 160 students with 62% successfully completing the program.

Table 1 shows the number of students that successfully completed the program at the end of each year since it began in 2005. For 2012, we currently have 24 students enrolled in the program. By the end of Summer 2012, the program will have reached 184 students.

Cohort Year	Students That Completed Program
2005 - 2007	17
2006 - 2008	22
2007 - 2009†	32
2009 - 2010‡	13
2010 - 2011	15

† In Fall 2008, program changed from a two year program to a one year program

‡ During 2010, the participant enrollment was limited to 15 students

Table 1: Number of Students that Completed the Program by Cohort Year

For apprentice participants, the demographics of the Triangle region, particularly that of Durham and Orange Counties, provides us with an ample recruiting pool, especially for students historically underrepresented in STEM fields. Table 2 demonstrates that the SUCCEED Apprenticeship Program [6] has successfully attracted a diverse population of students.

Ethnicity	Percentage
African-American	27 %
Asian	21 %
Bi-racial	4 %
Caucasian	42 %
Hispanic	4 %
Indian	2 %

Table 2: Participant Ethnicity Demographics

In addition, Table 3 shows that the program has attracted a high percentage of female participants.

Gender	Percentage
Male	58 %
Female	41 %

Table 3: Participant Gender Demographics

4. STRUCTURE AND CURRICULUM

Significant work has proceeded throughout the years of the program to develop and evaluate the project methodology of bridging the excitement-expert gap opportunities for upper middle and high school students in the local area. The overall structure of the program is based on providing apprentices with a collaborative, mentor-centered environment modeled after academic research teams, combining classroom-style skills training with project-based work experience. Through the SUCCEED Apprenticeship Program [6], Shodor has demonstrated that the combination of appropriate structure and meaningful work content can provide students with knowledge in computational science, technical and communication skills, personal interest and motivation, and more pragmatically, a resume-documented depth of experience to pursue an IT-intensive career path.

When the program started in 2005, it was a two-year program with apprentices focusing on an independent learning, self-paced instructional model. Apprentices were provided with written tutorials [5] for learning modeling and simulation as well as other IT skills. Students worked through the tutorials independently with help as needed from peers or mentors. As each tutorial was completed, or when the student felt ready, he/she was given a Challenge problem, to demonstrate attainment of the skill. Although many apprentices thrived in this environment, only about half of them attained the number of IT skills that had been expected.

Based on the stagnant progress of the apprentices, the program was redesigned in 2007 to a one year program to provide more structure, require firmer time commitments, group projects and require that each apprentice spend more time with their mentors. As a result, the current structure, which is based on the program redesign, is a successful model involving workshops, individual assignments and team projects to promote the development of understanding of concepts as well as technical proficiency that cannot be reached in short, intensive courses. The current structure has proven that continuous involvement over a period of time reinforces learning and provides sufficient practice in computational science and IT skills for apprentices to reach a level of expertise.

4.1 Program Structure

With the current design, each apprentice spends approximately 360 hours in the program over the course of one year (20 hours a month during the school year and 6 weeks during the summer). The 20 hours per month consist of attending 3 workshops per month (Saturdays from 9:00am-2:00pm) and 2 hours of in-office time during the week. In the workshops, computational science and STEM skills are taught in the context of a problem to be solved using hands-on activities and the latest in advanced technology.

Apprentice schedules are organized into 3 modules, ranging from two to three months in length. Apprentices must complete assignments to help them learn and practice using new skills and group projects to demonstrate their knowledge of a given skill set. They must also complete quality assurance through verification and validation testing for all assignments and projects. Participants are asked to re-work any assignments or projects that are incomplete. Our goal is for each student to not only learn but also to become proficient in a given skill set.

The structure of the program focuses on the following five areas:

1. **Teaching and supporting the appropriate and authentic use of IT-related technologies, techniques, and tools, with a particular focus on computational science and its associated areas.**
   Throughout the program, apprentices attend workshops, complete assignments and work in collaborative groups on projects. Apprentices attended workshops to gain experience and develop expertise in one or more areas of computational science and associated uses of the technologies, techniques, and tools of IT within the context of STEM. Topics involved general uses of computational science, basic numerical methods, scientific programming, model validation and verification and research methods incorporating computational science. Apprentices have a structured curriculum to learn IT skills such as agent modeling, web design, programming and graphics. Their work is organized into modules of two to three month durations. After a module is completed, apprentices worked in teams to complete assigned projects.

2. **Providing mentors to define individual goals and timelines as well to provide guidance through technical difficulties.**
   In addition to students attending classes, the SUCCEED Apprenticeship program continues to focus on mentoring students. The SUCCEED Apprenticeship program seeks to implement a 'true' apprenticeship program where young people learn from working with and learning from those with more experience. Shodor has approximately 18 staff who have a range of expertise in computational physics, biology, chemistry, and math as well as computer science, system administration, graphics and web design. Each staff member is assigned to mentor 2-3 students. Mentors monitor work progress as well as skill development for individual apprentices. In addition, mentors are responsible for overall research team dynamics, distribution of work and project oversight. Students are required to meet one hour per month with their mentor. Communication between mentors, program coordinator and parents is also ongoing throughout the program.

   In addition to monthly meetings, mentors track progress and skill development of apprentices by reviewing the apprentices weekly reflections (questions students have to answer weekly), progress on individual assignments, and projects. Many students receive additional mentoring from staff when they need help understanding and/or completing out-of-class assignments and projects.

3. **Providing opportunities for apprentices to work on meaningful projects - local, regional and nationally funded projects.**
   Throughout the program, apprentices are provided with the opportunity to work on local, regional and nationally funded projects. We continue to partner with local organizations to provide real world experiences for our apprentices' projects. These projects range from working with with local organizations to learning and developing skills for building Shodor's award winning resources such as the Computational Science Education Reference Desk (CSERD) [8] and Interactivate [2] projects. The projects are done as a learning process and thus require intensive guidance to ensure quality workmanship.

4. **Providing instruction and opportunities to practice a wide variety of communication skills, including working effectively in a group, interacting with customers and clients, teaching younger students about the technologies, and exercising leadership.**
   The SUCCEED Apprenticeship program provides opportunities for apprentices to practice a wide variety of communication skills, including working effectively in a group, interacting with clients as well as teaching STEM workshops for younger students. Apprentices have to prepare presentations and present several projects they worked on to demonstrate their knowledge of the skills they learned. These presentations help the apprentices improve their communication skills.

   Some apprentices are also given the opportunity to teach workshops about computational science technologies and tools to younger students taking Shodor workshops. In the past, apprentices have taught workshops from system and agent modeling to system administration for both middle and high school students. In addition to teaching, Shodor apprentices are encouraged to exercise leadership by learning how to be effective mentors to other students in their own peer group.

5. **Providing formal and informal opportunities in critical thinking, including data retrieval, data organization and analysis, application of evidence-based reasoning, problem-solving, creative thinking and decision making.**
The SUCCEED Apprenticeship Program provides many opportunities for apprentices to use and demonstrate critical thinking skills. Apprentices must attend workshops to learn computational science and STEM skills. Apprentices are given assignments to practice and hone new skills. In addition, apprentices complete group projects to demonstrate a given skill set. Apprentices work in project teams to integrate various technical skills needed for the completion of their projects. These projects require that each apprentice use a variety of innovative skills as well as problem solving. Apprentices learn and use skills such as modeling, PHP, MySQL, CSS, HTML, Javascript and Google Maps for their projects.

4.2 Curriculum

The SUCCEED Apprenticeship Program [6] workshops cover a wide range of topics such as Unix, computational modeling, operating system basics, and web design using HTML and Cascading Style Sheets. In addition, during the summer, apprentices work on projects such as the Computational Science Education Reference Desk (CSERD) [8] and Project Interactivate [2] as well as projects for community organizations. Apprentices also document and assist in teaching SUCCEED [3] workshops at Shodor.

For the Apprenticeship Program [6], Shodor has developed a curriculum that consists of materials and lesson plans designed specifically for use in training middle and high school students in the uses of computational science and technology. The materials are divided into 50-minute lesson plans. The materials are free, adaptable, and available to teachers and students outside of Shodor's programs. The program curriculum and materials are located online [7].

The apprentice workshop curriculum covers four basic areas:

1. **General Skills**

   - *How Do You Know?*: Students are challenged to think about how they know what they know as a foundation for all other classes in our curriculum. This is the first introduction to Verification (solving the problem right) and Validation (solving the right problem).
   - *Math and Verbal Skills*: Students choose an activity from Interactivate or the National Science Digital Library and learn how to use this activity and the mathematics that it covers. Each activity has a set of Exploration Questions, which the apprentices complete and submit for assessment.
   - *Office Ethics*: Role playing, lectures, presentations, guest speakers, live illustrations, and student interactivity all help the student grasp the concept of office ethics to experience how to work in a business office with integrity and loyalty.

2. **Scientific Modeling**

   - *AgentSheets and NetLogo*: An introduction into agent modeling, using two complementary approaches. Students are introduced to agent behaviors and emergent properties while building a population-dynamics model where agents exchange colors.
   - *Vensim*: An introduction to the concepts of systems modeling. After building a model "on the board" and thinking about the flow of the model away from the computer, students are led through the steps to build and run the model on the computer.
   - *Excel and Spreadsheets*: An introduction to some of the basics of Excel and how to use formulas, slider bars, and graphs to create a model of population growth. This model takes into account factors such as birthrate and carrying capacity. Then students explore how to build a Susceptible, Infected, Recovered (SIR) Model of the spread of a disease, by finding and graphing the number of susceptible, infected, and recovered people in the model over time.

3. **Graphics and Web Design**

   - *Introduction to HTML*: Provides students with a background in Hyper Text Markup Language, and knowledge of basic markup languages and tags. Students learn how to make minor edits to the source of an existing HTML page, and view and understand the results. Students then explore HTML tags and create a simple web page. They learn how to use formatting tags, lists, images, and link tags.
   - *Introduction to CSS*: An introduction to Cascading Style Sheets and why it is useful. Assuming a previous knowledge of HTML, students explore a variety of web designs that use CSS for content presentation and management, learning how to add CSS to the header of an existing page.
   - *Introduction to Visualization*: Students first learn GIMP and the concept of raster graphics. It covers basic image manipulation concepts, filters, and saving for the web. Students will learn what a raster/bitmap graphic is as opposed to a vector graphic. They manipulate images in GIMP, selecting, moving, retouching. Then Inkscape and vector graphics are presented. Students explore simple vector graphics, create basic shapes and text, as well as use the Fill and Stroke palette to change the color of an object.

4. **Programming and System Administration**

   - *Introduction to Programming*: After mastering modeling and visualization, the Shodor approach then introduces the high level idea of programming concepts. It is not specific to a particular computer language, environment, or other such limitations. The goal is to teach the ideas behind programming, not to teach programming itself.

- *Language Specific Workshops*: Multiple workshops are offered for a range in languages including scripting (Python, PHP, Perl), modeling (NetLogo), and more general-purpose languages (C, Fortran, C++, Java). Each workshop serves as a review of modeling while being an introduction into scientific programming. Students are introduced to agent behaviors and emergent properties while building a variety of models across various scientific areas. The goal is to emphasize similarities of the languages while recognizing underlying comparative strengths and weaknesses of procedural and object-oriented approaches.
- *System Administration*: This workshop introduces students to the idea of operating a computer via the Unix command line, as well as some of the basic concepts required to understand the Unix way of doing things. Command language, scripting, security, data management, remote monitoring, and batch execution are covered.

In addition to learning computational science and IT skills, a math component has been incorporated into the Apprenticeship Program curriculum [7]. Finding that many students struggled with certain aspects of the program because of weak math skills, Shodor dedicated one Saturday a month to teaching and helping apprentices learn and enhance their math skills. The math component not only focuses on helping apprentices improve their math skills but also helps them learn math as an integral part of their education in computational science. For example, the apprentices learn the mathematics behind a disease spread model, such as probability, functions, and data analysis. The program helps apprentices gain a deeper understanding of mathematics as they learn to apply it to real world situations through the use of computers.

5. EVALUATION

Evaluation of the SUCCEED Apprenticeship Program [6] focuses primarily on providing authentic experiences for middle and high school students to develop computational science and IT skills and the assessment of each student in the use of those skills. Evaluation is an ongoing process embedded into all aspects of the program and is related to the pedagogical approach of engaging students in experiences directed toward acquiring IT skills and knowledge. Evaluation tools used required that each apprentice maintain a journal, completed assignments, and participated as part of a team in completing a succession of projects that required application of learned skills. Apprentices are required to rework assignments and projects until they produced work that met standards set by their mentors.

Since the program began in 2005, an extensive evaluation process has helped us continually improve the effectiveness of the program's structure and curriculum. Evaluation data includes skills assessments, students participation, completion of assignments, personal journals, responses to routine surveys, and feedback from staff. Additionally, periodic interviews with students are conducted to track their attitudes and career plans. Shodor staff continually monitor, observe, and interact with apprentices seeking ways to improve the program.

As a result of their participation in the SUCCEED Apprenticeship Program [6], students showed an improvement in their IT and soft (presentation, teamwork, teaching, etc.) skills. In addition to informal assessments and feedback from the Shodor staff mentors, apprentices performed self-assessments at the beginning of the program, beginning of summer, and end of summer. The apprentices rated themselves on a scale of 1-5 on a range of IT skills. A rating of 4-5 is deemed an expert and able to teach the skill to others. A rating of 3 indicates competent use of the skill in projects assigned and completed. A lower rating indicates that an apprentice can use the skill with help but has not yet become competent to apply it independently. Table 4 shows that students in the 2010-2011 cohort had an improvement of between 16% to 75% in their computational science and IT skills when they completed the program.

Skill	Percentage Improved
HTML/CSS	25 %
Graphics	25 %
PHP	16 %
MYSQL	50 %
Subversion	75 %
Object-Oriented Programming	75 %
AgentSheets	33 %
Netlogo	33 %
Excel	33 %
Vensim	16 %
Technical Writing	25 %
Unix Commands	42 %
Team Work	16 %
Presentation Skills	16 %
Self Confidence	16 %
Teaching	16 %

Table 4: Percent of Students that Improved from Level 3 or Lower to Level 4 or Higher

6. SUMMARY

The SUCCEED Apprenticeship Program [6] provides a structured learning environment that evolved as experience in managing the program was gained over the program's six year period. During the program, apprentices learn skills such as computer modeling and simulation, HTML, CSS, Graphics, PHP, MYSQL, AgentSheets, NetLogo, Excel, System Administration, as well as Technical Writing and Presentation skills. The structure includes classes, homework, teamwork, and independent work. Apprentices are held accountable for time on a task, attendance at classes, project completion, and development of workplace skills.

An evaluation of the value added and the measurability of the Apprenticeship program [6] is assessing the extent to which the appropriate structure and meaningful work content effectively develops students to become an expert in the areas of computational science and associated IT components. We continue to evaluate the program's success and to improve the effectiveness of the program's structure and curriculum.

Since the program began in 2005, the evaluation process has helped us continually improve the effectiveness of the program's structure and curriculum. The Apprenticeship program [6], as it has developed over the last six years, enables students of diverse backgrounds to gain computational science and IT proficiency. We have shown that students of diverse racial and ethnic backgrounds have gained proficiency in a wide range of skills. The percentage of underrepresented groups remains high among the apprentices who persist through the entire program. In addition, many students who have participated in the program expect to have IT-related careers and thus will take their places in IT-intensive workplaces as they move into the adult workforce. An interesting and unusual aspect of the population of apprentices in the program is the high percentage of African American females. We believe that interaction between students of diverse ethnicities and genders in a learning environment where respect for all students, as well as mentors and instructors, is the norm and is expected and required has been an important aspect of the program. The results of the assessment by mentors who have been actively engaged with apprentices indicates that, with very few exceptions, all apprentices who persisted in the program are able to use IT skills learned at an expert or practical level in any environment where these skills are required.

The following summarizes the evaluation results and effectiveness of the SUCCEED Apprenticeship Program [6] methodology:

- The program was notably successful in recruiting students of diverse racial and ethnic backgrounds and both genders. Since the program began in late 2005, a total of 184 students, including the current cohort, have been admitted into the program and a total of 123 students will have successfully completed the program by Summer 2012. The project has been successful not only in recruiting a diverse group of students but also in maintaining diversity in the groups who have persisted through the full program.

- The majority of those who completed the program have maintained a strong interest in science, technology, engineering and math as evidenced by their plans to pursue careers in one of the STEM areas. Persistence in intention to pursue a career in STEM was used as an indicator of enthusiasm and was measured by three successive interviews or surveys of apprentices conducted (a) soon after entry into program, (b) midway through program, and (c) at completion of the program. In addition, there were others who reported that they had decided to pursue a career in STEM or had been prompted to rethink their career plans.

- Apprentices who completed the program are equipped to continue their education and eventually enter the workforce with unusual proficiency in such IT skills as Java, Unix, HTML/CSS, Graphics, PHP, Object-Oriented Programming, MYSQL, AgentSheets, NetLogo, Excel, and Vensim. Those who complete the program become proficient in a number of more advanced IT skills that require a higher level of planning, thinking, and execution.

- In addition to acquiring technical skills, apprentices also acquired significant skills in professional ethics, technical writing, presentation skills, team work and self confidence.

- Data and observations collected over the six years have allowed us to identify certain elements of the program that have promoted and enabled attainment of proficiency in IT by middle and high school students who have an initial interest in IT and STEM. The main elements are: the long term involvement of participants, the recruitment policy of requiring only interest in learning, using structured learning experiences, and mentoring of students with ongoing evaluation.

Shodor [4] programs provide opportunities for students to progress from excitement to experience to expertise through workshops, apprenticeships and internships. Our students gain skills that are equipping them for the 21st Century workforce. Many students from the SUCCEED Apprenticeship Program [6] have continued their interest in STEM by becoming interns at Shodor and accomplishing great achievements. Here we will mention several students who progressed from workshop student to apprentice and then to intern. It should be noted that all of these students have received accolades for their achievements and are also all from groups historically underrepresented in STEM.

In 2010, The National Center for Women and Information Technology (NCWIT) selected Shodor intern, Ada Taylor, as a recipient of the National NCWIT Award for Aspirations in Computing and Krista Katzenmeyer, another Shodor intern, was a semi-finalist for the award. Both Ada and Krista started as workshop students that developed into apprentices in the SUCCEED Apprenticeship Program [6] at Shodor [4]. Additionally, Ada participated in the Conrad Spirit of Innovation Competition with her team, called Unisecurity, from the North Carolina School of Science and Math, taking first place in the competition. Held at the NASA-Ames Research Center, the competition was intended to develop mobile applications based in security.

Yet another Shodor intern, Alex Revelle, received the Student Leadership Award in Science, Mathematics, and Technology Education by the North Carolina Science, Mathematics, and Technology (SMT) Education Center. Alex began his work at Shodor as a workshop student, completed Shodor's Apprenticeship Program [6] and ultimately became an intern. As an intern, Alex became involved in Shodor's outreach efforts by helping to teach workshops in the surrounding community. Like Alex, intern Cameron Aviles became involved with teaching Shodor outreach workshops. Cameron and another apprentice from his cohort year, Maya Gouw, both won the Duke University Durham Student of the Week award for their outstanding work.

The SUCCEED Apprenticeship model is a straightforward application of good educational principles, carried out in a work-oriented environment by staff with experience and expertise in information technology as well as a deep understanding in the areas of science and mathematics. This model does not offer a quick and easy solution to training middle and high school students to be competent in technical and/or workplace skills but it attests to the value and efficacy of making a long term commitment to a group of students diverse in both background and ability and helping them develop the skills they will need for the 21st century workforce.

7. ABOUT SHODOR

Shodor [4] is a non-profit education and research organization located in Durham, N.C. Shodor is dedicated to mentoring and providing hands-on learning for students to master 21st century workforce skills, building high quality on-line interactive tools and curriculum for mathematics and science, and supporting the professional development of educators by providing training on the appropriate and effective use of computer technologies and content in the classroom. Special emphasis is placed on hands-on learning for students to gain experience in computational technology by assisting in the development of models and simulations used by educators.

Shodor's online educational resources and materials reach upward of 4 million page views per month. In addition to developing interactive models, simulations, and educational tools, Shodor serves students and educators nationwide directly through workshops and other hands-on experiences. Shodor offers innovative workshops helping faculty and teachers incorporate computational science into their own curricula or programs. For students from middle school through undergraduate, Shodor offers workshops, apprenticeships, internships and outreach programs that explore new approaches to math and science education through computational science.

8. ACKNOWLEDGEMENTS

The funding support for Project SUCCEED [3] and the SUCCEED Apprenticeship Program [6] was provided by the National Science Foundation's Innovative Technology Experience for Students and Teachers program (ITEST 05-25178), the Burroughs Wellcome Fund, the Department of Education, and the Cyberinfrastructure Training, Education, Advancement, and Mentoring program (CI -TEAM 1041391). The Project PI, Dr. Robert Panoff, is founder and Executive Director of Shodor and has been a consultant at several national laboratories. He is also a frequent presenter at NSF-sponsored workshops on visualiztion, supercomputing, and networking, and continues to serve as consultant for the education program at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign. He has served on the advisory panel for Applications of Advanced Technology program at NSF. Dr. Panoff has dedicated his life to promoting and improving science and mathematics education in North Carolina and throughout the country. The Project Evaluator, Dr. Ann Howe, now an independent evaluation consultant, retired as Chair of the Department of Curriculum and Instruction at the University of Maryland after serving as a head of the Department of Mathematics and Science Education at North Carolina State University and Professor of Science Teaching at Syracuse University. She has had extensive training in both quantitative and qualitative research methods which she used in her own research and as advisor to doctoral students. Over the past twenty years, she has had broad experience as evaluator on numerous projects in science, mathematics and technology education sponsored by NSF and other agencies and foundations.

9. REFERENCES

[1] Computing MATTERS: Mentoring Academic Transitions Through Experiences in Research and Service. http://www.computingmatters.org/.

[2] Project Interactivate. http://www.shodor.org/interactivate/.

[3] Project SUCCEED: Stimulating Understanding of Computational science through Collaboration, Exploration, Experimentation, and Discovery. http://www.shodor.org/succeed.

[4] Shodor: A National Resource for Computational Science. http://www.shodor.org/.

[5] Shodor Tutorials. http://shodor.org/tutorials/.

[6] SUCCEED Apprenticeship Program. http://www.shodor.org/succeed/apprenticeships/.

[7] SUCCEED Apprenticeship Program Curriculum. http://www.shodor.org/succeed/curriculum/apprenticeship/.

[8] The Computational Science Education Reference Desk. http://www.shodor.org/refdesk/.

[9] R. Bybee, J. Taylor, A. Gardner, P. Scotter, J. Powell, A. Westbrook, et al. The BSCS 5E instructional model: Origins, effectiveness, and applications. Technical report, Office of Science Education, National Institutes of Health, 2006.

A Web Service Infrastructure and its Application for Distributed Chemical Equilibrium Computation

Subrata Bhattacharjee
San Diego State University
5500 Campanile Drive
San Diego, CA 92182-1323
+1 (619) 594-6080
subrata@thermo.sdsu.edu

Christopher P. Paolini
San Diego State University
5500 Campanile Drive
San Diego, CA 92182-1323
+1 (619) 594-7159
paolini@attila.sdsu.edu

Mark Patterson
San Diego State University
5500 Campanile Drive
San Diego, CA 92182-1323
+1 (858) 735-5769
markooo@gmail.com

ABSTRACT

W3C standardized Web Services are becoming an increasingly popular middleware technology used to facilitate the open exchange of data and perform distributed computation. In this paper we propose a modern alternative to commonly used software applications such as STANJAN and NASA CEA for performing chemical equilibrium analysis in a platform-independent manner in combustion, heat transfer, and fluid dynamics research. Our approach is based on the next generation style of computational software development that relies on loosely-coupled network accessible software components called Web Services. While several projects in existence use Web Services to wrap existing commercial and open-source tools to mine thermodynamic data, no Web Service infrastructure has yet been developed to provide the thermal science community with a collection of publicly accessible remote functions for performing complex computations involving reacting flows. This work represents the first effort to provide such an infrastructure where we have developed a remotely accessible software service that allows developers of thermodynamics and combustion software to perform complex, multiphase chemical equilibrium computation with relative ease. Coupled with the data service that we have already built, we show how the use of this service can be integrated into any numerical application and invoked within commonly used commercial applications such as Microsoft Excel™ and MATLAB® for use in computational work. A rich internet application (RIA) is presented in this work to demonstrate some of the features of these newly created Web Services.

1. INTRODUCTION

Numerical determination of the equilibrium state mass fractions of gaseous and condensed matter is frequently needed in combustion simulations that model chemically reacting flows. The numerical method most often used to calculate an equilibrium distribution is based on minimizing a system's Gibbs free energy function with constraints. Several techniques for Gibbs energy minimization have appeared in the literature such as the tangent line/plane procedure suggested by Michelsen [1, 2], the maximum area method developed by Eubank et al. [3] and Elhassan et al. [4, 5], and the equal area method of Eubank and Hall [6], Shyu et al. [7, 8], and Hanif et al. [9, 10].

The total Gibbs energy G of a system composed of m species is given by

$$G = \sum_{j=1}^{m} \mu_j n_j \tag{1}$$

where μ_j is the chemical potential of the jth species and a function of the system temperature T, pressure p, and number of moles of each component species, $n_{i \neq j}$, $\forall i$. From (1) and using the definition of the chemical potential for an ideal gas species j, we have

$$\mu_j = \mu_j^{\circ} + RT \ln\left(\frac{p_j}{p^{\circ}}\right) \tag{2}$$

where p° is the standard state pressure of 1 bar and p_j is the partial pressure of species j. The minimum stationary point of (1) will be the vector of species molar values $\vec{n}$ where dG vanishes. Differentiating (1), we obtain

$$dG = \sum_{j=1}^{m} n_j d\mu_j + \sum_{j=1}^{m} \mu_j dn_j \qquad (3)$$

From the isothermal, isobaric Gibbs-Duhem equation we know that

$$\sum_{j=1}^{m} n_j d\mu_j = 0 \qquad (4)$$

and so we seek the unique vector $\vec{n}$ such that

$$\sum_{j=1}^{m} \left[\frac{\mu_j^\circ}{RT} + \ln n_j - \ln n + \ln\left(\frac{p}{p^\text{o}} \right) \right] dn_j = 0 \quad (5)$$

where n_j is the number of moles of species j and $n = \sum_{j=1}^{m} n_j$ is the total number of moles in the equilibrium composition. Solving (5) amounts to solving a nonlinear constrained minimization problem. The traditional numerical method used for solving optimization problems [11] of this type is the method of Lagrange multipliers using an iterative Newton-Raphson technique for solving the resulting set of nonlinear equations. Both NASA CEA [12] and STANJAN [13] employ this method of element potentials. Both of these codes were written in Fortran and cannot be easily incorporated into other codes. Paolini and Bhattacharjee [14] modified the CEA algorithm and developed an object-oriented Java code for computing an equilibrium distribution by minimizing a system's Gibbs function. Front-end Java applets with easy-to-use graphical user interface were developed to run this equilibrium code through a browser.

In this work, we extend our object oriented code into a publicly accessible Web Service based module that can be called from most contemporary programming environments and integrated into many applications written in a variety of computer languages. This service can be invoked as a reusable third party software component by a thermal science researcher when developing custom applications. As a result, the researcher is freed from having to worry about implementing his or her own code for computing chemical equilibrium. When a desired equilibrium distribution is needed, the developer need only insert the requisite code to remotely discover and dynamically invoke the Web Service.

As a demonstration of the power of this newly created service, we present a few case studies: (a) Use of this service from third party applications such as Microsoft Excel or MATLAB, and (b) Use of this service to create a user friendly rich internet application (RIA) to simulate a steady-state combustion chamber.

Web Services: Web Services extend the paradigm of object-oriented programming to the network whereby a single software application is composed of loosely-coupled modules that execute on autonomous networked systems. Recent efforts by Dong et al. [15], Truong et al.[16], and Paolini and Bhattacharjee[17] have shown the strength of using Web Service technology in chemoinformatics applications to facilitate the organization and retrieval of chemical data. Furthermore, advances in collaborative cyberinfrastructure for developing predictive models for chemically reacting systems have been spearheaded by Frenklach et al. through the open-source Process Informatics Model (PrIMe) project [18]. Additionally, the Cantera [19] package by Goodwin provides an open-source, object-oriented suite of software tools to aid in simulating problems in combustion and can be called to solve equilibrium problems from within FORTRAN, MATLAB, and Python scripts. Conceptually, Web Services can be thought of as a middleware technology that provides platform independent methods of facilitating machine-to-machine communication. This communication is accomplished through the use of Web application servers that deliver software services to client computers over a computer network. Web application servers are different than traditional Web servers in that an application server will manage and invoke user supplied code when requested to do so from a client system. Application servers publish a set of publicly available or exposed operations available to client applications using a standardized interface language called WSDL or *Web Services Description Language*. A WSDL specification is a structured XML document, often publicly accessible on the Internet, that client applications access remotely to determine the exact name, argument specification, and return type of a particular operation. The interface of a Web Service is separate from its actual implementation and the practice of separating interface from implementation is a core characteristic of all Web Services. While the interface of every Web Service is specified in standardized WSDL format, the implementation can be in any programming language. This separation allows Web Services to be platform-independent and provide transparent and modular access to pre-existing software services.

As a result of our work, there is now no need for a researcher to design his or her own equilibrium solver or locate a suitable third-party library. Our equilibrium Web Service can essentially be thought of as code that "plugs in" to existing software and takes advantage of distributed computational resources over the Internet. This style of software development based on orchestrating loosely-coupled and distributed software services is called a *Service Oriented Architecture* or SOA. Adopting an SOA approach to building combustion applications will have a sweeping impact on research and teaching in the thermal sciences as developers are able to construct new software tools that build upon an ever expanding collection of independent and modular Web Services.

2. EQUILIBRIUM COMPUTATIONS

We developed a chemical equilibrium Web Service that exposes an operation to calculate and return the equilibrium distribution of the products of an arbitrary reaction at a defined temperature and pressure. The input parameters are the reaction temperature in Kelvin, pressure in kilopascals, a comma and colon delimited list of reactants, and a comma delimited list of allowable products. The list of reactant species is specified using the format *moles:formula* where *formula* is a chemical formula of a reactant species using the Hill naming system and *moles* is the quantity of the respective species in the reactant mixture. To illustrate how our Web Service can be used, consider the standard problem of ammonia synthesis by means of the well known Haber process,

$$N_2(g) + 3H_2(g) \rightleftharpoons 2NH_3(g) \qquad (6)$$

The Haber process is carried out at about 520°C and 500 atm in the presence of an iron-molybdenum catalyst. The catalyst increases the rate of the reaction but does not affect the reaction stoichiometry. At equilibrium, the mole fractions of nitrogen, hydrogen, and ammonia are approximately 17%, 50%, and 33%.

