Submitted on: 08.08.2016

Source Notes: developing a news storage and research system

Eric Johnson
Data Librarian, Center for Digital Scholarship at Miami University, Oxford Ohio, U.S.A.
johnsoeo@miamioh.edu

Greg Reese
Senior Research Computing Specialist, Research and Computing Support at Miami University, Oxford Ohio, U.S.A.

Andrew Offenburger
Assistant Professor of History, Miami University, Oxford Ohio, U.S.A.
offenba@miamioh.edu

Abstract:

Finding, retrieving and annotating digital news items is almost as important as their generation and storage. Our university is developing new tools for digital news research. To aid in the collection and close reading of news documents, partners from diverse units across campus have collaborated to develop a software tool that allows researchers to code and search for a vast array of subjects, names, dates, direct quotes and interpretative notes within large digital news collections.

Each news item can be annotated by multiple people with accurate tracing of each person's contribution. Designed to be simple for undergraduates to use, but powerful enough for professional research, the text, annotations and other metadata are fully searchable. The results of researchers' queries can then be processed by further analytical tools including text mining or individual reading. This tool's theory, background, design and operation is described along with plans for the future.

Keywords: newspaper database, digital humanities, history research, research tool, crowdsource

Overview

SourceNotes is a tool being designed by our library that combines an interface for researchers to annotate newspaper documents, a database for storing notes and the full text of each article with a search engine to retrieve thematically connected notes and articles. This database is designed not just to store the digital information, but to allow scholars worldwide to better understand, share and critique those news resources.

To do this, we created a collaboration between Miami University's History Department, Research Computing Support, and the Library's Center for Digital Scholarship.

The Center for Digital Scholarship (CDS) is a support service of our university library system. We provide high level, high touch support to faculty, grad students and high performing undergraduates. This support takes on many forms. Beginning with an initial consultation, we determine the researcher's desires and needs. This initial reference interview can also be a supportive training event that helps new researchers refine their research agenda, identify and locate requisite data and clearly state their research question. We determine what resources and guidance can be provided by the CDS and set up a schedule of follow-up meetings as needed.

Many of our projects involve digital humanities and include a 3D gallery exhibit for the anthropology department, archives of historic student newspapers, scanning and organizing historical documents of the local chapter of the National Organization for Women and teaching workshops.

We also provide services such as helping faculty write data management plans for grant applications, quick turnaround purchasing of research data for undergraduates, scanning and organizing, and guidance in visualizing data using a variety of software tools.

Background

In September, a new History Department faculty member Andrew Offenburger approached the Research Computing Support (RCS) department for help with a database and software project. He had an idea for the design of a research note taking system, but with no programming or database skills, he needed the collaboration of a team. The people in RCS knew about Eric Johnson because he had worked with them on a prior project.

They called together a meeting of five individuals including Andrew to discuss his project and how we could support him. The resulting team of three included Greg Reese from RCS to provide computer code writing, Eric Johnson from CDS to design the database and Andrew Offenburger from the History department for vision and project goals. We detailed a timeline and began meeting to refine the project's details.

Andrew had created a system for personal note taking while he was writing his dissertation. He didn't want to be like a grad student ending up with a lot of Word documents that they can't make sense of. So he used a Microsoft Access database to store comments, direct quotes, subject categories and other metadata about newspapers he was researching.

These papers were focused on the American West region near the United States – Mexico border during the late 1800's. He would read every article in a run of papers, recording comments about anything he found interesting. He documented the names of all individuals mentioned as well as historically significant events and activities in his database. He was replicating the method used by many historian researchers, but using electronic storage instead of index cards.

A few decades ago, researchers could use sortable punch cards for recording and retrieving categorized information.

