FRIENDING THE HUMANITIES KNOWLEDGE BASE:
EXPLORING BIBLIOGRAPHY AS SOCIAL NETWORK IN ROSE

WHITE PAPER FOR THE NEH OFFICE OF DIGITAL HUMANITIES:
ROSE DIGITAL HUMANITIES START-UP GRANT (LEVEL 2) HD-51433-11
(9/1/2011 TO 9/30/2012)

Alan Liu (ayliu@english.ucsb.edu), Rama Hoetzlein, Rita Raley, Ivana Anjelkovic, Salman Bakht, Joshua Dickinson, Michael Hetrick, Andrew Kalaidjian, Eric Nebeker, Dana Solomon, and Lindsay Thomas

University of California, Santa Barbara

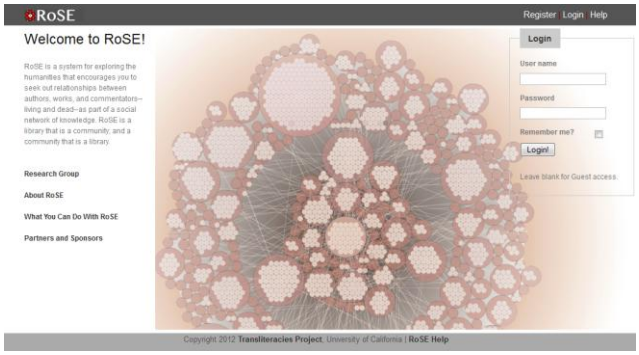

Figure 1: RoSE home page. (See Appendix A for larger screenshots from RoSE.)

WHAT IF?

• What if bibliographies of past authors and works could be modeled as a dynamic, evolving society linked to today's scholars and students?

• What if scholars and students could add data about biographical, historical, and intellectual relationships to the bibliographical entries, thus using present-day "crowdsourcing" to make more socially meaningful the crowds of history?

• What if humanities resources--sometimes with non-conformant metadata from the distant past--could especially benefit from this process?

• What if we could change the nature of initial research from "searching" to participatory "making"?

• What if visualizations could help us actively "storyboard" intellectual movements and not just spectate them?

• And, recursively, what if such a system for active learning with interactive technology could mirror the way the system developers themselves collaborated--integrating the humanities, arts, and engineering to explore the humanistic issues in technological problems, and the technological issues in humanities problems?

1. INTRODUCTION AND SUMMARY

We report in this paper on a project we advanced from an initial prototype to beta stage in 2011-12 with a NEH Digital Humanities Start-up Grant (Level 2). We aim not just to narrate grant objectives, activities, and results but also to surface some of the larger digital humanities issues--inextricably humanistic and technological, theoretical and practical--that we engaged.

The project is called RoSE (Research-oriented Social Environment), an online knowledge exploration environment for humanities scholars and students developed in the Ruby on Rails programming environment on top of a MySQL database. Accessed through a Web site (http://rose.english.ucsb.edu), the system includes the following main content and interface features: • an extensive set of bibliographical metadata (but no full texts) machine-harvested from Project Gutenberg, YAGO, and SNAC (Social Networks & Archival Contexts); • an initial set of user-entered metadata (including "relationships" and "keywords") added to the pre-existing data; • a user interface with search and editing functionality modeled as a social network site with "profile pages" for each author, work, and user; • interactive visualizations in several styles to facilitate navigation and understanding; • "history"-tracking and "collections"; • "storyboards" to shape visual arguments; • and user documentation, including a "Quick Start Guide" and demo video.

ACCESSING ROSE

Explore RoSE (rose.english.ucsb.edu) as a guest user by leaving the login fields blank and clicking "login." Or request access as a registered user with a profile page who is able to add to our knowledge base (contact: ayliu@english.ucsb.edu) Currently, the RoSE beta is open on a limited basis (by request) for registration from scholars, teachers, students, and others.

with data for 11,964 people and 34,077 documents, but with no keywords or information about relationships. YAGO provided data for 7,557 people and 11,395 documents, with keywords offering opportunities for clustering studies and also "influence" identifiers to establish relations between people.

To initiate our new work with SNAC data we engaged with Daniel Pitti, Director of SNAC, and other SNAC lead developers, who generously provided data from their project. We worked with 125,000 individual SNAC XML files to harvest information useful for RoSE. The process began with parsing SNAC entries into selected component name, date, document, and keyword metadata and removing parsing errors to produce a clean, relatively compact set of XML files compatible with RoSE metadata format. Then, because the particular nature of SNAC's data sourced from finding aids produced a disproportionately large amount of material not well suited to RoSE (e.g., singleton items not connected to other items; many corporate authors or collecting entities), we undertook a second stage of processing. We identified principles for selecting well-connected items from the SNAC data that would help users see networks of people and works; and we also honed further methods for filtering, parsing, and cleaning up the information to match RoSE's mission. In doing so, we concentrated on the approximately 74% of the above-mentioned 125,000 SNAC files that are entries for people, and concentrated in addition on the 3% of these SNAC files representing highly connected people.