Our Web Service *solve* operation returns the unique distribution of product species that corresponds to a reaction at a fixed temperature T in Kelvin and pressure p in kilopascals. To invoke the solve operation, a SOAP message body is constructed by a client process that includes a definition of the four required input parameters shown in Figure 0. In Figure 0 we see four XML elements with names that correspond to the four parameters required to invoke the solve operation. Note the format used to specify reactants and products: reactants are given as a comma delimited list of number-formula pairs, each delimited by a colon, and the products are given as a comma delimited list of just chemical formulas. The solution returned by the Web Service upon calling the solve operation is encoded in JavaScript Object Notation (JSON).

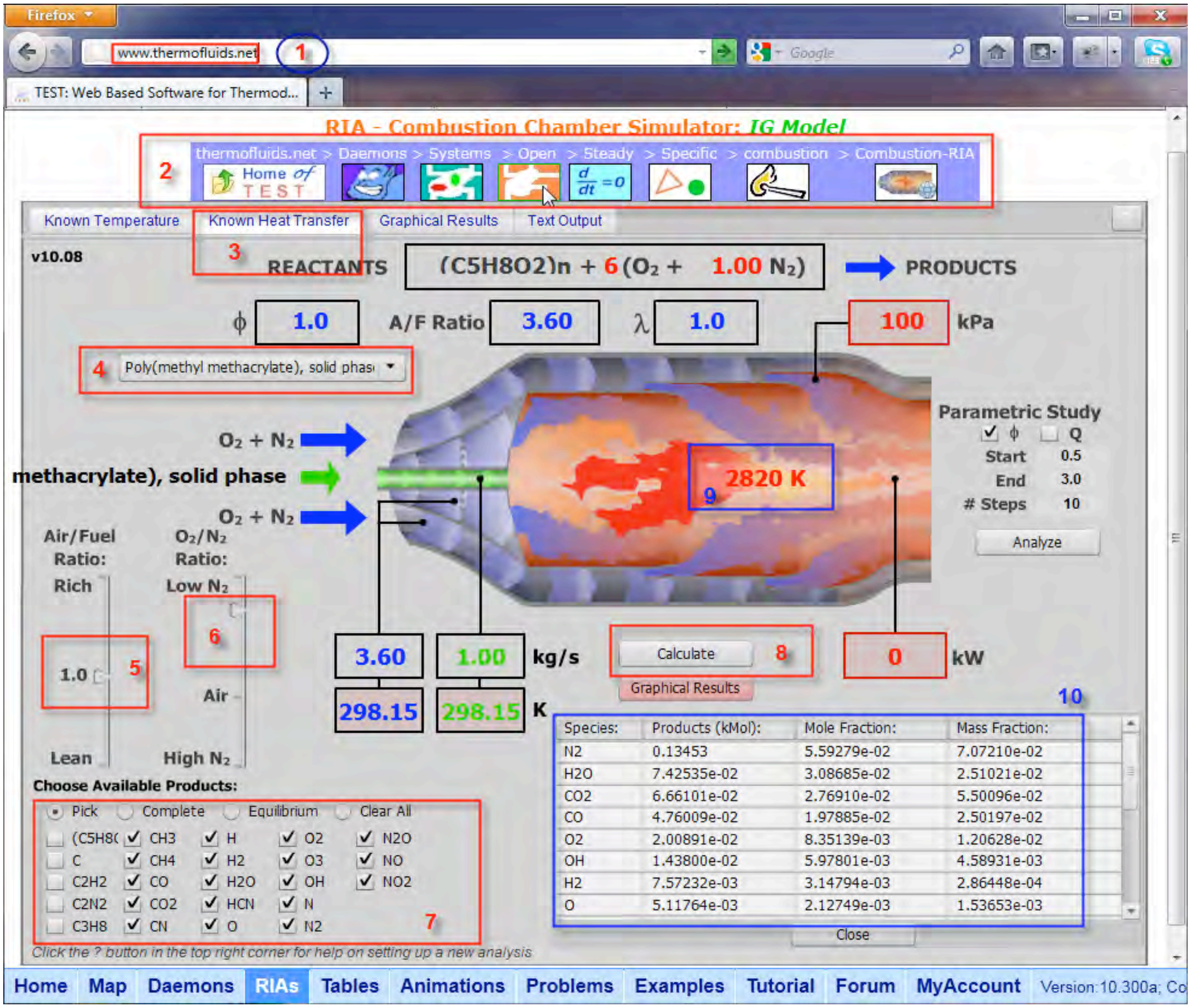
```
double solve(double T, double P,
             string reactants,
             string products)
```

Figure 0. The chemical equilibrium Web Service exposes an operation named solve that returns the equilibrium state distribution of product species in JSON format.

```
<equ:solve
 xmlns:equ="http://cheqs/">
 <temperature>773.15</temperature>
 <pressure>50662.5</pressure>
 <reactants>1:N2,3:H2</reactants>
 <products>N2,H2,NH3</products>
</equ:solve>
```

Figure 0. The body of a SOAP message corresponding to the Haber process for ammonia synthesis given in reaction (6).

Web Services can be invoked within most contemporary programming languages such as FORTRAN, C++, and Java. Programs written in these languages invoke the solve operation to compute the equilibrium distribution. The code uses the dynamic dispatch interface (DDI) provided by JAX-WS. Complete example code showing how to parse the SOAP response using the Simple API for XML (SAX) is available from the Tools section of http://cheqs.sdsu.edu/. As discussed in the previous section on Web Services, the DDI can be used in conjunction with a service registry to dynamically discover Web Services and invoke them on-the-fly. In the Java environment, Web Service discovery is frequently accomplished using the Java API for XML Registries (JAXR). Complete Java code that shows how software can dynamically discover our chemical equilibrium Web Service using a classification name and then call the service using the DDI technique is available from the Tools section of http://cheqs.sdsu.edu/.



(C5H8O2)n + 6 (O2 + 1.00 N2)

ϕ 1.0 A/F Ratio 3.60 λ 1.0 100 kPa

Poly(methyl methacrylate), solid phas

2820 K

Parametric Study: ☑ ϕ □ Q; Start 0.5; End 3.0; # Steps 10

3.60 kg/s, 1.00 kg/s; 298.15 K, 298.15 K; 0 kW

Air/Fuel Ratio: 1.0

Choose Available Products:

● Pick ○ Complete ○ Equilibrium ○ Clear All

□ (C5H8(☑ CH3 ☑ H ☑ O2 ☑ N2O
□ C ☑ CH4 ☑ H2 ☑ O3 ☑ NO
□ C2H2 ☑ CO ☑ H2O ☑ OH ☑ NO2
□ C2N2 ☑ CO2 ☑ HCN ☑ N
□ C3H8 ☑ CN ☑ O ☑ N2

7

Click the ? button in the top right corner for help on setting up a new analysis

Species:	Products (kMol):	Mole Fraction:	Mass Fraction:
N2	0.13453	5.59279e-02	7.07210e-02
H2O	7.42535e-02	3.08685e-02	2.51021e-02
CO2	6.66101e-02	2.76910e-02	5.50096e-02
CO	4.76009e-02	1.97885e-02	2.50197e-02
O2	2.00891e-02	8.35139e-03	1.20628e-02
OH	1.43800e-02	5.97801e-03	4.58931e-03
H2	7.57232e-03	3.14794e-03	2.86448e-04
O	5.11764e-03	2.12749e-03	1.53653e-03

Figure 0. Screenshot of the combustion chamber simulator calculating the equilibrium flame temperature of PMMA combustion.

As previously stated, Web Services can be invoked from other languages besides Java such as FORTRAN, C, and C++. Because many present day software packages developed for combustion applications are written in FORTRAN, there is much interest in being able to invoke our chemical equilibrium Web Service from within a preexisting FORTRAN program. The two basic ways this can be accomplished is to use a FORTRAN compiler that provides native support for SOAP/XML Web Services or use a third party SOAP library. The former is provided by compilers such as the popular Lahey/Fujitsu Fortran v7.1 [20] while the later can be accomplished through the use of libraries like gSOAP [21].

The Combustion Chamber Simulator: An interactive rich internet application is created to demonstrate the usefulness of the Web Service in making complex calculations simple and user friendly. The front-end of the simulator is written in Adobe^{TM} flash, which calls Web Services in the background. The simulator is integrated with our thermodynamic Web portal TEST (The Expert System for Thermodynamics) www.thermofluids.net, is freely accessible, and used by more than 25,000 registered users.

To illustrate the use of the RIA, suppose we are interested in the combustion of PMMA (poly methyl metacrylate), specifically, the stoichiometric equilibrium flame temperature when condensed PMMA is burned with a 50-50 mixture (by volume) of oxygen and nitrogen at 1 atm. Using any browser (IE, Firefox, Chrome, Safari, etc.) equipped with the Flash plug-in, launch the simulator (see Figure 0) by clicking the RIAs tab on the task bar and selecting the Combustion Chamber RIA (Label-1 in Figure 0). A Flash animation of a combustion chamber is loaded with its address displayed in a hierarchical manner relative to other resources (Label-2). For calculation of adiabatic temperature, we select the *Known Heat Transfer* tab (Label-3). Note that the pressure in the combustion chamber is set at 100 kPa by default and the heat transfer is set at 0, each of which can be edited as desired. From the fuel selector (Label-4), choose PMMA (solid phase). The reactants side of the reaction is displayed for the default reaction of the fuel burning with theoretical amount of air. Using the first slider bar (Label-4) you can change the equivalence ratio and see its effect on the reaction as it is made leaner or richer. Using the second slider bar (Label-5) you can change the makeup of the oxidizer from pure oxygen to very low level of oxygen in an oxygen-nitrogen mixture. The stoichiometric coefficient of nitrogen or the oxidizer (in red) can also be directly edited. For each settings, reaction parameters such as the equivalence ratio Φ, oxidizer/fuel ratio, percent theoretical oxidizer used λ, amount of oxidizer (in kg/s) entering the combustion chamber for a fuel mass flow rate of 1 kg/s, etc., are dynamically updated.

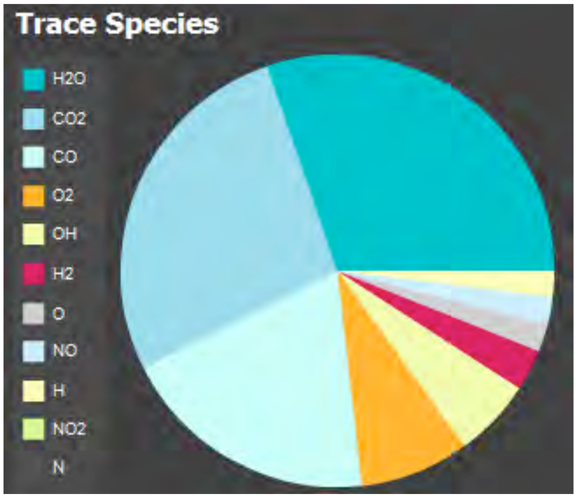


Figure 0. Graphical distribution of the trace species in the equilibrium composition of Figure 0.

Once the reactants mixture is set – in this case PMMA reacting with a theoretical amount of a 50-50 oxygen-nitrogen mixture, we need to specify the nature of the reaction. The RIA offers several options (see Label 7): complete combustion whereby all carbon and hydrogen atoms are fully oxidized, equilibrium combustion where most common combustion products are automatically selected, and a customized selection of products species. If the *Complete* radio-button is selected, the adiabatic flame temperature is calculated as 3998 K. When the *Equilibrium* button is selected, the 23 most common combustion species are automatically selected. By clicking the *Pick* button, the list can be customized. With the selected species shown in Figure 0, when the *Calculate* button (Label 8) is pressed, a Web Service call is made to the CHEQS server. Based on the number of products species chosen, the calculations take anywhere between 5 s to 20 s. The calculated equilibrium flame temperature, 2820 K, and the composition (mass and mole fraction) of the products are displayed (Label 9 and 10). Clicking the *Graphical Results* button displays the composition in a graphical pie chart as shown in Figure 0.

One of the strengths of the RIA approach to computing is the simplicity with which one can explore fundamentals of chemical equilibrium. For example, the chamber pressure can be changed to 1 MPa and the equilibrium temperature can be recalculated as 3033 K without having to set up the problem again. Similarly, the effect of a specific reaction, say participation of N_2 can be evaluated by de-selecting all species that contains the N atom, except N_2. The equilibrium temperature can be shown to remain relatively unchanged, changing slightly from 2820 K to 2831 K, showing the relatively weak effect of nitrogen participation on the energetics of PMMA combustion. In fact, it can be shown that including only the eight components - N_2, H_2O, CO_2, CO, O_2, OH, H_2, and O_2 - produces sufficient accuracy in the prediction of equilibrium temperature in this specific case.

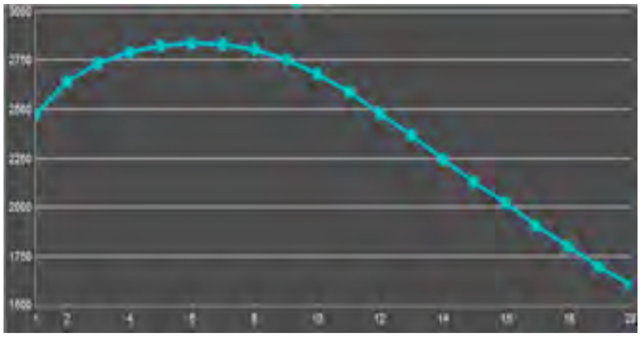


Figure 0. Equilibrium temperature against run number (equivalence ratio) is updated in real time as the web service calls returns.

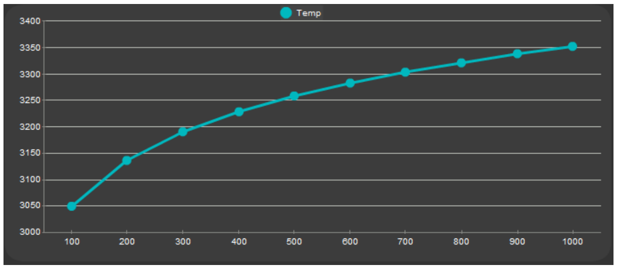


Figure 1. Effect of increasing pressure on the adiabatic flame temperature of methane combustion. Pressure is varied from 1 bar to 10 bar.

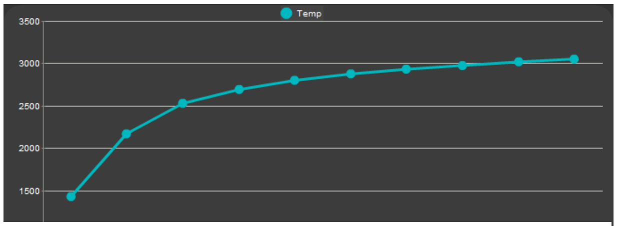


Figure 2. Parametric study of diluent (N_2) addition on adiabatic flame temperature in constant-pressure methane combustion. O_2 varies from 10% (by volume) in the oxidizer to 100%.

Arguably, the most valuable tool the RIA offers is its one-click approach to performing a parametric study. Simply select the range and number of steps in the parametric study block and click the *Analyze* button. The equilibrium solver Web Service is called repeatedly and the products composition table is updated continually. On the *Graphical Result*s panel, the equilibrium temperature is plotted against the run number in real-time as the equivalence ratio is varied from 0.5 to 3 through 20 steps (see Figure 0). Detailed output for each set of calculations for a given equivalence ratio is produced in the *Text Output* panel. In addition to varying the equivalence ratio, the parametric study facility allows one to vary chamber pressure and oxygen mole fraction in the oxidizer. For example, one can evaluate the effect of diluent (N_2) addition on adiabatic flame temperature in the constant-pressure combustion of methane by choosing *Oxygen Mole Fraction* in the parameter drop-down menu of the *Known Heat Transfer* tab. By default, the volume of O_2 is configured to vary from 10% (mole fraction of 0.1) in the oxidizer to 100% (mole fraction of 1.0), although these limits are user adjustable by modifying the parameter *Start* and *End* values. As one can see from Figure 2, diluent addition decreases adiabatic flame temperature, with a maximum temperature of 3049 K reached from an oxidizer consisting of pure oxygen. A similar analysis can be conducted by varying pressure. In Figure 1 we see how adiabatic flame temperature in methane combustion changes as pressure is varied from 1 bar to 10 bar. From the resulting graph, adiabatic flame temperature is shown to increase with pressure.

The *Known Temperature* panel can be used to calculate the heat transfer if the exit temperature is known. For example, the enthalpy of combustion for PMMA(s), Δh_C^{o}, can be calculated by setting the exit temperature as 298 K and pressure as 100 kPa. For complete combustion, the heat transfer for a 1 kg/s fuel flow rate is calculated as -25,983 kW, that is, $\Delta h_C^{\mathrm{o}} = 25.98\ \mathrm{MJ}$.

Although illustrated for a specific fuel, the RIA can be used with any of the 24 fuels listed in the fuel selector. One of the advantages of the Web Service architecture is that the fuel list can be expanded on the server side as more data is available without having to recompile and redeploy the front-end application. The thermochemical data for fuels and combustion product species accessible through CHEQS Web Services is an aggregation of data provided by NASA [22], NIST [23], and Professor Alexander Burcat [24].

Emission Calculations through Web Services: The combustion chamber simulator allows up to 23 pre-selected species. For a more general emission calculation, many more species are necessary. An applet based interface [6] has been created for this purpose and integrated with our thermodynamic Web portal: www.thermofluids.net at the location Daemons> Systems> Open> Steady> Specific> Combustion> ChemEqlWS. To illustrate the use of our Web Service to solve a rather complex equilibrium problem, consider the combustion of isooctane and air at p = 50 bar and T = 3000K producing the 18 possible products for a given equivalence ratio ϕ and molar stoichiometric ratio s. The equilibrium reaction is

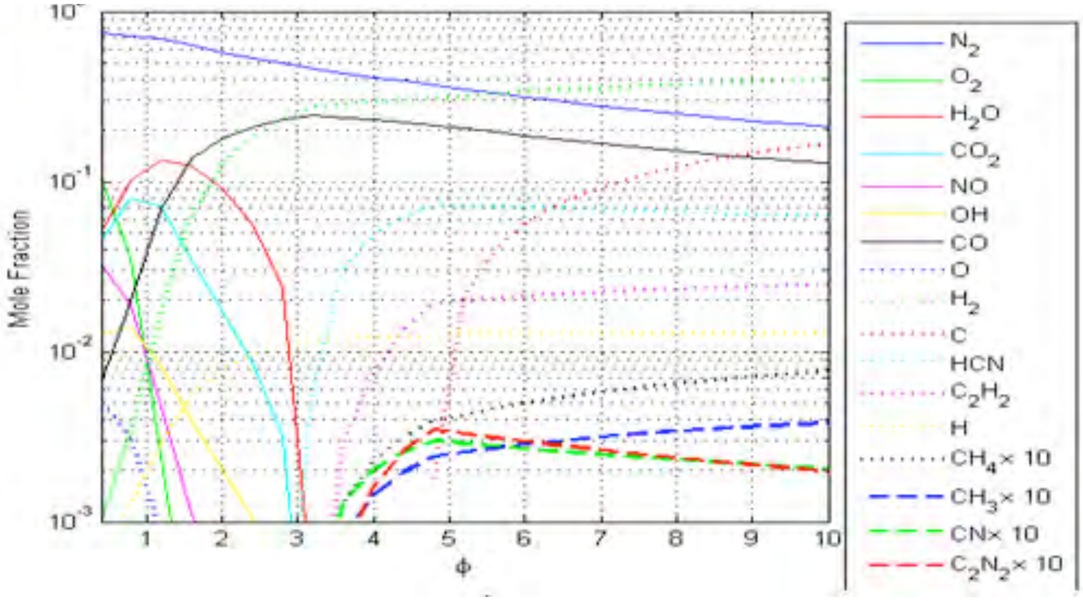


Figure 2. Equilibrium distribution of emissions from isooctane burning with air at 50 bar and 3000K for $0.4 \le \phi \le 10$.

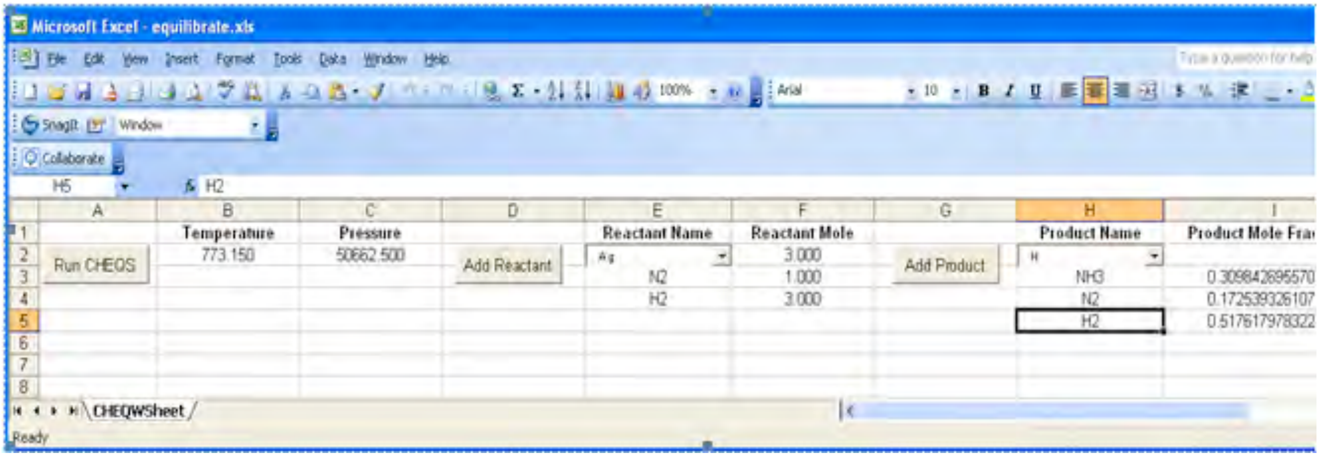
$$C_8H_{18} + \frac{5}{\phi}\left(O_2 + 3.76N_2\right) f$$

$$\left\{\begin{matrix} N_2, O_2, H_2O, CO_2, NO, OH, CO, O, H_2, C(s), \\ HCN, C_2H_2, H, CH_4, CH_3, CN, C_2N_2, C_2H \end{matrix}\right\}$$

The molar stoichiometric ratio s for octane is 12.5. A plot of equilibrium distributions obtained from invoking our Web Service for values $0.4 \le \phi \le 10$ is shown in Figure 2. The results presented compare very well with the results of NASA CEA calculations [25].

Similar comparisons for other published data have produced consistent validation for the Web Service based equilibrium computation.

Integration with Third Party Applications: One of the primary advantages of standardized Web Services is platform independent accessibility. This means one should be able to invoke a Web Service as if it were a third party software component. To demonstrate this concept, we have developed a Microsoft Excel™ macro package and a MATLAB™ toolbox, both downloadable via the Tools section of



`fx` H2

	A	B	C	D	E	F	G	H	I
1		Temperature	Pressure		Reactant Name	Reactant Mole		Product Name	Product Mole Fra
2	Run CHEOS	773.150	50662.500	Add Reactant	Ag	3.000	Add Product	H	
3					N2	1.000		NH3	0.309842695570
4					H2	3.000		N2	0.172539326107
5								H2	0.517617978322
6									
7									
8									

Figure 2. Solving the Haber process example using the Excel interface to our chemical equilibrium Web Service.

http://cheqs.sdsu.edu/ that accesses our chemical equilibrium Web Service to compute the equilibrium distribution of an arbitrary multiphase reaction. Figure 2 shows a screenshot of running our Excel™ macro package to solve for the mole fraction of NH_3 in the Haber process for ammonia synthesis given by equation (6). Using the Excel spreadsheet, one can quickly specify additional reactants and products using the drop-down list and perform *what-if?* scenarios quite easily. For example, one might be interested to know if any monatomic species are present once equilibrium is reached during the Haber process. To find out, one simply adds N and H to the products list using the *Add Product* button and then recalculates the distribution by clicking the *Run CHEQS* button. As shown in Figure 4, one can see no appreciable amount of monatomic species are produced.

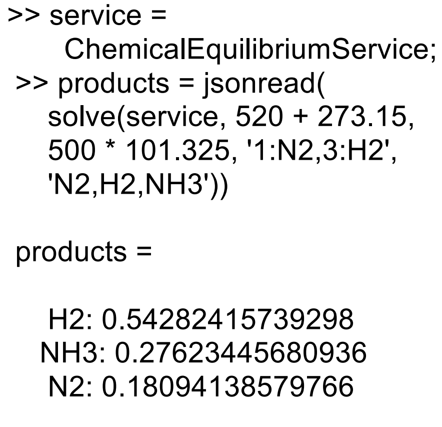
In addition to Microsoft Excel™, the numerical computing package MATLAB is often used by researchers to perform computational work. MATLAB possesses a powerful interpretive scripting engine that allows researchers to develop computational codes with relative ease, compared to development using a compiled language such as FORTRAN or C. Introduced in version 7, MATLAB contains built in functionality to generate SOAP messages and invoke Web Services. This capability allows researchers who use MATLAB for solving combustion problems to interface with our equilibrium Web Service.

To illustrate, Figure 3 shows an example using the MATLAB chemical equilibrium toolbox we developed to solve the ammonia synthesis problem from the MATLAB command line. Example MATLAB scripts are downloadable from our chemical equilibrium Web Service JSP cover page via URL http://cheqs.sdsu.edu:8080/CHEQS/.

3. CONCLUSIONS

The current trend in software development is to make use of distributed software components hosted on remote systems accessible through the Internet. Combustion applications can make use of these distributed components by calling Web Services in client code. In this paper we presented a modern alternative to software applications frequently used in combustion research, such as STANJAN and NASA CEA, for performing chemical equilibrium analysis. Adapting custom applications to libraries supplied with tools like STANJAN and CEA is possible but cumbersome and time-consuming. Instead, we have proposed a unique and different approach that essentially "outsources" computational responsibility to network accessible Web Services. We developed one such Web Service for computing the equilibrium distribution through Gibbs free energy minimization and provided several examples of how this service can be invoked in Flash, Java, MATLAB, and Excel. Our Web Service vastly expands the scope of equilibrium computation and is especially advantageous in the simulation of chemically reacting flows. Educators from universities around the world have registered in our web portal to use this rich internet application in graduate level combustion classes.

Work is currently underway to demonstrate how Web Service based equilibrium calculations can be loosely coupled to existing CFD codes. Capabilities such as automatic load-balancing and failover protection being introduced in newer Web application servers like Glassfish will allow us to distribute the invocation of Web Services in parallel without the need for an additional high performance computing infrastructures. We will also investigate how a clustered configuration of Web application servers can facilitate parallel equilibrium computation. We will be extending our equilibrium Web Service to allow users to specify temperature and volume *(T,V)*, entropy and pressure *(S,P)*, or entropy and volume *(S,V)* and then calculate the equilibrium composition by minimizing the corresponding thermodynamic potential.



```
>> service =
     ChemicalEquilibriumService;
 >> products = jsonread(
    solve(service, 520 + 273.15,
    500 * 101.325, '1:N2,3:H2',
    'N2,H2,NH3'))

 products =

    H2: 0.54282415739298
   NH3: 0.27623445680936
    N2: 0.18094138579766
```

Figure 3. Invoking our chemical equilibrium Web Service from the MATLAB environment.

4. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0753283

5. REFERENCES

[1] Michelsen, M. L., Fluid Phase Equilib. 9, pp. 1–20, 1982.

[2] Michelsen, M. L., Fluid Phase Equilib. 9, pp. 21–35, 1982.

[3] Eubank, P. T., Elhassan, A. E., Barrufet, M. A., Whiting, W. B., Ind. Eng. Chem. Res., 31, pp. 942–949, 1992.

[4] Elhassan, A. E., Lopez, A. A., Craven, R. J. B., J. Chem. Soc., Faraday Trans., 92, pp. 4419–4433, 1996.

[5] Elhassan, A. E., Tsvetkov, S. G., Craven, R. J. B., Stateva, R. P., Wakeham, W. A., Ind. Eng. Chem. Res., 37, pp. 1483–1489, 1998.

[6] Eubank, P. T., Hall, K. R., AIChE J., 41, pp. 924–927, 1995.

[7] Shyu, G. S., Hanif, N. S. M., Alvarado, J. F. J., Hall, K. R., Eubank, P. T., Ind. Eng. Chem. Res., 34, pp. 4562–4570, 1995.

[8] Shyu, G. S., Hanif, N. S. M., Hall, K. R., Eubank, P. T., Ind. Eng. Chem. Res., 35, pp. 4348–4353, 1996.

[9] Hanif, N. S. M., Shyu, G. S., Hall, K. R., Eubank, P. T., Ind. Eng. Chem. Res., 35, pp. 2431–2437, 1996.

[10] Hanif, N. S. M., Shyu, G. S., Hall, K. R., Eubank, Fluid Phase Equilib., 126, pp. 53–70, 1996.

[11] Kramlich, K., General Gibbs Minimization as an Approach to Equilibrium, Course notes for ME 430 Advanced Energy Systems, University of Washington, August 2005.

[12] Gordon, S. and McBride, B. J., Computer Program for Calculation of Complex Chemical Equilibrium Compositions and Applications – I. Analysis, NASA Reference Publication 1311, October 1994.

[13] Reynolds, W. C., The Element-Potential Method for Chemical Equilibrium Analysis: Implementation in the Interactive program STANJAN. Technical Report A-3391, Department of Mechanical Engineering, Stanford University, Palo Alto, CA, 1986.

[14] Paolini, C. P. and Bhattacharjee, S., An Object-Oriented Online Tool for Solving Generalized Chemical Equilibrium Problems, Proceedings of the 2008 ASME International Mechanical Engineering Congress and Exposition, Boston, Massachusetts, October 31 – November 6, 2008.

[15] Dong, X., Gilbert; K.E., Guha, R.; Heiland, R.; Kim, J.; Pierce, M.E.; Fox, G.C.; and Wild, D.J. Web Service Infrastructure for Chemoinformatics. J. Chem. Inf. Model., 2007, 47(4), 1303 – 1307.

[16] Truong, T.N.; Nayak, M.; Huynh, H.H.; Cook, T.; Mahajan, P.; Tran, L.T.; Bharath, J.; Jain, S.; Pham, H.B.; Boonyasiriwat, C.; Nguyen, N.; Andersen, E., Kim, Y.; Choe, S.; Choi, J.; Cheatham, T.E.; and Facelli, J.C.; Computational Science and Engineering Online (CSE-Online): A Cyber-Infrastructure for Scientific Computing, J. Chem. Inf. Model., 2006, (46), 3, 971 - 984.

[17] Paolini, C. P. and Bhattacharjee, S., A Web Service Infrastructure for Thermochemical Data, *J. Chem. Inf. Model.* **2008**; 48(7); 1511-1523.

[18] Frenklach, M.; Packard, A.; Seiler, P.; and Feeley, R. Collaborative Data Processing In Developing Predictive Models Of Complex Reaction Systems. Int. J. Chem. Kinetics, 2004, (36), 57 – 66.

[19] Goodwin, D. G.; CANTERA: An Open-Source, Object-Oriented Software Suite for Combustion, NSF Workshop on Cyber-based Combustion Science, National Science Foundation, NSF Headquarters, Arlington, VA, April 19-20, 2006.

