Practical Realities of Computer-based Finding Aids: The NARS A-1 Experience

ALAN CALMES

THE NATIONAL ARCHIVES and Records Service (NARS) A-1 System is a computer-assisted procedure for compiling all the series level inventories of the National Archives record groups into one master *file*. Two main purposes of the system are: (1) to provide the Office of the National Archives with administrative control over the record groups (allocation to custodial units and quantity control); and (2) to compile all series descriptions into one *machine-readable* file according to a standard *format* and a hierarchical *addressing* scheme. The NARS A-1 System came about after an in-depth background study between 1972 and 1974 found the NARS administrative and descriptive programs inadequate for controlling archival records.[1] The design of the A-1 system took place in 1974–75, and the system went into production in 1976.[2]

System Study

At first, the NARS program analysts asked whether gaining administrative control over record groups and compiling series descriptions required the creation of a new system. The A-1 system study consisted of an analysis of existing procedures for describing and controlling archival records. Program analysts determined the adequacy of the various procedures, compiled a list of problems, identified types of information of interest to archivists and researchers, and assembled a variety of finding aids. In the process of evaluating existing series descriptions, the program analysts identified strengths and weaknesses of each and outlined methods to bring all the series descriptions together into one system. It was recognized that the format (layout of information and the kinds of information) varied widely from one series de-

[1] Claudine Weiher, "Control and Description of Records in the National Archives," unpublished, NARS–NAA, 4 vols. (1974).

[2] Alan Calmes, "NARS A-1 System Documentation," unpublished, NARS–NNB (15 August 1978).

scription to the next and that a standard format ought to be adopted. In the final analysis, the study envisioned a system which would incorporate desirable new features, such as format control, into the existing system.

The main consideration then became how to assemble and compile the formatted series descriptions into a master file. Subject indexing was considered impractical because a dictionary of terms would have to be developed and applied systematically to all series descriptions. It would make the previously described series descriptions inadequate and would require a considerable amount of additional time for all future series descriptions. Indexing would require that an archivist identify appropriate *index* terms for each series description. This would slow down the decision-making process during series description writing. It was estimated that such a process would triple the series description processing time. The A-1 system was supposed to accelerate the production of series descriptions, not retard their production. Furthermore, the A-1 system was thought of as a means of gaining administrative rather than intellectual control over the records. The analysts recommended that subject retrieval receive serious attention only after the problems of administrative control were solved.

Initially, automation was seen as the second of two alternatives; for at first the problem was conceptualized in terms of manual processing. The analysts estimated the time needed to reproduce old series descriptions according to a standard format and to process new series descriptions. The cost-benefit analysis evaluated the relative worth of manual processing as compared to automated processing and concluded that costs would be comparable. However, this was a moot point since creation of the file would represent about 60 percent of the total cost, and the Archives administration could not hire a large enough number of clerk typists and provide them with enough office space for typing and maintaining an active file of complete inventories of all the record groups. Therefore, to assist the manual operations and reduce the need for additional personnel and office space, NARS management looked toward automation. The A-1 system study, which cost $70,000,[3] concluded that a computer-assisted system mainly for the purpose of text editing—sometimes called "wordprocessing"—should be implemented instead of a computer centered system designed for information retrieval by subject, which would require the use of index terms.

System Design and Implementation

System design of the A-1 system involved two different activities: the preparation of input forms, and computer processing. The following description of the A-1 system will not deal with the preparation of input forms.[4] Computer processing consisted of the conversion of statistical and descriptive data into *machine-readable* form, the manipulation of the machine-readable file, and the production of computer printouts. Because the A-1 system study called for a computer assisted system rather than a computer centered one, *batch processing*

[3] It took four and one-half man-years to complete the system study.

[4] GSA Form 6710A, Change of Status Record, was designed for the collection of series description information. Form design was carried out independently of the system design.

was viewed as the main mode of operation.

During the design stage of the A-1 system, system analysts first considered the purposes and objectives of the computer printouts. NARS management decided that the traditional NARS inventory format should be the main output product. Batch processing of the record group descriptions in inventory format required the machine-readable file to be arranged sequentially according to a hierarchical numbering scheme (an addressable control number) which reflected the placement of archival series within the originating agency's organizational and/or functional structure. The addressable control number was a combination of record group number, subrecord group numbers (eight possible layers), series number, subseries number, and sub-subseries number.

