CLIR Essay 1 Prompt:

Briefly describe a recent research project you undertook, then explain how you organized and managed the evidence that you used to support your research. Evidence might include library, museum or archival resources; data you collected; or data produced by others. If anyone helped with your research and/or organizational strategies, please explain how they helped. In retrospect, how would you have organized and managed your evidence differently? What kinds of outside advice and training might have enhanced your project?

Essay 1: Data Management
Hannah Alpert-Abrams

The *Reading the First Books* project is a two-year effort to develop tools for the automatic transcription of early modern printed books. The project originated as a collaboration between myself and two computational linguists working to develop an automatic transcription (OCR) tool called Ocular. Together we modified the tool to handle the unique orthography and multilinguality of books from the early modern period.

As a student of comparative literature, I have systems in place for organizing my own research, mostly in the form of notes on secondary sources. But developing Ocular required an entirely different way of thinking about information. We collected language data in the form of previously transcribed documents for every language that we work with --- currently seven, and counting. We collected scanned pages to use for training and testing. We created gold-standard transcriptions of test pages against which we measured our results. And we collected results, in the form of automatically-produced page transcriptions and numerical evaluations.

The computational linguists built tables for organizing our numerical results. But keeping track of the hundreds of .jpg and .txt files has been my responsibility. Because the project developed organically and exponentially, keeping track of the various kinds of information didn't seem like it would be a significant problem until, all of a sudden, it was. Over time, I learned to maintain a permanent database structure that distinguished between files that needed processing, files that were available for testing, and temporary output files. This made it easy to reuse test files and to locate and discard temporary files. After accidentally duplicating work several times, I wrote a csv file to keep track of the files we had, their location on my hard drive, their status, and their source. After hiring an assistant in the second year of the project, we worked together to restructure the data in a way that could be understood and used by outside project participants.

At stake in our data management plan were scientific consistency and efficiency. Scientific research depends on regular, replicable results, but these can only be achieved when data remains consistent across experiments. This is fundamentally different from literary study, which prioritizes close reading and subjective interpretation over shared experimental results. In addition, consistent data makes it easier to extend experiments and conduct research on new features without building new data sets each time a change is implemented. Indeed, we ultimately made our test data available on GitHub so that other researchers could draw on our data to replicate --- or improve on --- our results.

To develop the organizational systems that our project uses, I worked primarily with the computational linguists who ran experiments on our data. I also spoke with the digital humanities coordinator associated with our project about more complex approaches to data management, like MySQL or OpenRefine, but we found that the learning curve was too steep for the relatively simple kinds of data that we were working with. With the benefits of hindsight, I can see the advantages that the early implementation of a data organization system would have had for the long-term organizational consistency of the project; were I to repeat a project of this nature, I would be sure to use a scalable system that could handle project expansion.

The first phase of the *Reading the First Books* project was experimental. In the second phase, we took our work and integrated it into a much larger, pre-existing project hosted by a library at another institution. At this point, the simple data management methods that we were using were no longer sufficient. Fortunately, our collaborators had years of experience building

database structures to manage precisely the kind of data that our project depended on. With the help of their team, we were able to make a relatively easy transition into larger-scale production, though we did have to make special accommodations for some of our project's more unique features, like special characters and multiple languages. We have also been able to take advantage of management and preservation specialists in the libraries in order to plan for the long-term preservation of our workflow, documentation, and results.

I was fortunate to be able to depend on librarians with specific knowledge about data management and preservation throughout the *Reading the First Books* project. As a CLIR fellow, I hope to learn more about the systems and procedures for managing data, especially in projects based on large corpora or more complex datasets. This will better situate me to propose large-scale projects, support faculty proposals, educate students and faculty, and see projects to completion.

CLIR Essay 2 Prompt:

Describe some ways that research methodologies and/or the dissemination of scholarship in your field have changed in the past 25 years. What factors prompted these changes? How do you think libraries, cultural heritage institutions, publishers, and/or universities should respond to these changes in order to support the advancement of knowledge in your field?