[20] Lahey/Fujitsu Fortran Enterprise v7.1, Lahey Computer Systems, Inc., P.O. Box 6091, Incline Village, NV 89450 USA.

[21] van Engelen, R. A. and Gallivan, K., The gSOAP Toolkit for Web Services and Peer-To-Peer Computing Networks, in the proceedings of the 2nd IEEE International Symposium on Cluster Computing and the Grid (CCGrid2002), pages 128-135, May 21-24, 2002, Berlin, Germany.

[22] Gordon, S. and McBride, B. J., Thermodynamic Properties of Chemical Substances to 6000 K, NASA Report SP-3001, NASA Glenn Research Center, Cleveland, OH, 1963.

[23] NIST Chemistry WebBook, NIST Standard Reference Database Number 69, National Institute of Standards and Technology. http://webbook.nist.gov/chemistry/, June 2005 Release.

[24] Burcat, A., Third Millennium Thermodynamic Database for Combustion and Air-Pollution Use with updates from Active Thermochemical Tables, ftp://ftp.technion.ac.il/pub/supported/aetdd/thermodynamics/, http://garfield.chem.elte.hu/Burcat/burcat.html.

[25] Ferguson, C. R. and Kirkpatrick, A. T., Internal Combustion Engines Applied Thermosciences, 2nd Edition, John Wiley & Sons, Inc., New York, 2001, p. 71.

Cloud-enabling an Evolutionary Genetics Tool and Computational Methods for Invigorating STEM Learning and Research

Bina Ramamurthy
CSE Department
University at Buffalo
345 Davis Hall
Buffalo, NY 14260
bina@buffalo.edu

Jessica Poulin
Department of Biology
University at Buffalo
109 Cooke Hall
Buffalo, NY 14260
jpoulin@buffalo.edu

Katharina Dittmar
Department of Biology
University at Buffalo
109 Cooke Hall
Buffalo, NY 14160
kd52@buffalo.edu

ABSTRACT

We present a cloud-enabled comprehensive platform called *Pop!World* for experiential learning, education, training and research in population genetics and evolutionary biology. The major goal of *Pop!World* is to leverage the advances in cyber-infrastructure to improve accessibility of important biological concepts to students at all levels. It is designed to empower a broad spectrum of users with access to cyber-enabled scientific resources, tools and platforms, thus, preparing the next generation of scientists. *Pop!World* offers a highly engaging alternative to currently prevalent textual environments that fail to captivate net-generation [11] audiences. It is also more mathematically focused than currently available tools, allowing it to be used as a basic teaching tool and expanded to higher education levels and collaborative research platforms. The project is a synergistic inter-disciplinary collaboration among investigators from Computer Science & Engineering and Biological Sciences. In this paper, we share our multi-disciplinary experience (CSE and BIO) in the design and deployment of the *Pop!World* platform and its successful integration into the introductory biological sciences course offerings over the past two years. We expect our project to serve as a model for creative use of advances in cyber-infrastructure for engaging the cyber-savvy net-generation students and invigorating STEM (Science, Technology, Engineering, and Mathematics) education.

Categories and Subject Descriptors

C.2.4 [Distributed Systems]: Client/server and Distributed applications; D. [Software]: D.1.3 [Concurrent Programming]: Distributed programming; Parallel programming; D.1.7 Visual Programming; H.3.4 [Systems and Software]: *Distributed systems;* H.5.1 [Multimedia Information Systems]: Animations; H.5.2 [User Interfaces] *Graphical user interfaces (GUI);* J. [Computer Applications]: *Education;* J.3 [Life and Medical Sciences] Biology and genetics.

General Terms

Algorithms, Performance, Design, Experimentation.

Keywords

Cloud computing, cyber-infrastructure, tools, genetics, evolutionary biology, net-generation, education.

1. INTRODUCTION

The primary motive behind the *Pop!World* project is to build a comprehensive tool by leveraging the emerging cyber-infrastructure to inspire learners at all levels to engage in Biological Science study and research. There are three primary questions that motivated us, all of which are relevant in any educational setting: (i) How can we address pedagogical (didactic) challenges in learner-centric teaching? (ii) How can we provide a sustainable and scalable technology infrastructure for a learning tool? and (iii) How can we use a cyber-infrastructure enabled delivery model to effectively teach students the mathematics behind evolutionary biology and engage them in critical and computational thinking?

Delivery models for education have undergone quite dramatic changes since the introduction of the Internet technology. One significant advance has been the move from teacher-centric to learner-centric models, and incorporation of learning styles and pedagogic models into teaching and learning [12,14]. Providing flexibility on how often and when students access educational interventions is especially important for the cyber-savvy, multi-tasking generation [11]. We need an education delivery model that is learner-centric, provides continual formative assessment, effectively monitors the students' progress, appeals to broad range of audiences, and is easy to implement for the user/educator. This need is all the more critical for the

STEM areas such as Biological Sciences and Computer Science and Engineering.

We wanted to create a learning environment that is usable by a wide variety of people without any technological barrier, while at the same time minimizing financial involvement. Specifically, we want to combat the extremely short cycle of obsolescence prevalent in technology products. In many cases, this is particularly detrimental in poor school districts, or situations of economic crisis, as it may not be possible to continuously buy new infrastructure, or to support labs or administrative staff to accommodate modern tools and technologies.

Another important motive is to foster the student's understanding of the interdependency of disciplines (e.g. to do biology, one needs to know math), and thus increase their grasp on interdisciplinary research. A comprehensive solution was developed to address all the issues discussed above.

2. BACKGROUND

2.1 Evolutionary biology: is the central guiding principle of all biological sciences [6]. Evolution shapes the living world we see in both adaptive and non-adaptive ways [5]. Population genetics is the field of biology that allows scientists and students to make predictions and trace the outcomes of evolutionary forces [9]. Yet, as population genetics is highly mathematical, these concepts are difficult to convey, even to advanced students. As evolution is rarely suited to a traditional bench lab environment, these topics may be best demonstrated through computer simulation.

2.2 Cloud computing: Cloud computing is commonly defined [13] as a cyber-infrastructure that provides dynamically scalable, and often virtualized resources as a service. Powerful working examples for this approach are, for instance, GoogleApps (software) [7], Microsoft Azure (infrastructure) [15] and Amazon's EC2 (platform) [1]. In each of the cases, resources are provisioned entirely by an external platform (the cloud), which provides the service on demand and at the configuration specified by the user. Most importantly, the user need not have expert knowledge in, or control over the technology infrastructure in the cloud. Apart from commercial clouds, there are a few emerging clouds [3, 4, 10], which are mostly used to address computer-intensive problems. This project will demonstrate an innovative use of cloud computing for supporting scientific learning environments. Specifically, we will explore the flexibility of the cloud computing approach to enable convenient and on-demand access to a shared pool of configurable learning/research applications that can be released with minimal management effort and learning curve for the end-user.

3. Project *Pop!World*

The *Pop!World* project was initiated by the Biological Sciences department primarily to address the attrition in their introductory Biology courses. The specific goal of the project is to develop a professional education and research platform for high-fidelity simulations of evolutionary processes (e.g.: selection, mutation, migration, genetic drift, and non-random mating) to engage net-generation students and improve their interest and understanding of these concepts. Specific technical goals involved design and development of a comprehensive and end-to-end set of features, including simple interaction, high fidelity simulation, appealing visualizations, pop-up annotations as explanation for results and processes, parameterization and equation editor, split screen comparison of real world and simulated outcomes, automatic testing, personal preference for students, system configuration for teachers, experimental setup for hypothetical research, innovative collaboration setup for research and storage of historical and experimental results and data with load and playback facilities.

The project began with a simple "proof of concept" prototype of the tool in the Fall 2009. This preliminary stage helped immensely in building the truly diverse team of researchers and students, and in clearly understanding the requirements from the Biological Sciences (BIO) end, and the limitations on the Computer Science and Engineering (CSE) end. Rapid prototyping methods were used to complete the concept tool in about 6 months. The tool was used in a classroom setting in the summer of 2010. Cyber-savvy students were not only the users of *Pop!World* but also served as the creative force behind the design and implementation of the tool.

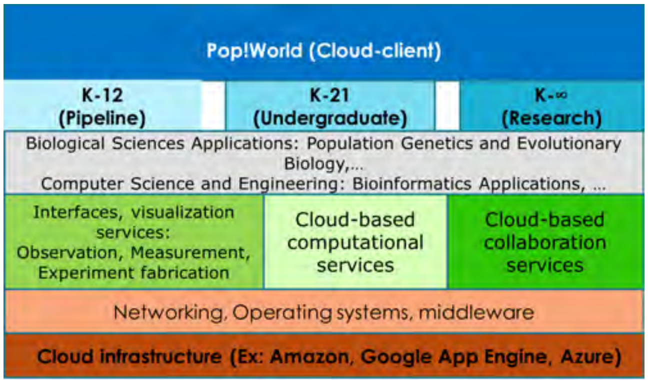

Pop!World (Cloud-client)
K-12 (Pipeline)
K-21 (Undergraduate)
K-∞ (Research)
Biological Sciences Applications: Population Genetics and Evolutionary Biology,...
Computer Science and Engineering: Bioinformatics Applications, ...
Interfaces, visualization services: Observation, Measurement, Experiment fabrication
Cloud-based computational services
Cloud-based collaboration services
Networking, Operating systems, middleware
Cloud infrastructure (Ex: Amazon, Google App Engine, Azure)

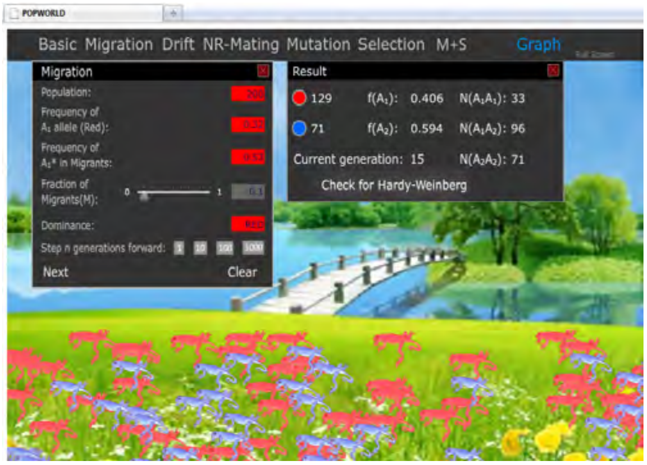
Figure 1: A High-Level Architecture of Pop!World

The student feedback from the use of the prototype and the lessons learned during its development and use helped us design a completed revamped version of the tool with three levels: K-12, undergraduate (teaching K-21), graduate (research K-∞) for use in formal and informal settings. The design and development of *Pop!World* is guided by the cyber-infrastructure approach explained in the report advisory panel on cyber-infrastructure. This report titled

"*Revolutionizing Science and Engineering through Cyberinfrastructure*" is frequently referenced in the context of cyber-infrastructure research as the Atkins Report [2]. We modeled our infrastructure on the core diagram given in the report, and modernized/updated it to reflect the recent advances such as cloud computing, and service-orientation to offer a truly state-of-the art environment in *Pop!World*.

3.1 Three Levels of Pop!World

Pop!World was conceptualized at three different levels of complexity: K-12 (***pipeline or gateway***: middle school – high school instruction), K-21 (***teaching***: college level undergraduate courses, esp. Biological Sciences), K-∞ (***research:*** undergraduate and graduate level research and experimentation) as shown in figure 1. These three levels are available as three different tools and the focus of this paper is on the teaching (K-21) version of the *Pop!World* tool. It has a highly visual presentation with capabilities for entering the experimental data (parameters) representing the various evolutionary forces (Selection, mutation, etc.) working through generations of evolution, engaging display of hypothetical red and blue lizards to present the population's changing traits, graph output that summarizes the evolutionary trend, and a computation windows that display the computed numbers. A representative screen shot of the *Pop!World Teaching* environment is shown in figure 2. The *Pop!World Gateway* environment is meant for high school students to teach them the fundamentals of Mendelian and population genetics. A screen shot of this interface is shown in figure 3.


POPWORLD
Basic Migration Drift NR-Mating Mutation Selection M+S Graph
Next
Clear
Check for Hardy-Weinberg

Migration

Population: 200

Frequency of A_1 allele (Red): 0.32

Frequency of A_1* in Migrants: 0.53

Fraction of Migrants(M): 0 — 1 0.1

Dominance: RED

Step n generations forward: 1 10 100 1000

Result

● 129 $f(A_1)$: 0.406 $N(A_1A_1)$: 33

● 71 $f(A_2)$: 0.594 $N(A_1A_2)$: 96

Current generation: 15 $N(A_2A_2)$: 71

Figure 2: Pop!World Discovery Interface

3.2 Deployment on the cloud

Pop!World was developed using Adobe Flash and the action script language strictly following software engineering principles in the design and object-oriented programming. The K-21 Teaching module has about 10000 lines of modular code understandable by BIO professors and students. The .swf (shockwave file) generated is then deployed on the Google App Engine (GAE) [8]. GAE is a comprehensive cloud platform that allows for deployment of applications irrespective of their size. GAE was chosen since the basic quota on this environment is free and more resources (processing power, bandwidth, storage etc.) beyond the free usage tier can be added by enabling billing. It offers the same reliability, availability and scalability at par with Google's own applications. Also GAE cloud offers excellent monitoring features for observing the load and for load balancing on demand.

GAE allows for multiple instances of the tool to be deployed. We deployed about 10 instances and load balanced the incoming requests among the instances of the tool. We expected about 200-300 simultaneous access request across the 10 instances.

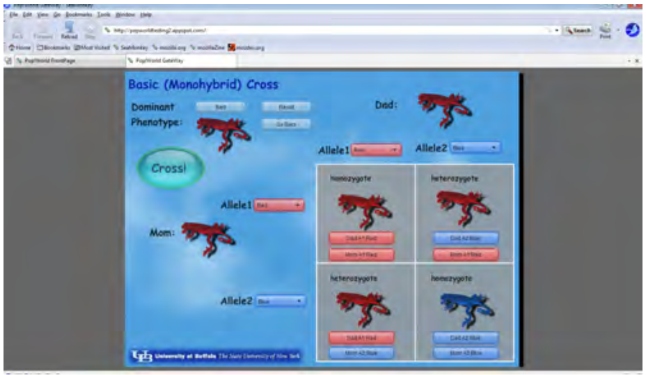


Figure 3: *Pop!World* Gateway (K-12)

4. *Pop!World* IN BIO COURSES

Entering undergraduates use the *Pop!World* K-21 tool to have an immersive experience of each evolutionary force (migration, non-random mating, genetic drift, mutation, and selection). By entering the parameters that affect the population, students will "derive" the mathematical laws governing evolution by each force, allowing them to have a more intuitive understanding of how each microevolutionary force works to change a population. This can be paired with a more concrete mathematical presentation of the ideas, which is typical in a standard lecture format. Our experience has been thatthe graphical interface of *Pop!World* leads to a visceral understanding of these processes even if students do not have the mathematical background to use the actual formulas that govern these forces. Early simulation experiences aim to be directive about how to manipulate the simulation outcomes, but the format of the program allows students to be increasingly independent, leading to self-discovery of evolutionary patterns and trajectories.

When students control the parameters that govern the simulation, they are more inclined to engage with the program and understand the impact of the results. *Pop!World* is an experiment the student is doing – in much the same way an evolutionary modeler would use more advanced versions of the software. Such experiences are invaluable in encouraging students to remain in the STEM disciplines.

The *Pop!World* K-21 module has been fully integrated into the instruction and curriculum of the introductory course (BIO 200: Introduction to Evolutionary Biology) in the Biological Sciences department of University at Buffalo. The course is typically offered in the Summer and in the Fall semesters. The course is taught by the co-author on this paper, Dr. Jessica Poulin. About 1200 students were enrolled in the course in the Summer and Fall 2010. In general the use of the tool was well received and liked by the students. For example, when a lab work was assigned to the class of more 1200 students, the due dates were staggered so that we could manage the traffic and load. However when we monitored the load on the first day of the assignment we found that more than half the students in the class were on *Pop!World* working on their assignment! The bandwidth quota on GAE was exceeded and we had to increase the bandwidth capacity of GAE by enabling billing as well as by deploying more instances of the tool. It was good to observe the eagerness of the students in learning science through *Pop!World*. The cloud deployment of *Pop!World* helped in a quick response to the surge in the number of users (students) and to deploy additional instances of the resources on demand.

4.1 Outcome Assessment

The most valuable outcome of the project is the formation of a highly productive multi-disciplinary team of Computer Scientists and Biologists.

We designed survey instruments to assess the effect of *Pop!World* and were approved by the Internal Review Board at University at Buffalo. The evaluations discussed here are from Fall 2010.

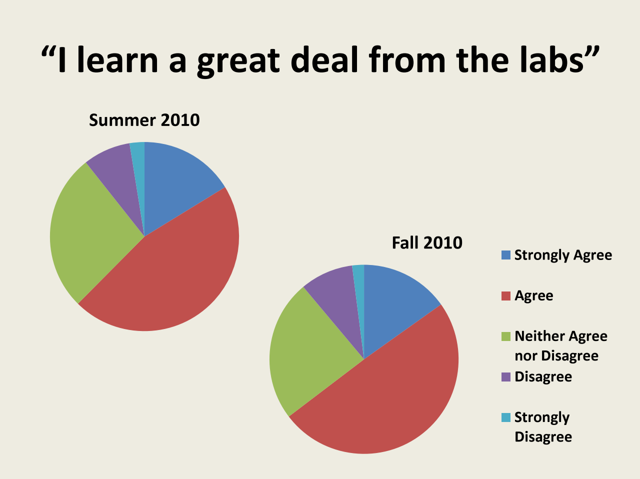


Figure 4: Learning through Pop!World

Figures 4-7 provide a series of representative evaluations from the first year of use of the *Pop!World* learning environment. Figure 4 shows about 65% of students agree that they learned a great deal from the lab. There is a slight improvement from the Summer 2010 to Fall 2010. It is in this Fall offering that we had the bandwidth overload for the cloud access. Since that time we have added a load balancer and we did not experience a problem with load during the Fall 2011 offering of the course.

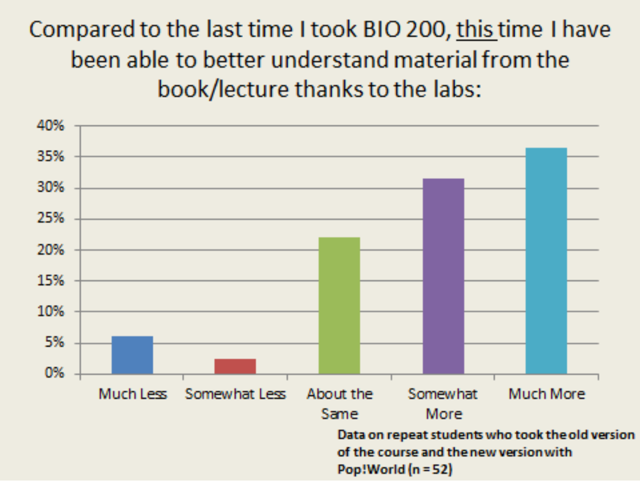


Figure 5: Understanding of Course Material

The introductory biology course is an important course for many science (STEM) majors including pre-med students. The attrition rate in the traditional offering of the course is as high as 20%. It is not unusual to find many students repeating the course for various reasons. We used this fact to evaluate the "understanding of the course material". The majority of repeat students clearly agreed that *Pop!World* (which they had not used in the first offering of the course) helped them understand the course material (Figure 5). This statistic is all the more significant since this is on the students repeating the course.

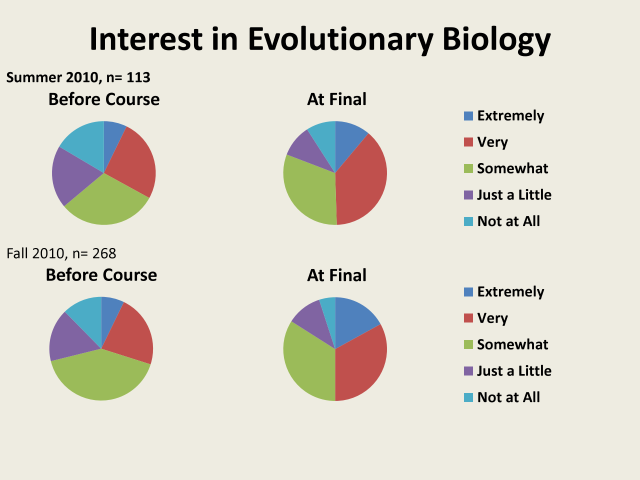


Figure 6: Interest in Evolutionary Biology

We evaluated the improvement in students' interest in Evolutionary Biology before and after the course, using *Pop!World* (Figure 6). The majority of the students responded positively. Clearly more than 80% agree that their interest increased though by different degrees (extremely, very, etc.). The significance of this outcome is that we have been able to engage the majority of the students in the course, whereas before many of them were dropping out of the course.

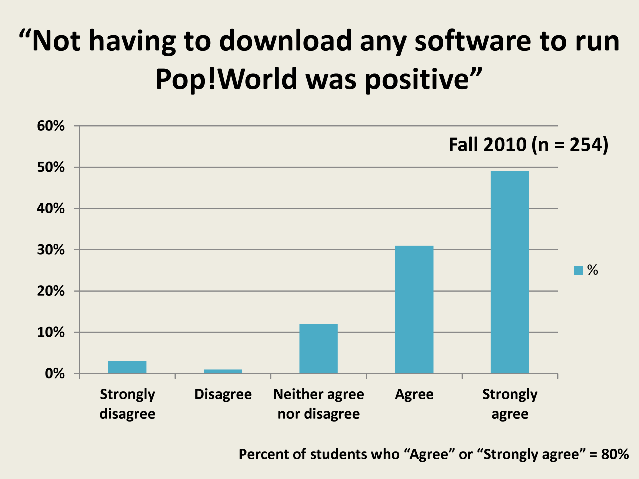


Figure 7: Pop!World Access Model (Cloud)

Figure 7 shows the students' evaluation of the cloud access model. Students simply access the cloud-deployed *Pop!World* without any need for downloading or installation. Students really like this aspect of *Pop!World*. Indirectly cloud-enabling also provided the 24X7 availability and access from anywhere that may not be given with departmental or university-wide servers.

Figures 4-7 provides validation for the introduction of *Pop!World* into the introductory Biology course BIO 200. Besides these, attrition in the course is down to 5-10%. We expect further improvement in the recent offering (2011) of the courses. There is a lot of interest in the *Pop!World* environment from students who are not in the course. The Chronicle of Higher Education had a report on the software. We experienced quite a response from the readers (general public, probably educators) since many of them accessed the cloud-enabled version when the article was published. Figure 8 shows the cloud monitored for the access patterns of *Pop!World*. The load curve at the top of the screen shows the intense activity we experienced after the publication of the *Pop!World* article. We attribute this traffic to the interest in the readers of the newspaper since there was no *Pop!World*-related work assigned to the students during that time frame.

5. SIGNIFICANT CONTRIBUTIONS

A technical success of the project is that of rapid prototyping and cloud-deployment of a highly usable Biology learning environment to invigorate STEM learning and research. This project is a success model for synergistic inter-disciplinary collaboration between Computer Science and Biological Sciences. Students from the CSE and BIO department worked together, learned from each other and accomplished the goals of the project. The *Pop!World* environment is available for anyone to use at http://popfrontpage.appspot.com/.

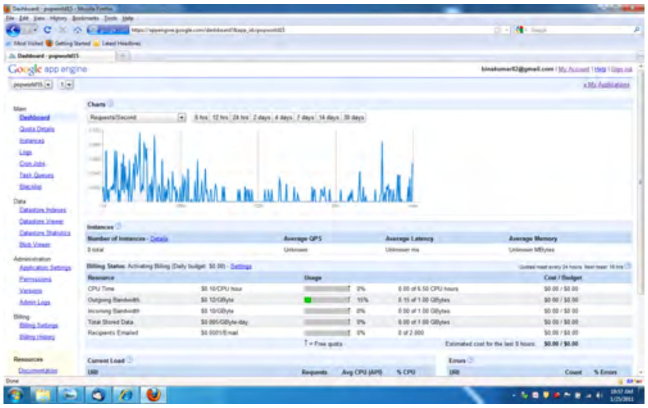


Figure 8: Pop!World Monitoring

6. FUTURE DIRECTION

We have nearly completed work on a K-12 model of *Pop!World* (Pipeline/Gateway). By summer 2013 we will have a *Pop!World* Research version available. The three versions each have distinct contents without much overlap, but they feed into each other. We plan to filter out the experiences from this project to publish general guidelines and best practices for cloud-enabling STEM tools and methods.

7. ACKNOWLEDGMENTS

This project is funded by the NSF OCI-CI-TEAM 1041280 awarded to B. Ramamurthy, J. Poulin, and K. Dittmar. Current version of (K-21) Pop!World was developed with the help of Tyler Taylor (undergraduate, Biological Sciences), and Timothy J. Hartnett, and Shuang Wu (graduates, Computer Science and Engineering).

8. REFERENCES

[1] Amazon Elastic Computing Cloud. Amazon EC2, http://aws.amazon.com/ec2/, last visited January 2012.

[2] D. E. Atkins, K. K. Droegemeier, S. I. Feldman, H. Garcia-Molina, M. L. Klein, D. G. Messerschmitt, P. Messina, J. P. Ostriker, M. H. Wright. Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, January 2003, http://www.nsf.gov/od/oci/reports/atkins.pdf, last viewed January 2012.

[3] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katzh et al, "Above the Clouds: A Berkeley View of Cloud Computing" http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf

[4] K. A. Delic and M. A. Walker. 2008. Emergence of the Academic Computing Clouds. Ubiquity 2008, August, Article 1 (August 2008), 1 pages.

DOI=10.1145/1414663.1414664 http://doi.acm.org/10.1145/1414663.1414664

[5] J.S. Freeman and J.C. Herron. Evolutionary Analysis, 4th Edition. Pearson/Benjamin Cummings, San Francisco, CA, 2007.

[6] D.J. Futuyma. Evolutionary Biology, 3rd Edition. Sinauer Associates, Inc, Sunderland, MA, 1998.

[7] Google apps. http://net.educause.edu/ir/library/pdf/ELI7035.pdf and http://www.google.com/apps, last viewed January 2012.

[8] Google App Engine. http://code.google.com/appengine/, Last viewed January 2012.

[9] P.W. Hedrick. Genetics of Populations, 4th Edition. Jones and Bartlett, Sudbury, MA, 2011.

[10] N. Leavitt, “Is Cloud Computing Really Ready for Prime Time?”,Computer Volume 42, Issue 1, Jan. 2009 Page(s):15 – 20, http://www.computer.org/portal/web/computingnow/0209/theme/computer

[11] D. G. Oblinger and J. L. Oblinger. *Educating the Net Generation*. Educause Publication, 2005, http://www.educause.edu/EducatingtheNetGeneration/5989, last visited January 2012.

[12] The Project Kaleidoscope (PKAL), http://www.pkal.org/, last viewed January 2012.

[13] B. P. Rimal, E. Choi and I. Lumb, "A Taxonomy and Survey of Cloud Computing Systems" ncm, pp.44-51, 2009 Fifth International Joint Conference on INC, IMS and IDC, 2009 http://www.computer.org/portal/web/csdl/doi/10.1109/NCM.2009.218

[14] E. Soloway, "How the Nintendo Generation Learns," Communications of the ACM, 34(9), pp. 23-26, 1991.

[15] Windows Azure Platform. http://www.microsoft.com/windowsazure/, last visited January 2012.

First Steps in Transforming the Primary Research Process through a Virtual Linguistic Lab for the Study of Language Acquisition and Use: Challenges and Accomplishments

María Blume[†]
University of Texas at El Paso
Department of Languages and Linguistics
Liberal Arts Bldg., Room 119
El Paso, TX 79968
1-915-747-6320
mblume@utep.edu

Barbara Lust[†]
Cornell University
Department of Human Development
G57 Martha Van Rensselaer Hall
Ithaca, NY 14853
1-607-255-0829
bcl4@cornell.edu

ABSTRACT

This project involves both the development of a community of scholars committed to cross-institution, interdisciplinary and cross-linguistic collaboration (a Virtual Center for Language Acquisition, VCLA) and the creation of a web-based infrastructure through which a new generation of scholars can learn concepts and technologies empowered through this CI environment. These technologies, constituting a Virtual Linguistic Lab (VLL), provide the student with the structure for data creation, data management and data analysis as well as the tools for collaborative data sharing. This infrastructure, informed and executed through computational science, involves the coherent integration of an open web-based gateway (The VCLA website), linked to a specialized web-based VLL portal which includes not only real world examples and visualizations of data creation and analyses, but several cybertools by which these data can be managed and analyzed. This infrastructure subserves both the beginning student and the researcher pursuing calibrated methods and structured data sharing for collaborative purposes. Students continually engage in the development of the cybertools involved and in the scientific method involved in primary research. In this paper we summarize our objectives, the challenges we face and the solutions we have developed to these challenges. At this point, the project is just completing an implementation stage, and the first steps in creating a virtual community of practice, and is being readied to move to a diffusion stage.

Categories and Subject Descriptors

J.5 ARTS AND HUMANITIES. *Linguistics*

General Terms

Management, Documentation, Design, Experimentation, Security, Human Factors, Standardization, Languages, Theory, Legal Aspects.

Keywords

Community of practice, language acquisition, language use, language documentation, research, education, database, data management, standardization.

1. INTRODUCTION

Recent developments in cyber-infrastructure offer new possibilities to scientists for advancing research questions and methods [2, 4, 5, 11, 12, 13, 25, 29, 30, 33], opening possibilities for interdisciplinary collaborative research and empowering cross-linguistic and cross-cultural research in a global perspective. These developments can empower the study of the language sciences as they have empowered other areas of science.

However, these new possibilities challenge the field of linguistics and the language sciences to develop (1) an infrastructure of collaboration that will allow us to create a virtual community of practice [2, 3, 4, 12, 34, & 38]; (2) standardized tools and best practices which can be shared while at the same time allowing unique methods by individual researchers; (3) infrastructure for data storage, management, dissemination and access, including means for interfacing databases that differ in both type and format [5, 13, 15, 25, 31, and 32]; (4) preservation and ‘portability’ of data and related materials [6 & 37]; and (5) changes to the ways in which we educate our students and train new researchers in scientific methods.

As noted by King [24] for the social sciences:

> “The potential of the new data is considerable, and the excitement in the field is palpable. The fundamental question is whether researchers can find ways of accessing, analyzing, citing, preserving, and protecting this information.” (p. 719)

Our purpose in this project is to train students in new methods of primary research which exploit new cyberinfrastructure-enabled possibilities to collect and manage complex data in a collaborative scientific environment and to develop cyber-infrastructure for documentation and accessibility of an ever-growing set of shared complex data, allowing data use, reuse and repurposing.