In Figure 1, the term "Level" identifies the depth of the hierarchically located *record.* Level 1 represents the record group number level. A level 1 record contains information describing the record group as a whole; for example, an organizational history. Level 9 is the subgroup which is most embedded in the record group's hierarchical structure. Subgroups thus represent such organizational structures as "office," "division," and "branch." Levels 10, 11, and 12 are series, subseries, and sub-subseries, respectively; they may be immediately associated under a record group as a whole (level 1) or with any subgroup (levels 2–9). Figure 2 illustrates the operation of the control number illustrated above.

The resulting control number shown in Figure 1 identifies the address for the series description of "LETTERS RECEIVED."

In order to avoid the problem of *variable length records*,[5] all the data belonging to the same hierarchical address are linked sequentially by a set identification code (i.d.) and subset number. A set is a repeatable line of the same type of information, such as

Figure 1, Control Number

Level	1	2	3	4	5	6	7	8	9	10	11	12	
	RG	SUBGROUPS								SERIES	SUB-S	SUB-SUB-S	LEVEL
CON-TROL		A	B	C	D	D	E	F	H				
No.	127	1	1	1	1	1	1	0	0	18	0	0	10

[5] Variable length records require special handling called "list-processing," which involves a computer check for the end of each record each time it processes the file. Multiply one variable length record times two hundred thousand records to make up the entire file, and it is easy to see that the computer's job would have been slowed down considerably and been costly. For an example of list-processing see Alan Calmes, "A PL/I Free-Field Handling System," in *Historical Methods Newsletter* 8 (December 1974): 39–47.

Figure 2, Record Hierarchy

RG 127 Records of the U.S. Marine Corps
A.1 Textual Records
B.1 Headquarters U.S. Marine Corps
C.1 Office of the Commandant
D.1 General Records
E.1 Correspondence
S.1 REGISTER OF LETTERS RECEIVED
.
.
.
S.18 LETTERS RECEIVED

in Figure 3. Each line of information is one subset of a periodic data set, sometimes called a "repeating group." In the control number example above, a set i.d. such as "05," may be attached to the end of the control number to indicate that the following information consists of one narrative line. Subset number one attached to the set i.d. indicates that the information represents line number one of a paragraph. A set may be repeated up to 999,999 times by means of the subset number. Figure 3 outlines the descriptive narrative and location register part of the computer file format.

A line of text narrative may exist alone or be followed by an indefinite number of additional lines up to 999,999. For output-page-width reasons, each line of text consists of a maximum of seventy-six characters.