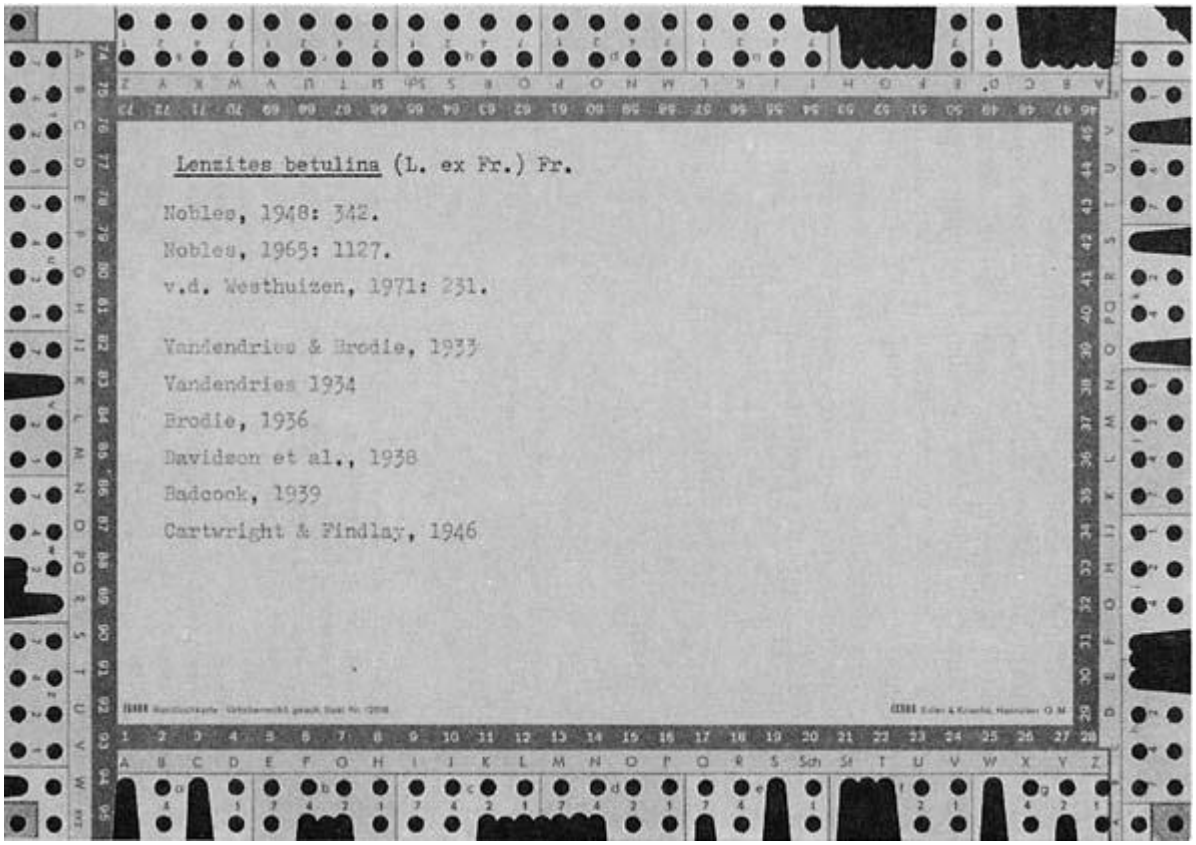

Figure 1: Generic research punch card

This card has space in the middle for recording free-form information as well as holes and clipped notches around the edge. Cards like this were bought from local university bookstores. The researcher would enter whatever data was important in the middle and then using a self-created methodology, cut notches around the edge (there was a special notching cutting punch available) to indicate subject categories and other metadata.

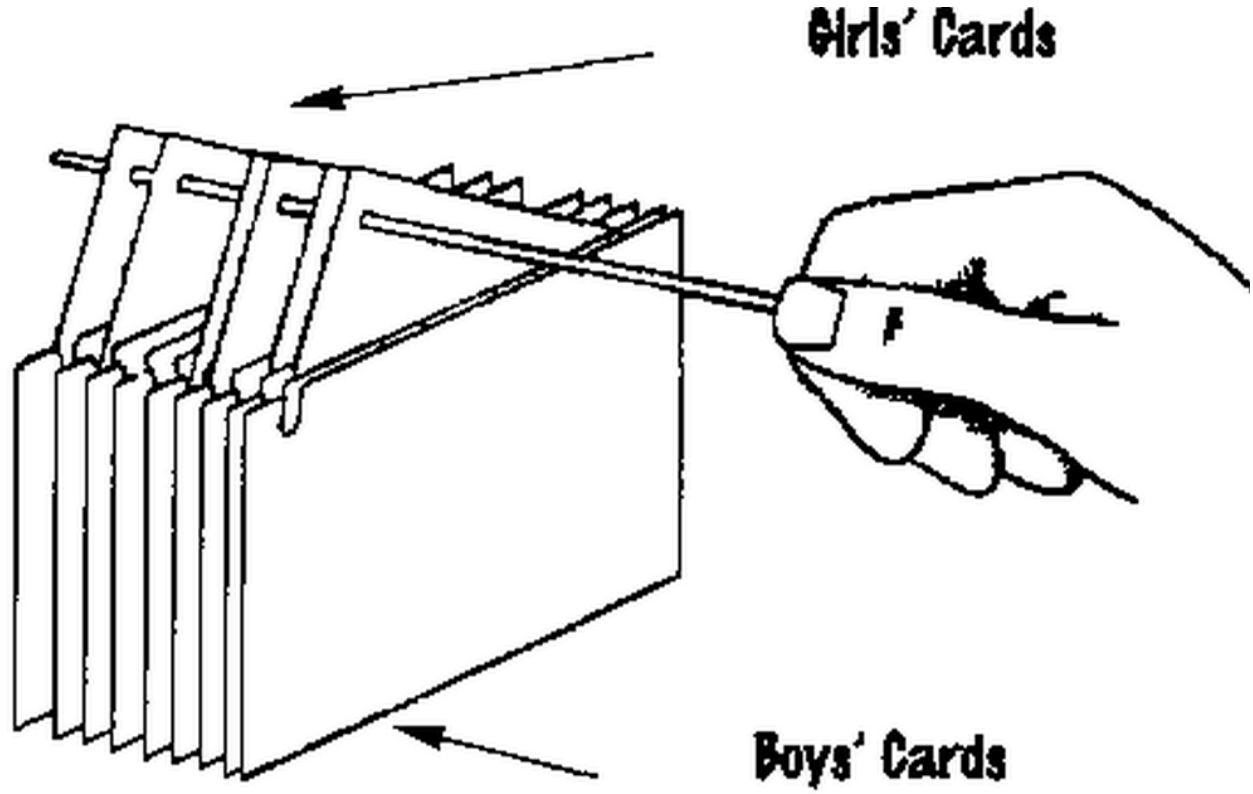

Figure 2: Rod sorting of cards

By using a rod and selecting the notch for a particular subject, notes relevant to that category could be extracted. Refinement of the category, or selection of type of annotation within the results could be achieved by using the rod again on each subset of cards. This resulted in a handful of cards that would be topically relevant for the desired section of a research paper or dissertation.