To this end, we created a new Virtual Learning Environment for the language sciences, through development of a Virtual Linguistics Lab (VLL).[1]

[1] http://clal.cornell.edu/vll

† With the collaboration of the founding members of the Virtual Center for Language Acquisition

In section 1 we introduce our project, audience and objectives. In section 2 we describe our multiple interrelated challenges. In section 3 we present the components of the VLL and then in section 4 we explain how they approach solution to the challenges we face. In section 5 we summarize our educational achievements to date. Section 6 describes the broad impact of the project. Section 7 presents future challenges and lessons learned. A description of the VLL especially with regard to its role in language documentation can be found in Lust et al. 2010 [27]. A description of cybertool development in the VLL can be found in Blume et al. 2012 [8]. Here we focus on the educational mission of our project.

1.1 Audience

Our program mobilized faculty from a new community of practice across eight diverse US Universities and one initial international extension (Peru).[2] Project members were interdisciplinary VCLA founding and contributing members who are linguists, developmental and cognitive psychologists, and neuroscientists.[3] Members come from different fields, institutions and countries.

Courses involved in this project were addressed to undergraduate and graduate students from linguistics, psychology, human development, and computer sciences and to researchers across the world wishing to collaborate on shared data and/or learn best practices for scientific methods in the study of language acquisition and use.

1.2 Objective

We seek to "transform the primary research process" by providing a systematic infrastructure and cybertools to foster and support scientific collaborative data collection and management starting from the initial stages of a research project and throughout to its report of results.

Thus, this project seeks to educate a new generation of interdisciplinary undergraduate and graduate students and future researchers –with diverse geographical and cultural backgrounds– who will gain a solid formation on language documentation, standardization and cybertool use through our courses[4]. It also seeks to support interdisciplinary researchers[5] interested in international and cross-institution collaboration that need to create and share data but do not have the means or training to do so. As they and their students use our virtual learning environment they can both give us feedback and later train other researchers or students at their institutions.

2. CHALLENGES

Our project faced several challenges related to education (2.1.), data complexity (2.2.), and cultivation of researcher collaboration (2.3).

2.1 Educational Challenges

2.1.1 Interdisciplinarity of the language sciences.

The major questions in Cognitive Science — *is the brain programmed for language knowledge and acquisition? What are the universals of language structure? What is innate and what is learned with regards to language? How is new linguistic knowledge developed over time?* — require us to be able to study all languages (of which 6,000-7,000 have been estimated) and all developmental stages of language acquisition. Language acquisition is therefore, by its very nature, a multidisciplinary area, which must be studied by linguists, developmental psychologists, educators, language pathologists, human development researchers, and computer scientists who have means for collaboration. Both researchers and students need to be able to collect, analyze, and compare large amounts of cross-linguistic data in interdisciplinary forms (e.g., brain images and language utterances)[6]. Our scientific enterprise thus requires cross-institution and international collaboration, in addition to a well-designed computational platform for its development.

2.1.2 Language documentation and data management

Students need training to manage language data in a scientifically-sound way. They must also be taught methods of data sharing. For example:

> "Finally, universities and individual disciplines need to undertake a vigorous programme of education and outreach about data. Consider, for example, that most university science students get a reasonably good grounding in statistics. But their studies rarely include anything about information management —a discipline that encompasses the entire cycle of data, from how they are acquired and stored to how they are organized, retrieved, and maintained over time. That needs to change: data management should be woven into every course in science, as one of the foundations of knowledge." [15][7]

This needs to be accompanied by education in a culture of collaboration and data sharing, as highlighted by King [24]:

> "[…] More importantly, when we teach we should explain that data sharing and replication is an integral

[2] MIT, Boston College, Rutgers University at New Brunswick, Rutgers University at Newark, California State University at San Bernardino, Southern Illinois University at Carbondale, Cornell University, University of Texas at El Paso, Pontificia Universidad Católica del Perú.

[3] FOUNDING MEMBERS: Suzanne Flynn (MIT), Claire Foley (Boston College), Liliana Sánchez (Rutgers University, New Brunswick), Jennifer Austin (Rutgers University, Newark), YuChin Chien (California State University at San Bernardino), Usha Lakshmanan (S. Illinois University at Carbondale), Barbara Lust, Claire Cardie, James Gair, Marianella Casasola, and Qi Wang (Cornell University), María Blume (University of Texas at El Paso), and Elise Temple (NeuroFocus). CONTRIBUTING MEMBERS relevant to this project: Jorge Iván Pérez Silva (Pontificia Universidad Católica del Perú), Gita Martohardjono (CUNY Graduate Center/Queens College), Cristina Dye (Newcastle University), Yarden Kedar (Ben Gurion University at the Negev).

[4] The courses had different learning objectives since they were taught at different institutions on different semesters; some focused on first language acquisition, others on bilingual acquisition, and one was focused on the acquisition of Spanish. However, they all incorporated this learning objective as a main objective.

[5] At this point, faculty at the nine institutions that helped us develop this project.

[6] See section 2.2 below.

[7] See also [1].

> part of the scientific process. Students need to understand that one of the biggest contributions they or anyone is likely to be able to make is through data sharing".

The ability to manage and share complex data, in its turn, depends on students being trained in basic computational skills needed for data management and analysis.

2.1.3 *Student background*

Students interested in language acquisition come from different fields, and may come to our courses without much of the necessary background. In particular, students from psychology, human development, and computer science need to learn linguistic theory and terminology; students from linguistics, human development and computer science need to learn about research design, and experimental methods in developmental psychology; students from psychology, human development, and linguistics need to receive basic training in computer science in terms of using and understanding complex databases, as well as in basic computational thinking to be able to create their own searches in the database; computer science students need to learn to apply their computing skills to language data. All students need to be trained in transcription of language data and in conducting basic linguistic analyses of natural language. Computer science will be necessary both for modeling and analyzing large data sets, but also for the students' contribution to the development of cybertools themselves.

2.1.4 *Research with human subjects*

To conduct research on natural language students require extensive training to work with human subjects and access to human subjects is tightly controlled. Students must be taught procedures to ensure confidentiality and informed consent that are set by individual Institutional Review Boards in conjunction with new mandates by federal funding agencies (e.g., National Institutes of Health (NIH).[8] Students must be taught that all records regarding human subjects must become part of the complete language documentation process.[9]

2.2 Data challenges

2.2.1 *Data Complexity*

In the fields of language acquisition and use, data are multi-linguistic, multi-modal (i.e., audio, video, transcripts in different formats, etc.), multi-formatted, and derive from multiple methods of data collection (i.e., observational and experimental, cross-sectional or longitudinal). In addition, they involve multiple aspects of data provenance (e.g., age and/or developmental or cognitive stage of speaker, social and pragmatic contexts, culture). These features result in a complex set of databases. The scientific use of any single record requires access to many levels of data, ranging from raw (establishing provenance) to structured and analyzed data (establishing intellectual worth).[10] The computational science necessary to accomplish analyses and interoperability over representations of such large, diverse and expanding data sets is challenging to students and researchers not trained in computer science.[11]

Various linguistic theories are invoked for data description and analysis, creating a need to interface theoretical vocabularies. The variety of languages needs to be described in language typology, while we search for language universals by the creation of uniform formats for cross-linguistic comparisons.[12] Audio or audiovisual samples provide the authoritative archival form of language data creating technical challenges [23]. Generating transcriptions of language requires a time consuming, cognitive and analytic process with variation expected across individual transcribers [16 & 17]. At every moment, different points of data creation must be linked and sound methods of data documentation must be applied. Language data collections are infinitely expandable and should be merged, used, reused and, when possible, repurposed. Continual data-driven computation and statistical analysis is required as is theoretical modeling through computational methods.

2.2.2 *Data Documentation and Standardization*

These features of language data result in a complex set of databases often appearing in diverse formats as different labs generally practice distinct forms of data collection and management. Therefore, there is a need in the field for standardization of data collection, labeling and storage methods that will allow for preservation and portability of such data.

Once the data are collected, researchers must develop ways to link diverse data sources, calibrate them and make sure they are subjected to the same reliability standards so that data can "speak to" data" [25; see also 5, 13, 30, & 32].

There is also a need for databases that provide access to all the information related to a project, from PI information, to project design, batteries, results, as well as the actual data from each subject. These background data are fundamental both as a teaching tool and as a prerequisite for data reliability and researcher collaboration. Data must be described and preserved with systematic and significant metadata, which include general concepts recognized across fields and linguistic concepts for specific inquiries (see [27] for further description of issues related to language documentation).[13]

2.2.3 *Data Management and Dataset Design*

Data must be stored so that relationships can be discovered within and across data sets. The more each data singleton can be significantly connected or "interlinked", the more powerful and useful it becomes [5 & 13]. Such links can be cross-disciplinary (e.g., connecting brain images with behavioral experimental results testing language comprehension or production), or specific

[8] http://grants2.nih.gov/grants/policy/data_sharing/

[9] In addition to collecting data and comparing data on multilingual populations, students and researchers need to be able to determine whether the multilingual populations of any two different projects are homogeneous [22], since numerous variables affect a speaker's language knowledge, dominance, and patterns of use.

[10] For example, data from more than 20 languages and cultures and thousands of child subjects[10] and adults exist in the Cornell Language Acquisition Lab and Virtual Center for Language Acquisition alone.

[11] See the introduction and papers collected in [13], for other efforts regarding open data in linguistics aided by a strong computational component.

[12] This challenge is being confronted by the General Ontology for Linguistic Description (GOLD) project [21] in the Electronic Metadata for Endangered Languages Data (EMELD) enterprise [19].

[13] See also [6, 31, & 32].

to linguistics. Data from any one language must be comparable to that in another if one pursues a hypothesis concerning linguistic universals or variation linked to language typology.

An effective data management infrastructure must not only provide a powerful database that can handle both experimental and naturalistic data, but, at the same time, it must structure the primary data creation process from its initial stages, providing a way to represent new data so that it can be analyzed subsequently in a standardized and theory-neutral way which ensures data comparability. At the same time, this representation must allow researchers to create theory-specific coding screens allowing multiple types of analyses in their own data or linking data across projects.

2.3 Cultivating Researcher Collaboration

2.3.1 Researcher training

The scientific study of language acquisition and use not only requires researchers to conduct field work and collect data according to sound scientific methods but also to manage international collaboration and to be trained in shared principles of data documentation, database use, and collaboration through cybertools. Researchers need a resource that allows them to compare data across datasets and projects, and to reuse previously collected data. As noted in *Nature Biotechnology* [14]

> "[…] More often, though, a failure to share simply reflects the considerable time and effort associated with formatting, documenting, annotating and releasing data. In this regard, the availability of new tools, […] should prove helpful"[14]

2.3.2 Intellectual property

Finally, intellectual property rights must be addressed in the case of language data as for research data in general. Language data painstakingly collected and created by individual scientists belongs primarily to the researcher and to the institution in which they work. Principles for sharing data or scientific materials must be developed in a manner that respects this premise [1, 3, 7, 8, 11, 12, 14, 15, 18, 27, 33, 34, 35, 38, & 39]. Such agreements must also become part of comprehensive language documentation where language is to constitute scientific data.

2.3.3 Cross-institutional and international collaboration

For active researcher collaboration to expand, our academic institutions need to standardize the ways in which IRB permissions are complied with across institutions. IRBs ordinarily do not have common rules and common standards for cross-institution research. In some countries, comparable IRBs do not exist.

3. COMPONENTS

Our project pursued solution to these challenges by building an infrastructure that includes two main components integrating a VCLA[15] with a Virtual Linguistics Lab whose elements are provided through a structured VLL portal which can be used in both synchronous and asynchronous courses collaboratively across institutions.

3.1 The Virtual Center for Language Acquisition (VCLA)

The VCLA unites a series of research labs across the country and the world. A set of founding members collaborated to build an infrastructure for its mission: to foster collaborative research among scientists working in the area of language acquisition, collaborations which are potentially interdisciplinary, which may be at a distance geographically and which may involve the comparative study of multiple languages, interactions on shared data, and a variety of lab methods.

3.2 The Virtual Linguistics Lab (VLL)

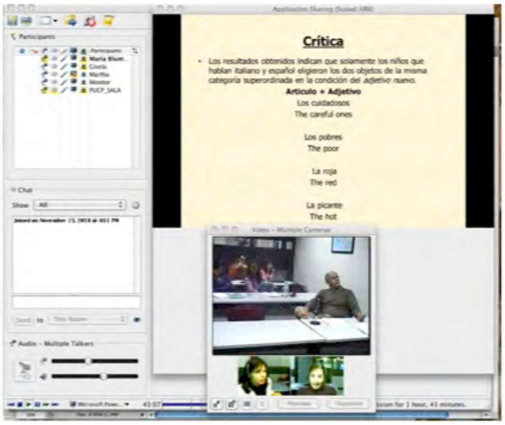
The VLL portal,[16] which is now accessible in English and Spanish, provides structured access to the components of a virtual linguistic lab, which are:

- Materials constituting a series of web-based courses, integrating synchronous and asynchronous forms of interactive information distribution that teach them the specific procedures for investigating language knowledge. Each topic contains:
  - PowerPoint presentations that can be used in class or for review.
  - Audio/visual samples that may be used as part of the lessons to demonstrate a particular method or issue in language acquisition.
  - Published or unpublished papers.
  - Specific exercises/homework, linked to the audio/visual materials so that students can practice data transcription and analysis with real research examples.
- A Research Methods Manual [10] explicating best practices for the scientific study of language acquisition. It provides students and researchers with standard methods for data collection as well as with the background knowledge the DTA tool presupposes.
- A glossary [26] as part of the general Research Manual further helps students from different fields learn the terminology and concepts of the other fields.
- A set of materials to assist in data collection, data management and data analyses, e.g., a multilingualism questionnaire for assessment of degree and nature of multilingualism.
- A series of web conferences (as exemplified in figure 1) through which students can participate in discussions with students and researchers at other institutions synchronically during courses, or later review recordings of these conferences asynchronically.
- A discussion board, with a blog and a wiki, to share ideas and post assignments, presentations and research.

[14] See also [24 & 28]

[15] http://vcla.clal.cornell.edu

[16] Programming for the VLL portal was created by Tommy Cusick, then a Cornell undergraduate student in computer science, now at Google.



Crítica

- Los resultados obtenidos indican que solamente los niños que hablan italiano y español eligieron los dos objetos de la misma categoría superordinada en la condición del *adjetivo nuevo*.

Artículo + Adjetivo

Los cuidadosos
The careful ones

Los pobres
The poor

La roja
The red

La picante
The hot

Figure 1. Webconferences join institutions for discussions.

In addition, the VLL portal provides access to cybertools developed to structure the primary research process in the area of language acquisition and use.[17] These cybertools are:

- The Data Transcription and Analysis Tool that provides a structured interface for metadata and data collection, [8]. It not only guides researchers and students in the primary research process but also results in a web-based calibrated database of continually expanding cross-linguistic data, and an Experiment Bank.
- Virtual Workshops, and a technical user's manual [9] provide training for students on the use of the DTA tool and the Experiment Bank.

These materials are integrated into a university-supported cyberinfrastructure to underwrite potential courses and to ensure the high availability needs of a distance-learning program.

4. MEETING THE CHALLENGES

4.1 Educational Challenges

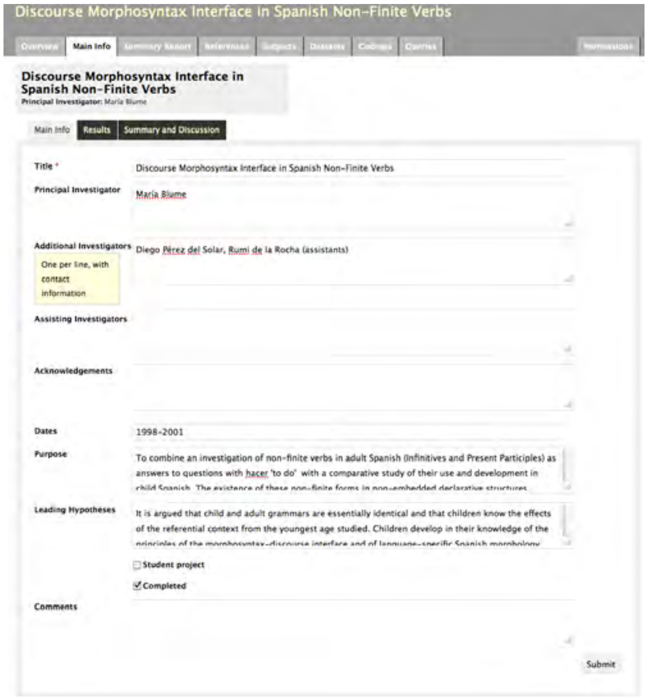
4.1.1 Interdisciplinarity

All VLL materials are developed to serve interdisciplinary access. Students across disciplines are encouraged to collaborate in original research projects, once they complete training, and they read about and discuss the challenges of collaboration [1, 2, 11, 12, 14, 18, 27, 28, 33, 34, 35, & 39]. Through the VCLA website, which lists projects by interdisciplinary VCLA members in order to give undergraduate and graduate students and other researchers ideas for future research and collaboration, our students have access to researchers and projects from outside their institution, country, and discipline, who they can contact for advice, or collaboration.

The students meet with interdisciplinary students and professors at other institutions through Elluminate/Blackboard Collaborate.

4.1.2 Language documentation and data management

Our VLL materials provide students with lessons on scientific methods and best practices for data collection and management in primary research and provide web available tools for learning and practicing these. In particular, the DTA tool guides students and/or researchers through the steps of data creation and management, including metadata representation (cf. 4.2 below). Through its Experiment Bank component, it collects all information related to a study (experimental or observational) in the same location, and makes it available to researchers seeking to replicate it, criticize it or consult it when reading a particular scientific research paper reporting an experiment's results. The first sets of screens in the DTA tool guide students and researchers to save/access the metadata information for a research study, thus providing an entry in an Experiment Bank. Main areas include project investigators, purpose and leading hypotheses, subjects, and results and discussion. The DTA tool also tracks publications, presentations, related studies, and bibliography related to a research project. At several different points, documents can be attached. Figure 2 shows the first screen a researcher completes when starting a new project.



Discourse Morphosyntax Interface in Spanish Non-Finite Verbs

Discourse Morphosyntax Interface in Spanish Non-Finite Verbs
Principal Investigator: María Blume

Title *	Discourse Morphosyntax Interface in Spanish Non-Finite Verbs
Principal Investigator	María Blume
Additional Investigators One per line, with contact information	Diego Pérez del Solar, Rumi de la Rocha (assistants)
Assisting Investigators	
Acknowledgements	
Dates	1998-2001
Purpose	To combine an investigation of non-finite verbs in adult Spanish (Infinitives and Present Participles) as answers to questions with hacer 'to do' with a comparative study of their use and development in child Spanish. The existence of these non-finite forms in non-embedded declarative structures
Leading Hypotheses	It is argued that child and adult grammars are essentially identical and that children know the effects of the referential context from the youngest age studied. Children develop in their knowledge of the principles of the morphosyntax-discourse interface and of language-specific Spanish morphology
	☐ Student project
	☑ Completed
Comments	

Figure 2. DTA Project Information

4.1.3 Student background

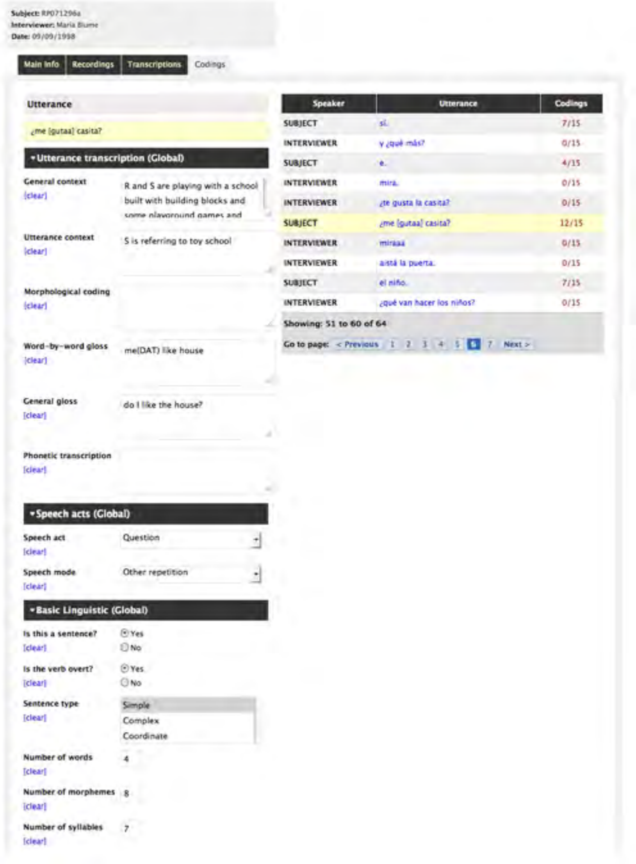
Students in need of linguistic background, get it through our course presentations, readings, manual, and glossary. Students without background in psychology or experimental methods receive training in research design and particular methods for data collection through our learning topics dedicated to basic concepts on scientific research, and through several others dedicated to particular methods. Students also read relevant chapters of our VCLA Research Methods Manual [10] and assigned papers, use the Experiment Bank component of the DTA tool to see detailed examples of previous research, do assignments that have them extract the research design of a published paper and enter it in the DTA tool, and complete a final project in which they are required to design their own study. Finally, interactive assignments teach

[17] These tools will be explained in more detail in section 4 when we describe their educational and research features.

all students to transcribe (cf. 4.2.2.1), reliability check[18], and analyze previously collected language data, using the DTA or the original project's format. For example, in the Elicited Imitation learning module in the VLL, students get access to information on this research task and to a set of research articles using elicited imitation as their primary method. In the assignments they have the option to focus on different projects related to the articles they read. They can then see samples of the original session recordings for the projects which they can score using systematic scoring guidelines and materials of the project. Then they can compare their own results to those reported in the article. They can also look at all the details of the relevant research project by looking at the project information in the DTA Tool. This hands-on experience with real research material is fundamental to help students understand all steps of research development, from creating a researchable question to designing a research project, to collecting data, testing hypotheses, and relating results to previous research.

The DTA tool provides coding sets that train students in first steps of linguistic analysis in a theory-neutral design.[19] Students and researchers coding natural speech data are expected to use all these basic codings, so that the data are calibrated across projects. Figure 3 exemplifies basic coding of an utterance of natural speech data of a Peruvian monolingual Spanish-speaking child from the "Spanish Natural Speech Corpus-Blume", such as the ones that are coded by our students in their assignments. Such codings render the data ready for further analyses in connection with specific research questions.

In this way, the student or new researcher can create their own dataset and begin asking questions regarding how the child acquires the knowledge of question formation.



Subject: RP071296a
Interviewer: Maria Blume
Date: 09/09/1998

Utterance

¿me [gutaa] casita?

▾Utterance transcription (Global)

General context [clear]: R and S are playing with a school built with building blocks and some playground games and

Utterance context [clear]: S is referring to toy school

Morphological coding [clear]:

Word-by-word gloss [clear]: me(DAT) like house

General gloss [clear]: do I like the house?

Phonetic transcription [clear]:

▾Speech acts (Global)

Speech act [clear]: Question

Speech mode [clear]: Other repetition

▾Basic Linguistic (Global)

Is this a sentence? [clear]: ● Yes ○ No

Is the verb overt? [clear]: ● Yes ○ No

Sentence type [clear]: Simple / Complex / Coordinate

Number of words [clear]: 4

Number of morphemes [clear]: 8

Number of syllables [clear]: 7

Speaker	Utterance	Codings
SUBJECT	sí.	7/15
INTERVIEWER	y ¿qué más?	0/15
SUBJECT	e.	4/15
INTERVIEWER	mira.	0/15
INTERVIEWER	¿te gusta la casita?	0/15
SUBJECT	¿me [gutaa] casita?	12/15
INTERVIEWER	miraaa	0/15
INTERVIEWER	aistá la puerta.	0/15
SUBJECT	el niño.	7/15
INTERVIEWER	¿qué van hacer los niños?	0/15

Showing: 51 to 60 of 64

Go to page: < Previous 1 2 3 4 5 6 7 Next >

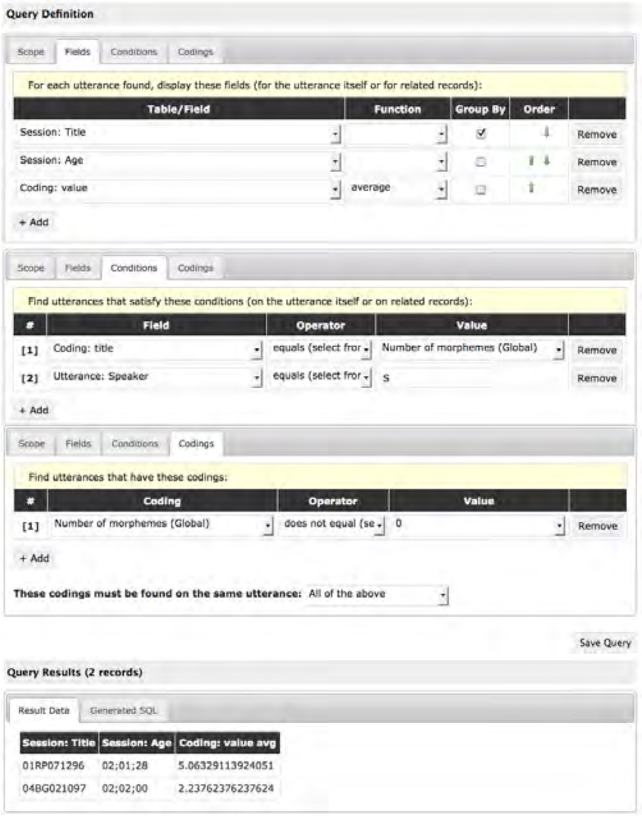
Figure 3. Basic linguistic coding screen.

A set of basic queries, which is essential to calibrating language data, is available in the DTA tool. Queries can be run on all sessions that have been coded for the relevant features in all projects in the DTA tool, thus linking across sessions, subjects, and projects. These basic queries are common queries in the field that are ready-made for the researcher, and also serve as teaching examples for students who can use them to complete assignments on selected samples of language data. The query screens are designed to guide the student or researcher through the different necessary steps to do the computation that would answer their search question. For example, students are asked to compute the subject's Mean Length of Utterance or MLU, a common measure of a child's language development, and compare the MLU they find with the MLUs for the different developmental stages reported for the subject's language. To do this they are required to code manually the number of morphemes in each of the utterances produced by the subject. Then they run a simple query that adds the utterance counts (looking only at the utterances produced by the subject) by the total number of utterances produced by the subject. Figure 4 shows one such query run on two children of comparable age from two different corpora[20]

[18] Reliability checking is the process by which a researcher's transcription or coding is cross-checked with that of another researcher to establish its reliability.

[19] These basic linguistic codings (analyses) include: an utterance-level coding set (i.e., literal gloss, general gloss, and pragmatic context), a speech act coding set (e.g., speech act and speech mode), and a basic linguistic coding set (e.g., sentence codings and syllable, morpheme and word counts).

[20] Spanish Natural Speech Corpus-Blume and English Natural Speech Corpus-Lust.



Query Definition

For each utterance found, display these fields (for the utterance itself or for related records):

Table/Field	Function	Group By	Order	
Session: Title		☑	↓	Remove
Session: Age		□	↑ ↓	Remove
Coding: value	average	□	↑	Remove

+ Add

Find utterances that satisfy these conditions (on the utterance itself or on related records):

#	Field	Operator	Value	
[1]	Coding: title	equals (select fror	Number of morphemes (Global)	Remove
[2]	Utterance: Speaker	equals (select fror	S	Remove

+ Add

Find utterances that have these codings:

#	Coding	Operator	Value	
[1]	Number of morphemes (Global)	does not equal (se	0	Remove

+ Add

These codings must be found on the same utterance: All of the above

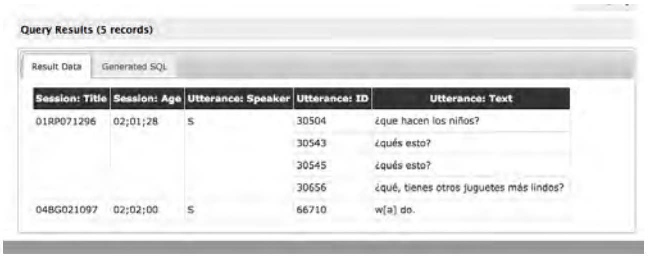
Save Query

Query Results (2 records)

Session: Title	Session: Age	Coding: value avg
01RP071296	02;01;28	5.06329113924051
04BG021097	02;02;00	2.23762376237624

Figure 4. MLU query example.

Students are also asked to do queries searching for particular speech acts and sentence types and subtypes; for example, all utterances produced by the subject that have a question speech act, that are at the same time *wh*-questions[21] and simple sentences. Figure 5 shows the results of one such query run on the same two subjects of the query on figure 4.



Query Results (5 records)

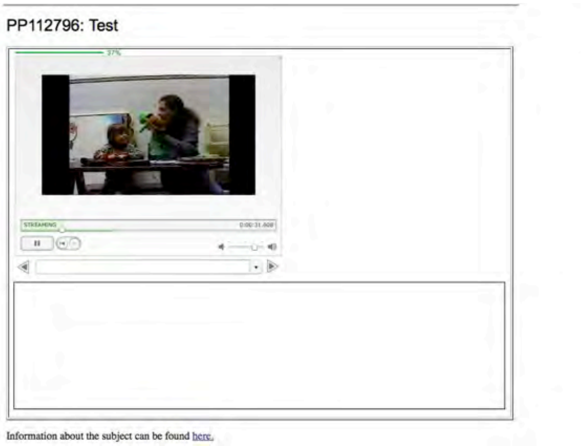
Session: Title	Session: Age	Utterance: Speaker	Utterance: ID	Utterance: Text
01RP071296	02;01;28	S	30504	¿que hacen los niños?
			30543	¿qués esto?
			30545	¿qués esto?
			30656	¿qué, tienes otros juguetes más lindos?
04BG021097	02;02;00	S	66710	w[a] do.

Figure 5. *Wh*-question, simple sentence query.

Students are also asked to create some queries of their own invention so that we know they understand the logic behind the search engine and also to check that they are able to generate new research questions for an existing dataset. Thus all students with all backgrounds are taught methods of data management and analysis as well as research inquiry.

4.1.4 Human subjects

Various topics provide the student and/or researcher with the virtual experience of working with human subjects through audio-visual examples of research sessions in each learning module. They allow students to learn a method to use in their own research before going into the field. Figure 5 demonstrates an experimental study using the Elicited Imitation task done with a 2-year-old in Peru (Discourse Morphosyntax Interface in Spanish Non-Finite Verbs-Blume).



PP112796: Test

Information about the subject can be found here.

Figure 5. A video showing a particular research study that exemplifies the Elicited Imitation method.

Initial VLL topics integrate the IRB training and tests to ascertain that all students comply with their institution's regulations in this respect and learn Human Subjects requirements in general.

4.2 Data Challenges

4.2.1 Data complexity

The DTA cybertool in the VLL provides a structured annotation scheme for the representation of layers of metadata related to language data (i.e., language utterances). It does so in addition to providing structure for representation of reliability-checked language transcriptions and analyses of the utterances in those transcriptions. It thus helps to make data complexity tractable. Figure 6 provides an overview of the tool's structure showing the major area of data and metadata entry from a user's perspective.