Figure 3, File Format

	Control #	+ *Set i.d.*	+ *Subset #*	+ *DATA*
Repeating	(full control #)	05	000001	(narrative line)
group	(full control #)	05	000002	(narrative line)
	(full control #)	05	000003	(narrative line)
Repeating	(full control #)	06	000001	(location)
group	(full control #)	06	000002	(location)

```
                                                      S
                                                      E  SUB-
RG  |A|B|C|D|E|F|G|H|  S  |SS |  SSS |L|X|T| SET |              DATA
0075034002022012010000001000000001001P7000010                         7/OHS
0075034002022012010000001000000001001P6000020chronological order.
0075034002022012010000001000000001001P6000030          Correspondence, reports, receipts for payments, vouchers,
0075034002022012010000001000000001001P6000040requisitions, bonds, and miscellaneous records relating to the Chickasaw
0075034002022012010000001000000001001P6000050tribe.  The bulk of the correspondence is between the U.S. Indian Agent and
0075034002022012010000001000000001001P6000060officials of the Bureau of Indian Affairs, officers of the Chickasaw tribal
0075034002022012010000001000000001001P6000070government, and the general public.  The records have been "sorted and
0075034002022012010000001000000001001P6000080cataloged" by the staff of the Society and each item has been assigned an
0075034002022012010000001000000001001P6000090identifying number.  There is a list of items which serves as a finding aid
0075034002022012010000001000000001001P6000100to these records.
0075034002022012020000000000000000801FS1NNFN              001848    0019090000000
0075034002022012020000000000000000801P10000107NCN  0000000000000000000000000000000000 0000000
0075034002022012020000000000000000801P301Miscellaneous Records
0075034002022012020000000000000000801P401MISCELLANEOUS RECORDS
0075034002022012020000000000000000801P6000010          Records of the Chickasaw Senate and House; minutes of various
0075034002022012020000000000000000801P7000010                         7/OHS
0075034002022012020000000000000000801P6000020Commissions and committees; copies of acts of the National Council; records
0075034002022012020000000000000000801P6000030of Chickasaw District courts; records of the National Treasurer; permit
0075034002022012020000000000000000801P6000040records for non-citizens; reports of the Chickasaw Attorney General; and
0075034002022012020000000000000000801P6000050records relating to applications for citizenship, operation of schools,
0075034002022012020000000000000000801P6000060appropriations, and property.
0075034002022012020000000000000000801P6000070
0075034002022012020000000000000000801P6000080          There are 106 bound volumes (12 linear feet) in this subgroup.
0075034002022013000000000000000000701FS1NNFN              001831    0019090000000
0075034002022013000000000000000000701P10000107NCN  0000000000000000000000000000000000 0000000
0075034002022013000000000000000000701P301Records Relating to Choctaws
0075034002022013000000000000000000701P401RECORDS RELATING TO CHOCTAWS
0075034002022013010000000000000000801FS1NNFN              001831    0019090000000
0075034002022013010000000000000000801P10000107NCN  0000000000000000000000000000000000 0000000
0075034002022013010000000000000000801P301General Records
0075034002022013010000000000000000801P401GENERAL RECORDS
0075034002022013010000001000000001001FS1NNFN              001831    0019090000000
0075034002022013010000001000000001001P10000107NCN  0000000000000002900000008840000000 0000000
0075034002022013010000001000000001001P401SUBJECT FILE
0075034002022013010000001000000001001P6000010          Arranged alphabetically by subject and thereunder in rough
0075034002022013010000001000000001001P7000010                         7/OHS
0075034002022013010000001000000001001P6000020chronological order.
0075034002022013010000001000000001001P6000030          Correspondence, reports, receipts for payments, vouchers,
0075034002022013010000001000000001001P6000040requisitions, bonds, and miscellaneous records relating to the Choctaw
0075034002022013010000001000000001001P6000050tribe.  The bulk of the correspondence is between the U.S. Indian Agent and
0075034002022013010000001000000001001P6000060officials of the Bureau of Indian Affairs, officers of the Choctaw tribal
0075034002022013010000001000000001001P6000070government, and the general public.  The records have been "sorted and
0075034002022013010000001000000001001P6000080cataloged" by the staff of the Society and each item has been assigned and
0075034002022013010000001000000001001P6000090identifying number.  There is a list of items which serves as a finding aid
0075034002022013010000001000000001001P6000100to these records.
0075034002022013020000000000000000801FS1NNFN              001859    0019070000000
0075034002022013020000000000000000801P10000107NCN  0000000000000000000000000000000000 0000000
0075034002022013020000000000000000801P301Records of the Choctaw Supreme Court
```

Figure 4, Example of NARS A-1 Machine-readable Data File

Within each subset of a set the data occupies fixed *fields* of descriptive information. Most fields contain uncoded information. A limited amount of coding reduces the length of some fields; for example, a "2" might equal the phrase "still pictures."

The General Services Administration (GSA) insisted that NARS obtain computer services from GSA's Office of Data Systems. Furthermore, GSA's systems analysts had to design systems more from the point of view of their machinery than according to the needs of the system. Therefore, in order to avoid the limitations imposed on the system by GSA and to accomplish as much of the A-1 system requirements at the National Archives as possible, NARS decided to buy a mini-computer to handle data input and to handle the relatively small statistical file on each record group. Because the mini-computer is programmable, part of the A-1 system is handled *on-line*. The Office of the National Archives maintains statistical control of the record groups (by quantity allocation to custodial units) by means of the mini-computer; batch processing of the entire inventory file for the compilation of series descriptions, however, takes place at a large computer installation controlled by GSA.

