

White Paper Report

Report ID: 103929

Application Number: HD-51343-11

Project Director: Lucia Binotti (lbinotti@email.unc.edu)

Institution: University of North Carolina, Chapel Hill

Reporting Period: 9/1/2011-2/28/2014

Report Due: 5/31/2014

Date Submitted: 3/7/2015

- Type of report: White Paper
- Grant # HD 5134311
- Title of Project: Gnovis: Flowing Through the Galaxy of Knowledge
- Project Director: Lucia Binotti
- Institution: The University of North Carolina at Chapel Hill
- Date Submitted: March 7, 2015

GNOVIS, Flowing Through The Galaxy of Knowledge

NEH Start-Up Grant Final White Paper

During the fall semester 2011 Prof. Lucia Binotti taught a graduate seminar titled “The Intellectual History of Spain” as a workshop in which the students gathered and organized materials to create a database of selected works of Renaissance and Golden Age Spanish Literature. The seminar attracted nine graduate students from different disciplines in the Humanities: Romance Languages, Comparative Literature, and Information and Library Science.

During a couple of initial sessions computer programmer Chris Smith (NCSU Dept. of Statistical Biology) taught the students how to encode the knowledge they would be gathering using Google spreadsheets to take advantage of their existing skills. They created spreadsheets by author and work with a wealth of detail: biographical information about the author (i.e. profession, birthplace, major life events -- serving in the military, or being imprisoned, etc.); geographical ties between the author and different places in Spain, Europe or the rest of the world; information about the most famous works by the author (for each author we primarily focused on one work); main characters and themes in the work; genre and subgenre of the work; literary influences on the author or the work; literary movements to which the work is related; influence of the work or author on other works or authors; publication dates of the work; early editions of the work (both place and publisher); early translations of the work into other Romance Languages and English; and, finally some useful modern annotated editions and scholarly resources, such as names of articles and specialists on certain authors.

As the seminar proceeded, Mr. Smith showed the students several possible visualization solutions for the network model and the students were then asked to create simple images of how they imagine Gnovis might look. These images were presented to the class for discussion, and have been posted on the project website <http://gnovis.net>. Throughout the semester we documented our progress on the site, and, at the end of the seminar, the students posted their final projects and their reflections on the experience of developing a computer application. A round table discussion on Gnovis was presented by the students and the project participants on March 22nd 2012 at the 18th Annual Carolina Conference on Romance Literatures at the University of North Carolina at Chapel Hill.

During the second period of the grant we started to experiment with new visualizations based on the designs produced by the students, going beyond the simple timeline and world map displays already available. We started working on two interactive visualizations, one that uses segmented wheels and one that uses rectangles to represent works and people. This investigation led us to start discussing the exact details of how the user will interact with these representations of the entities, as explained in the narrative above.

Mr. Smith added a new tool to the world map visualization that allows the user to restrict the date range of the entities displayed in the visualization. This is just one example of the user interface elements for filtering, searching, and selecting that will be an important component of our implementation phase.

During the fall semester 2012 Prof. Lucia Binotti taught the undergraduate course "Survey of Medieval and Renaissance Spanish Literature" as an "applied research" course in which the students contributed original research to the Gnovis database. Every fourth session a group of students presented a new database spreadsheet to the rest of the class. In addition, each student worked on an individual spreadsheet for her final project. Emily Clark, one of the alumni of the Gnovis graduate seminar directed the students' projects and introduced them to the parameters for gathering information for the Gnovis database. During the spring semester of 2013 Prof. Lucia Binotti taught the undergraduate course again. This time Ruben Rosales, another one of the alumni of the Gnovis graduate seminar directed the students' projects. Although at that time the project was beyond the Start-Up Grant funding period, data gathering for the sample database of Renaissance Spanish literature continued to be collected.

In July 2013, Dr. Lucia Binotti presented Gnovis, in its Spanish Literature version, at the Annual Conference of the Asociación Internacional de Hispanistas held in Buenos Aires, Argentina. The project received enthusiastic endorsements and aroused interest among scholars, especially from the editors of the digital medieval and early modern journal *E-Humanista*

During the summer of 2013 Ms. Elena Casey worked on the project's website graphics, focusing specifically on ways to visualize individual authors' connection.

During the duration of the grant we started to develop a novel flexible data harvesting and visualization tool designed specifically to meet the requirements needed for a useful visualization of data relevant to the humanities. In particular, we wanted to expand the visualization paradigm to incorporate qualitative characteristics of data entities and their linkages. Additionally, each individual user would be provided with as much control as possible in deciding which characteristics are most important to their particular search. We saw user extensibility and the ability to add and modify data to the tool as critical to the widespread acceptance of its use. We envisioned Gnovis as a tool that in its fully developed stage would allow the visualization of large volumes of information of a general nature as opposed to information of a technical nature. We started designing the tool to be able to accommodate the large variation in the types of data entities encountered and the possible linkages between them.