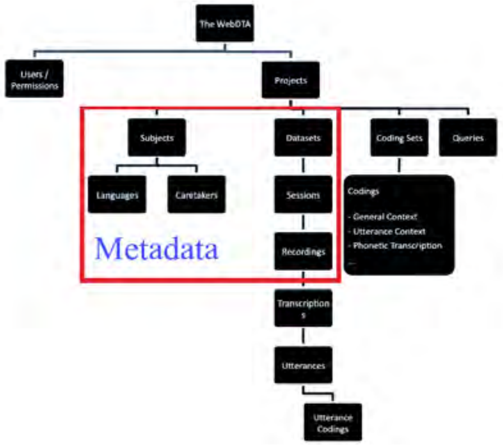


Figure 6. DTA structure diagram.

[21] Questions that in English start with a *wh*-word such as *what, why, how,* etc.

The DTA tool is based on 10 tables with the following basic markup categories: Project, dataset, subject, session[22], recording,[23] transcription, utterance, coding set, coding, and utterance coding.[24] Metadata codings involve the project and subject levels (figure 7) and the datasets themselves (figure 8) leading to transcribed utterances and related linguistic codings.

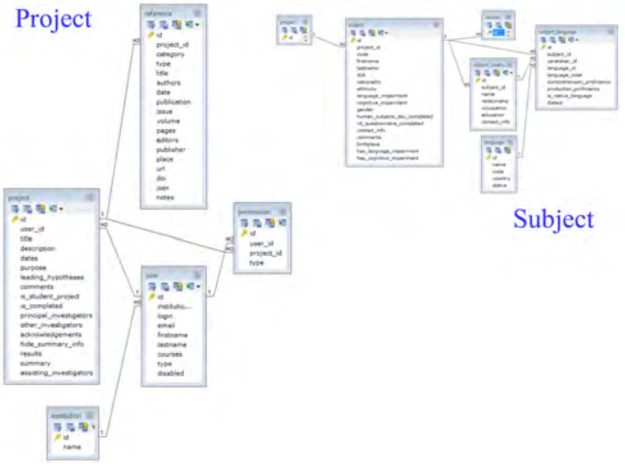


Figure 7. Project and subject metadata

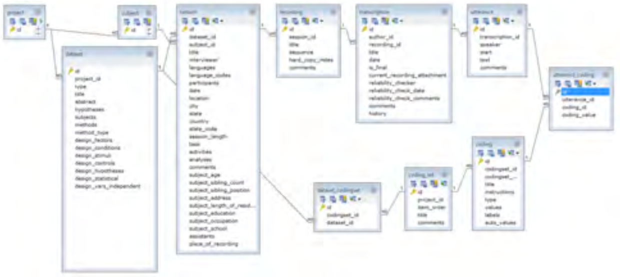
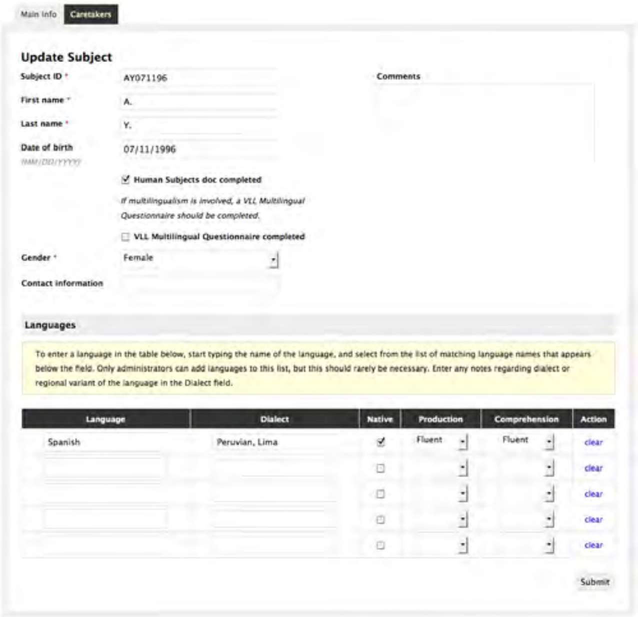


Figure 8. Dataset metadata.

4.2.2 *Data documentation and standardization*

The DTA tool provides the user with a web interface that guides him/her through steps for generating, storing and accessing both metadata and data. It trains researchers and students on how to organize research data. Users contribute data in a structured, uniform manner and they access calibrated data from a shared relational database. Therefore diverse data become comparable at many levels.

After entering the project's background information (cf. section 4.1.2), the student and/or researcher enters subject information, exemplified in Figure 9.[25]



Update Subject

Subject ID *	AY071196	Comments
First name *	A.	
Last name *	Y.	
Date of birth MM/DD/YYYY	07/11/1996	
	☑ Human Subjects doc completed	
	If multilingualism is involved, a VLL Multilingual Questionnaire should be completed.	
	☐ VLL Multilingual Questionnaire completed	
Gender *	Female	
Contact information		

Languages

To enter a language in the table below, start typing the name of the language, and select from the list of matching language names that appears below the field. Only administrators can add languages to this list, but this should rarely be necessary. Enter any notes regarding dialect or regional variant of the language in the Dialect field.

Language	Dialect	Native	Production	Comprehension	Action
Spanish	Peruvian, Lima	☑	Fluent	Fluent	clear
		☐			clear
		☐			clear
		☐			clear
		☐			clear

Submit

Figure 9. The subject screen.

Then, a series of screens help students and researchers represent the research study's design. For each dataset, the user provides the main information (experiment/investigation, topic, abstract, related WebDTA projects/datasets), hypotheses, general subject description, methods, design, stimuli, procedures, and scoring procedures. These screens have an important educational purpose in teaching students how a particular experiment is designed. When a research project is completed, results, and conclusions can be linked.

Next, a session information screen guides the user to enter information for every time a subject was recorded for a given dataset. Each session has associated to it a recordings screen, a transcription screen, and a coding screen. The recordings screen houses information on all available primary data for a given session (audio or video files or previous transcripts in a number of formats) plus an inventory of the location of such files, and their backups.[26] The user then moves to a transcription screen where he/she can watch and listen to all available recordings (switching from one to the other as needed), transcribe and manually set timings to align the transcript with the recordings. The transcription screen is shown in figure 10.

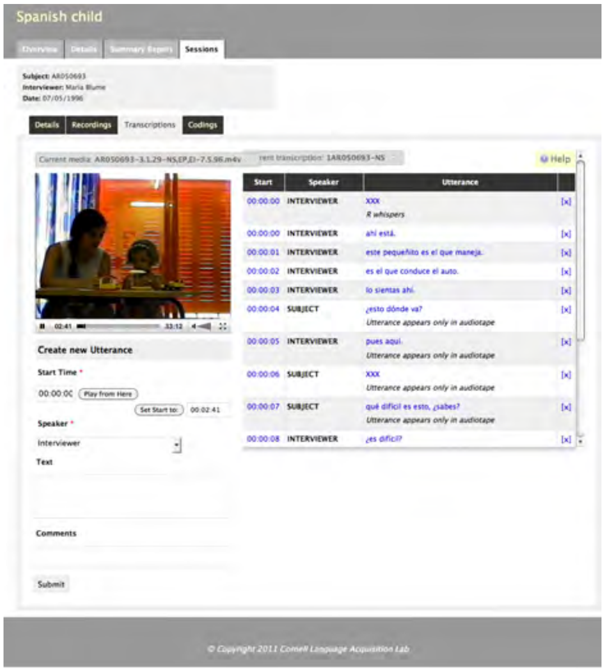
[22] A 'session' refers to a particular time in which a particular set of language data is recorded.

[23] Digital audio or video file, an electronic document (e.g. a Word, Excel, or PDF file), or information in a non-digital format such as a tape recording or paper transcription.

[24] An utterance coding records specified linguistic values (which the DTA tool refers to as 'codings') for a given utterance.

[25] In this figure, the confidential information of the subject has been eliminated. Only this screen provides confidential identifying data from subjects. The rest of the screens refer to subjects by an anonymous ID (initials and date of birth). Confidential data are only open to project PIs and selected primary collaborators. As the project moves to a diffusion stage, permission levels will structure this access.

[26] This screen supports all files supported by the JW Player,[26] *QuickTime* player, PDF, HTML, and image files, and, with additional software, other file formats such as *Microsoft Office* files.



Subject: AR050693
Interviewer: Maria Blume
Date: 07/05/1996

Current media: AR050693-3.1.29-NS,EP,EI-7.5.96.m4v
rent transcription: 1AR050693-NS

Start	Speaker	Utterance	
00:00:00	INTERVIEWER	xxx *R whispers*	[x]
00:00:00	INTERVIEWER	ahí está.	[x]
00:00:01	INTERVIEWER	este pequeñito es el que maneja.	[x]
00:00:02	INTERVIEWER	es el que conduce el auto.	[x]
00:00:03	INTERVIEWER	lo sientas ahí.	[x]
00:00:04	SUBJECT	¿esto dónde va? *Utterance appears only in audiotape*	[x]
00:00:05	INTERVIEWER	pues aquí. *Utterance appears only in audiotape*	[x]
00:00:06	SUBJECT	xxx *Utterance appears only in audiotape*	[x]
00:00:07	SUBJECT	qué difícil es esto, ¿sabes? *Utterance appears only in audiotape*	[x]
00:00:08	INTERVIEWER	¿es difícil?	[x]

02:41 33:12

Create new Utterance

Start Time *
00:00:00
00:02:41

Speaker *
Interviewer

Text

Comments

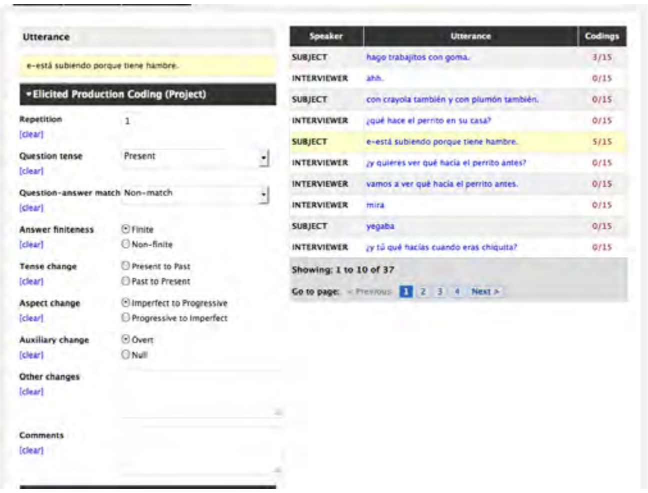
Figure 10. Transcription screen.

The utterances in these transcriptions, once they are reliability checked, constitute the basis for linguistic analyses.[27] Finally, researchers and students can code their data.

Through this tool we achieve systematization and calibration of metadata and data, thus allowing collaborative research programs and addressing the challenges of data documentation and standardization.

4.2.3 Data management and dataset design

As mentioned above, we wanted to design a structure that was open enough to accommodate both experimental and natural speech data from various types of populations[28] because such a resource did not exist in our field. The creation of this database, thus, provides previously unavailable resources for collaboration among researchers. The DTA tool structures data creation and analysis but allows the researcher to create project-specific coding sets and queries. Figure 11 illustrates a specific coding set created for an utterance from a Peruvian child participant in the experimental Project, "Discourse Morphosyntax Interface in Spanish Non-Finite Verbs-Blume" where language production is being systematically elicited from a child by the experimenter following an experimental design.



Utterance
e-está subiendo porque tiene hambre.

▾Elicited Production Coding (Project)

Repetition [clear]: 1
Question tense [clear]: Present
Question-answer match [clear]: Non-match
Answer finiteness [clear]: ● Finite ○ Non-finite
Tense change [clear]: ○ Present to Past ○ Past to Present
Aspect change [clear]: ● Imperfect to Progressive ○ Progressive to Imperfect
Auxiliary change [clear]: ● Overt ○ Null
Other changes [clear]
Comments [clear]

Speaker	Utterance	Codings
SUBJECT	hago trabajitos con goma.	3/15
INTERVIEWER	ahh.	0/15
SUBJECT	con crayola también y con plumón también.	0/15
INTERVIEWER	¿qué hace el perrito en su casa?	0/15
SUBJECT	e-está subiendo porque tiene hambre.	5/15
INTERVIEWER	¿y quieres ver qué hacía el perrito antes?	0/15
INTERVIEWER	vamos a ver qué hacía el perrito antes.	0/15
INTERVIEWER	mira	0/15
SUBJECT	yegaba	0/15
INTERVIEWER	¿y tú qué hacías cuando eras chiquita?	0/15

Showing: 1 to 10 of 37

Figure 11. Project specific linguistic coding set.

At the same time that the DTA tool provides a primary research tool, it automatically provides a rich, continually growing archive allowing present and future collaboration on shared data, potentially long distance and potentially interdisciplinary. In general, with external linkages, through Linked Data formats [5, 8, 13, it can be linked to a wide intellectual knowledge base, e.g., linking published forms of research to the actual data and data methods used to create the results reported.[29]

4.3 Researcher Collaboration

4.3.1 Researcher training

Membership in the VCLA provides researchers a new medium for accessing peers at other institutions and countries to engage in collaborative projects under considered principles for collaboration and data sharing, and our web-conferencing allows them to have online meetings. The VLL portal materials provide essential readings on issues related to distance collaboration. The DTA tool helps researchers find detailed information about other researchers and student projects and helps organize collaborative research materials for a specific project. Researchers, like students, can get trained in cybertool use through the virtual workshops and the DTA User manual.

In addition, the tool allows continual generation of new queries on data based on codings that derive from a particular research question, so that each researcher or student researcher can get the results they are looking for in their specific projects. Data analyses are cumulative, as they are stored in the resulting database. For more detail on this aspect of the DTA tool see [8].

[27] The DTA tool keeps a record of all recordings and transcriptions available for a particular session. The user can easily switch between recordings and transcriptions to view all versions of the raw and primary data. In addition, it keeps a record of who was the transcriber and when was the date of the first transcription, as well as of any subsequent reliability checked versions of a transcript, including reliability checker and date information.

[28] e.g. child vs. adult, first vs. second language acquisition, child vs. adult second language acquisition, normal vs. disabled populations.

[29] In order to maximize generalizability across fields of our tools, The DTA tool is designed to maximize the possibility for linked data by integrating with field standards. For example, the application uses the UTF-8 encoding to store text, which can represent any language. For this, the application adopts ISO 639-3 standard language codes [36], which lists over 7000 languages, developed by Ethnologue/SIL (http://www.ethnologue.com/codes/default.asp). It links with GeoNames (http://www.geonames.org/) [20] in geographic reference.

4.3.2 Intellectual property

Founding members of the VCLA have, through a series of video conference meetings, begun to design principles of agreement by which to assure the protection of each VLL member's intellectual property rights, while at the same time allowing for collaboration in new projects and the repurposing of previously collected data. Although this is still work in progress, a first summary of our vision and principles can be found on the VCLA website.[30]

4.3.3 Cross-institutional and international collaboration

To begin to address the issues we identified above which challenge cross-institutional and cross-country collaboration, VCLA founding members have begun meeting collectively with representatives of the IRB committees at their institutions and have begun collecting cross-institutional data to determine commonalities and differences across them.

5. EDUCATIONAL ACHIEVEMENTS TO DATE

A structured series of synchronous cross-institutional courses, including two with Peru, at the undergraduate and graduate levels have introduced the components of the VLL through our structured VLL web portal. Students from Computer Science, Human Development, Linguistics, and Psychology participated in these courses and web conferences.[31] These several courses took students through initial introduction to scientific methods for data collection and management, followed by advanced cybertool learning through specialized research agendas. A series of cross-institutional web-conferences supplemented these courses in order to cultivate collaboration among students and faculty.

Our cross-institutional and international courses gave students new perspectives. For example, in a course on bilingualism three different professors and three groups of students (University of Texas at El Paso, Rutgers University, New Brunswick, and Pontificia Universidad Católica del Peru) participated and shared information on three different multilingual situations: New Jersey: dominant English, Spanish minority language, El Paso: border city with majority Hispanic population and strong ties to Mexico, and Peru: Spanish dominant language and various indigenous languages.

Students, through use of the VLL, engaged in several stages of original data analyses, culminating in original experimental research proposals. [32]

The synchronous courses provided accumulated syllabi, materials and assignments that were used asynchronously as well.

Our main educational achievement has been to train students from the beginning on documenting data in such a way that will spare them on having to do what previous generations of researchers, us included, spend too much time having to do, i.e., find our old data, find our old records, find our old tapes, try to connect them all, hunt up metadata, experimental designs and stimulus sentences, etc. In this exploratory stage of our project, we have educated a small section of a new generation of students which can now pass such information to colleagues and future students, so that our efforts will, hopefully, payoff in the future.

We have also exposed these students to all the arguments in favor of large-scale data sharing and research collaboration, as well as to several of the problems surrounding such collaborative projects so that they can avoid pitfalls in the future. This is a topic that is not traditionally discussed in our field. We have, thus, planted a seed and achieved some beginning collaborative projects; whether students will embrace this new culture is yet to see, but we have at least given them the opportunity, the technology, and the computational skills to do so.

Student surveys conducted during synchronic cross-institutional courses to date have indicated high satisfaction with the course and in particular with its cybertools.[33] Critically, students asked for more interaction with students at other universities, indicating a positive inclination towards collaboration.[34] Figure 12 shows the overall results of the surveys.

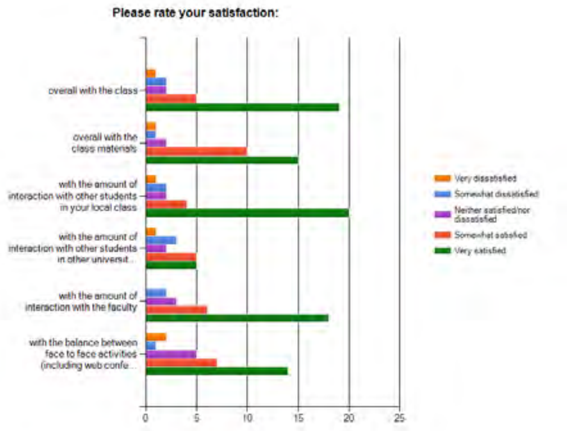


Figure 12. Student satisfaction survey.

[30] http://vcla.clal.cornell.edu/en/principles#collaboration

[31] At this point we have no information available for some semesters in which we taught the courses due to changes in server administrators and a lack of recognition on our part that we had to keep a record of all users before deleting their accounts. We do have information on some semesters that were not different from other semester in terms of access. Fall 2010, 67 new users accessed our materials and courses. For 2011, 35 new accounts were created, but it is important to bear in mind that users from previous semesters/years do not need to get new access each semester, so more students from previous semesters were actually still using our VLL.

[32] Examples of collaborative research projects including students and researchers are at Cornell (Barbara Lust): "SAQL Phase 1: Expert Evaluation and Validation of a New Child Multilingualism Questionnaire", Newcastle University (Cristina Dye) and Boston College (Claire Foley): "Acquisition of VP ellipsis in mono and bilingual children"; MIT (Suzanne Flynn), Massachusetts General Hospital (Janet Cohen Sherman) and Cornell (Barbara Lust): "Alzheimer's language project" with Jordan Whitlock.

[33] We provided all students the opportunity to answer these end-of-semester online surveys. Only a small proportion (n. 29) of the students who took the courses synchronously answered the survey. We are looking at alternative ways of distributing surveys in the future to improve the number of respondents.

[34] In general, this part of the project suffered from scheduling and language barrier problems. We are discussing how to improve this in future courses.

6. BROAD IMPACT

This is the first project of its kind in the social sciences and humanities and thus can serve as a model for other Social and Cognitive Sciences, as well as other STEM sciences.

It empowers a wide array of collaborative and interdisciplinary research and teaching agendas. It incorporates sound scientific principles and structured data management in a cybertool that provides a distributed infrastructure for collaborative learning and research in the study of language, bilingualism, and language development.

It creates new learning and research possibilities for Hispanic students, usually underrepresented in the sciences, at University of Texas at El Paso, Rutgers University, New Brunswick, and Peruvian students at Pontificia Universidad Católica del Perú.

7. FUTURE CHALLENGES AND LESSONS LEARNED

Our main challenges now, as we approach the diffusion stage of our project, concern the dissemination of our infrastructure and materials to a broader community of practice, one that is both interdisciplinary and cross-linguistic.

In this, we face issues of sustainability. In order to open the VLL materials to a wide audience, we must build a sustainability model that includes licensing and/or subscription options. For this we have now initiated correspondence with Cornell's E-Cornell program (eCornell.com). To ensure long-term sustainability, we must also negotiate and fund long-term storage and maintenance of the DTA tool and its database. We must develop an infrastructure for long-term management of the tool and its access and use. In our view this would ideally be some form of a distributed infrastructure rather than a localized one.[35]

To extend the DTA tool to new users we must establish a set of user principles and agreements involving shared materials and data. This must involve establishment of a leveled set of permissions, e.g., read only, etc. Founding Members of the VCLA are currently addressing this challenge.

Some practical issues also create challenges for the development of a project such as ours. Ironically, one of them is language. To teach our first international class with Peru, we had to translate most materials to Spanish, and since the class was taught in Spanish, not all US sites were able to participate in our class discussions. Coordinating schedules across US and international time zones for joint courses also proved to be a challenge for those who wanted to participate synchronously in our courses. These challenges will arise with each new language and country added to the project (e.g., India, Korea, Israel planned for extensions). One of our next and continuous challenges includes translating materials to other languages and possibly providing interpretation to allow better collaboration across countries. Technical and administrative challenges in cybertool development required additional costs and time beyond that first expected.

Among the lessons learned is, hence, the fact that cross-institutional collaboration demands precisely planned infrastructure. Another issue that confronted us was how hard it is to foster cross-institutional and international collaboration, even when tools for collaboration are in place and the desire of collaboration exists in all parties. While students were relatively easily encouraged to collaborate, time constraints and previous commitments on faculty,[36] plus a lack of real support for collaborative work by academic institutions, as observed in *Nature* [39], will require additional support for faculty time and commitment, if collaborative projects such as this are to flourish.

8. ACKNOWLEDGMENTS

8.1 Founding

National Science Foundation

CI-TEAM program. "Transforming the Primary Research Process Through Cybertool Dissemination: An Implementation of a Virtual Center for the study of language acquisition" (Award No. OCI-0753415, María Blume and Barbara Lust)

Planning Grant: A Virtual Center for Child Language Acquisition Research (BCS-0126546) (Barbara Lust and Founding Members) NSF Development and Learning Sciences Program

Small Grant for Exploratory Research. Planning Information Infrastructure through a New Library-Research Partnership. (NSF#0437603) (Janet McCue, director of Cornell Albert Mann Library and Barbara Lust) NSF Science and Engineering Information Integration and Informatics.

Cornell:

CISER Seed Grant: Building an Interdisciplinary Infrastructure to Foster Collaboration among Research Scientists in the Language Sciences: Proposal to Create Instrument-based Virtual Linguistics Laboratory for the International Study of Language Acquisition and Neural Plasticity.

FABIT (Faculty Innovation in Teaching) Award. "Integrating Digital Multimedia resources in two interdisciplinary language development courses" and "A VRE (Virtual Research Environment) for Language Acquisition and Cognitive Science. "

New York State Hatch grant. "Cornell Language Acquisition Lab: Building an Electronic Library of Words of the World's Children".

Seed Grant from Cornell Einaudi Center for International Studies: Exploring Activation of an International component in a Virtual Center for the Study of Language Acquisition.

8.2 Collaborators

We would also like to acknowledge the contributions to this project of:

VCLA members and affiliates: Elise Temple (NeuroLab) Gita Martohardjono (CUNY), Cristina Dye (Newcastle University), Isabelle Barrière (CUNY), Yarden Kedar (Ben Gurion University at the Negev), and Sujin Yang (Tyndale University College and Seminary)

Application developers: Ted Caldwell and Greg Kops (GORGES). Susan Mehringer (Cornell University Center for Advanced Computing) and Tommy Cusick (Google, cf. footnote 16 above), Cliff Crawford for early web DTA tool programming, Alex Immerman for portal and cybertool development.

[35] [8] lists further challenges specific to the DTA tool.

[36] "[…] As Gibbons and anthropologist Nancy Fried Foster observed in their 2005 postmortem, «The phrase 'if you build it, they will come' does not apply to IRs [institutional repositories].»" [28, p. 160].

Creators of previous versions of our DTA tool: Katharina Boser, David Parkinson, and Shamita Somashekar, then graduate students at Cornell.

Student RAs: Darlin Alberto, Gabriel Clandorf, Natalia Buitrago, Poornima Guna, Jennie Lin, and Jordan Whitlock at Cornell, and Marina Kalashnikova. Martha Rayas Tanaka, Lizzeth Pattison, María Jiménez, and Mónica Martínez at UTEP.

Undergraduate and graduate students who have participated in the courses and helped with their input at UTEP, Cornell, Rutgers New Brunswick, MIT, and Pontificia Universidad Católica del Perú.

9. REFERENCES

[1] "A fair share." *Nature* 444, 7120 (2006): 653-654.

[2] Atkins, D. 2005. CyberInfrastructure and the Next Wave of Collaboration, D. E. Atkins, Keynote for EDUCAUSE Australasia, Auckland, New Zealand, April 5-8, 2005.

[3] Bender, E. 2004. Rules of the Collaboratory Game. Science of Collaboratories papers: *MIT's Technology Review*. November 23, 2004

[4] Berman, F. and H. Brady. 2005. Workshop on Cyberinfrastructure for the Social and Behavioral Sciences: Final Report. http://*ucdata.berkeley.edu/pubs/**CyberInfrastructure_FINAL**.pdf*

[5] Berners-Lee, T. 7/26/2006. Linked data. http://www.w3.org/DesignIssues/LinkedData.html

[6] Bird, S. and G. Simons (2003). Seven dimensions of portability for language documentation and description. *Language*, vol. 79, 3. (557-582)

[7] Birney, E., T.J. Hudson, E.D. Green, C. Gunter, S. Eddy, J. Rogers, J.R. Harris, and S. Dusko Ehrlich. 2009. "Prepublication data sharing" *Nature* 461, 7261: 168-170.

[8] Blume, M., S, Flynn, and B. Lust. 2012. Creating Linked Data for the Interdisciplinary International Collaborative Study of Language Acquisition and Use: Achievements and Challenges of a new Virtual Linguistics Lab. In C. Chiarcos, S. Nordhoff, and S. Hellmann (Eds.) *Linked Data in Linguistics: Representing Language Data and Language Metadata*. Berlin/Heidelberg: Springer, pp. 85-96.

[9] Blume, M. and B. Lust, 2012. Data Transcription and Analysis Tool User's Manual. (with the collaboration of S. Somashekar and T. Ogden). http://webdta.clal.cornell.edu

[10] Blume, M., S. Yang, and B. Lust. (with the collaboration of T. Ogden, S. Somashekar, Y. Chien, L. Sánchez, C. Foley, M. Kalashnikova, M. Rayas, and N. Buitrago)(in prep) Cornell University Virtual Linguistics Lab (VLL) Research Methods Manual: Scientific Methods for Study of Language Acquisition.

[11] Borgman, C. 2007. *Scholarship in the Digital Age*. Cambridge: MIT Press.

[12] Bos, N., A. Zimmerman, G. Olson, J. Yew, J. Yerkie, E. Dahl, and G. Olson. 2007. From shared databases to communities of practice: A taxonomy of collaboratories. *Journal of Computer-Mediated Communication*, *12* (2), article 16. http://jcmc.indiana.edu/vol12/issue2/bos.html

[13] Chiarcos, C., S. Nordhoff, and S. Hellmann (Eds.) *Linked Data in Linguistics: Representing Language Data and Language Metadata*. Berlin/Heidelberg: Springer.

[14] "Credit where credit is overdue" *Nature Biotechnology* 27, 7 (2009): 579.

[15] "Data's shameful neglect" *Nature* 461, 7261 (2009): 145.

[16] Edwards, J. A. 1992a. Transcription of discourse. *International Encyclopedia of Linguistics*, ed. by William Bright, 367-370. Oxford: Oxford University Press.

[17] Edwards, J. A. 1992b. Computer methods in child language research: four principles for the use of archived data. *Journal of Child Language* 19.435-58.

[18] Elkins, K. 2012. Tiptoeing through Minefields: Launching Collaborations. *Association for Women in Science*, Winter 2012*,* vol. 43, no. 1, pp. 21-23.

[19] E-MELD: Electronic Metastructure for Endangered Languages Data http://www.emeld.org/index.cfm

[20] GeoNames: http://www.geonames.org/.

[21] GOLD Community: http://www.linguistics-ontology.org/

[22] Grosjean, F. 1998, 2004. Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and Cognition*, 1, 131-149. And in T. K. Bhatia & W. C. Ritchie (Eds.). *The Handbook of Bilingualism* (pp. 32-63). Oxford, England: Blackwell Publishing.

[23] Grotke, R. W. 2004. Digitizing the world's largest collection of natural sounds: key factors to consider when transferring analog-based audio materials to digital formats. *RLG DigiNews*, Vol. 8, Number 1. Online: http://worldcat.org/arcviewer/2/OCC/2009/08/11/H1250010262952/viewer/file2.html

[24] King, G. 2011. "Ensuring the Data-Rich Future of the Social Sciences" *Science*, 331, 1: 719-721.

[25] "Let data speak to data." *Nature* 438, 7068 (2005): 531.

[26] Lust, B., M. Blume, Y., Kedar, S. Yang, and S. Callahan. A Glossary of Language Acquisition.

[27] Lust, B., S. Flynn, M. Blume, E. Westbrooks, and T. Tobin. 2010. Constructing Adequate Language Documentation for Multifaceted Cross-Linguistic Data: A Case Study from a Virtual Center for Study of Language Acquisition. In L. A. Grenoble and N. L. Furbee (eds.). *Language Documentation: Practice and Values*. pp. 89-107. Amsterdam and Philadelphia: John Benjamins.

[28] Nelson, B. (2009) "Empty archives" *Nature* 461, 7261: 160-163.

[29] NSF. 2003. *Revolutionizing Science and Engineering through Cyberinfrastructure*.

[30] NSF. 2007. *Cyberinfrastructure Vision for 21st Century Discovery*. March NSF 07-28.

[31] OLAC: Open Language Archives Community, http://www.language-archives.org/

[32] OLWG: Open Linguistics Working Group, http://linguistics.okfn.org.

[33] Olson, G., A. Zimmerman, and N. Bos. (eds.) 2008. *Scientific Collaboration on the Internet.* MIT Press.

[34] Pfirman, S., J. Collins, S. Lowes, and A. Michaels. 2005. February 11. Collaborative Efforts: Promoting Interdisciplinary Scholars. The Chronicle Review. *The Chronicle of Higher Education*. Vol. 51, issue 23, page B15. http://chronicle.com.

[35] Schofield, P.N., T. Bubela, T. Weaver, L. Portilla, S.D. Brown, J.M. Hancock, D. Einhorn, G. Tocchini-Valentini, M. Hrabe de Angelis, N. Rosenthal, and CASIMIR Rome Meeting participants. 2009. "Post-publication sharing of data and tools" *Nature* 461, 7261: 171-173.