The systems analysts designed the series description part of the system around batched text editing and magnetic tape storage. Batched text editing is designed to eliminate the need for retyping whole series entries after each change. (An archivist submits a draft, it is *keyed* into the system, and a printout is returned to the archivist for revision. Input operators key-in the changes; a batch update cycle replaces or deletes, or inserts data.) With this method, the design and implementation of which cost about $60,000,[6] analysts estimate that it will take the Archives twenty years to input approximately 200,000 series descriptions of the previously described, the undescribed, and the newly accessioned archival records, and be up-to-date with the latter.

Hardware

The most attractive aspect of computer assistance, from the point of view of the A-1 objective, is fast, accurate data-entry (the process of typing information and creating machine-readable data).[7] About 60 percent of the entire system's estimated cost consisted of data-entry. NARS looked for *hardware,* i.e., computer machinery, for an in-house, data-entry facility to carry out some limited programmable processing, especially data verification, editing, and reformatting. GSA's Office of Procurement studied the problem of whether to rent or to buy. On the basis of the A-1 long-term needs, the office decided it would be cost-effective to buy.[8] Over a twenty year period, the annual cost of the in-house, mini-computer, data-entry equipment, plus maintenance, will be $10,350 per

[6] This figure does not take into account all the GSA, Office of Data Systems, costs which could only roughly be tabulated because GSA did not bill NARS directly for all costs.

[7] Reliable sources of information on mini-computers may be found in *Data-Entry Awareness Reports,* published monthly by Management Information Corporation, 140 Barclay Center, Cherry Hill, NJ 08034; and *Datapro,* special reports published by Datapro Research Corporation, 1805 Underwood Boulevard, Delvan, NJ 08075.

[8] NARS bought a Four Phase, Data IV/90 computer for $147,000. It included a 96K byte processor/memory, 67.5 million byte direct access disk, 1600 BPI tape drive, 300 lines per minute printer, and eight CRT/KEY stations, capable of displaying and processing both upper and lower case alphabetic

year. Rental would have been $41,000 per year.

Software

The A-1 system uses data-entry *software* (a collection of computer programs) that came with the hardware.[9] The software allows the input operator to see projected on a television-like screen questions (prompts) which ask for specific information to be read off the input form and typed into the computer. The typed data appear on the screen and the software checks to see if the data meets a set of criteria previously programmed into the machine. If the data fail the validation-check, the machine sounds an alarm. With the machine in the "search" mode, the operator makes the computer locate a particular record stored on disk by a unique identification code, field, or logical search strategy, after which the operator inserts, corrects, changes, or deletes one character at a time. Because several input operators may be doing a variety of tasks, the software/hardware combination has to allow for simultaneous data-entry into different formats and into the same format, while background processing takes place and data is being transferred to the tape-drive or printer. Automatic editing is another desirable feature—checking for alphabetic or numeric characters, checking for limits of a range, and for comparisons to stored values. Automatic field generation for repetitive and incremented fields, automatic insertion of constants, automatic duplication of data, and a key-verify mode[10] with immediate, direct access correction are other required features. In summation, the hardware/software combination provides for quick and accurate data-entry.

Data Entry Process

A key-entry operator takes the source document (GSA Form 6710A, Change of Status Record) containing archival descriptions of records, and keys the data into a television-like screen, field by field. Some fields are repetitive and do not require rekeying. A format control program edits the data—right-justifies and zero-fills numeric fields—and restarts the display for the next source document. The work-cycle consists of data-entry, verification, and batch input to the master *data-base.*

With four input operators, the A-1 system averaged 6,000 finished series descriptions per year during its first three years of production. There was an average of ten lines per series consisting of an average of fifty keyed characters per line. Each series, with identifying subrecord group titles and text, averaged 500 characters. Thus

characters. A short-term five year project would be better off with rented equipment. Due to rapid changes in computer technology, especially in the mini-computer field, and the decreasing cost of computer parts, it may be best to rent for two or three years and then obtain a new contract or a new mini-computer.

[9] Four Phase provided a wide variety of software including data-entry formatting, editing, and reformatting. It also provided COBOL programming, tape read-and-write instructions, and format instructions to the printer. Four Phase trained the data-entry operators and the supervisory computer operator, and distributed manuals.

[10] Key-verify mode consisted of a second typing. The input operator typed over the previously typed data which was disguised by scrambled letters; and, when the keyed character disagreed, the machine alerted the operator that an inconsistency existed. The operator would than sight verify the particular character in question and correct the mistake.