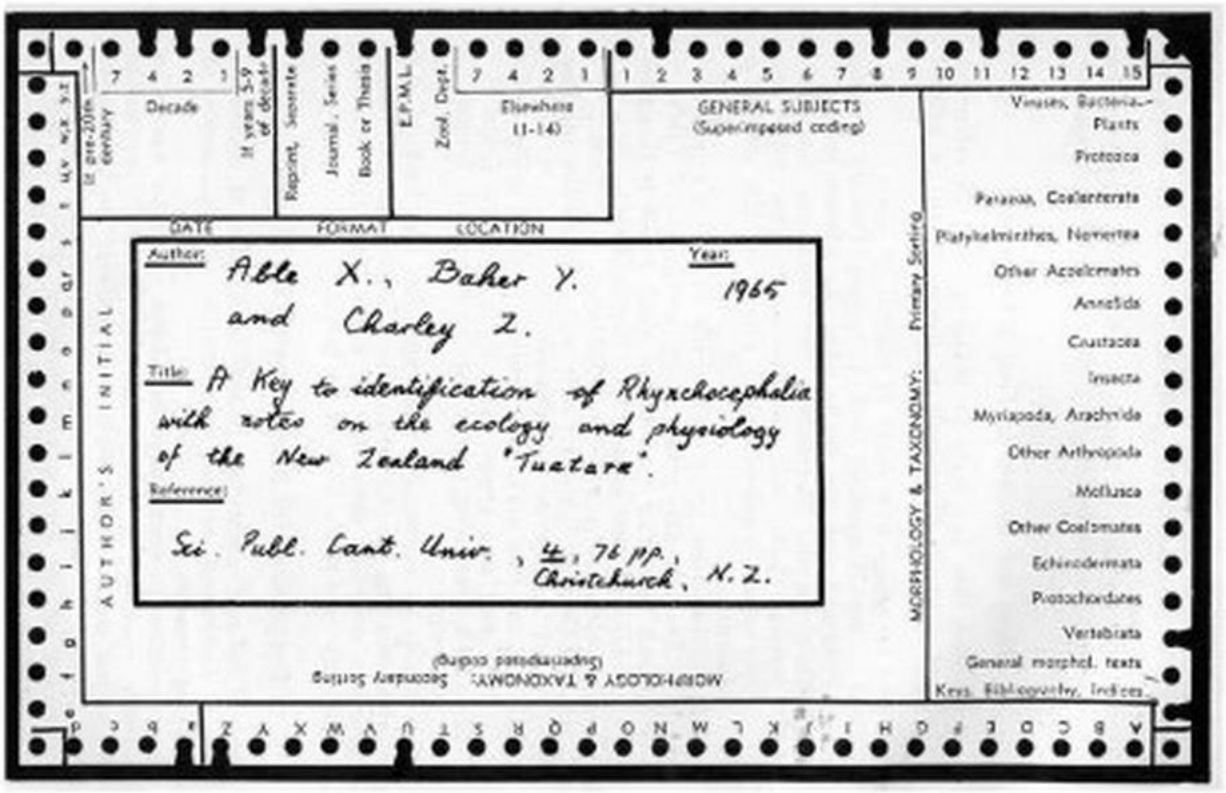

Figure 3: Research card with designated metadata

Customized cards were also available with some of the metadata schema pre-encoded on the face of the cards.

Using a modern electronic database, Andrew was able to record his information in a more flexible format. To extract information, he created queries that resulted in sorted spreadsheets of data. He would then painstakingly go through the many rows of data to find pieces necessary for his dissertation. While laborious, it allowed him quicker and more refined access to his research than other methods.

This system has some drawbacks. It is designed for a single user with the database installed on a single computer. It was not easily scalable for larger runs of newspapers or other primary source documents. The procedure to extract and refine information from the database into spreadsheets and then find the relevant references was multi-stepped with each procedure initiated by hand.

But, there were advantages. Each newspaper article was analyzed only once. Itdidn't need to be revisited for each new facet of the research process or dissertation section. It was easy to sort and filter by subject, people in the articles, geographic area covered, time frame, etc. The results of a search produced everything that had been recorded: researcher's comments, direct quotes, and full names of the individuals. This made it easy to write the final dissertation.

Environmental Scan

One of the first things we did was review software tools from the digital humanity and other communities that were already available. While there are many existing data collection and analysis tools they don't have all the features needed. Andrew said, "They were missing the keyword and content angle that would be useful for historians."

Design Goals

Andrew wanted to provide a system similar to what he had used, but it had to be easier to use. "Easier than Wikipedia" he said. He has a dream of teaching this research method to his students and creating a global database that researchers around the world can contribute to and use for their research. This meant there must also be a method to track who added each annotation or comment to the database and when. Even if a great historian creates a note, it may not be right for what you need, so each researcher should be able to add their own comments to an article. This in turn could become a rich source of marginalia for additional research. Eventually the system may add an up-voting process where higher quality annotators and annotations are promoted.

For teaching, comparing the comments from different individuals can show how varied people can be in their interpretation of the same article and which elements they found most important.

The dream for SourceNotes has more design constraints and requirements than his original process. While Andrew has a vision for the project, he recognized that he didn't have the requisite programming and database design skills, so he reached out to others.

He also wanted the system to be "flexible enough for professional historians so they can use the system to develop their notes and then after publication make them available to others.", instead of the current system where "historians write notes that they donate to archives after they die." There needs to be a method for researchers to hide their annotations from others before publication and then flip a switch to make the source of their article available to everyone.

Andrew had ideas about the interface layout that arose during his dissertation writing process. The initial hand drawn interface pages were developed into prototype interfaces. Subsequent software development and user feedback will further refine the interfaces and work flow.

