1983

Optical Scanning and CINDEX: Tools for Creating a Cumulative Index to the Laurens Papers

David R. Chesnutt
University of South Carolina

Optical Scanning and CINDEX: Tools for Creating a Cumulative Index to the Laurens Papers

DAVID R. CHESNUTT

When the Papers of Henry Laurens project published its first volume in 1968, computers were still very much the domain of our scientific colleagues. Only a handful of humanists like Wilhelm Ott in Germany and Eric Boem in the United States had begun to realize the computer's potential for eliminating some of the drudgery associated with publishing scholarly materials. Today, computer assistance has become the sine qua non for almost every large-scale edition. Older editions like the Laurens Papers have had to automate their procedures gradually—moving step-by-step to replace traditional methods with computer-assisted methods.

The Laurens Papers began that process in 1975 and the process continues today. The first task identified for computer assistance back in 1975 was the making of single-volume indexes and ultimately the ability to create a cumulative index. By the end of 1976, a computer-assisted indexing system had become a reality. That indexing system was given the name CINDEX—an acronym for Cumulative INDEX. Although the system could produce only single-volume indexes, the acronym was chosen to reflect the project's objective of creating a multi-volume index.

CINDEX was a major step forward because it made the creation of the next four single-volume indexes much faster and improved indexing accuracy. During that same period refinement of the CINDEX programs continued, and in 1981, CINDEX reached the point of being able to merge single-volume indexes and to produce from that merge the project's first cumulative index. At that point the cumulative index included only Volumes 6–9; access to Volumes 1–5 was still limited to the individual indexes for those volumes. As we found the cumulative index more and more useful in preparing the next volume, we turned our attention to the question of integrating those first five indexes into a full-scale cumulative index. Several solutions to the problem seemed feasible.

The most straight-forward method of converting the old indexes was to have someone retype the old indexes at a computer terminal. The old indexes totaled about 200 printed pages or something like 600,000 characters. Estimated entry time was about 150 hours, or approximately $1500. A second method (and the method ultimately chosen) was to capture the old indexes in machine-readable form through optical scanning of the printed text. Although the initial cost estimate for scanning was $1200, the actual cost proved to be $384.

The optical scanning of the printed indexes was handled by a commercial service bureau using a Kurzweil data entry system. The Kurzweil is one of the more sophisticated scanners in use today because of its ability to "learn" almost any printed character font. For modern printed materials, the learning time on a Kurzweil often is less than 30 minutes when done by a skilled operator.

The decision to scan the indexes rested not only on cost, but on accuracy as well. The first utilization of optical scanning for the Laurens project had taken place in 1980 when about 6,000 pages of typescript were scanned by a commercial service bureau. In that first experience the accuracy rate of 75% had been cost effective, but had created more work for the staff than initially anticipated. When the Kurzweil scanner subsequently became available, tests on recently published books provided a recognition rate of better than 99% accuracy. With that kind of accuracy rate optical scanning became very attractive. Let me emphasize, however, that the 99% accuracy rate applies only to modern printed materials; our tests with nineteenth-century books and newspapers have been very disappointing.

The final factor which led to the decision to optically scan the indexes was the ability to reformat them so that they could be processed by CINDEX. The critical issue was whether or not the reformatting could be handled by computer processing once the machine-readable index files had been created. A careful analysis of the old indexes revealed that a computer program could be devised to reformat the indexes for CINDEX. (The Laurens project is somewhat unique in having a full-time programmer as a member of its staff. The staff programmer not only provides support for computer applications at the Laurens project, but for other projects which use CINDEX.) Once the general assessment had been completed, the work began.

The procedures used in creating the cumulative index for the first nine volumes can be broken down as follow:

1. Development of a test program to convert the scanned indexes into files which could be read by CINDEX.

2. Development of scanning specifications to retain characteristics which allowed the conversion program to format the indexes for CINDEX.
3. Scanning of the old indexes by a service bureau.
4. Proofreading of the scanned files.
5. Development of the final conversion program.
6. Processing of the scanned files and merging through CINDEX to produce the cumulative index.

In essence, the approach used here was almost circular. The system design started with the final product and worked backwards to the input. Programs were then written to accomplish the specific tasks. Finally, the work itself was done. Although this may sound complicated, only one new program was required—the conversion program. The actual time involved in creating this new application was less than a day. The system design took about two hours; the new program took about four hours.

The key factor which made it possible to convert the printed indexes into files which could be processed by CINDEX was the regular format of the indexes. In the case of the Laurens indexes, the printed format is a run-on style with a hanging indent. Main entries begin at the left margin and subsequent lines are indented. A semicolon is used to separate a main entry and its subentries; a semicolon is also used as a separator between two subentries. In other words, every subentry is preceded by a semicolon. Thus, main entries could be identified because of their unique position and subentries because they are preceded by unique punctuation. The conversion of the printed files to CINDEX files is perhaps most easily seen by referring to examples.

The process was not without its pitfalls. One was the failure to communicate adequately with the service bureau which scanned the files. Although explicit instructions for retaining font changes from roman to italic were given by telephone, the vender's representative failed to give those instructions to the person who operated the Kurzweil scanner. Thus, text to be set in italics had to be marked during the proofreading stage. Nor did the contractor understand that spacing used in the printed version was to be retained. The original files returned to the project made it impossible to determine when a main entry began. Fortunately, the contractor still had the original scanner files on hand and was able to easily rectify that mistake. Obviously, instructions regarding the scanning of files like these need to be transmitted in writing and then verified orally.