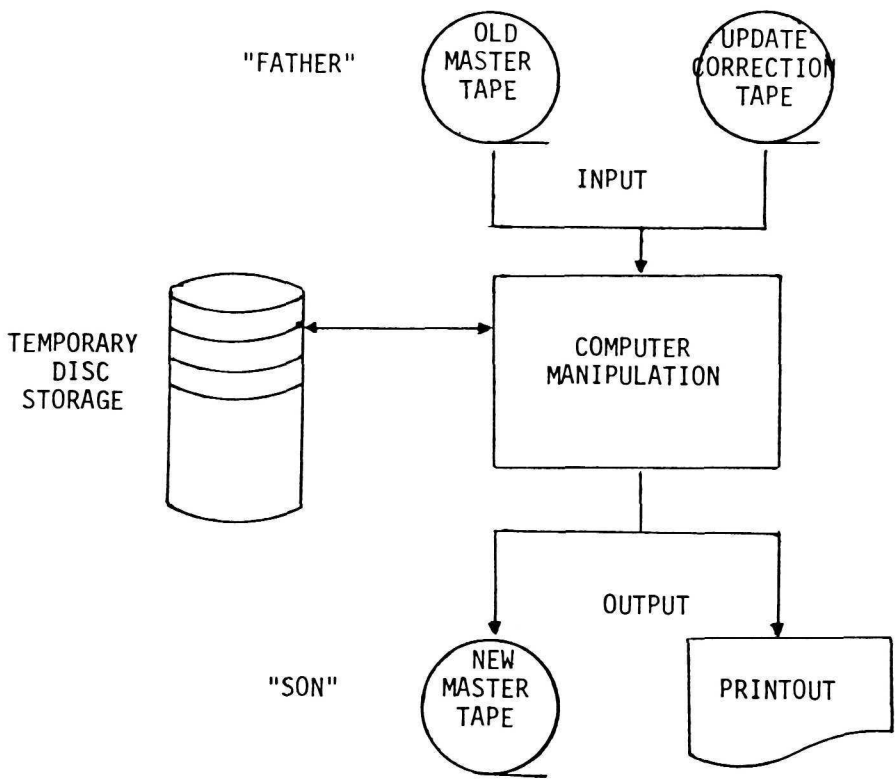
ten lines times fifty characters, times six thousand series, equals three million characters.

Data Base Maintenance

The master machine-readable file of record group inventories, stored as a data base on magnetic tape at a large computer installation, receives a monthly update of new and revised data in batch processing mode. As illustrated in Figure 5, the update transactions cause the data-base tape to be read into temporary disk storage. The insertions, replacements, and deletions take place on the disk. Then the program creates a new, updated, sequential file on tape for storage.

The data on the correction/update tape are replaced and/or merged into the addressed sequential slots of the "father" data-base tape. A new, revised tape called the "son" tape, including the revisions and additions, becomes the latest revision of the data-base. During the process, the computer also produces printout reports of the actions taken. The old data-base tape remains unchanged as a "father" tape. Backup tapes go back to the "grandfather" generation. A "great grandfather" tape is recycled into a

Figure 5, Flow Chart



new, current, master, data-base tape (the "son" tape). After an update cycle, a listing of valid and invalid transactions instructs the input operators to make corrections. For example, a new line of narrative description might be entered with a used subset number. This transaction would be rejected as an invalid transaction, and the operator would have to change the subset number to an unused number. The transaction would have to be reentered with the next update cycle. This error could be avoided by the operator's check of the current, master file. The master file, however, is so large that printouts are impractical and microfiche must be used instead. The example of the error above results from misreading one of the thousands of lines of data identified by the forty-nine character control-number; misreading one of the digits causes the mistake.

The A-1 data-base will grow in size to about one hundred million characters over the next twenty years. This will require the machine-readable file to be segmented by record group. It will become inefficient to keep the entire file in one sequential series of tapes which will have to be read into the computer and put on disk. Instead, it will be better to have a separate tape for each record group or group of record groups. Segmentation of the file will require some program changes. This will have to be done by GSA programmers, however, and will constitute an extra cost and some trial and error programming and testing; this will result in some "down time." Furthermore, the GSA programmers will have to make constant adjustments to the programs after each hardware enhancement and after each new software release. With the development of the mini-computer base of operations in the Archives building, however, eventually it may be possible to maintain the data in-house and to have on-line, control number access to each series description.