[36] SIL. 2006. ISO/DIS 639-3. Dallas: SIL International. http://www.sil.org/iso639-3/.

[37] Simons, G. 2004. Ensuring that Digital Data Last. The priority of archival form over working form and presentation form. Paper presented at Symposium on Best Practice. Linguistic Society of America Annual Meeting, Boston, Mass.

[38] Wenger, E. and J. Lave. 1998. *Communities of practice: Learning, Meaning, and Identity*. N.Y.: Cambridge University Press.

[39] "Who'd want to work in a team?" *Nature* 424, 6944 (2003):1.

Institutional and Individual Influences on Scientists' Data Sharing Practices

Youngseek Kim
Syracuse University
221 Hinds Hall
Syracuse, NY 13244
+1-315-443-4508
ykim58@syr.edu

Jeffrey M. Stanton
Syracuse University
206 Hinds Hall
Syracuse, NY 13244
+1-315-443-2879
jmstanto@syr.edu

ABSTRACT

Many contemporary scientific endeavors now rely on the collaborative efforts of researchers across multiple institutions. As a result of this increase in the scale of scientific collaboration, sharing and reuse of data using private and public repositories has increased. At the same time, data sharing practices and capabilities appear to vary widely across disciplines and even within some disciplines. This research sought to develop an understanding of this variation through the lens of theories that account for individual choices within institutional contexts. We conducted a total of 25 individual semi-structured interviews to understand researchers' current data sharing practices. The main focus of our interviews was: (1) to explore domain specific data sharing practices in diverse disciplines, and (2) to investigate the factors motivating and preventing the researchers' current data sharing practices. Results showed support for an institutional perspective on data sharing as well as a need for better understanding of scientists' altruistic motives for participating in data sharing and reuse.

Keywords

Data Sharing, Data Reuse, Data Repository, Institutional Theory, Theory of Planned Behavior, IT Capability, Altruism

1. INTRODUCTION

As the scope and scale of science has increased, sharing and reuse of data have become essential to many scientific and engineering activities. In the 2003 report entitled, "Revolutionizing Science and Engineering through Cyberinfrastructure," members of a blue ribbon National Science Foundation (NSF) panel wrote, "We envision an environment in which raw data and recent results are easily shared, not just within a research group of institution but also between scientific disciplines and locations" [2]. Years later researchers are realizing this vision in some disciplines and sub-disciplinary areas such as high energy physics [6], climate change [19], and proteomics [22]. In other fields, however, and particularly the social sciences [34], progress has been very slow. Although scientists in these areas generate considerable amounts of valuable data every year, disciplinary traditions, institutional barriers, intellectual property concerns, and other factors appear to impede the sharing and reuse of data.

For example, even though the American Psychological Association (APA) mandates data sharing for researchers who publish articles in their flagship journals, Wicherts et al. [35] found it difficult to convince 103 out of 141 research teams who had published with APA to fulfill this responsibility, despite repeated attempts and extensive assurances that the requested data would not be publicly released or reused. While it is tempting to attribute this failure to particular characteristics or situations in that discipline (e.g., long publication lags), Savage and Vickers [25] experienced an even worse failure rate when requesting data from researchers who had published in two PLoS (Public Library of Science) journals – PLoS Medicine and PLoS Clinical Trials. Note that the PLoS journals reflect the new trend of "open access" in journal publishing and have explicit requirements in their editorial policies that require researchers who publish there to share their data freely with the research community. It seems evident from these examples that the idea of data sharing and reuse as a strategy to accelerate scientific discovery is appealing, but the impediments to doing so across a range of disciplines are still substantial.

Several prior studies, such as Wicherts et al. [35] and Savage and Vickers [25] have sought to document the extent of the problem in the context of different disciplines. For this paper, we take as a given that data sharing and reuse is highly variable across disciplines, and we sought to explore why this was the case. We also began with the assumption that data sharing and reuse practices were not a matter of whimsy for individual researchers, but rather that the decisions whether or not to share data for reuse (outside of one's own research group) reflected choices among communities of colleagues embedded within their universities and disciplines. More explicitly, we asked what combination of individual and contextual factors influenced scientists' decisions to share data for reuse. Because relatively little is known about this question, we elected to use the rich, qualitative data collection method of one-on-one semi-structured interviews to explore the landscape. We were guided in this exploration by a few promising theoretical perspectives that consider individual decision makers in their institutional contexts in order to understand their decisions. In the next section, we provide a brief overview of these perspectives prior to a presentation of our interview data.

2. BACKGROUND

Contemporary collaboration in nearly all of the Science, Technology, Engineering, and Mathematics (STEM) fields requires a three way combination of technological infrastructure,

Send correspondence concerning this article to: Youngseek Kim (ykim58@syr.edu) at Syracuse University's School of Information Studies.

institutional support, and interpersonal interactions. John Taylor, the former Director General at the Office of Science and Technology in Great Britain, focused the attention of that office on the development of sensors, networks, and computing infrastructure: "e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it" [16]. Using examples from the World Health Organization and the Large Hadron Collider, Sonnenwald [29] highlighted the powerful social and interpersonal aspects of scientific collaboration. Avery [3] reported the history and challenges of creating the multi-institutional Open Science Grid and drew particular attention to the institutional context that allowed this large scale cyberinfrastructure collaboration to emerge.

Although data sharing and reuse is only one facet of collaboration in the STEM fields, it represents a microcosm of these same three areas: institutions, infrastructure, and people. For example, to have a well functioning data repository with lots of raw data going in and lots of other users tapping into that data, one or more institutions must have the financial wherewithal to establish the infrastructure, publicize its existence within the community, work with the community to enhance and support the systems, and maintain the infrastructure over time. Meanwhile, individual contributors to that data repository must see personal and/or professional advantage – again in part set by the context of their home institutions, disciplinary training, and professional organizations – to contributing data into that repository. Individual contributors must also have a certain degree of mastery of the tools involved in preparing and submitting the data. Training and personnel support provided by their host organizations can lower the barriers to using these tools and preparing the data for reuse. As this scenario suggests, however, the institutions, infrastructure, and people are intimately connected in ways that are not easy to subdivide.

One perspective from sociology and organizational studies that may help to weave together the intertwined forces of institutions, infrastructure, and people arises in an area called institutional theory. While the traditional center of attention in institutional theory has been on the organizational level of analysis, neo-institutional theories add the proviso that macro-level influences affect micro-level behaviors [14]. Contemporary perspectives on institutional theory consider individual beliefs concerning proper social behavior, specifically when those beliefs arise from organizational rules, structures, and practices [5, 7, 10]. This idea meshes nicely with individual-level motivational theories (e.g., the Theory of Reasoned Action) that describe behavior as jointly influenced by attitudes, norms, and intentions.

In fact, institutional theory posits three kinds of institutional influences on behavior: coercive, normative, and mimetic pressures [8, 9, 27]. Coercive influence arises from the rules that the organization and its leaders set for desirable behavior of organizational members. Normative pressures refer to those typical behavioral patterns that are established historically either by organizations or by members of relevant professions. Newcomers to an organization or the profession must follow these patterns to succeed within the organization (or more broadly, within the industry, sector, or profession). Finally, mimetic pressures result from the observation of how other comparable organizations accomplish key tasks. Generally speaking, the leaders of one organization will observe the activities of another organization that is performing well and will seek to adopt those activities or methods for use within their own organization. Such imitation is often cast as a form of risk reduction: by following the lead of an apparently successful peer, one avoids the risks involved in alternative, novel activities that may be untested or may have unforeseen consequences.

These three forces map plausibly onto the data sharing role of the individual STEM researcher in the context of his or her organization and profession. First, institutions may have coercive pressures that they apply to foster the desired behavior by the individual. For example, funding organizations that support a researcher's work may stipulate that funding is conditional on the researcher's agreement to participate in data sharing. Savage and Vickers [25] documented such a condition in the editorial policies of the PLoS journals. Second, research disciplines (professions) may have historically-rooted practices that encourage or discourage data sharing. Particle physicists, for example, with their expensive, large scale experiments, pioneered the practice of large-scale scientific collaboration in the 1950s and 1960s [18], and thus were arguably the first discipline to actively exploit the Internet for collaborative scholarship (e.g., with arXiv, which was deployed prior to the availability of the World Wide Web). Finally, with respect to mimetic pressures, the National Science Board documented an explosion of genomic repositories – each focused on a different organism – following the success in the late 1980s of the European consortium that sequenced and published the genome of *Saccharomyces cerevisiae* (budding yeast [21]).

In classic institutional theory, coercive, normative, and mimetic pressures legitimize certain organizational structures and practices in a given sector. In turn, this legitimacy tends to foster isomorphism across many organizations within the sector. In other words, laws and regulations, acceptable practices, and copying of methods diffuse through the community of organizations, causing them to become more alike. In this paper, we are less concerned with the effects of coercive, normative, and mimetic pressures on *organizational* isomorphism and more concerned with how these pressures trickle down to influence the behavior of *individual* STEM researchers. Legitimacy and isomorphism arguably map on to the individual decision making level through their influence on individual researchers' motivations to share data. Prior research has frequently made such a linkage: For example, Shi, Shambare, and Wang [28] connected institutional theory and the Theory of Reasoned Action [1, 12, 13] to examine the adoption of Internet banking. Teo, Wei, and Benbasat [30] took a similar strategy in predicting the development of inter-organizational electronic data interchanges.

The Theory of Reasoned Action and its successor, the Theory of Planned Behavior (TPB), are well-established perspectives from social psychology that describe how salient beliefs influence behavioral intentions and subsequent behavior [1, 13]. TPB explains an individual's behavior based on his or her behavioral intention, which is influenced by his/her attitude toward the behavior, perception of the subjective norms regarding the behavior, and perceived behavioral control to conduct the behavior. Behavioral intention refers to a person's aim to perform a particular behavior. An attitude is a cognitive and emotional evaluation of an object. A subjective norm is a person's belief that people who are important to her expect that she should or should not perform a particular behavior. Perceived behavioral control is an individual's perceptions of his/her ability to perform a given behavior easily [1]. Each of the determinants of behavioral intention is in turn influenced by underlying belief structures (including behavioral, normative, and control beliefs [1, 13]).

Using a perspective such as the Theory of Planned Behavior, we suggest that individual researchers' data sharing intentions and behaviors emerge from their formation of attitudes about: 1) their beliefs about the "outcomes" of data sharing, and 2) their understanding of the normative behavior of other scientists in their field [20, 27]. In effect, as institutional and disciplinary pressures on data sharing increase due to increased data sharing among colleagues within a scientific community, individual researchers will respond to these pressures with some consideration of the merits of participating in the trend [27, 31].

Data sharing behavior may also be influenced by the controllability of the behavior. Perceived controllability is similar to a construct proposed by Bandura [4] – self-efficacy – that reflects judgments of one's own capabilities to enact a behavior successfully. With respect to data sharing behavior, a sense of perceived behavioral control may arise from their expertise (or lack thereof) in using the tools and technologies that facilitate data sharing. Likewise, a researcher's judgments about the availability of IT support within a team or organization, and the existence of data sharing standards, procedures, and data repositories may influence how likely they are to engage in data sharing [17].

In summary, this study adopts a neo-institutional theory perspective that incorporates the influence of individual motivations (e.g., as considered in the Theory of Planned Behavior) to examine how institutional factors impact individual researchers' attitudes and decisions about data sharing and reuse. Schein [26] argued that the institutional environment shapes participants' shared beliefs and, eventually, their attitudes towards certain behaviors in the environment. By focusing on STEM researchers' perceptions of the benefits and costs of data sharing, this study seeks to expose what combination of individual and contextual factors influences scientists' decisions to share data for reuse.

3. METHOD

3.1 Overview

We conducted a total of 25 individual interviews to understand STEM researchers' current data sharing practices. The main focus of our interviews was two-fold: (1) to explore domain specific data sharing practices in diverse disciplines, and (2) to investigate the factors motivating and impeding STEM researchers' current data sharing. Our Institutional Review Board (IRB) provided approval of a plan to conduct the individual interviews within three research universities in the eastern U.S. We sent a recruiting email message directly to the STEM researchers, and we also contacted department chairs to distribute the recruiting email message to their STEM researchers. We received 28 responses in total from STEM researchers in three research universities, and we ultimately interviewed 25 interviewees. The remaining three respondents could not be scheduled in time to complete data collection. In order to understand the domain specific data sharing practices in diverse disciplines, we tried to include at least one or two researchers in each research discipline (see Table 1).

All the interview sessions were audio-recorded and subsequently transcribed. All the interviews were conducted in English except one interview, which was conducted in Korean for the convenience of the interviewee. The first author transcribed the interview in Korean and then translated into English for the data analysis. Each interview took 25-35 minutes. We used an open-ended semi-structured interview method by asking similar structured interview questions to all the interviewees including STEM researchers' current data sharing methods, types of data generated and shared, their motivations and barriers of data sharing, and lastly interviewees' demographic information and work environments. An example of our interview questions was: "What motivates researchers (including you) in your field to share their data?" During the interviews, the participants were asked to answer the questions based on not only their own experience but also their observations in their research disciplines in general.

The 25 participants for the interviews include 11 tenured (full and associate) professors, eight assistant professors, one emeritus professor, one professor of practice, two post-doctoral research associates, and two doctoral candidates from three major research universities in the eastern U.S. (17 men and eight women). Given the goals of this research, we mainly interviewed professors rather than graduate students, but the two post-docs and two senior doctoral students provided perspectives that seemed complementary to the other data, so we retained them in the corpus. Table 1 shows the research disciplines of the 25 interview participants. There were a few minor differences between the names of the departments the interviewees belonged to versus their disciplinary affiliations.

Table 1. Research Disciplines of Interviewees

Discipline	Number of Interviewees
Biology	2
Chemistry	3
Computer Science	2
Ecology	5
Electrical Engineering	1
Environmental Engineering	4
Mathematics	1
Mechanical Engineering	2
Physics	3
Radiation Oncology	1
Science Education	1
Total	*25*

3.2 Data Analysis

The transcribed interviews were imported into "QDA Miner," a qualitative data analysis tool optimized for coding, annotating, and analyzing textual information. QDA Miner is designed to analyze interview or focus-group transcripts, documents (e.g. journal articles), and even images, and it also provide statistical data analysis along with content analysis and text-mining. The coding scheme was developed by using both deductive and inductive approaches. We started with ideas arising from neo-institutional theory and individual motivation perspectives to create our coding scheme. As we processed the data, we also used an inductive approach to create more specific codes. The basic coding scheme included institutional theory based constructs (coercive, normative, mimetic pressures), individual motivation based constructs (benefits and costs), and perceived controllability constructs (internal and external capabilities). The interview corpus contained 837 utterances overall; we applied codes to 276 of these utterances regarding the factors both motivating and preventing researchers' data sharing (Table 2 reports the number of respondents out of 25 interviewees whose interview contained one or more instance of each code.)

4. RESULTS

The codes and verbatims revealed STEM researchers' work environments, the types of data they commonly generated in their work, current data sharing methods (if any), and their motivations for and barriers to data sharing. In the following sections, we report on each of these topics by providing a holistic overview of what the codes and their underlying utterances revealed. Table 2 shows the coding scheme we used for the motivating and impeding factors of data sharing, provides a brief explanation of each code, and the numbers of respondents out of 25 interview participants in each code.

Table 2: Content Code Explanations and Counts

Cate-gory	Code Name	Brief Explanation	# of Responses
Coercive Pressures	Funding agency push	Funding agencies (e.g. NSF and NIH) require researchers to share their data	16
	Journal's requirement	Journal publishers require researchers to publish their data before their articles are published	9
	Special funding restrictions	Sharing private companies' and military data is restricted	6
Normative Pressures	Professionalism in the fields	Data sharing is a part of their professional mission to develop science	13
	Colleagues' expectations	Feel social pressures by colleagues (being expected to share their data)	7
Mimetic Pressures	Colleagues' performance	Observed other colleagues who use shared data improve their research performance	3
Perceived Benefits	Demonstration of quality work	Shared data indicates the quality of your work; improve the overall research quality	6
	Credits and reputation	Expect credits (e.g. authorship, citations, acknowledgements), reputation, and recognition	15
	Research performance	Conduct a comparative study or large-scale study (novel scientific finding); save time and effort in replicating and collecting data	14
Perceived Costs	Data annotation	Need to annotate data with their own metadata schemes (no standardized metadata scheme)	10
	Data organization	Takes time to organize data for more understandable, compatible, interoperable formats	11
	Data set location and interpretation	Takes time to find appropriate data sets and understand the data exactly	4
	Technical problems	Being involved with compatibility and interoperability issues with data	9
Perceived Risks	Losing publication opportunities	Have less opportunities for future publications; make more exclusive publications if data is not shared	15
	Getting Scooped	Worried about data theft; cannot trust others	8
	Misinterpretation and scrutiny	Worried about having different results by not being analyzed properly or being criticized by others because data is not reliable or low quality	13
Internal IT Capability	IM/IT expertise	Have technology expertise to manage data	5
	IM/IT support	Have internal IT/IM supports from their organizations	11
External IT Capability	Data repository	Have data repositories or enough space to share data	9
	Data standard	Have data sharing standards (metadata schemes) and systematic procedures	13
Altruism	Altruistic motivation	Allow other researchers to find something interesting that the first people missed; contribute to scientific developments; help others to save time and effort	12

4.1 Research Environment and Data Generated

Most of our interview participants worked in team-based research environments or a mixture of team-based and individual work; only two scholars, a mathematician and theoretical physician, mainly worked solo. The research teams usually included a lead professor, one or two post-doctoral research associates, and a few doctoral and masters' students.

The researchers reported that they generated a large amount of domain-specific original data including experimental data (e.g. genome sequencing data, compound data), field data (e.g. soil measurement, animal behavior, tree counts), and computational data (e.g. software code, computer simulation data). Most of the interviewees felt that they have limited *individual authority* to share their data by acknowledging that sometimes they need to seek permission from others for any collaboratively collected data. Only two interviewees (one post-doc and one doctoral candidate) felt they had *no authority* over sharing the data they collected.

Researchers reported different perceptions of the *importance* of data sharing in their fields. The researchers in biology, chemistry, and ecology agreed that data sharing is critical for novel scientific findings, but the researchers in computer science, electrical engineering, mechanical engineering, mathematics, and radiation oncology disagreed with this belief. Researchers in environmental engineering and physics reported a mixture of both perspectives.

4.2 Data Sharing Methods

Researchers in different disciplines reported different data sharing methods. Most researchers reported *internal* data sharing within their research teams or among collaborators; they usually used email, FTP servers, and website as the major internal data sharing methods. We assumed from the start that this type of internal sharing was occurring, and did not further investigate beliefs or motivations in this area.

Researchers also reported diverse forms of *external* data sharing with the researchers outside their research team or collaborators. First, researchers asserted that they share their data upon request; they use email or website upload as method of fulfilling such requests. Researchers also reported contacting other researchers individually to gain access to their data sets from published articles. Across different disciplines, this data sharing method was common, and it was the only data sharing method in the disciplines that do not have any informal or formal data repositories.

Second, some researchers who do not have any formal data repositories in their disciplines used a personal website to share their data with other researchers. A group of scholars in a similar research subject develop an informal or ad hoc data repository and share data with other researchers in the research subject area.

Third, some disciplines, including biology, chemistry, and ecology, use a range of external repositories (e.g. Dryad), and domain-specific data repositories (e.g. GenBank, Protein Data Bank, Computational Chemistry Database, Crystallography Open Database, Long Term Ecological Research Data Repository). These researchers reported well-developed data sharing protocols including data repository and data standard. In these same disciplines, most of the journals require researchers to publish their data in data repositories.

Finally, researchers in certain disciplines such as chemistry – where there are small, but highly structured data sets – share their data as an electronic supplement through the journals' websites. For example, some scholars in chemistry share their compound data through their journals' online supplements.

Some researchers reported an explicit expectation of various types of professional credits for data sharing including co-authorship, citation, and acknowledgement when their data are used by other researchers. There was insufficient information to judge the differences for these expectations among different disciplines, but we noted that the researchers whose disciplines have well established data sharing practices expected less credit than the researchers who do not have any formal way of data sharing. Additionally, we noted that junior researchers had higher expectations for credit (i.e., by means of co-authorship) than senior researchers; they mentioned strengthening the tenure case as the primary motivation for this. Senior researchers seemed to have less desire for credit, as well as more altruistic motivation for other researchers.

Roughly one third of our interviewees reported that the researchers in their field generally share their data after publication. The researchers in the disciplines that do not have formal data sharing mechanisms almost always share their data only after publication. For example, researchers in the engineering fields reported sharing their data only after publication. Another third of our interviewees reported that the researchers in their disciplines shared their data right after their data collection or after a fixed embargo period, regardless of publication status. For example, the researchers in molecular biology and genetics shared their data to a data repository right after data collection. These particular researchers reported a strong sense of trust that their colleagues would not "scoop" them using the shared data.

Lastly, where data sharing was a journal requirement, researchers in chemistry and biology and some researchers in ecology shared their data along with their publications. As noted above, these were cases where journals support a simultaneous publication of relatively small, structured data sets as supplements.

In terms of types of data shared, the researchers in some disciplines (e.g., biology, ecology, environmental engineering) shared raw data, but the researchers in other disciplines (e.g., chemistry, physics) share more refined or processed data. Additionally, the researchers in computer science, computational chemistry, and physics were likely to share both software and simulation results.

4.3 Factors Influencing Data Sharing

The primary focus of this research was on the factors influencing researchers' current data sharing practice. Based on the coding we did, we confirmed specific factors both motivating and preventing researchers' data sharing. In the material below, we explain these factors in three separate groups including institutional influences, individual influences, and IT capabilities.

4.3.1 Institutional Factors

Pressures by funding agencies, journal publishers, and private funding organizations influenced researchers' data sharing practice. First, the single most significant motivation for scientists' data sharing (giving) is a push by funding agencies to make data from funded projects available. Scientific funding agencies in the U.S. including NSF and National Institutes of Health (NIH) require their awardees to share the research data from projects they fund. Second, journals' requirement of data sharing is another factor. The journals in biology, chemistry, and some in ecology require their researchers to publish their data in any types of data repositories. Third, private and certain government funding agencies restrict researchers' data sharing. For example, some pharmaceutical companies and military agencies typically do not allow their awardees to share their data.

Disciplinary influences also affected researchers' data sharing. In many disciplines, data sharing is considered part of the professional responsibility; researchers believe that data sharing is one of their missions, and that it will help the development of their research disciplines. In these same disciplines, researchers reported that they are *expected* to share their data; they feel pressure from their colleagues to do so. Researchers reported observing what other researchers do, and they indicated that they tried to follow colleagues' practices that they saw as useful. A few researchers reported a belief that the research performance of other researchers who use the shared data would improve.

4.3.2 Individual Motivation Factors

Researchers also gave evidence that they carefully examined pros and cons of data sharing before they committed to sharing data. First of all, some researchers reported a belief that data sharing could highlight the quality of their work in research. For some, data sharing provided professional "credit" including co-authorship, citation, and acknowledgement, and reputation. In terms of using the shared data, researchers also believed that data sharing would improve their research (e.g. time saving in collecting the same data, replicating data for another research, conducting diverse comparison studies and large scale research).

In contrast, researchers also believed that data sharing imposes costs for them. In some scientific disciplines (e.g. ecology and environmental engineering) researchers saw the importance of data sharing, but they saw data sharing as very costly in time and effort. Due to a lack of established metadata standards and data preparation procedures, they saw the processes of organizing and annotating their data as very expensive. These same researchers also reported technical problems in the data sharing such as data compatibility and interoperability issues. This was a similar finding across each discipline that did not have well-established data sharing standards (metadata), procedures, and repositories. Researchers in those disciplines also reported that it took substantial time to locate and understand other researchers' data since the data do not have any established data repositories and standardized metadata.

Certain perceived risks by researchers also prevented them from sharing their data with other researchers. Many researchers worried about losing publication opportunities by sharing their data. It took a lot of time and effort to collect data, and they desired having as many publications as possible from their data. These researchers also worried about getting scooped on innovative findings when they shared their data with other researchers. Two scholars in environmental engineering mentioned that "data sharing is a little bit of a threat to our science because it is less incentive (sic) to collect your own data when all data is freely shared." Additionally, several researchers considered that misinterpretation and heightened scrutiny of their data would be possible risks if they shared their data.

4.3.3 Perceived Controllability: IT Capability Factors

IT capabilities were found to be important factors influencing researchers' data sharing practice. We focused our questioning on two distinct areas: an individual's self perceived capability to work with the relevant IT tools, including local support (internal capability), and the availability of appropriate community tools and infrastructure (external capability). Internal capability included researchers' own expertise in information and technology management in sharing their data, and also included any information management and/or IT support from within their own research team or host organization. Researchers with strong expertise and internal support in these areas also reported more extensive data sharing and reuse.

External IT capability referred to supports for researchers to share their data provided by the research community at large. In this area, researchers reported data repositories, data standards (i.e., metadata standards), and established data sharing procedures as key features. Biologists and chemists reported that they could easily share their data because they have well-developed data repositories, standards, and procedures to share their data with other researchers. Researchers in engineering fields generally did not report any central or domain data repositories. These engineers also reported needing to spend a lot of time to annotate, organize, upload, and manage their data on subject-specific or ad hoc data repositories. Researchers in ecology reported that they are aware of the importance of data repositories and standards and they have developed domain specific repositories and subject specific repositories. Since their data were unstructured, however, they reported that they still needed to develop better metadata standards and data sharing procedures.

4.3.4 Altruism

Unexpectedly, altruism emerged in about half of the interviews as a factor influencing researchers' data sharing. Some researchers reported a strong desire to help their colleagues to save time in collecting data and to avoid replicating experiments unnecessarily. Additionally, these researchers believed that their colleagues could exploit the data in ways that would extend the original findings and thereby benefit the scientific area where they collectively worked. These researchers reported a sense of personal satisfaction coming from sharing their data. A couple of our interviewees mentioned the importance of data sharing across disciplines not only within a discipline. A biologist mentioned that "it is also critical to improve [data] sharing across disciplines because a lot of research nowadays is becoming more multi-disciplinary so for example you have engineers working with biologists or physicists working with engineers and especially in my field in tissue engineering its very multidisciplinary field… If scholars in different disciplines could share that information, then the field of tissue engineering would progress a lot faster."

4.4 Changes in Data Sharing

Our interviewees reported that during recent years they had observed changes in their data sharing practices. Many of our interviewees reported that researchers' awareness, funding agencies' push, journals' requirements, technological improvements, and increased availability of data repository as changes they had experienced within recent memory. Just a few mentioned the emergence of data sharing standards as another recent change.

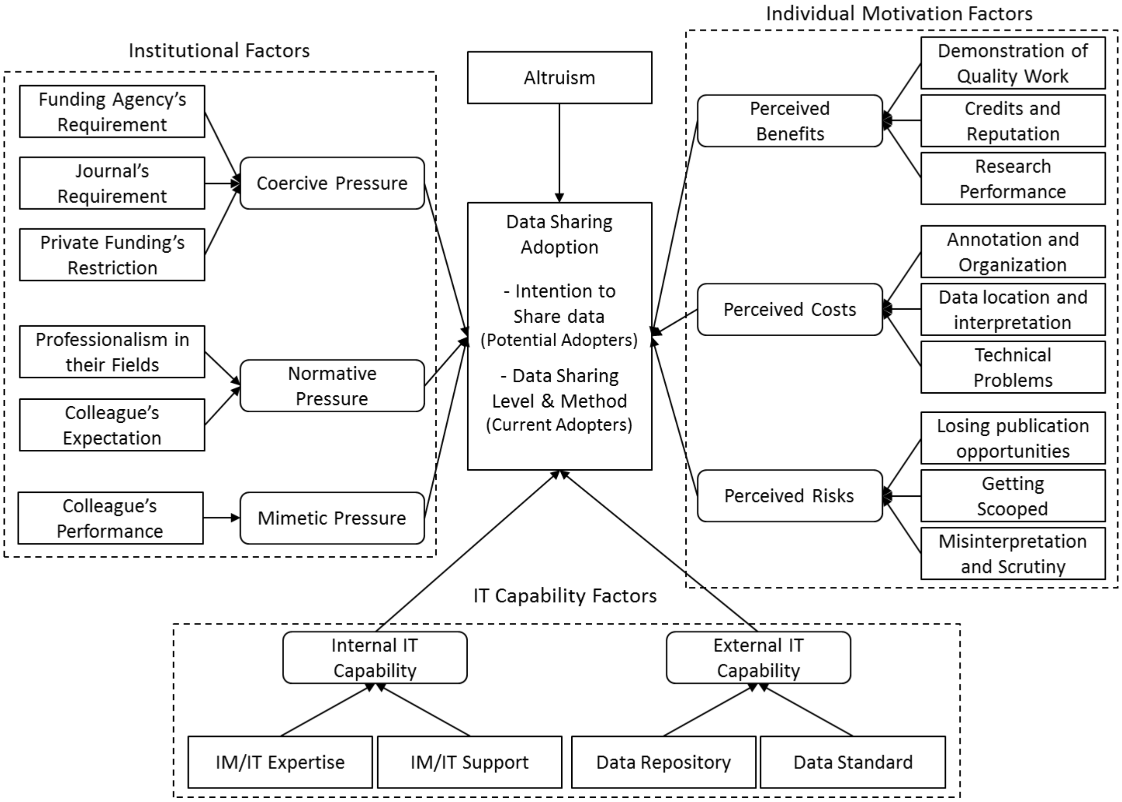
4.5 Supports Needed for Data Sharing

We asked our interviewees what kinds of additional supports they needed to facilitate data sharing. Ten of our 25 interviewees mentioned they do not need any supports since they are satisfied with their current data sharing practices. One biologist and one chemist said that they can easily share their data because they have well-established metadata standards, data sharing procedures, and data repositories. However, the remainder of our interviewees mentioned that metadata standards and data repositories are the main concerns of their current data sharing practice. Additionally, two researchers mentioned that they desired a data portal site where they could search available data sets. Several interviewees indicated that they needed better technology support. In particular, they reported that they needed professionals who could manage data sets, databases, storage, and other IT infrastructure.

5. DISCUSSION

Neo-institutional theory provided a productive lens for reviewing our interview data. Recall that some newer forms of institutional theory incorporate a cross-level perspective by linking institutional forces together with the motivations and behaviors of individual actors. We began this paper by framing the situation of the researcher as an individual actor embedded within his or her discipline as well as within the host institution and a variety of external institutions (e.g., funding agencies). Coercive, normative, and mimetic forces acting on institutions may trickle down to influence the decisions and behaviors of individuals who work within those institutions. Figure 1 provides an overview of our findings.

Figure 1. Factors Influencing STEM Researchers' Data Sharing Practices.



To have well-established data sharing practices, researchers need to have supportive institutional environments (e.g. data sharing structures, norms, policies), sufficient IT capability (e.g. data standards and repositories), and positive attitudes toward data sharing (e.g., perceived benefits, costs, risks). The combination of these can lead to more proactive data sharing practices among researchers.