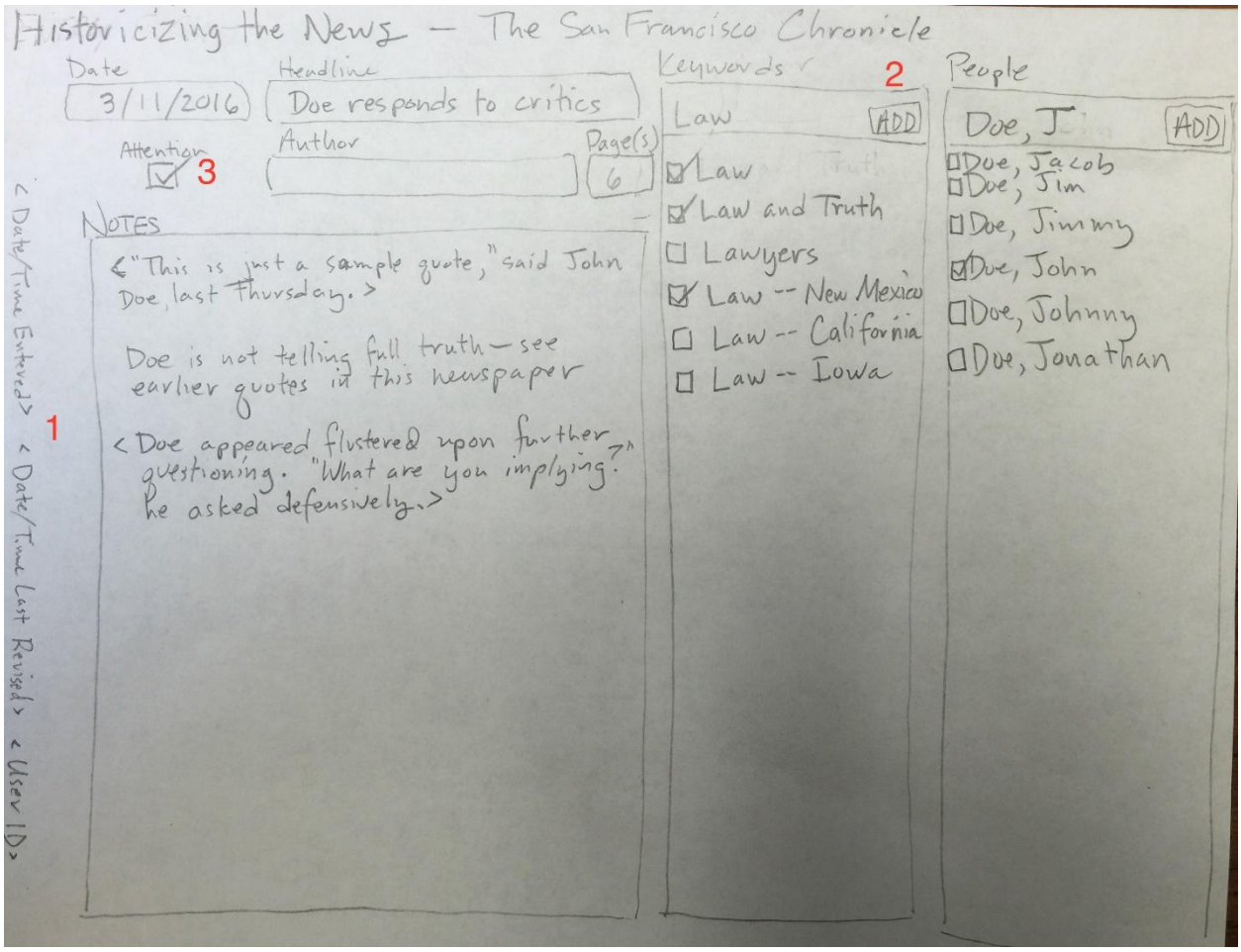

Figure 4: Initial sketch of user interface

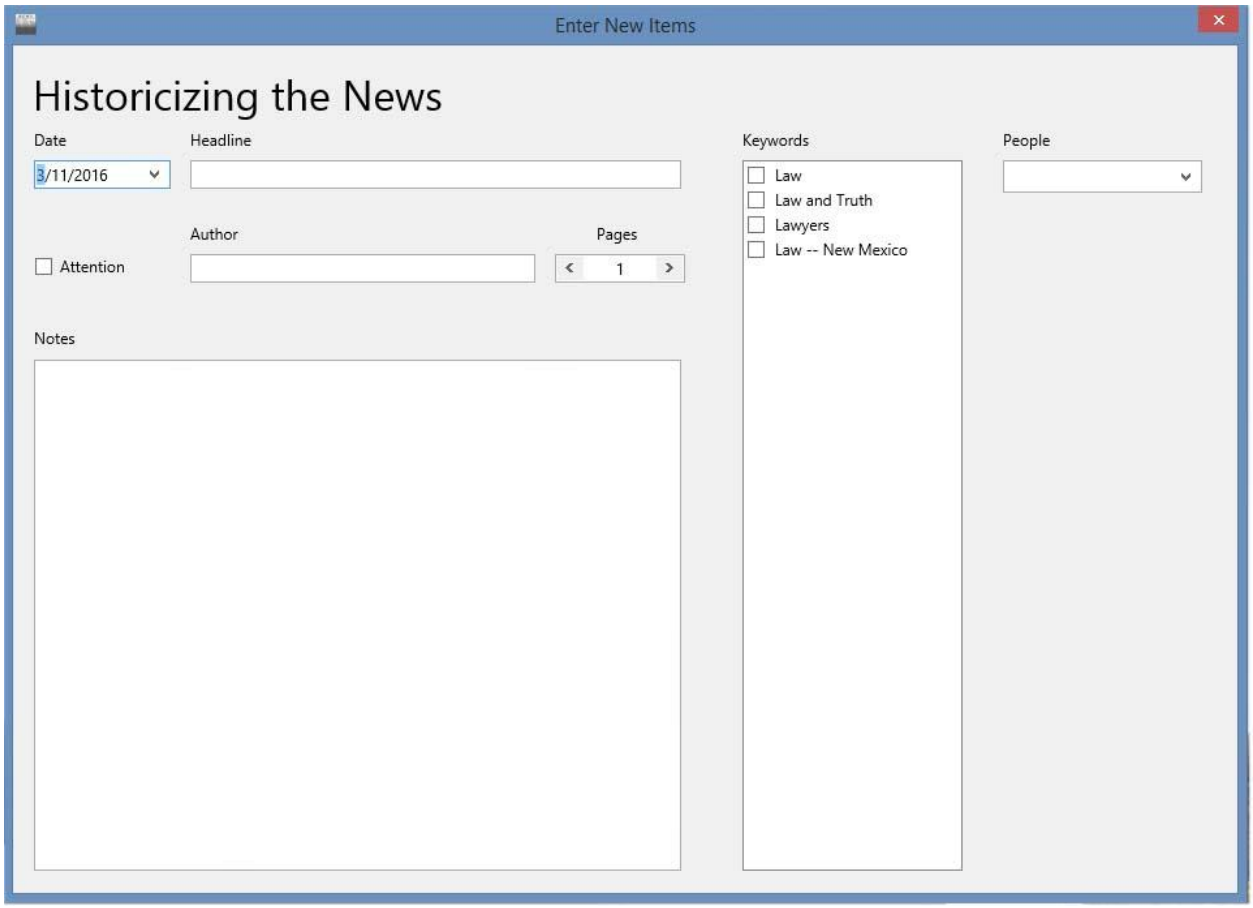

Figure 5: Current data collection user interface

We also showed Andrew a “Style Viewer” so that he could try out different layout styles to see how each of the controls would appear and pick that which suited his purpose best.