Output

For the most part, output means printout pages and microfiche. The batch mode computer programs at the GSA-controlled computer produce printout pages or place print-page images on magnetic tape. The Archives personnel may either print the magnetic tape on paper at the in-house mini-computer or hire a service contractor to produce computer output microfiche (COM). The entire file is placed into inventory format about every three months. Microfiche is made twice a year. Smaller reprints are run locally by the mini-computer, where a great deal of diversity of format is possible. Together, the various forms of output cost about $5,000 per year, including paper and microfiche.

Personnel

An overwhelming part of the A-1 system involves people. An in-house task force of program analysts carried out the initial study. The system received a wide range of participation and input from archivists throughout the National Archives. Several system analysts worked on the system design. Computer programmers and operators wrote programs for the GSA-controlled computer. At the Archives, four key-entry operators and one supervisor are carrying out the task of coding and keying-in the data.

After three years experience building the machine-readable file of series descriptions, the rate of production

appears to be very slow.[11] An average rate of 375 characters in final form per hour is based on the production of six thousand series descriptions per year by four input operators during a two thousand hour work-year. The rate of production is low because the data are keyed twice (once entered and once verified); furthermore, the update cycle requires re-input of invalid transactions each time a line of data fails to go on the machine-readable file. The A-1 data-base is constantly being changed; some series descriptions are revised several times a year. Finally, personnel turnover and training also slow the rate of production, as does computer down-time.

The A-1 personnel costs were $44,500 per year, broken down as follows: $8,000 per entry operator, times four operators, equaling $32,000 per year. The supervisor cost $12,500 per year.

Total Operating Costs

The Archives personnel cost of $44,500 annually amounted to nearly 60 percent of the total annual costs. The cost of a series entry was computed by dividing the six thousand series placed into the system each year by the total annual cost, producing $12.39 per series. Based on an average of 500 characters per series, each final character costs the government almost two and a half cents ($.0248). Each character was keyed at least twice (entry and verify) and some three or four times for revision and correction.

Comparison of the A-1 Experience with the A-1 Study Projections

The A-1 experience costs were fairly close to the A-1 study projected costs.[12] The total estimated costs, though aligned somewhat differently from actual costs, are shown in Figure 7.

Figure 6, Annual Cost

System Study*	$ 3,500
System Design and Implementation*	3,000
Hardware/Software*	10,350
Data Entry Personnel	44,500
Data-base Maintenance	8,000
Output	5,000
Total Annual Cost	$74,350

*Total cost distributed over a twenty-year period.

[11] A formula for estimating time and personnel needed to accomplish a comparable job would be: number of lines per average entry (a) times number of characters to be keyed in each line (b) times number of entries (c) divided by the constant 750,000 equals the number of man-years required (d) to produce a finished product. ($a \times b \times c \div 750{,}000 = d$). The constant 750,000 was derived by dividing the three million characters of the six thousand series descriptions by four input operators.

[12] Weiher, vol. 4, chapter 5.

Figure 7, Estimated Costs

	TOTAL	*ANNUAL*
System Study	$ 70,000	$ 3,500
Program development	103,000	5,150
Data Entry	868,000	43,400
Data base maintenance & output	127,725	6,386
	$1,168,725	$58,436 (or $9.75 per series)

There was no allowance in the projection for data-entry hardware. There was mention of an unspecified amount for "other costs," which covered the projected cost of some sort of typewriter to produce machine-readable data. The cost of data-entry equipment left out of the study may acount for the difference of $2.65 per series between the study and the actual implementation cost of the A-1 system.

Conclusion

Though the automated part of the NARS A-1 system is limited to computer assistance, the overall cost of the system is high. If a fully automated system with on-line retrieval by index terms had been implemented, the cost would have been excessively high, and the production rate so slow that it would have taken sixty years to catch up. The A-1 system is a fair warning, therefore, that automation of archival finding aids must be approached carefully. The size of the data-base and the cost per keystroke must be calculated in advance to see if there is enough time, enough people, and enough money to convert the finding aid information into machine-readable form.