One surprising finding arose from the spontaneous reports of altruistic motivations for sharing data. Typical formulations of institutional theory do not explicitly account for altruism among individual actors or groups embedded within institutions; when altruism appears, researchers use other theories to account for it [33]. Yet one essential and, arguably, widely shared value in contemporary science lies in the sharing of scientific resources for the common good [24]. Unlike commercial organizations, which generally use competition in an attempt to succeed in the marketplace, or public agencies, that ostensibly serve the common good as their central mission, STEM researchers (at least the ones working in universities) work within a middle ground that is marked by both competition and service to the common good – what some researchers [32] have termed "coopetition" and others call "competitive altruism" [15].

While institutional theory often focuses on risk reduction (institutional isomorphism is typically a strategy for avoiding risks by conducting activities in the "generally accepted" manner), perspectives on both coopetition and competitive altruism tend to focus on performance, both at the individual and the collective levels [23]. From either an evolutionary or a game theoretic perspective, behaviors that help others and thereby increase the overall performance or fitness of a group can also have benefits to individual performance and fitness. Interestingly, in situations where an individual's reputation is important, competitive altruism appears to be a powerful strategy [11]. This idea seems to map quite neatly onto the typical contemporary situation of a STEM researcher who seeks to enhance his or her reputation through publications, presentations, and other acts of sharing with the community. Possibly, future analyses of data sharing behavior among STEM researchers should incorporate some of the theoretical elements emerging from the altruism literature.

6. LIMITATIONS

Our sample included only a subset of the range of STEM disciplines, only one or two researchers from each of these disciplines, and only researchers from eastern U.S. research universities. Each interviewee reported observations and own experiences from their own research careers, so it is likely that the results are idiosyncratic for certain disciplines – and particularly those where there is substantial variation in sub-disciplinary practices. In future research, we need to include a more representative range of scholars and a more deliberate effort to obtain participants from a representative set of sub-disciplinary areas. Although the interview method provides rich data, future research should also include mixed methods (e.g., surveys) in order to triangulate on the findings offered here. In addition, an objective snapshot of available repositories and data standards for presentation to informants could elicit more specific responses to why a researcher uses or does not use a particular data sharing resource. In addition, we focused in this study primarily on the motivations and challenges to *sharing data* rather than those associated with using deposited data. Although certain questions assessed both sides of the data sharing equation, we found that using other researchers' data is still new to many researchers.

7. CONCLUSIONS

Under the assumption that data sharing and reuse can help in the overall advancement of the scientific endeavor, we sought to understand STEM researchers' data sharing. The institutional perspective seems helpful in this regard. In the disciplines of biology and chemistry as well as within some areas of physics, researchers seem to have well-established data sharing methods covering the data lifecycle. These methods are supported by many of the institutions in which they are embedded, mainly through the availability of data sharing standards and repositories.

Contrasting biology or chemistry with the discipline of ecology, many ecologists realize that data sharing is critical for their research, but they have difficulties in data sharing because they have few well-established metadata standards and domain-specific data repositories. For those who do share data, this means spending more time and effort to annotate and organize their data with their own metadata and format. Relatedly, because they do not have well-established central or domain specific data repositories, they share their data through ad hoc mechanisms such as Web servers and email exchanges among their collaborative group members. One ecologist mentioned that "[they] should have the official protocol for [data they collected] … those should be peer reviewed and approved and archived just like our data documentation … [they need to] share the procedures, not the data only." Researchers also mentioned the importance of having access to information professionals who can support their data sharing in terms of information and technology management. The information professional can help not only share their data, but also use other researchers' data by locating and interpreting the data.

In addition, it seems important to have a central data search mechanism so that researchers can find appropriate data sets for their research. Some researchers mentioned that they have difficulties in locating and interpreting other researchers' data, and they mentioned the necessity of a central data search mechanism. Even in areas where researchers are very good at sharing their data with other researchers, many researchers still do not actively seek other researchers' data sets. Data sharing is a two-way process of providing their own data and using other researchers' data. In order to achieve the promise of data sharing, researchers need to not only provide their data, but also use other researchers' data more actively.

Finally, and perhaps most importantly, our data indicated the importance of aligning institutional pressures with individual motivations for professional achievement. The most frequently mentioned driver of data sharing behavior was the "push" by the funding agencies that support research to ensure that data from the projects they support are made available to other researchers. This force, together with pressure exerted from scholarly journals, can have a strong influence over time on the choices and activities of individual researchers. Ultimately, the advocacy of funders and journals will also need to reflect on universities' policies and mechanisms for promotion and tenure in order to have a more direct influence on the data sharing activities of researchers. When sharing (and reuse) of data leads directly to an improvement of professional reputation and resulting career rewards, researchers will have strong individual motivations to participate in data sharing and reuse.

Taken together, our results support the idea that when institutional forces, infrastructure, and individual motives converge, the behavior of individual researchers will change in response. Many of the researchers we interviewed reported having seen this convergence and these changes during the course of their own careers. Further research efforts are needed to examine the role that altruistic motivations may play in establishing a virtuous cycle of data sharing and reuse that can increase the collective benefits obtained from societal investment in science and engineering.

For future research, it would be valuable to conduct a study to understand how the factors depicted in Figure 1 may influence scientists' data sharing and reuse in different science communities. A multi-level model including individual and institutional variables may serve as a useful research strategy to understand the dynamics of different factors and their cross-level relationships. A broad-based survey study that incorporates a representative sample of scientists from several different disciplines may help us to compare STEM researchers' data sharing and reuse by validating and confirming the multi-level research model.

8. ACKNOWLEDGMENTS

This research was supported in part by an award from the National Science Foundation (OCI-0753372). The National Science Foundation does not necessarily endorse any of the findings or conclusions of this report.

9. REFERENCES

[1] Ajzen, I. 1991. The Theory of Planned Behavior. *Organizational Behavior and Human Decision Process.* 52, 2, 179-211.

[2] Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., et al. 2003. Revolutionizing science and engineering through cyberinfrastructure: Report of the National Science Foundation blue-ribbon advisory panel on cyberinfrastructure.

[3] Avery, P. 2007. Open science grid: Building and sustaining general cyberinfrastructure using a collaborative approach. *First Monday.* 12, 6.

[4] Bandura, A. 1986. *Social Foundations of Thought and Action: A Social Cognitive Theory.* Englewood Cliffs, NJ: Prentice-Hall.

[5] Barley, S. R., & Tolbert, P. S. 1997. Institutionalization and Structuration: Studying the Links between Action and Institution. *Organization Studies.* 18, 1, 93-117.

[6] Becla, J., & Lim, K. T. 2008. Report from the first workshop on extremely large databases. *Data Science Journal.* 7, 1-13.

[7] Daniels, K., Johnson, G., & de Chernatony, L. 2002. Task and Institutional Influences on Managers' Mental Models of Competition. *Organization Studies.* 23, 1, 31-62.

[8] DiMaggio, P. J., & Powell, W. W. 1983. The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review.* 48, 2, 147-160.

[9] DiMaggio, P. J., & Powell, W. W. 1991. Introduction. In W. W. Powell & P. J. DiMaggio (Eds.), *The New Institutionalism in Organizational Analysis (*pp. 1-38). Chicago: The University of Chicago Press.

[10] Duxbury, L., & Haines, G. 1991. Predicting alternative work arrangements from salient attributes: A study of decision makers in the public sector. *Journal of Business Research.* 23, 1, 83-97.

[11] Fehr, E., & Fischbacher, U. 2003. The nature of human altruism. *Nature.* 425, 6960, 785-791.

[12] Fishbein, M. 1979. A theory of reasoned action: Some applications and implications. Nebraska Symposium on Motivation. 27, 65-116.

[13] Fishbein, M., & Ajzen, I. 1975. *Belief, Attitude, Intention, and Behavior.* Reading, MA: Addison-Wesley.

[14] George, E., Chattopadhyay, P., Sitkin, S. B., & Barden, J. 2006. Cognitive understandings of institutional persistence and change: A framing perspective. *Academy of Management Review.* 31, 2, 347-365.

[15] Hardy, C. L., & Van Vugt, M. 2006. Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin.* 32, 10, 1402.

[16] Hey, T., & Trefethen, A. 2003. e-Science and its implications. *Philosophical Transactions of the Royal Society A.* 361, 1809, 1809-1825.

[17] Hsu, M.-H., & Chiu, C.-M. 2004. Predicting electronic service continuance with a decomposed theory of planned behaviour. *Behaviour & Information Technology.* 23, 5, 359-373.

[18] Kevles, D. J. 1995. *The physicists: The history of a scientific community in modern America*: Harvard Univ Pr.

[19] Meehl, G., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J., et al. 2007. *The WCRP CMIP3 multimodel dataset. Bull. Am. Meteorol. Soc*, 88, 1383–1394.

[20] Meyer, J. W., & Rowan, B. 1977. Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology.* 83, 2, 340-363.

[21] National Science Board. 2005. NSB-05-40, Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. http://www.nsf.gov/pubs/2005/nsb0540/

[22] Nesvizhskii, A., & Aebersold, R. 2004. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug discovery today.* 9, 4, 173-181.

[23] Padula, G., & Dagnino, G. B. 2007. Untangling the rise of coopetition: The intrusion of competition in a cooperative game structure. *International Studies of Management and Organization.* 37, 2, 32-52.

[24] Resnik, D. B. 1998. *The ethics of science: an introduction*: Psychology Press.

[25] Savage, C. J., & Vickers, A. J. 2009. Empirical study of data sharing by authors publishing in PLoS journals. *PloS one*. 4, 9, e7078.

[26] Schein, E. H. 1996. Culture: The missing concept in organization studies. *Administrative Science Quarterly.* 41, 2, 229-240.

[27] Scott, R. W. 2001. *Institutions and Organizations*, 2nd Edition. Thousand Oaks, CA: Sage Publications.

[28] Shi, W., Shambare, N., & Wang, J. 2008. The adoption of internet banking: An institutional theory perspective. *Journal of Financial Services Marketing.* 12, 4, 272-286.

[29] Sonnenwald, D. H. 2007. Scientific collaboration: a synthesis of challenges and strategies. *Annual review of information Science and Technology.* 41.

[30] Teo, H. H., Wei, K. K., & Benbasat, I. 2003. Predicting intention to adopt interorganizational linkages: An institutional perspective. *Mis Quarterly.* 19-49.

[31] Tolbert, P. S., & Zucker, L. G. 1983. Institutional Sources of Change in the Formal Structure of Organizations: The Diffusion of Civil Service Reform, 1880-1935. *Administrative Science Quarterly.* 28, 1, 22-39.

[32] Tsai, W. 2002. Social structure of "coopetition" within a multiunit organization: Coordination, competition, and

intraorganizational knowledge sharing. *Organization science.* 13, 2, 179-190.

[33] Vandenabeele, W. 2007. Toward a public administration theory of public service motivation. *Public Management Review.* 9, 4, 545-556.

[34] Wicherts, J., & Bakker, M. 2009. Sharing: guidelines go one step forwards, two steps back. *Nature.* 461, 7267, 1053-1053.

[35] Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. 2006. The poor availability of psychological research data for reanalysis. *American Psychologist.* 61, 7, 726.

Sustain City – A Cyberinfrastructure-Enabled Game System for Science and Engineering Design

Ying Tang
College of Engineering
Rowan University
201 Mullica Hill Rd.
Glassboro, NJ 08028
tang@rowan.edu

Sachin Shetty
College of Engineering
Tennessee State University
3500 John A Merritt Blvd
Nashville, TN 37209
sshetty@Tnstate.edu

Talbot Bielefeldt
International Society for Technology in Education
180 West 8th Ave.
Eugene, OR 97401
talbot@iste.org

Kauser Jahan
College of Engineering
Rowan University
201 Mullica Hill Rd.
Glassboro, NJ 08028
Jahan@rowan.edu

John Henry
Education Information Resource Center
107 Gilbreth Parkway
Mullica Hill, NJ 08062
jhenry@eirc.org

S. Keith Hargrove
College of Engineering
Tennessee State University
3500 John A Merritt Blvd
Nashville, TN 37209
skhargrove@Tnstate.edu

ABSTRACT

The emergence of transformative technological advances in science and engineering practice has necessitated the integration of these advances in engineering classrooms. In this paper, we present the design and implementation of a virtual reality game system that infuses cyberinfrastructure (CI) learning experiences into the Project-Lead-The-Way (PLTW) pre-engineering classrooms to promote metacognition for science and engineering design in context. The CI features, metacognitive strategies, context-oriented approaches as well as their seamless integration in the game system are elaborated in detail through two game modules, *Power Ville* and *Stability*. Both games involve students in the process of decision-making that contributes to different aspects of city infrastructures (energy and transportation). The evaluation of *Power Ville* deployment in a PLTW classroom is also presented. The preliminary assessment confirms the usability of CI and metacognitive tools in science and engineering design.

Categories and Subject Descriptors

I.3.7 [**Three Dimensional Graphics and Realism**]

General Terms

Design, Human Factors, Experimentation, and Verification

Keywords

Virtual Reality Game, Metacognition, Science and Engineering Design

1. INTRODUCTION

Political, social and economic advances in the United States during the 21st century will be possible only if the intellectual potential of American's youth is developed now. However, a number of recent reports make it clear that the United States is losing ground on key indicators of innovation and progress because of its poor performance in teaching math and science [3, 19]. Pre-college education, in particular, is lagging well behind its mandate to educate all children to higher standards, especially in areas that prepare students for science, technology, engineering, and mathematics (STEM). This eliminates many of the best and brightest schoolchildren from the ranks of future scientists and engineers. Many students who do undertake science and engineering studies in college are unprepared and drop out in frustration, while other potentially capable students never consider these subjects in the first place. Therefore, developing educational practices and settings in our K-12 classroom becomes extremely important; especially the ones that promote 21st century skills and help learners build up their "habit of mind" [4] for scientific reasoning and inquiry.

The radical and transformative technological revolution has resulted in fundamentally new ways of science and engineering practice. This paradigm shift has a significant impact on the skills needed for a diverse science and engineering workforce that is capable of designing and deploying cyber-based systems, tools and services. However, engineering and science education has not kept pace with this evolution, especially at the K-12 level. There is a growing need to incorporate cyberinfrastructure (CI) learning experiences into classrooms of secondary education. Two key CI-based technologies which have tremendous impact on education and training are 1) networked computing technologies; and 2) virtual learning environments, including games, simulations and modeling. Networked computing technologies enable new forms of collaborative learning to meet different learner's requirements. Virtual games allow interactions within immersive digital worlds that promote learning through authentic and engaging play. Simulations and models help provide insights into scientific phenomenon making difficult abstract concepts and large data sets accessible in ways that are more visual, interactive, and concrete. As such, infusing virtual reality (VR) games with simulations and models into a classroom setting becomes essential.

Living through the hurricane Katrina and its aftermath and reflecting on these experiences from the technical and humanist standpoints has led us realize the importance of bringing the perspectives of humanities and social sciences into design education [9]. Although some institutions have taken steps to meet this need, most of those efforts have focused on either a capstone design sequence [11], undergraduate research experience [5], or developing a full-fledged degree program [16]. Research indicates that such design education should start earlier for prospective and beginning science and engineering students, encouraging them to optimize their design activities not only in technical aspects, but also in social and environmental areas [18].

Motivated by these general remarks, this project, as collaboration between Rowan University and Tennessee State University (TSU), developed and implements a VR game system, called "*Sustain City"*, which is closely tied to the Project-Lead-The-Way (PLTW) curriculum [22] and provides CI learning experiences to pre-engineering students. In particular, our design carefully balances engagement and learning with the following unique aspects: (a) visual modeling and simulation tools in the games provide insights into scientific concepts and phenomena, and help analyze data in a more visual and interactive way; (b) the networked educational environment transcends the boundaries of school-based education to leverage learning taking place anytime and anywhere, and promotes learning through collaboration; (c) metacognitive strategies and problem-based learning advance learners' strategic thinking and enhance their social, methodological and professional competence for a broader perspective on design; and (d) each game module is self-contained with a focus on particular fundamental science and engineering concepts, so it can be used as stand-alone contributions to a typical science, technology-based, or pre-engineering course or used in a coordinated manner through the PLTW curriculum. In these games, students are central and important participants of a virtual world in which they can become environmental scientists, bridge construction engineers, and traffic engineers. They learn how to investigate, design and pose solutions that have an impact on the world. More importantly, students often find a passion for curricular content while navigating through the games and begin to see themselves as problem solvers. Such engagement allows students to better appreciate the content's value.

The focus of this paper is to illustrate the design and developmental aspects of our approach through an example of two VR games, *Power Ville* and *Stability*. Our assessment of the games when deployed in a high school classroom provides insights into effectiveness of the games in providing CI learning experiences. The rest of the paper is organized as follows. Section 2 gives the project overview with an emphasis on the integrated approach and three metacognitive interventions used in the design of the VR game. Section 3 provides the design and implementation of *Power Ville* and *Stability*. Section 4 outlines the evaluation results and findings from the assessment of the *Power Ville* game, followed by the conclusions in Section 5.

2. PROJECT OVERVIEW

2.1 Integrated Approach

Recent reports have indicated that many STEM classes rely heavily on textbooks but are weak on examples, such that students are exposed to encyclopedias of fact without ever engaging in the science and engineering process [6]. The perceived dullness or complexity of the material, a lack of concrete applications, and individual preconceptions further make introductory science and engineering classes difficult for students, leading to lower recruiting and retention rates of science/engineering majors. Thus, it is crucial to design a fun learning environment that engages students in the exploration of real science and engineering applications and promotes strategic, constructive, and big-picture thinking and problem solving. As for the "big-picture" strategy, Bordogna [1] has well expressed the need for curricular integration as, "*Most curricula require students to learn in unconnected pieces - separate courses whose relationship to each other and to the engineering process are not explained until late in a baccalaureate education, if ever. ... The content of the courses may be valuable, but this view of engineering education appears to ignore the need for connections and for integration - which should be at the core of an engineering education.*"

In our VR game system design, integration refers to a series of VR games in a given context (like "*Sustain City"*) that individually have a focus on particular fundamental science and engineering concepts, can be used as a replacement of traditional laboratory settings of courses at different levels of the PLTW curriculum, and eventually proceed to an increasingly complex open-ended capstone project at the senior level of the curriculum. Such integration enables students to understand that their courses are part of a flow that contributes to the design of a system rather than being separate bodies of knowledge. This idea is clearly presented in Fig. 1 with the explanation below.

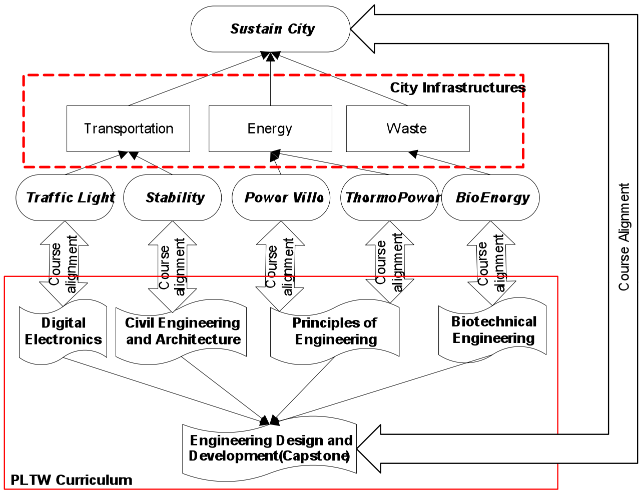


Fig. 1: The curriculum alignment and integration

A sustainable city is a city designed to improve the quality of life, including ecological, cultural, political, institutional, social, and economic components without leaving a burden on the future generations [23]. It is also an exquisite combination of interacting systems (infrastructures) that can be designed and analyzed using multidisciplinary engineering and scientific principles. With the future sustainable city as a broader context and the city infrastructures as the themes, *Sustain City* consists of a series of VR games (e.g., *Power Ville, Traffic Light,* and *Stability*, etc.) that provide students an opportunity to learn what it means to be a scientist, engineer, or mathematician who helps design and maintain an eco-city. This opportunity also brings their content knowledge and skills learned in the traditional classroom environment to a contextual reality. As shown in Figs. 1 and 2 (the screenshots from *Sustain City* for different game theme environments), our games align with the curriculum and academic subject matter - including math, circuit design, or persuasive writing – and each game taps into subject knowledge. The eventual integration of individual game components will be performed by teams of students in the senior-level capstone course, resulting in a fully-functional virtual eco-city. Through demonstration, explanation, and practice in different aspects of

Sustain City, students are motivated to see the interconnection between their courses as a progression of increasing design complexity.

2.2 Metacognitive Interventions

The process of learning is a very complex cognitive task that requires a lot of effort and motivation from learners. Research indicates that the more students are aware of their learning process, the more they can control such matters as goals, dispositions, and attention, and the better they become successful learners. Such awareness and monitoring processes are often referred to as metacognition –" the processes in which the individual carefully considers thoughts in problem solving situations through the strategies of self-planning, self-monitoring, self-regulating, self-questioning, self-reflecting, and or self-reviewing" [7]. In light of this, it benefits education to not only create an interesting and stimulating learning environment for students (i.e., VR games), but also to naturally integrate important learning tools (i.e., metacognitive interventions) into the interactive game activities.

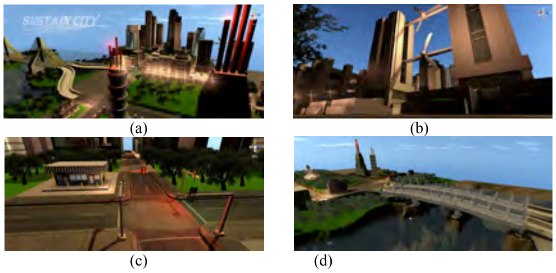
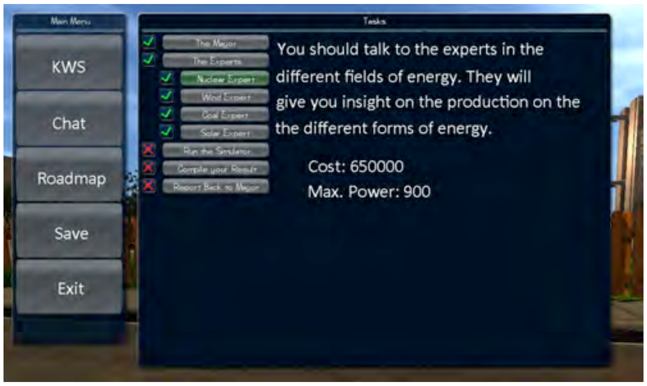


Fig. 2: Screenshots from *Sustain City*: (a) Overview of Sustain City; (b) Power Ville; (c) Traffic Lights; (d) Stability

- Learning Road Map – Learning roadmap provides study guides that endow students with the capability to find relevant information and to capture key concepts in the study materials [13]. Depending on game content, road map in our game might be a task list that guides students to navigate through game assignments and retrieve important information (Fig. 3); or it might be a set of suggestions designed to lead students through a problem solving process by directing attention to key ideas and suggesting the application of proper skills (Fig. 4).



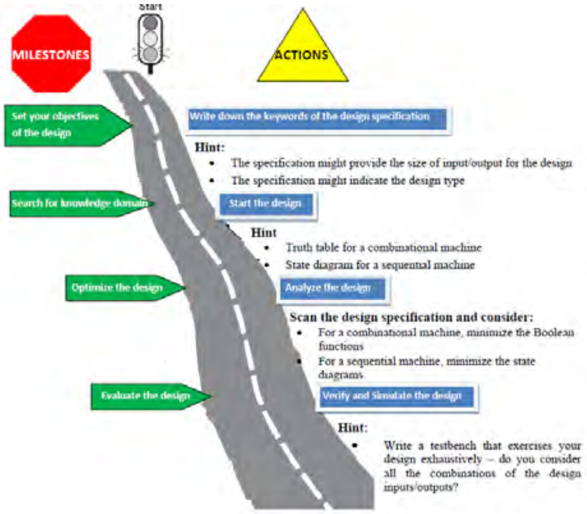
You should talk to the experts in the different fields of energy. They will give you insight on the production on the the different forms of energy.

Cost: 650000

Max. Power: 900

Fig.3: The roadmap in *Power Ville* game

- What I **K**now-What I **W**ant to Know- What I have **S**olved (KWS) training – KWS is adapted from a well known reading strategy, What I **K**now-What I **W**ant to Know- What I Have **L**earned (KWL) [12]. It typically provides a three-column chart structure to activate students' prior knowledge by recalling what students know about a problem (K), to motivate students to read/think by asking what they want to know (W), and finally to review what part of the problem has been resolved and what is yet to be solved (S). This type of intervention is usually implemented in the traditional classroom environment with facilitation from instructors. However, in a virtual game environment, students are often left alone with the responsibility to explore and to figure out problems themselves. The lack of guidance makes it difficult to implement such an intervention because not all students are motivated to use it without facilitation. The solution we devised uses a series of progressive prompts at key game stages, as exemplified in Fig. 5 (a). The information provided by students at different prompts is automatically recorded into the corresponding portions of their KWS chart (Fig. 5 (b)). More importantly, the What-I-Know portion of the KWS chart will be used to compose the player's final report, which will deliver to the city Mayor and ultimately to his or her instructor for grading. Therefore, the player might need to modify his or her KWS as often as needed, since his or her knowledge progresses with the game (this part will be elaborated in detail in the next section).


Start
MILESTONES
ACTIONS
Set your objectives of the design
Search for knowledge domain
Optimize the design
Evaluate the design
Write down the keywords of the design specification
Start the design
Analyze the design
Verify and Simulate the design

Hint:

- The specification might provide the size of input/output for the design
- The specification might indicate the design type

Hint

- Truth table for a combinational machine
- State diagram for a sequential machine

Scan the design specification and consider:

- For a combinational machine, minimize the Boolean functions
- For a sequential machine, minimize the state diagrams

Hint:

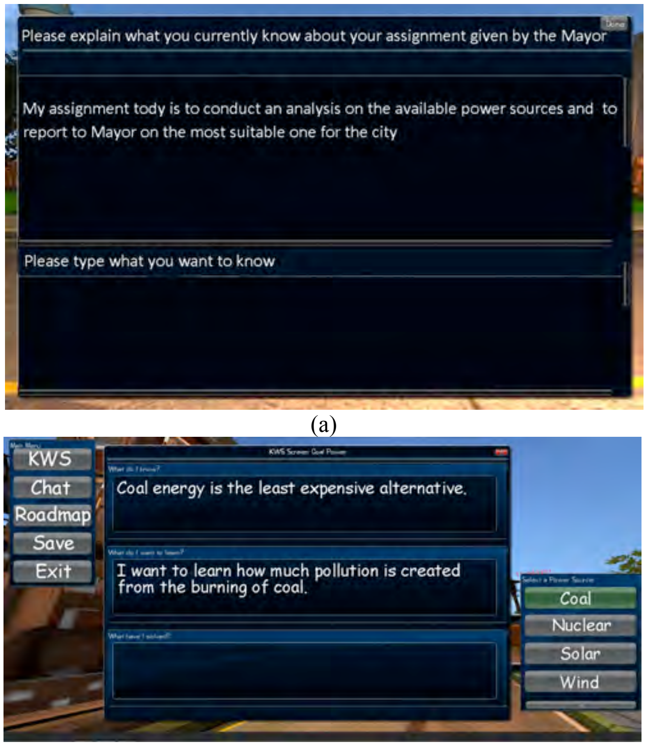
- Write a testbench that exercises your design exhaustively – do you consider all the combinations of the design inputs/outputs?

Fig. 4: A sample road map for a traffic light design

- Think-Aloud-Share-Solve (TA2S) training – As Vygotsky pointed out, learning is an inherently social and cultural rather than individual phenomenon [17]. The interactions among peers allow intellectual synergy of many minds to bear on a problem, and promote the social stimulation of mutual engagement in a common endeavor. TA2S, implemented in our game system through online chatting, is a variation of the collaborative learning strategies, Think-Aloud [8] and Pair Problem-solving [10].

2.3 Cyberinfrastructure Tools

According to the National Science Foundation, CI describes an environment in which computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people are all linked together by software and high performance networks to improve scholarly productivity and to enable breakthroughs not o possible. Although *SustainCity* has much in common with the *GreenCity* [20] and the *Mobility* [21], our design explores way beyond the scope of these commercial games with seamless integration of key elements of cyber infrastructure, namely collaboration and communication in a visualization environment. For example, in *Power Ville*, students are hired by the city Mayor to analyze viable energy sources for the future of the city. An interactive simulator as shown in Fig. 6 is designed in the game to help students visually view environmental and economic impacts of four available energy options (coal, nuclear, solar and wind) for the city. In some other games, such as *Stability*, students are introduced to a 3D bridge model (Fig. 2 (d)), where they can visually analyze the structure impact of gravity and other loading effects on the bridge visually, and provide mitigation solutions to reinforce the bridge structure. To complete this task, students need to apply their structural systems fundamentals, including forces, loads, beams, and columns, to vary model parameters, and to derive their correlations to bridge performance. We view games like these as environments that make curricular content a necessary tool and that position the learner as a leader who transforms a virtual world


Please explain what you currently know about your assignment given by the Mayor

My assignment tody is to conduct an analysis on the available power sources and to report to Mayor on the most suitable one for the city

Please type what you want to know

(a)

KWS Screen Coal Power

What do I know?

Coal energy is the least expensive alternative.

What do I want to learn?

I want to learn how much pollution is created from the burning of coal.

What have I solved?

(b)

Fig. 5: (a) a prompt after a player visited the coal building; (b) the student's inputs to the prompt is recorded in his or her KWS

To enable the network functionalities that allow students to share ideas and knowledge with their on-line group members, our design uses a client/server architecture. Only the instructor has access to the server program, where he or she can set up the group size and group password (Fig. 7 (a)). At the client site, all games are a group assignment that requires each student in a group to log into the system with a security password. While logged in, the players can participate in group discussions (Fig. 7(b)) on the problem/solutions through a synchronous chat function. Group members are not necessarily present at the same place and time. The discussions as well as each player's actions are recorded in the system. The data is only accessible to instructors and researchers, providing a resource to analyze student performance and game effectiveness in promoting learning.

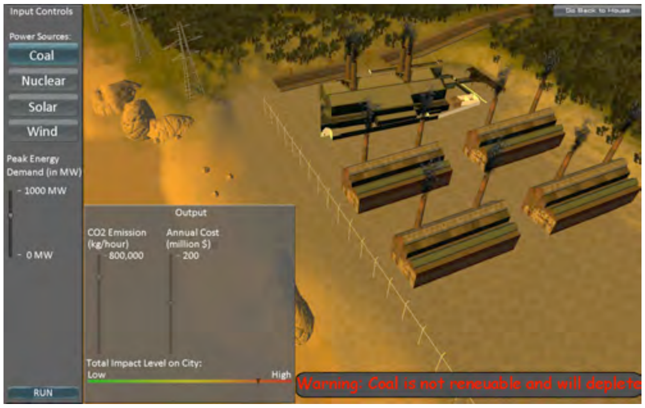

Warning: Coal is not renewable and will deplete

Fig. 6: The interactive simulator in *Power Ville*

3. Sample Game Theme and Design

In this section, we describe the educational gaming process in two modules, *Power Ville* and *Stability*. These modules exemplify the seamless integration of fun, metacognitive interventions and engineering problem-solving in a well-balanced engagement and learning process.