Essay 2: Changing Methodologies and the Advancement of Knowledge
Hannah Alpert-Abrams

When reading a scholarly book or article in a printed edition, I have to fight the urge to command-f my way to key sections, to copy-paste a key passage, to highlight or filter or email the results. The computational processes enabled by the shift to digital publication have changed, in subtle but fundamental ways, the approach that scholars across disciplines take to research in both primary and secondary sources. Students of my generation expect information to be discoverable, in the modern sense; to be delivered, without much labor, to our screen.

While the digitization of scholarly research and the establishment of searchable catalogues, online repositories, and open-access publications are relatively recent inventions, textual replication, circulation, and organization have shaped research for a very long time. In my dissertation, I write about the history of textual replication, and the ways that it has shaped the legibility of historical documents. In the nineteenth century, for example, printed books and manuscripts were copied by hand and circulated among historians across nations and oceans. Copying increased access to the historical record, but it also changed the shape of history: by improving handwriting and modernizing spelling, scribes civilized the past. (In other cases, a poor copy could mark the downfall of a civilization.) Access to history was mediated by the form of its re-inscription. This has not changed in the case of digitization, where the imperfections of automatic transcription (OCR), and the interpretive frameworks of xml, html, and Dublin Core shape the legibility of a scholarly resource.

Like legibility, accessibility was and remains fundamental to the circulation of information, but universal accessibility has never been the only priority for collectors of information. In my dissertation I describe how private collections have long been kept behind closed doors, restricted to individuals with particular kinds of social status. Public collections, too, have used complex and costly entry requirements or limited hours of operation to keep people away. Limiting discoverability by using obscure cataloguing systems or maintaining minimal records has served a similar function in controlling the kinds of people that have access to historical information. This remains true today. The libraries that I have visited for my dissertation research often require a letter of recommendation and an entrance fee. (Other kinds of collections, like the parochial archives in Mexico, may restrict access to foreigners altogether.) While this may seem like an effort to maintain elitist control over information, it can serve other purposes. In one case that I describe in my dissertation, the leaders of an indigenous community in Mexico restrict access to a historical map as a means of resisting municipal efforts to take control over land that has long been part of indigenous cultural practice.

My research shows that while the changes wrought by digitization seem radical, the problems they pose have long histories. The liberating potential of digitization, then, is found in its ability to expand the legibility, access, and discoverability of information, breaking down the historical barriers put in place by elite individuals and institutions. This certainly feels like freedom to researchers based at academic institutions in the United States. Many of the historical texts I study are available as PDFs online, through Google Books, Hathi Trust, or other repositories. In my own work, I have helped to build tools that increase the discoverability of these documents, and I am proud of my work to make it easier to connect students and researchers with historical texts. Furthermore, with access to paywalls provided by libraries, I can read more secondary sources than ever before. If scholars publish in open-access journals, then their work, too, can be made free.

But legibility, accessibility, and discoverability are not the only priorities for the caretakers of information. As a student of Latin American history and literature, I have found that the work of my international colleagues often fails to rise to the top of search engines and library catalogues; the Anglophone bias of search engines, combined with the limited financial resources of Latin American journals, conspire to push these works to the margins. It is the responsibility of developers and librarians to work against these biases by creating spaces for underrepresented texts. Through involvement in the development of search engines, metadata schemas, and other information structures, librarians can ensure that access to information is not dictated by cultural prejudices. I have prioritized this in my own work of developing tools to transcribe historical documents by focusing on the transcription of historically marginalized languages and ways of writing.

At the same time, librarians can take a leading role in protecting information that should not be free. We have seen the importance of information privacy in cases of identity theft and other breaches of privacy. In the realm of historical research, these concerns come to the fore when documents have particular cultural or religious value or are politically sensitive. As in the case of the indigenous Mexican map described above, preventing access can be an important task for the caretakers of documents and artifacts. This is true in the case of sacred objects and in the case of politically sensitive documents, as Kim Christen Withey and Kirsten Weld have argued, respectively. In these cases, justice is facilitated by balancing increased access with increased protection for valuable documents. Again, libraries and other institutions can lead the way in making ethical decisions about database structures, metadata, and online access that take seriously the risks posed by both keeping this information secret, and making it discoverable online.

Digitization has changed the legibility, accessibility, and discoverability of historical documents, posing new challenges and creating new opportunities for caretakers of historical information. Advancement of knowledge, in this context, requires not only increasing access to information, but also developing structures that are more subtle and responsive to the conditions of individual collections or forms of information.