At the end of our grant, we had met our objective by selectively harvesting data from SNAC for 3,412 people and 28,036 works. The inclusion of SNAC enabled us to make interesting global scale comparisons between our three machine-harvested datasets. For example, in Figure 3 we compare author data we harvested from Yago, Project Gutenberg, and SNAC by historical birth year. (Although only a fraction of author data included birth dates--87% for Yago, 67% for Guttenberg, and 22% for SNAC--we project the actual number of authors per year in the data based on the ratio of those with dates to the total number of authors.) One finding we made is that these datasets tend to be complementary rather than redundant in coverage. Yago, based on Wikipedia, and Project Gutenberg, for example, overlap much less in authors than we originally expected, due primarily to the different periods of time they cover. By corollary, another finding we made was that YAGO, Project Gutenberg, and SNAC show a distribution of birth dates peaking around the years 1950, 1860, and 1900, respectively. Other characteristics of the collections such as the ratio of the number of unique surnames to the number of unique person entries seem to indicate a "familial preference" within some of the datasets, although this theory will have to be explored in more detail before any hard conclusions can be stated. Nonetheless, the statistical comparison of high-level inter-dataset features such as these

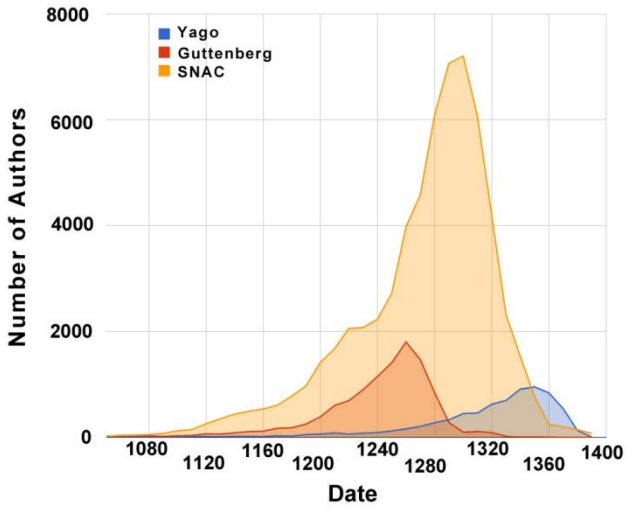

Figure 3: Graph of number of authors in our Yago, Project Gutenberg, and SNAC datasets by year. Our research shows that, as scholarly sources, these datasets are complementary rather than overlapping.

could reveal otherwise hidden patterns and be the basis for future data-mining research.

We also benefited in other ways from working with SNAC data. First, we were in effect trying out the paradigm by which one digital humanities project, in this case, RoSE, could act as a "client" of another digital humanities project, SNAC, where the client is positioned as a data-receptor for the other project's data source. This not only allowed us to make a suggestion to SNAC (that they consider creating what amounts to a selectable set of "channels" for particular datastreams--e.g., "humanities," "history," "literature," or "poetry" content-channels, as opposed to "science" channels)--but also encouraged us to think about how our own RoSE project might in the future serve client-projects of its own. RoSE currently has limited interoperability through XML/RDF export on an individual record basis. Secondly, we also benefited from comparing notes with the SNAC developers about data visualization. (The SNAC team began work on visualizations during our grant period.)

• ***Improve Visualizations in RoSE; Add More Visualization Types*** *(original stated objectives).* Another goal was to improve and expand RoSE's repertory of visualizations, which originally consisted of social-network and timeline graphs. Creating new, and improving existing, visualization types was challenging because our visualizations are dynamically generated; filterable; interactive (allowing a user to change their "point of view" in a network, for instance, by clicking on a different node); and information-rich (optionally reporting navigation history and other metadata in a sidebar). Also, we expanded our goals for visualization to support the important new "storyboard" function in RoSE (see below).