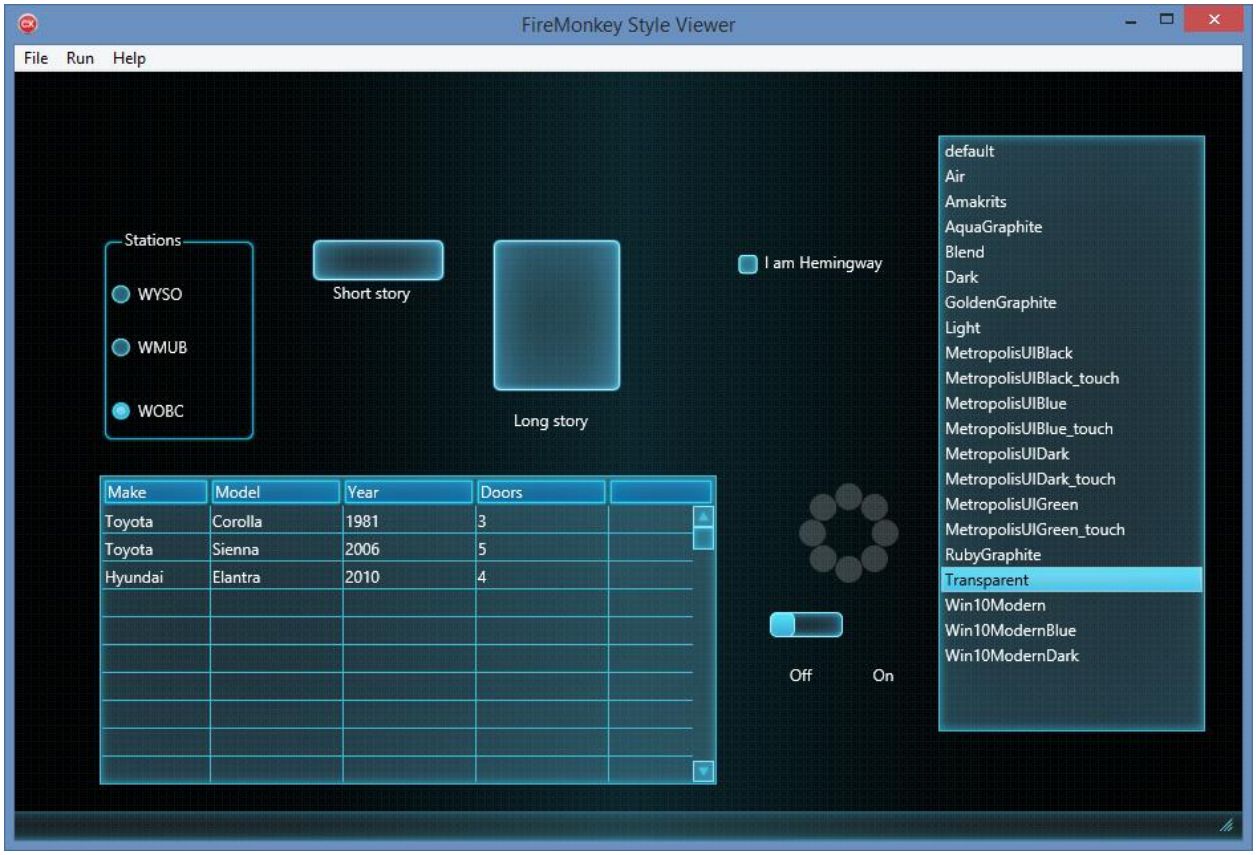

Figure 6: Style viewer

Work on the interface helped inform the database metadata design and focus the project's goals. By stepping through what we wanted to do, the data we needed to collect and the methods needed, we were able to come up with a database design that meets our needs.