During the startup phase we developed a prototype visual interface for a compact but detailed set of knowledge in the humanities, while concomitantly we explored a more fluid style of interface than the current web-page paradigm where a click can take you to a new page, more smoothly navigable even than the Google maps style that allows the user to drag information from out of the current view. We worked with the projection that a full implementation Gnovis would allow researchers to combine disparate data sources with local data to produce scholarship that puts local research in a larger context. The tool would be similar to Google Knowledge in scope, but would encourage an open ecosystem of data importers that will allow researchers to contribute new data sources. Since the project was to be open source, researchers would customize the tool to import their own data and utilize custom visualizations. The standard tools we envisioned would include a Wikipedia importer which would allow users to download all of Wikipedia's tagged data and combine it with their own institution's data using a

spreadsheet importing tool to provide visualizations which put their own research into a broader context, making wider connections.

The scope of the Gnovis prototype remains for now within the boundaries of academic research and teaching. During the start-up period its main audience was conformed of students (graduates and undergraduates) at the University of North Carolina at Chapel Hill. The prototype's development actively contributed to the fostering of independent research among undergraduates, extending the boundaries of work in the digital humanities to a future generation of scholars. A round table discussion on Gnovis was presented by the students and the project participants on March 22nd 2012 at the 18th Annual Carolina Conference on Romance Literatures at the University of North Carolina at Chapel Hill. In the spring of 2013 the project was presented to the newly created Digital Innovations Lab at UNC-CH and featured as one of the Carolina Digital Humanities Initiative Faculty Projects. When Binotti presented Gnovis at the Annual Conference of the Asociación Internacional de Hispanistas the project received enthusiastic endorsements and aroused interest among scholars, especially from the editors of the digital medieval and early modern journal *E-Humanista*.

The Gnovis prototype was never formally evaluated outside of the classroom. Among the students the response was very positive. Students write:

"The Gnovis project is an extraordinary and rare opportunity to pursue research in the humanities. While significant data and writing on these authors exists, the information is scattered throughout thousands of libraries, books and online articles—often in varying languages or with restricted access. Through the Gnovis project, we are creating a searchable, central database that connects seemingly independent authors and subjects. What I entered as an undergraduate Spanish literature class evolved into a vast research project revealing sweeping influences and interconnectivity of Spanish authors throughout the centuries." "There is currently no central database where one can access information on the most prominent works and their author's biographical information in a way that is both easy and interactive. When this database will be completed, a student, scholar or educator will be able to search for an author or work of medieval Spanish Literature, and locate on a map, where a given author was born, traveled to and which other authors he or she may have interacted with. Such digitalization will be important in drawing new insights about an incredibly important time in Spanish Literature, that is often limited by gaps in information due to age of information. My experience with research was incredibly fulfilling because it allowed me to take autonomy over my education. By working on individual project where I had to immerse myself in a specific work and author, I got a glimpse of the world of academia. It meant going through every study, criticism, review, book and text that I could get my hands on and forcing myself to make my own conclusions about what information to include in my spreadsheet of information, while remaining as objective as possible.

Our own assessment of the program identified Gnovis major strengths in its ability to recognize the need for a multiplicity of different data domains and to provide the means to navigate through and between these domains. Each domain is coded as a separate database and maybe as specialized (i.e., Spanish poets who were soldiers) or general (i.e., all poets with Wikipedia pages) as the user wishes. A common XML coding scheme allows Gnovis to know how to display the data as well as how to interface the two (or more) databases. This design permits the creation of reusable database units by Gnovis users, which may be recombined later for new results.

On the other hand, we identified Gnovis major challenge in regards of its extensibility. The most important aspect of the extensibility of the tool for the novice user is the ability to create new databases and link them in various combinations with other databases. However, for the tool to have any long-term viability, it must be able to respond to unforeseen changes in technology. We felt that the best way to overcome this hurdle is to provide the code for the project to the open source community. The computer coding done to date has been done with an eye towards future extension. It is hoped that in the years to come Gnovis will develop a following of users who are willing to provide new code to the project as needed.

Gnovis is a first attempt at producing an engine that can display corpora of knowledge in multi-dimensional environments, allowing for the visualization of extensive amounts of material organized semantically in nested clusters.