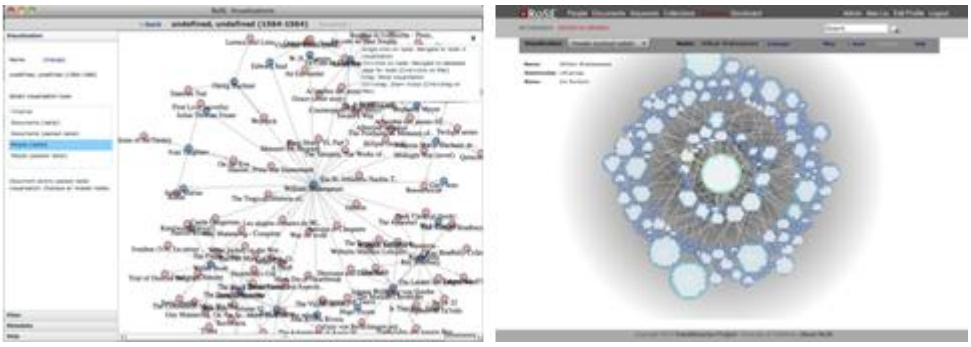

Figure 4: Social network and packed radial visualizations in RoSE

At the end of the grant period, we had met our objectives by improving our existing visualizations types (though the timeline graphs do not yet fully demonstrate the results we would like because of missing "date" information in many of our sources). We had also added (and continued to improve) a new radial "packed radial" visualization plus a means for users to produce, save, and export visual "storyboards." More detailed description of the design principles we employed in creating the RoSE visualizations may be found in **Appendix C**.

Finally, we undertook a major logical redesign of the algorithm generating visualizations from our database. To optimize real-time performance, we implemented a step-saving method of iteratively processing a whole "wavefront" of active nodes at a time (detailed explanation in **Appendix D**).

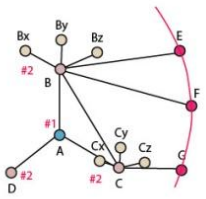

Figure 5: RoSE "wavefront" visualization algorithm

• ***Improve User Interface and General Usability*** *(ongoing objective in addition to stated objectives).*

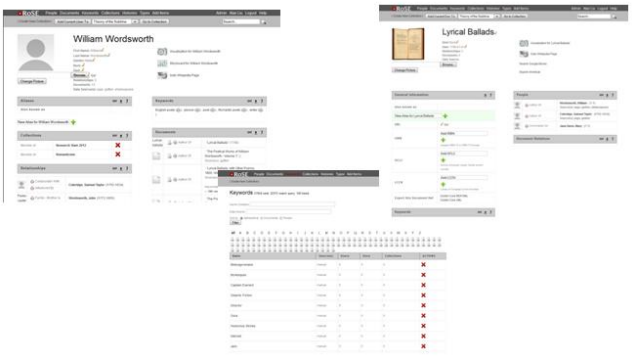

Figure 6: Examples of user interface screens in RoSE

During the year, we continuously and iteratively improved the RoSE user interface to simplify the presentation of data and search/edit functions; clarify features; improve the workflow by which users find, add, and edit data; integrate the database user interface with the visualization interface; and, of course, solve bugs and browser compatibility issues. To guide this work, we drew on feedback from our own developer group (who engaged in collective test sessions) and from students in our use-scenario study (see below).

However, we also knew we had to be realistic about how polished and smoothly functional our beta product could be, since otherwise the entire grant could have been spent on usability improvements to the detriment of other tasks. For this reason, we carefully prioritized usability issues and concentrated on those that were most important or relatively easy to fix, leaving additional improvements for a future implementation stage of the project.

At the end of the grant period, we had made significant progress on usability--especially in regard to issues that students identified in our use-scenario study. In addition, we created help documentation--including a Quick Start Guide, a demo video, and a "Learn More" suite of resources. However, a realistic assessment would be that we ended only at about 85% of the way toward "optimal" usability, where "optimal" means that a new user is unlikely to run at some point into a bottleneck (a confusing feature, a feature not yet fully debugged, or an inefficiency in work flow).

• ***Create "Collections," "Histories," and "Storyboard" Features*** *(new objective)*. Our work on adding metadata to our database, creating visualizations, and improving usability led us midway through the grant year--as a logical extension of functionality improvements--to develop a way for users to collect and save their findings in RoSE. We thus created a "history"-tracking feature, which can be toggled on to automatically save a record of items traversed; and also a "collection" feature, which creates named, shareable collections of items (to which other users can "subscribe").