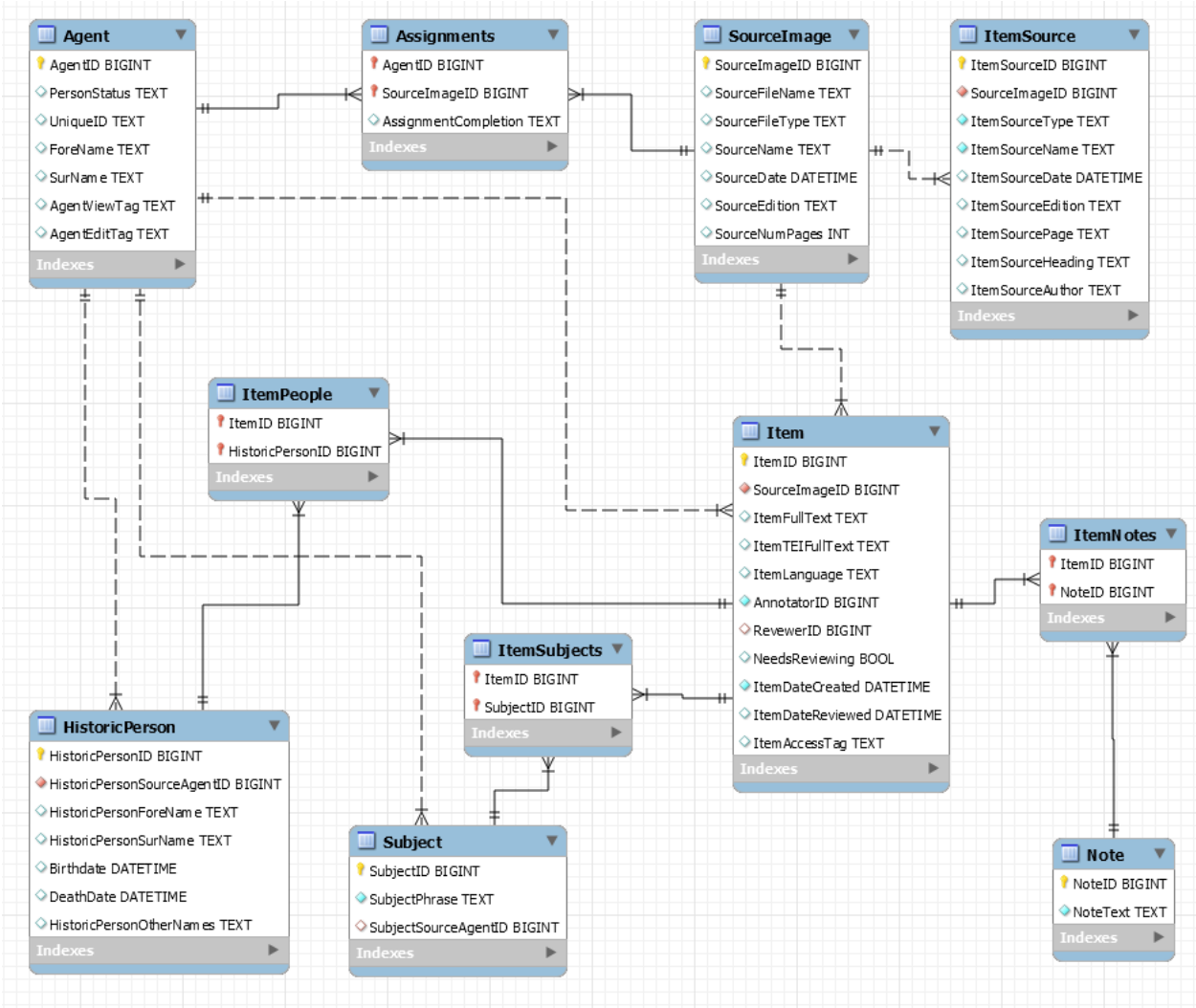

Figure 7: Database relationship diagram

One of the goals is for this project to enable students to receive critical feedback on their synthesis and analysis of historical sources. As part of a 400 level (senior undergraduate to entering graduate student) Fall 2016 class, "Historicizing the News", students will synthesize a five- or ten-year span of a particular newspaper. The entire class will work on the same title, with each student responsible for a few months of issues. Most history classes give feedback on a student's research abilities only through their grade on a final paper. With the use of a team-built database covering a single source, the instructor will be able to critique each student's handling of evidence as it is entered into the database. New questions can be asked related to the student's research skills such as did the student correctly synthesize the newspaper article, identify the appropriate keywords, and tag the correct people associated with the news item?

This educational process will foster an environment of collaboration among students, who will work as a group to analyze a single historical newspaper. Once the students create this collective data source, they will then write individual papers, using the aggregate data as a foundation for their argument.

The research model will then be expanded to include multiple classes and a variety of resources - not just newspapers but also archives, novels and born digital works. Eventually, we will apply for NEH (National Endowment for the Humanities) or other grant funding to build an open web interface and research database for collections such as the Library of Congress's Chronicling America collection.

We recognized that the system needed to be expanded to handle the unique requirements of students. They would be assigned articles or newspaper runs to process, so there needed to be an assignment queue that would track user's progress. Each student wanted to be able to access their finished work as well as see some example works while not having their work shared with the rest of the class until they were ready to write their research papers. The instructor is given access to all the student's data so that guiding comments can be given in a timely manner.

Since multiple researchers may eventually annotate the same article, the database tracks who was responsible for each entry. This allows the instructor to sift for each particular student.

Controlled Vocabulary

An important component for finding the correct article and comments when writing a paper is good subject metadata. Each researcher can add multiple subject tags to any article. But, what happens when one annotator uses the word "child" and another uses "kids" or "children"? The subsequent researcher must know and search for every possible term. A solution for this is to use a controlled vocabulary. This is a list of words that are used in preference to synonyms to describe particular topics.

