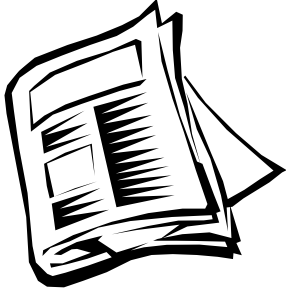


Digitization: What does it mean?

Digitizing the bound volumes means creating a new archive of *The Cavalier Daily* and *College Topics* in a digital format that can be stored, duplicated, distributed and viewed using personal computers and the Internet.



There are two options:

1.) A simple archive can be browsed by date. This is done by digitally “photographing” all the newspapers in the archive.

2.) A more complex archive is searchable based on names or words that appear in headlines, bylines or the stories themselves. To do this, the information on the pages must be converted to a text format, i.e. digitally “reading” all the newspapers in the archive.

How do we do it?

There are two ways to reach the end goal of a searchable archive. One way is to physically retype all of the articles and headlines, and possibly supplement this with scanned graphics and photography.

The CDAA investigated this approach and, while it is technologically feasible right now, it is so labor-intensive that the cost is too high.

An alternate method is to convert the pages in the bound volumes or on a set of microforms into scanned digital images, essentially taking a photograph of the entire newspaper page including the stories, headlines and graphic elements.

The scanning process can be largely automated, especially if we start from microform reels. This will produce an enormous archive of digital photographs that can be catalogued and displayed on a computer screen, as well as stored on compact discs. Before we scan, we will need to ensure that the existing microforms created by Alderman Library are all of good quality. Any issues that were poorly photographed in the past need to be re-transferred to microform.

To create a searchable archive, Optical Character Recognition (OCR) software can then be used to “read” the stories contained in the digital images. This will be a challenging task, as the individual elements for each story need to be grouped together (such as linking a story to its jump and its graphic elements). In addition, older OCR packages have difficulty with unusual fonts and poor contrast between the page and text.

However, the quality of OCR software is improving dramatically every year as it becomes more complex. The CDAA envisions completing this step with a future OCR package, though some demonstrations of the potential can be performed using today’s software.

How much will this process cost?

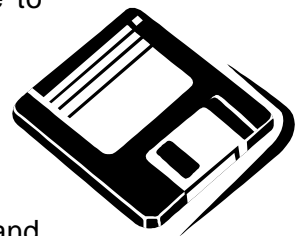
Estimating the cost accurately is difficult to do at this stage, which is why the CDAA will create a demonstration project, such as digitizing a portion of the archives, as a test run. The cost for a demonstration project will be several thousand dollars.

Based on current estimates, a complete browsable archive may cost \$50,000 to generate, though that figure could be off by a factor of two or more based on what we learn. The cost of converting a browsable archive to a searchable one is strongly dependent on the quality of the OCR package, and would become less expensive as better packages are developed.

This project can be attacked incrementally, with the costs spread out over a long period of time.

How will the project be financed?

The CDAA is currently in the information-gathering stage. However, once starting funds are generated we can then create a demonstration project. That experience will allow us to be realistic in our future goals, and provide us with the knowledge to write grant applications for other sources of funding. We are currently coordinating with Alderman Library to do this demonstration, and we plan to enlist the library’s support in pursuing grants to fund this project.



**Why preserving
The Cavalier Daily's
bound volumes in a
digital format is so important.**

1.

Paper and film archives deteriorate over time, while a digital archive could be duplicated without error in perpetuity.

2.

Accessing back issues of the newspaper is difficult and time-consuming unless you know the exact date an article appeared. Because of this difficulty, the back issues are almost completely ignored. Making them accessible would allow Cavalier Daily reporters to have the full history of any given topic at their fingertips and, subsequently, conduct better research.

3.

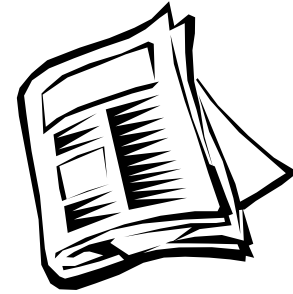
The pages of *The Cavalier Daily* and *College Topics* are a significant record of the history of the University over the past century. Making this history public will raise awareness and appreciation of the newspaper within the University community.

How you can help:

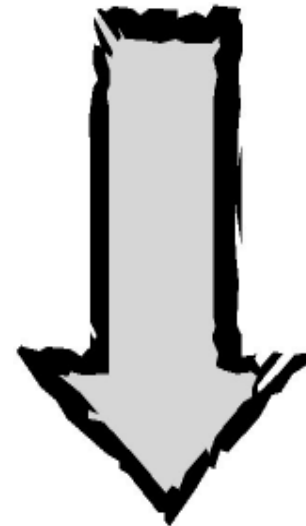
The Cavalier Daily Alumni Association relies on donations from its members for all its projects. Completion of this venture is no different.

Tax-deductible donations can be sent to:

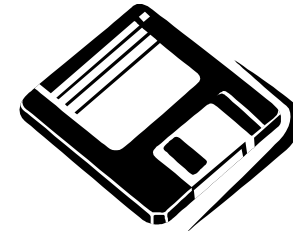
The Cavalier Daily Alumni Association
P.O. Box 4731
Charlottesville, Virginia 22905



dig·it·ize:



v. 1. the
process of
converting a
document to
electronic
format



An Explanation of
The Cavalier Daily Alumni Association's
Archive Digitization Project