Another problem was the amount of time required for proofreading. An average volume index required about 16 hours to proof, or about 80 hours for the five volume indexes. Had this been anticipated, it would have been better to have considered scanning the indexes twice. Statistically, a machine collation of the same files scanned by different contractors would probably have resulted in a greater accuracy rate than the tandem proofreading we used. (Peter Shillingsburg at the Thackeray edition has demonstrated that machine collation produces better proofreading results than a manual procedure.)

Abatement, 19
Abercrombie, Capt., 154, 161
Accounts, open, 174, 381
Act (English) for extending and improving the trade to Africa, 44n
Adams, Capt. (*Molly*), 262, 264, 265, 323
Adams, Capt. (*Two Brothers*), 169, 171
Adams, James, 87, 91, 213
Adams, William, 238
Addison, Benjamin, 3n, 54n, 60, 65, 88, 104, 196. *See* Laurens & Addison
Administrators, 59n, 241
Admirals, 300, 301, 313, 314
Adventure, 161, 185
Adventure, M.W., 26, 27, 39, 43, 61, 67, 73, 83, 94, 102, 127, 135, 137
Advertisements, 240–243
Africa, 115n, 201n, 202n, 212, 224n, 242, 245, 249, 252, 258, 264n, 271, 288n, 295n, 296n. *See* Angola, Bite, Bonny, Cameroon, Cape Mount, Gambia, Gold Coast, Grain Coast, Guinea, James Fort, Majumba Coast, Malimba, Mindinga country, Sierra Leone, Windward Coast
Africa, 288n, 348

Fig. 1. First page of printed index in Laurens edition. Main entries begin in Column 1; subentries are preceded by semicolon.

Although this particular activity was related specifically to the Laurens project, the process is one which can be adapted to other projects. For example, the staff of the Jefferson Papers at Princeton is now in the midst of using CINDEX to merge the indexes for the first twenty volumes of that series. The Jefferson project is more complex, however, because each of the indexes has been extensively revised. The major advantage to both the Laurens and Jefferson projects has been the ability of the computer to provide a correctly sorted cumulative index which can serve as the basis for further editing and refinement. The process of merging files to produce a multi-volume index usually takes less than five minutes for a 10-volume index.

```
Abatement, 19
Abercrombie, Capt., 154, 161
Accounts, Open, 174, 381
Act (English) for extending and
  improving the trade to Africa, 44n
Adams, Capt. (@Molly@), 262, 264, 265,
  323
Adams, Capt. (@Two Brothers@), 169,
  171
Adams, James, 87, 91, 213
Adams, William, 238
Addison, Benjamin, 3n, 54n, 60, 65,
  88, 104, 196; @See also@ Laurens & Addison
Administrators, 59n, 241
Admirals, 300, 301, 313, 314
@Adventure@, 161, 185
@Adventure@, M.W., 26, 27, 39, 43, 61,
  67, 73, 83, 94, 102, 127, 135, 137
Advertisements, 240-243
Africa, 115n, 201n, 202n, 212, 224n,
  242, 245, 249, 252, 258, 264n, 271,
  288n, 295n, 296n; @See also@ Angola, Bite,
  Bonny, Cameroon, Cape Mount,
```

Fig. 2. Scanned file created from first page of printed index in the Laurens edition. 'At' signs indicate italic font.

```
??VOL(1)
Abatement* 19 *
Abercrombie, Capt.* 154, 161 *
Accounts, Open* 174, 381 *
Act (English) for extending and improving the trade to
  Africa* 44n *
Adams, Capt. (@Molly@)* 262, 264, 265, 323 *
Adams, Capt. (@Two Brothers@)* 169, 171 *
Adams, James* 87, 91, 213 *
Adams, William* 238 *
Addison, Benjamin* 3n, 54n, 60, 65, 88, 104, 196*
          @See also@ Laurens & Addison **
Administrators* 59n, 241 *
Admirals* 300, 301, 313, 314 *
@Adventure@* 161, 185 *
@Adventure@, M.W.* 26, 27, 39, 43, 61, 67, 73, 83, 94, 102,
  127, 135, 137 *
Advertisements* 240-243 *
Africa* 115n, 201n, 202n, 212, 224n, 242, 245, 249, 252,
  258, 264n, 271, 288n, 295n, 296n*
          @See also@ Angola, Bite, Bonny, Cameroon, Cape Mount
          Gambia, Gold Coast, Grain Coast, Guinea, James Fort,
          Majumba Coast, Malicoba, Mindinga country, Sierra
          Leone, Windward Coast **
@Africa@* 288n, 348 *
African, coast* 296n*
          ships* 299, 313*
          trade* 14n, 210, 227, 256, 259, 273n *
Agent* 134n, 141n, 170, 266n, 268n, 340, 341n, 356*
          of New Hampshire* 184n*
          of S. C.* 2n, 7n, 13n, 344 *
Ague* 38 *
```

Fig. 3. Cindex input file created from scanned file of printed index. Format is in 'edit' file format, not basic input file format for CINDEX.