A comprehensive lexicon of compound subject headings has been created over decades of work and is constantly being updated by the Library of Congress. This is available either as complete subject headings or the simplified LC-FAST (Faceted Application of Subject Terminology) dataset. Researchers begin by entering terms into the subject field in the SourceNotes application. Any subject headings that relate to the set of terms entered is displayed in a list format. The researcher can then select a controlled subject heading to apply it to the article currently being annotated. An unlimited number of subject tags can be applied to a single article. Users can also enter their own terms if an appropriate term can't be found in the controlled vocabulary. This will be searched along with the controlled vocabulary when retrieving articles but will have the limitations and flexibility that come along with non-controlled vocabulary keywords.

System Requirements

With multiple users, the requirement of being able to track each user's entries and temporarily restrict access to source comments for yet-to-be-published works, the system must have a login method. It must also have tracking or restriction for each person's actions based on the level of authority that is granted to each person.

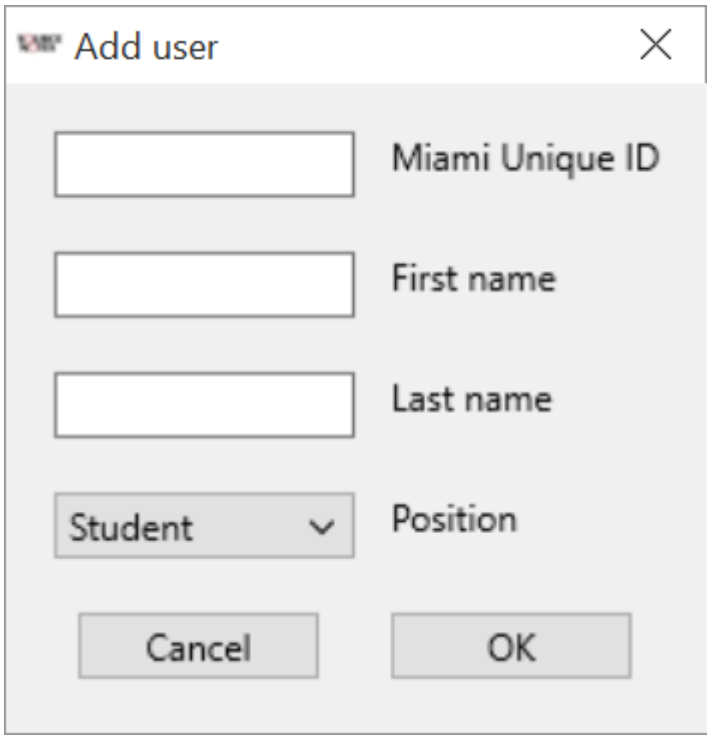

Figure 8: Login in screen

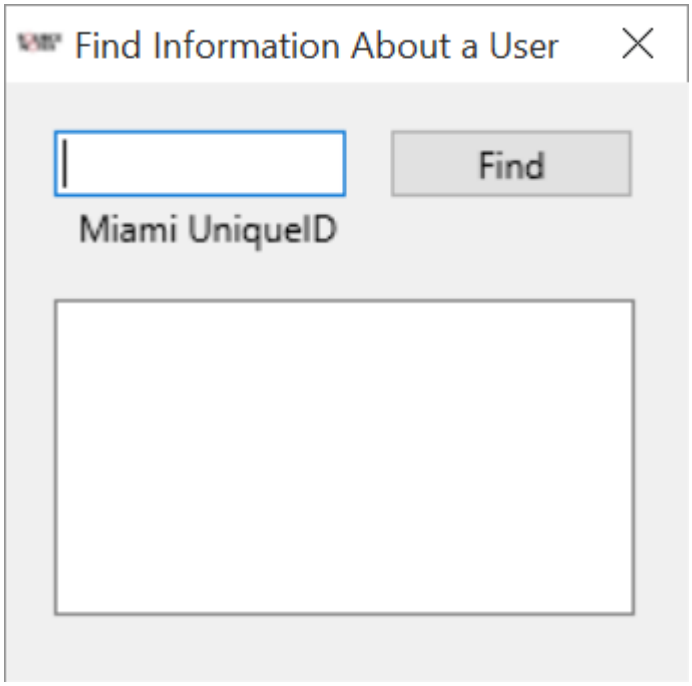

Figure 9: User information screen

Data Entry

After login and article selection, the system can display the image and full text of each article.

Eventually, there will be section where new article images can be uploaded along with their full text.

The interface has a section for free text comments. This is where notes and commentary about each article are entered. Verbatim quotes can also be recorded. Keywords or ideas that are "too wieldy" for tagging with a subject keyword but important enough to capture can be recorded in the comments field. When looking for articles, the search routine will index and search the full text of the comment field.

Ample metadata is collected about each article. This includes the title and date of the newspaper, the edition (some are published morning and night), page of the article, title of the article, author of the article, etc. Born digital content has slightly different metadata including a source url. Each different type of original content can be differentiated as needed.

Articles may mention individuals. These are recorded and linked between articles. To help disambiguate individuals with the same name, birth and death dates as well as other pertinent information can be recorded about each individual. As a researcher goes through the process of annotating a run of the same newspaper, the same individual may reappear with different forms of their name. Mayor Smith may also be known as Mayoral Candidate Jim Smith and later as the late James Oris Smith. Women often change their family name during marriage. Having the contextual knowledge acquired by reading consecutive or related articles together allows researchers to connect references together. This also allows for richer annotation commentaries. Once the different forms of each person's name are connected, a search for that person will return all articles that refer to the person no matter what name form was used.

After a set of data has been entered for several articles, there is a screen that allows the researcher (or teacher) to review the data entered.

Data selection

After all the information has been entered into the database, researchers need to be able to retrieve answers to targeted queries. There is an advance search interface which searches by subject keyword(s), individuals and full text of the annotations. Boolean terms OR and AND are available. The results are displayed in a row-by-row interface similar to the data entry review screen. Columns can be sorted and filtered to show subsets of the returned data.