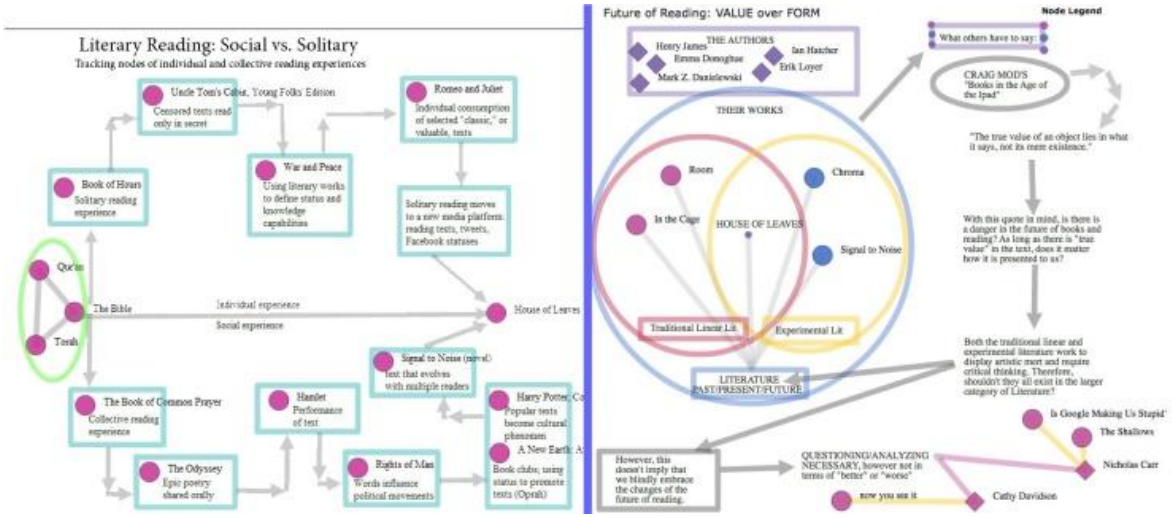

Figure 7: Examples of RoSE storyboards created by undergraduate students. (Courtesy of Kristin Crosier and Dani Williams.)

Then we had a breakthrough new idea: users should be able to work manually with their histories and collections so as to shape them into filtered, clarified, annotated, and otherwise interpreted proto-"arguments" or "narratives"--i.e., sketches of arguments about an intellectual topic that can be presented to others. An analogy might be the difference between communicating an idea as a bullet list of nouns and as a fully-formed, syntactical sentence. The logic of coordination and subordination required to form a sentence forces one to shape a loose set of ideas into an argument (or, as in a topic sentence, the beginnings of an argument).

We thus created an innovative "storyboard" feature that allows users automatically to populate a visual canvas with node-and-link representations of persons and works in their collections or histories. This is the equivalent in the analogy above of a bullet list. But then users can arrange, connect, color, annotate, and draw arrows and shapes around their data. This is the visual equivalent of writing a sentence. Finally, they can save their storyboard as an XML file and reload it later for revision (or load other people's storyboards). They can also export or print their storyboard as an image file (examples in Figure 7).

• ***Run Use-Scenario Studies*** *(original stated objective)*. Originally, we planned to study how RoSE operates in three real-life "use scenarios." Not the same as formal usability tests, these studies were appropriate to our prototype-to-beta development work because they could offer in-progress evaluation. Our intent was to identify opportunities and problems through observation of users and participant interviews and questionnaires.

The three use-scenario studies we planned included: (a) undergraduate classroom use, (b) professional conference or collaborative project use, and (c) individual-scholar use (e.g., researching a dissertation). This proved to be too ambitious, however; in part because we also had to apply for human-subjects protocols to work with human test subjects (an aspect of research not traditionally familiar to humanities scholars). We thus decided to simplify by concentrating on the use-scenario study from which we expected the most important results: undergraduate classroom use. For use scenarios "b" and "c" above, we only performed informal studies.

A. Undergraduate Classroom Use-Scenario Study

We chose for our use-scenario study an undergraduate course co-taught at UC Santa Barbara in spring quarter 2012 by two of our project team: Rita Raley and Dana Solomon. The course (with 23 students) was English 146, "Distracted Reading," which explored the topic of "reading" practices with special attention to different media environments (https://engl146.wordpress.com/about/). The instructors created one assignment for the course that required students to use RoSE. The assignment was described as follows (slightly abbreviated):

> Students will individually create what is called (1) a "collection" organized around a particular topic such as attention, distraction, and online reading, as well as (2) a "storyboard" that simplifies, narrates, or otherwise manually shapes their collection. These storyboards will visualize relations among authors and documents, such that each student will in effect be creating a network map that will both reflect the work we have done in the class (e.g. showing us how *House of Leaves* connects to Henry James) and produce new knowledge (e.g. how does your visualization show us something about the problem of distraction that we had not previously discussed). The objects you need for your collections may not all be in the RoSE system at present, so some of the work of this assignment may be data entry.
>
> We will demo the system in class and Dana Solomon will be holding regular lab hours in SH 2509, where you are