During the startup phase we developed a prototype visual interface for a compact but detailed set of knowledge in the humanities, while concomitantly we explored a more fluid style of interface than the current web-page paradigm where a click can take you to a new page, more smoothly navigable even than the Google maps style that allows the user to drag information from out of the current view.

With funding, we would be now ready to develop Gnovis to its full potential, as a tool that will be applicable to an extensive array of different corpora whenever connections among their elements can be made and encoded for programmatic access. When applied to the humanities a full implementation of such a tool will be extremely useful because it will permit the user to identify or define new connections that are not apparent when entities are classified in linear sequences. Gnovis will add to the body of visualization tools available to digital humanities researchers and it will improve visualization literacy amongst humanities researchers.

Although an application to the 2013/14 NEH-DH Implementation Grant was unsuccessful and the project has been idle for the last year, Binotti plans to seek further funding, as the start-up phase of the Gnovis prototype showed that our project addresses two important theoretical issues both at the core of the most current inquiry into the meaning and the future of digital humanities. These issues guide the two interrelated directions we would seek to pursue in the future.

1) As an innovative contribution to more theoretical research in the digital humanities, Gnovis is a tool that in its fully developed stage would allow the visualization of information of a general nature as opposed to information of a technical nature. We are designing the tool to be able to accommodate the large variation in the types of data entities encountered in the humanities and the possible linkages between them.

2) It is one thing to create a tool for *ad hoc* data sets designed for specific research projects in which a student or a scholar spends several weeks or months inputting data. It is quite another to take advantage of large *existing* data sets. The latter is far and away the more challenging task, but also the more useful. There already exist several databases that contain vast quantities of information of precisely the kind we would like to visualize. We would like to address the possibilities of scaling Gnovis up to handle the datasets of millions of objects.

A key component of the Gnovis project is its ability to visualize data previously stored. There are a plethora of libraries and document clearing houses that offer their data online. If the project were to be continued, we would collaborate with *PhiloBiblon* for

the purpose of developing a data importer module for Gnovis. With their assistance, we would create an interface to their web server via a 'web service' using the open source package such as Apache Axis that can extract data and format it into a form acceptable to Gnovis. The data connection to PhiloBiblon would serve as a template for future connections to other large data storage facilities. These facilities include ArtStore, JStor, and other large-scale humanities databases that already have precisely the kind of information that humanities scholars want to use.

PhiloBiblon, which we plan to use as a testbed for scaling up Gnovis to make use of existing large databases, is a bio-bibliographical database of the primary sources, manuscripts and early printed books, for the study of the medieval Iberian Romance literatures, Spanish, Portuguese, Galician, and Catalan. It began as an adjunct to the NEH-supported Dictionary of the Old Spanish Language at the University of Wisconsin, Madison, and has been led for the last thirty years by Prof. Charles Faulhaber (UC Berkeley). For over twenty years teams of scholars from Spain, Portugal, and the U.S. have dedicated thousands of hours of unpaid labor to enhancing it. It has been the recipient of nine NEH grants between 1989 and 2010. Currently, it contains over 300,000 records in four separate databases, Bibliografía Española de Textos Antiguos BETA, Spanish literature), Bibliografía de Poesía Áurea (BIPA, Golden Age Poetry), Bibliografía de Textos Antigos Galegos e Portugueses (BITAGAP, Portuguese and Galician), and Bibliografía de Textos Antics Catalans, Valencians i Balears (BITECA, Catalan).

We have chosen PhiloBiblon as our testbed because the data in the web version are already in XML format and because the Windows version of the underlying OpenInsight database has export capabilities that will allow us to create *ad hoc* data sets for test purposes. In addition, Prof. Faulhaber has agreed to serve on our advisory board.

Whatever will be the future of Gnovis, the Start-up period showed us that Gnovis can be used effectively as an innovative educational tool. Developing the potential of Gnovis as a tool for the teaching in the humanities has become an important focus of our sustainability goals for Gnovis.

Gnovis has a website at <http://anderson.statgen.ncsu.edu/gnovis/>

The data gathered to date has been compiled into a relational database in "triples" format. Chris Smith wrote scripts to automatically access the spreadsheets, parse the information contained in them, and create entries in the database from the content. The resulting triples can be seen on the Gnovis website (via the "Early Prototypes" menu entry), and the Spanish literature entities can be seen on a simple timeline, as well as on a world map.

The code behind the two prototypes has been consolidated into a single "Gnovis engine" with each prototype requiring only a thin layer of configuration code above the engine to produce the results seen on the website. Charlotte Lindemanis worked on the controlled vocabulary for the literature site.