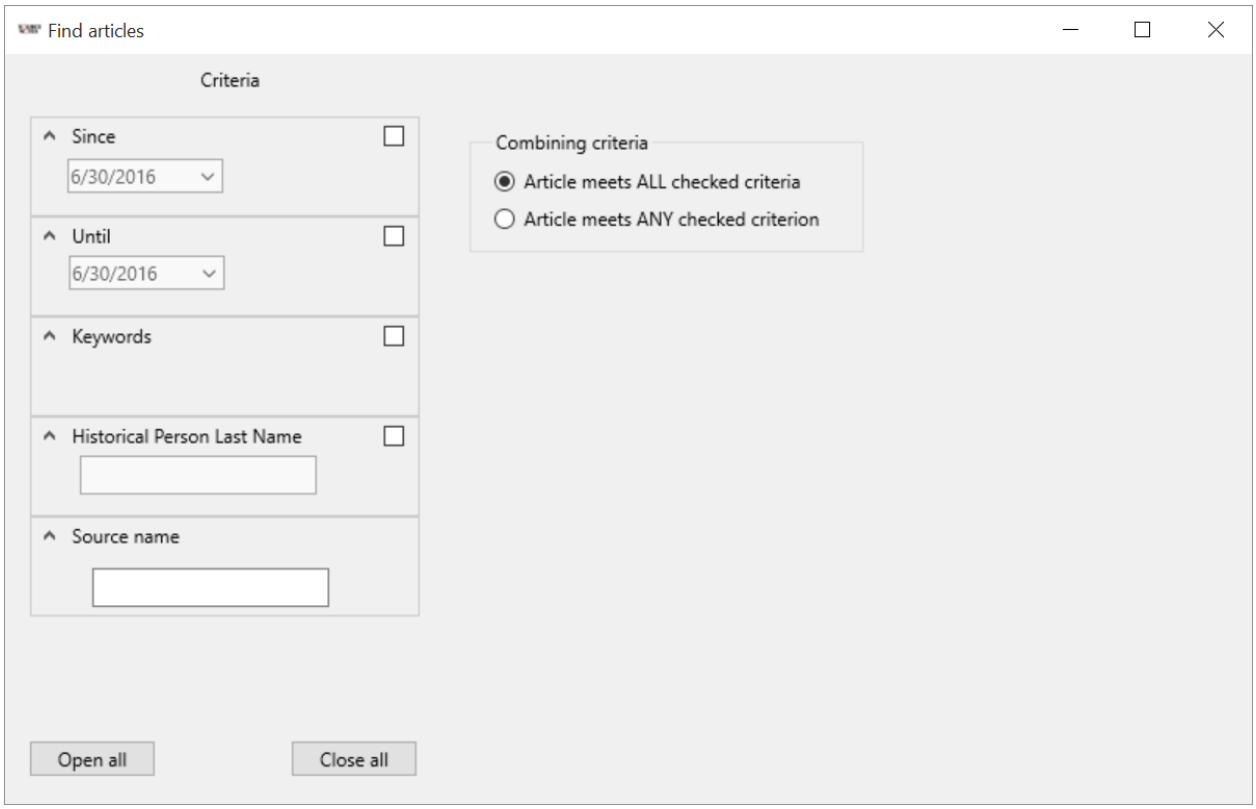

Figure 10: Search interface

Individual rows can be selected to be displayed in a page-by-page interface which allows easier reading of long comments. The page-by-page article interface can also display the full text of the article or the page image so the researcher can see contextually where in the newspaper the article appeared.

Further development and snags

As we test out the system, we expect to incorporate changes to the visual layout, theme colors and logo.

One of the development problems we faced was with database access. While the software worked fine on our development database, when we tried to connect to the production database we were getting errors. The software would indicate it had connected, but no data would acctually go through the pipe to or from the database. It turned out that campus I.T. had decided to block the standard database port 3306. This meant that we had to write SSH routines into our software to overcome this hurdle.

As we move to a multi-campus model, we would like a method of interacting with the system which doesn't require the installation of a piece of software. This means the whole system will have a web interface which will allow any researcher to use the system from any web browser capable system. This will also remove problems that are caused by changing campus firewall policies since the very popular ports 80 (HTTP) and 443 (HTTPS) will remain open.

Academic crowdsourcing by citizen humanistscan help populate the database. The more people that are involved, the deeper the coverage. Training modules on the website will

provided guidance for faculty or grad students wanting to get started with the system for their own research. Additonal modules can provide guidance for teachers wanting to use the system for their own classes.

This fall will be our first multi-user test with a class of students testing the system and providing feedback. Next spring will be a redevelopment based on that and other feedback.

Acknowledgments

Thanks to the Miami University Libraries' Center for Digital Scholarship, Miami University Research and Computing Support and Andrew Offenburger

References

Figure 7: Generic research punch card
http://www.cbs.knaw.nl/publications/1016/content_files/image007.jpg

Figure 8: Rod sorting of cards
https://lh6.googleusercontent.com/ftbHk70vR8mWVIBaBD05LucbIjp85tzrvd56KGbVySYCKkdN7bvGkz9u5ZVfMSwB4E6gtjcUBr_U-c4BEVJjS0u7-kJ_hta_qwnsp2kmoN5h8EqsEXaFZMucU1zqN4roNrE

Figure 9: Research card with designated metadata
https://lh5.googleusercontent.com/JlQRawXlMZhN-TwUZNFR54BB42Voa7bX17YgCGx4_ms9mXXpnVJkyw4GdAcdJ-8r3AKPgkWKYAWi6-OpWjiW71Iabw1mOiPE4VeV-v7naZXQlbHUBCo31WzUNaiOWHZ8TN0