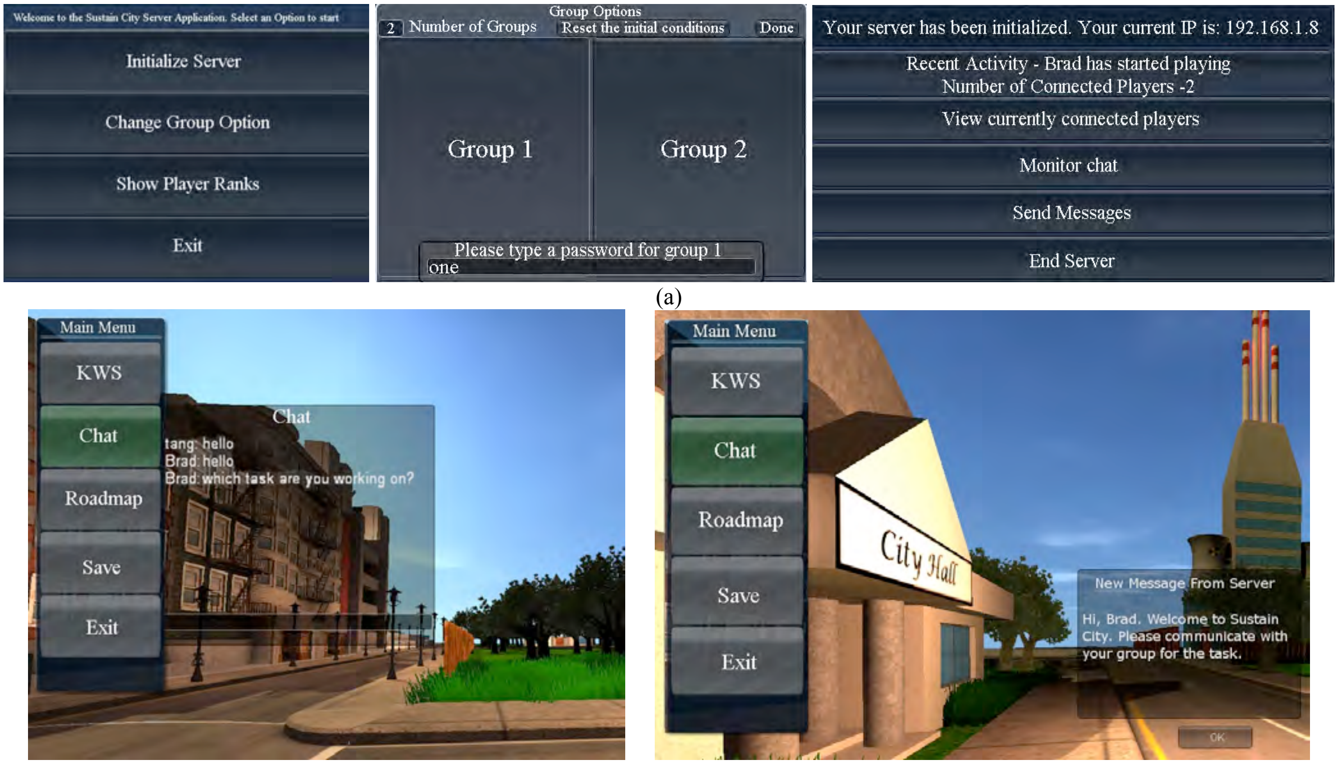
3.1 Power Ville

Cooking a dinner, heating a house, lighting a street, and running a factory - all of these need power. Energy is thus at the heart of everybody's quality of life. How to generate and use energy that satisfies increasing energy needs while combating climate changes at the same time becomes an unprecedented challenge for a sustainable city development. Bringing such real science and engineering design problem as well as involved societal and environmental issues into the *Principles of Engineering* course is the core of this game.

The goal of the Power Ville VR game is to educate students about four energy choices (coal, wind, solar, and nuclear) and the impact of those choices on the environment of the city. To implement this goal, the game incorporated the CI tools, which are gaming, simulation and networking, and the three metacognitive interventions.

Establishing a Meaningful Role

Goal identification is of importance in game playing to motivate players and promote a deep understanding of content [15]. In *Power Ville*, each player is introduced to the game by visiting the city hall and talking to the Mayor of the city. The conversation asks the player to take the task, as a consulting engineer, to conduct and report a thorough analysis on the most suitable form of energy for the future of the city with the given city budget and energy demands. Meanwhile, the player is urged to visit different facility buildings and talk to individual power system experts for vital information regarding the pros and cons of various energy sources. Succeeding in this role requires that the player understand and apply the knowledge about power and energy systems learned in both the classroom and the game environment, together with the writing skills to collect appropriate evidence and compose a persuasive piece of writing. In fact, the game is designed in the way that automatically composes a final report for the player by using every justification the player provides in the question prompts at different game stages.



Welcome to the Sustain City Server Application. Select an Option to start

Please type a password for group 1
one

Your server has been initialized. Your current IP is: 192.168.1.8

Recent Activity - Brad has started playing
Number of Connected Players -2

(a)

Chat
tang: hello
Brad: hello
Brad: which task are you working on?

New Message From Server
Hi, Brad. Welcome to Sustain City. Please communicate with your group for the task.

(b)

Fig. 7: The networking activities at both (a) the server side and (b) the client side

Exploring CI Tools for a Better Solution

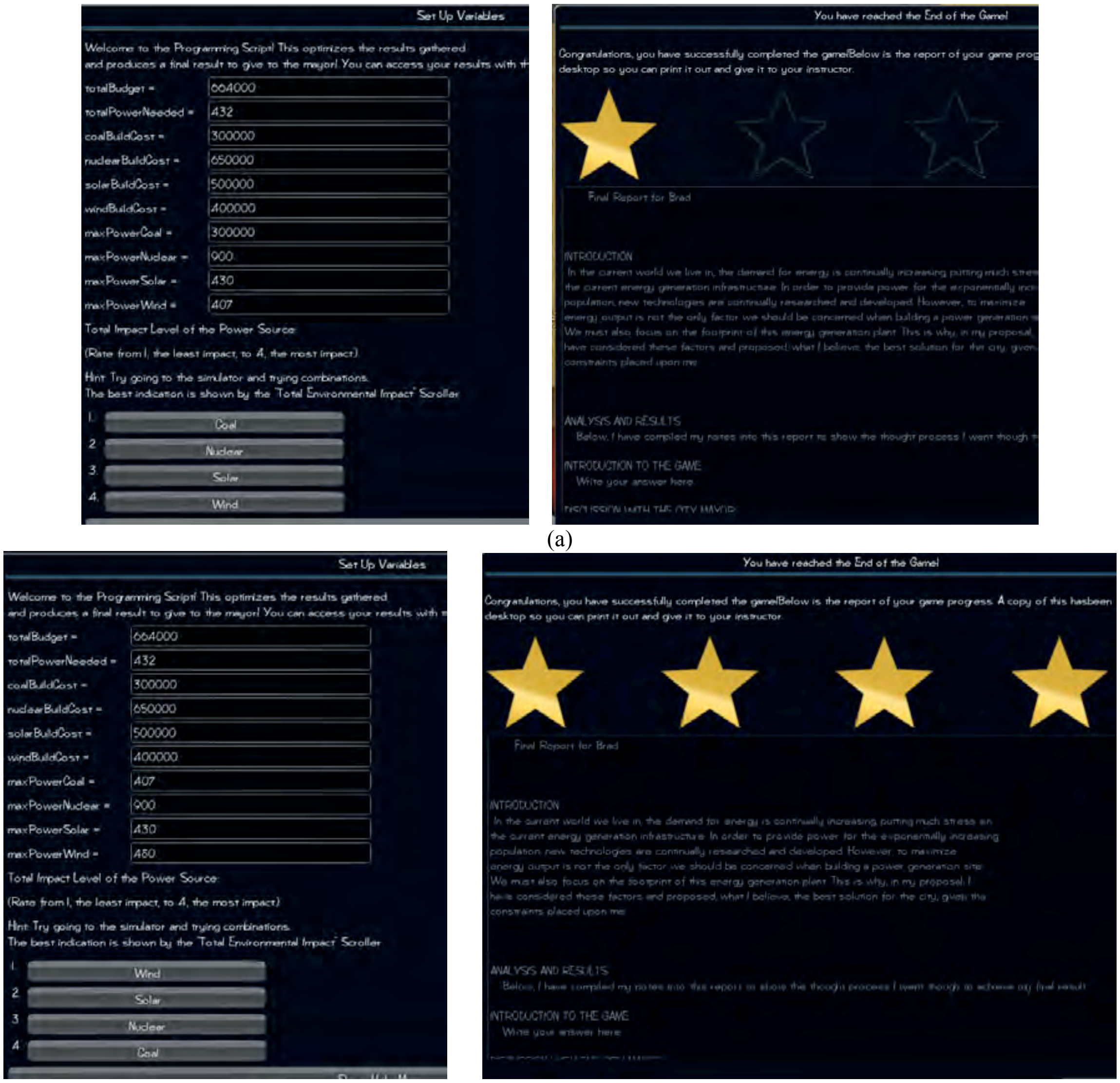
After players exit the city hall, they must talk to different power system experts located in offices spread across the city. The game environment is designed for easy navigation as each office building has identifiable and unique landmarks. For instance, the solar-power building contains racks of solar-panels and the wind power building has large turbines that tower over the streets below. In addition, the *Road Map* tool is always available on a game menu for players to retrieve the next task in the must-do list and to show the cost and the peak energy output of a particular power source. The game menu is launched through the TAB key. After players interact with individual experts, watch a video, and do a quiz, or play a mini-game on the energy production process they are exploring, they are prompted to write about stories from characters they talk to. These notes are recorded in KWS. Players can launch KWS via the game menu to modify their notes. As the logs for each player, particularly the What-I-Know section of the KWS, are automatically used to compose his or her final report, the player has to craft the notes using the most compelling support for their argument. Soon after visiting all experts and making their way to their apartment, players are provided with two other CI tools, *Simulator* and *Optimization Programming*, to help formulate their answers as to the best energy source for the future of the city. The *Simulator* allows players to select each of four different power sources with a given demand as well as other source-dependent constraints, such as sun exposure for solar power and wind speed for wind power. It then visually outputs the total amount of CO_2 emissions and the total environmental impact of the chosen power source. As shown in Fig. 6, the simulation run of the coal power source indicates the footprint of the power generation and warns the player the potential of coal being depleted as a non-renewable energy.

Discovering Consequences of different decision-makings

Computing has made possible profound leaps of innovations and imagination. This paradigm shift has a significant impact on the skills needed for a diverse science and engineering workforce who can bring the power of computing-supported problem-solving to an expanded field of endeavors. *Power Ville* provides students the ability to apply computational thinking through the use of the *Optimization Programming* tool. This tool takes inputs from a player (i.e., the data he or she collects through the game as for the cost and peak energy production of a particular power, and the city budget and energy demands), and his or her preliminary decision to rank four power sources based on their environmental impact. The choice each player makes at this point affects the final outcome of the optimization, ultimately determining his or her argument as for the best energy source. This decision is eventually brought to an in-game evaluator that uses a pre-programmed scoring method to assess the quality the decision. Fig. 8 shows two possible outcomes. When a player ranks coal as the power source with the least environmental impact, his or her preliminary understanding was criticized by *Optimization Programming* tool as a one-star decision (see Fig. 8 (a)). On the other hand, the correct understanding of the environmental impact of the four power sources was rated as a four-star decision (See Fig. 8 (b)).

3.2 Stability

There is a strong interrelationship between success and failure in engineering. When engineers properly anticipate the possible failure modes of a structure or system, they can obviate them by design [14]. *Stability* actually explores such nature of design – success through failure, providing students a virtual environment to reinforce the structure of a bridge to help prolong its lifetime. This game project fits nicely to the core of the "*Civil Engineering* *and Architecture*" course into the PLTW curriculum. Introducing a vision of engineers tackling real life problems to impact the quality of life generates tremendous enthusiasm and attracts more students, especially underrepresented groups, into engineering.



Set Up Variables

Welcome to the Programming Script! This optimizes the results gathered
and produces a final result to give to the mayor! You can access your results with th

totalBudget = 664000
totalPowerNeeded = 432
coalBuildCost = 300000
nuclearBuildCost = 650000
solarBuildCost = 500000
windBuildCost = 400000
maxPowerCoal = 300000
maxPowerNuclear = 900
maxPowerSolar = 430
maxPowerWind = 407

Total Impact Level of the Power Source:
(Rate from 1, the least impact, to 4, the most impact)
Hint: Try going to the simulator and trying combinations.
The best indication is shown by the 'Total Environmental Impact' Scroller

1. Coal
2. Nuclear
3. Solar
4. Wind

You have reached the End of the Game!

Congratulations, you have successfully completed the game!Below is the report of your game prog
desktop so you can print it out and give it to your instructor.

Final Report for Brad

INTRODUCTION
In the current world we live in, the demand for energy is continually increasing putting much stre
the current energy generation infrastructure. In order to provide power for the exponentially inc
population, new technologies are continually researched and developed. However, to maximize
energy output is not the only factor we should be concerned when building a power generation s
We must also focus on the footprint of this energy generation plant. This is why, in my proposal,
have considered these factors and proposed what I believe the best solution for the city, given
constraints placed upon me.

ANALYSIS AND RESULTS
Below, I have compiled my notes into this report to show the thought process I went though t

INTRODUCTION TO THE GAME
Write your answer here

DISCUSSION WITH THE CITY MAYOR

(a)

Set Up Variables

Welcome to the Programming Script! This optimizes the results gathered
and produces a final result to give to the mayor! You can access your results with t

totalBudget = 664000
totalPowerNeeded = 432
coalBuildCost = 300000
nuclearBuildCost = 650000
solarBuildCost = 500000
windBuildCost = 400000
maxPowerCoal = 407
maxPowerNuclear = 900
maxPowerSolar = 430
maxPowerWind = 450

Total Impact Level of the Power Source:
(Rate from 1, the least impact, to 4, the most impact)
Hint: Try going to the simulator and trying combinations.
The best indication is shown by the 'Total Environmental Impact' Scroller

1. Wind
2. Solar
3. Nuclear
4. Coal

You have reached the End of the Game!

Congratulations, you have successfully completed the game!Below is the report of your game progress. A copy of this hasbeen
desktop so you can print it out and give it to your instructor.

Final Report for Brad

INTRODUCTION
In the current world we live in, the demand for energy is continually increasing putting much stress on
the current energy generation infrastructure. In order to provide power for the exponentially increasing
population, new technologies are continually researched and developed. However, to maximize
energy output is not the only factor we should be concerned when building a power generation site.
We must also focus on the footprint of this energy generation plant. This is why, in my proposal, I
have considered these factors and proposed, what I believe, the best solution for the city, given the
constraints placed upon me.

ANALYSIS AND RESULTS
Below, I have compiled my notes into this report to show the thought process I went though to achieve my final result

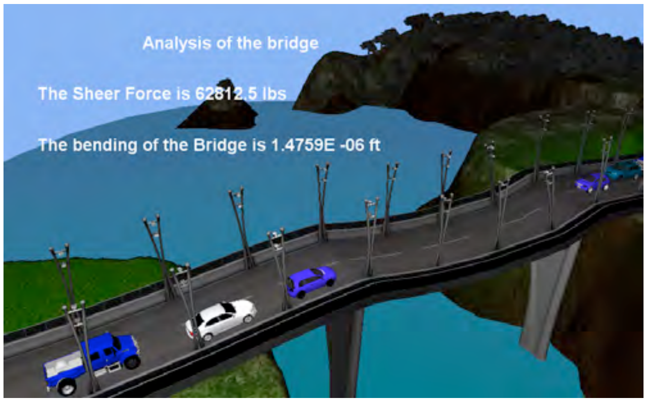
INTRODUCTION TO THE GAME
Write your answer here

(b)

Fig. 8: The optimization programming GUIs for (a) variable set-up and scoring. (a) and (b) present two different outcomes

In the introductory scene of *Stability*, where a busy bridge appears against the backdrop of the city, players are invited to conduct a what-if stability analysis of the bridge with the intention of estimating its life-time. To complete this task, players have to navigate back and forth between two game scenes multiple times to make different observations and to collect the best evidence in justifying their findings. First is the input scene, where players must provide the mechanical and structural specifications of the bridge (e.g., the length and width of the bridge), average amount of traffic per day (e.g., the number of cars per day), and climate situation (e.g., snow). The inputs each player provides at this point affect the kinds of bridges that the player interacts with, analyzes and draws a conclusion about.

Next comes with the analysis and mitigation scene. With the specifications being provided by the player, a simulator allows the player to analyze and visualize the bridge deflection (Fig. 9). Such experiential practice of visualizing bridge performance with changing loads, sizes and climate situations greatly helps players to draw correlations between bridge dimension, traffic, environment, and deflection. In addition, the analysis retrieves crucial facts, such as the amount of force and moments exerted on the bridge, and the area of beams required at the center and ends. The latter information is of significant importance in determining the number of beams and spacing required to mitigate bridge bending. Instead of recommending a solution, the game provides students a look-up table as shown in Table 1, where the cell values represent the required area of beams, and the column and row values correspond to the number and spacing of beams. There is no one-to-one correspondence between the value calculated from the analysis and the ones available in the table. Students must try different options that are close to the calculated value, observe individual reduction effects, and make an optimal decision for the best alleviation option. The game brings academic content, such as live loads, dead loads, beams and columns, into a contextual reality. Students not only hone their structural system design skills, but also find an appreciation for their content knowledge as well as the excitement of seeing themselves as problem solvers.



Analysis of the bridge

The Sheer Force is 62812.5 lbs

The bending of the Bridge is 1.4759E -06 ft

Fig. 9: The analysis scene of *Stability*

Table 1: Lookup table to identify number and spacing of beams

	Number								
SPACING	3	4	5	6	7	8	9	10	11
3	0.44	0.78	1.23	1.77	2.40	3.14	4.00	5.06	6.25
3.5	0.38	0.67	1.05	1.51	2.06	2.69	3.43	4.34	5.36
4	0.33	0.59	0.92	1.32	1.80	2.36	3.00	3.80	4.68
4.5	0.29	0.52	0.82	1.18	1.60	2.09	2.67	3.37	4.17
5	0.26	0.47	0.74	1.06	1.44	1.88	2.40	3.04	3.75
5.5	0.24	0.43	0.67	0.96	1.31	1.71	2.18	2.76	3.41
6	0.22	0.39	0.61	0.88	1.20	1.57	2.00	2.53	3.12
6.5	0.20	0.36	0.57	0.82	1.11	1.45	1.85	2.34	2.89
7	0.19	0.34	0.53	0.76	1.03	1.35	1.71	2.17	2.68
7.5	0.18	0.31	0.49	0.71	0.96	1.26	1.60	2.02	2.50
8	0.17	0.29	0.46	0.66	0.90	1.18	1.50	1.89	2.34
9	0.15	0.26	0.41	0.59	0.80	1.05	1.33	1.69	2.08
10	0.13	0.24	0.37	0.53	0.72	0.94	1.20	1.52	1.87
11	0.12	0.22	0.34	0.48	0.65	0.86	1.09	1.39	1.70
12	0.11	0.20	0.31	0.44	0.60	0.78	1.00	1.27	1.56
13	0.10	0.18	0.29	0.41	0.55	0.73	0.92	1.17	1.44
14	0.09	0.17	0.27	0.38	0.51	0.68	0.86	1.09	1.34
15	0.09	0.16	0.25	0.35	0.48	0.63	0.80	1.02	1.25
16	0.08	0.15	0.23	0.33	0.45	0.59	0.75	0.95	1.17
17	0.08	0.14	0.22	0.31	0.42	0.56	0.71	0.90	1.10
18	0.07	0.13	0.21	0.29	0.40	0.53	0.67	0.85	1.04

4. PRELIMINARY ASSESSMENT AND LESSON LEARNED

In fall 2011, *Power Ville* was piloted in *Principles of Engineering (POE)* course at Burlington County Institute of Technology – a vocational school in New Jersey. One focus of POE is types of energy (non-renewable and renewable) and energy distribution. After learning these concepts in lectures, the class of 15 students played *Power Ville* as part of their laboratory activities. An online survey, administrated by the International Society for Technology in Education (ISTE), was then given to the students upon their completion of the game. The survey instrument, as provided in Appendix, is particularly designed to assess the game's realism, the utility and usability of the metacognitive and CI tools, and students' impressions of what they had learned. It is important to note that this was formative data collection from students who are essentially beta testers. It is valid for suggesting game improvements, but not valid for making generalizations about any larger population.

One of the important findings from the survey is that some additional scaffolding may be necessary for students to get the most benefit from the game experience. For instance, the metacognitive tools were generally usable for students without much assistance, but the students reacted differently to Road Map, KWS, and chat. Road Map was the most popular. About half the students used KWS, but were not sure of its value. One student commented, "*I felt like the KWS was a bit unnecessary; however, if you were to create a notebook to take notes during the videos I believe that it would be used*." In fact, the note-taking function is present; but the student did not find it. Similarly, few students used chat (although, possibly because of personal computing experience, most predicted they would use it for future problem-solving). One student felt that KWS was "*redundant if you have chat*," as if the KWS structure of questions and the ability to ask questions were interchangeable. A larger issue is to help students understand that KWS is a simplified, specific instance of a general problem-solving framework that they will need to use throughout their careers in addressing novel challenges.

Minor interface adjustments might make KWS and chat more accessible and integrated, or the instructors might need to do some modeling of collaborative problem solving. For example, chat is currently designed to be launched through onscreen menu. A "You Got a Message" type of note will pop up on the top right corner of the game GUI whenever a group member initiates chat with the player. The current design raised a lot of suggestions from the students during their play as how to "tweak" the interface to improve its accessibility. For instance, a scrolling chat box at the bottom of the screen would be much better than going through hierarchical menu options.

The popularity of Road Map as a resource compared to KWS may be related to students having more experience with linear labs and projects, and thus being more familiar with this type of guidance. The students reported little experience with either tool prior to playing *Power Ville*. The larger question is how *Power Ville* and future games might be crafted to increase the emphasis on open-ended problem solving and the need to use appropriate tools for that purpose.

In terms of CI tools, videos and simulations were considered valuable by most students. Their impact on student learning was also partially reflected in student responses to other open-ended questions. For instance, students were able to provide important justifications when prompted to discuss energy sources with an advocate of a particular approach, such as "*You have to factor in the cost, the power it supplies, and the effectiveness over X amount of years.*" "*The best way to select an energy source is to focus on being environmentally friendly first. Then find the most cost effective that will produce enough energy for your needs.*" Students also commented on the most important things they learned through the game such as "*The most important thing that I learned was to be environmentally friendly rather than being the most cost and energy efficient*", and "*How money can decide on what energy source you can have to run your city*". Although about a third of the stu1dents had prior programming experience, the programming in the game was deemed hard, and was the tool most likely to require teacher assistance. Given that programming is to play a larger role in subsequent games, we plan to design additional instructions and/or internal resources to support use of this tool.

Overall, there was considerable variation in responses to most questions, indicating that the items were appropriate in addressing a range of student backgrounds and attitudes. Some felt the game was too elementary, others found it enlightening about aspects of energy generation. All but one student felt that the game was more realistic than textbook problems, although no students felt it was truly like an authentic job assignment. When asked how, after playing the game, they might discuss power options with an advocate for one or another energy source, students came away with a variety of opinions. A third of the students emphasized that "it depends" on balancing a variety of factors. Another third had pro-environmental or pro-energy-production positions. Others had more nuanced points of view, such as the need to have more than one energy source, or the need to find out what the people affected value most.

Future evaluation will involve interviews of students in addition to surveys. It will be important to know more about what specific features of the experience elicit their responses. For example, are students who find the game "elementary" responding to the engineering facts (which are covered by instructors prior to the game experience) or to the problem-solving context?

Student responses were mostly articulate and to the point, and some students seemed to have a career focus. One commented, "*Online chat is absolutely necessary in anything that involves engineering. Teamwork is very important and I use it so my fellow students and I can learn from each other or work together*." In general, however, the student responses were focused on the game context. It was not clear from the survey the extent to which students recognized that the game represented authentic professional activities: meetings with clients, preparation of complex project maps, formal brainstorming to identify knows and unknowns, collaboration with colleagues, and delivery of reports based on findings.

5. CONCLUSION

This paper addresses the inclusion of CI-learning experiences into the PLTW curriculum. In particular, the pedagogical approach develops a series of virtual reality games seamlessly integrated into different levels of the PLTW courses. The content and features of the games as well as their alignment with the curriculum are clearly presented through two interactive game examples, *Power Ville* and *Stability*. The preliminary assessment of *Power Ville* presents encouraging results as the students loved the game as an interactive and fun deliverable method of learning. The insights on how and why students responded to different tools provide a solid foundation to leverage our game system through iterative design-based research [2].

6. ACKNOWLEDGMENTS

This work is supported in part by the National Science Foundation grant #OCI-1041306 and #EEC-0935089. The authors would like to thank Aaron Johnson, Bradley Ebinger, and Christopher Franzwa for their helps in producing the screen shots presented in the work.

7. REFERENCES

[1] J. Bordogna, "Next generation Engineering: Innovation through integration," in *Proc. NSF Eng. Educ. Innovator's Conf.*, April 8, 1998, Keynote address. http://www.nsf.gov/pubs/1998/nsf9892/next.htm

[2] Dede, C., "Why design-based research is both important and difficult," *Educational Technology*, 45, 1, pp. 5-8, 2005.

[3] Douglas, J., Iversen, E., and Kalyandurg, C., "Engineering in the K-12 Classroom – An Analysis of Current Practice and Guidelines for the Future," http://teachers.egfi-k12.org/wp-content/uploads/2010/01/Engineering_in_the_K-12_Classroom.pdf.

[4] Duschl, R. A., and D. H. Gitomer, "Strategies and Challenges to changing the focus of assessment and instruction in science classrooms," *Educational Assessment*, 4:37-74, 1997.

[5] Harper, L. E., "The social consequences of design: PBL workshops for undergraduate researchers," *Proceedings of American Society for Engineering Education Annual Conference*, Session 3261, 2004.

[6] Herbert, B. E., "The role of scaffolding student metacognition in developing mental models of complex, Earth and environmental systems," *DFG-NSF International Workshops on Research and Development in Mathematics and Science Education*, Washington, D. C., Nov. 2003.

[7] Hyde, A. and Bizar, M. *Thinking in context*, White Plains, NY: Longman, 1989.

[8] Kim, B., Park, H., and Baek, Y., "Not just fun, but serious strategies: using meta-cognitive strategies in game-based learning," *Computers & Education*, 52, 2009, pp. 800-810.

[9] Lima, M., "Engineering education in the wake of hurricane Katrina," *Journal of Biological Engineering*, 2007, 1:6.

[10] Narode, R. B., "Pair Problem-Solving and Metacognition in Remedial College Mathematics," *Technical Report*, http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1c/53/3b.pdf.

[11] Neeley, K., Elzey, D., Bauer, D., and Marshall, P., "Engineering in context: a multidisciplinary team capstone design experience incorporating real world constraints," *Proceedings of American Society for Engineering Education Annual Conference*, Session 3461, 2004.

[12] Ogle, Donna M. (1992). KWL in action: Secondary teachers find applications that work. In E. k. Disher, T. W. Bean, J. E. Readence, & D. W. Moore (Eds.), Reading in the content areas: Improving classroom instruction (3rd ed., pp. 270-281). Dubuque, IA: Kendall-Hunt.

[13] Oliveira, M. and Serrano, J. A., "Learning roadmap studio: new approaches and strategies for efficient learning and training processes," http://www.elearningeuropa.info/files/media/media16938.pdf

[14] Petroski, H., S*uccess through failure: the paradox of design*, Princeton University Press, Feb. 2006.

[15] Sasha A. Barab, Melissa Gresalfi and Anna Arici, "Why Educators should care about games," Teaching for the 21st Century, Vol. 67, No. 1, pp. 76-80

[16] Schumacher, J. and Gabriele, G. A., "Product design and innovation: a new curriculum combining the humanities and engineering," *Proceedings of the 29th ASEE/IEEE Frontiers in Education Conference*, Session 11a6, San Juan, Puerto Rico, Nov. 10-13, 1999.

[17] Vygotsky, L. S. (1986). Thought and language. Cambridge, MA: MIT Press.

[18] Yasuhara, K., Morozov, A., Kilgore, D., Atman, C., and Loucks-Jaret, C., "Considering life cycle during design: a longitudinal study of engineering undergraduates," *Proceedings of American Society for Engineering Education Annual Conference*, AC 2009-1728, 2009

[19] "Cyberinfrastructure for Education and Learning for the Future: A Vision and Research Agenda," Technical report, Computing Research Association. http://archive.cra.org/reports/cyberinfrastructure.pdf.

[20] Green City game, http://ready2beat.com/entertainment/games/greencity-new-version-sim-city-games/linkout.

[21] Mobility game, http://www.mobility-online.com/en/informations/generalinformation.html.

[22] Project-Lead-The-Way, www.pltw.org

[23] What is a sustainable city?" http://archive.rec.org/REC/Programs/SustainableCities/what.html.

Appendix:

Power Ville Feedback Survey

1. To what extent did this assignment in the game seem like a realistic engineering problem?
   - Not Very. It was like any other textbook assignment.
   - Somewhat. It was more realistic than working problems in a textbook.
   - Very. It seemed like something I might do some day on a job.

2. Please rate the usability of the following game tools

Q2 response options	Unassisted on 1st attempt	Unassisted on 2-5 attempts	With peer help	With teacher's help	Did not use
KWS					
Road map					
Online chat with your group players					
Simulator					
Programming code					
Information acquisition tools					

3. In the game, you gathered information on power sources by talking to the Mayor and energy experts in the information acquisition stage. How often did you use the following tools during this stage?

4. In the game, you analyzed information for your final report in the analysis and decision making stage. How often did you use the following tools during this stage?

5. How helpful were the following tools for you to accomplish the assignment in the game?

Q5 response options	Not very. They made no difference in my game progress.	Somewhat. They helped, but I could have succeeded without them.	Very. They really helped me understand what I had to do next.
KWS			
Road map			
Online chat with your group players			
Videos on different energy sources			
Mini-quizzes			
Simulator			
Programming code			

6. If you could keep only one of these tools in the game, which one would you select?

7. If you could eliminate only one of these tools in the game, which one would you remove?

8. Before playing Power Ville, how often had you used each of these tools for solving problems in school work?

9. How likely are you to use these tools on your own in engineering or other subjects?

Q8-9 response options	Never	Once or twice	+3 times
KWS			
Road map			
Online chat with your group members			
Simulator			
Programming code			

10. You get in a discussion with a person who is totally for or against nuclear, solar, coal or wind energy. What would you tell them, as an engineering student, about the best way to select an energy source?

11. What is the most important thing you learned from playing Power Ville?

12. What is the most important thing you need to learn next for a clean power production system design?

Q3-4 response options	1-2 times	2-4 times	4+ times	never
KWS				
Roadmap				
Online Chat				

June 2012

Volume 3 Issue 1

TABLE OF CONTENTS

ISSN 2153-4136 (online)