University of Nebraska - Lincoln
DigitalCommons@University of Nebraska - Lincoln

Library Philosophy and Practice (e-journal) Libraries at University of Nebraska-Lincoln

7-13-2005

Using Sorted Lists for Error Checking

Robert L. Bolin
University of Nebraska--Lincoln, rbolin2@unl.edu

Follow this and additional works at: https://digitalcommons.unl.edu/libphilprac

 Part of the Library and Information Science Commons

Bolin, Robert L., "Using Sorted Lists for Error Checking" (2005). *Library Philosophy and Practice (e-journal)*. 2.
https://digitalcommons.unl.edu/libphilprac/2

Using Sorted Lists for Error Checking

Robert L. Bolin
Electronic Resources Librarian,
Associate Professor
University of Idaho Library
Moscow, ID 83844-2350

Introduction

In the course of preparing a bibliography, errors creep into the data. This paper describes the use of sorted lists for error checking. Often sorted lists are useful for spotting errors or inconsistencies in basic fields like the author and title fields. Sometimes special lists can be used to take advantage of patterns in the data.

Two bibliographic projects where this approach to error checking was used are described briefly. One project from the 1980s used mainframe technology; the other used PC technology. For the second project, "printing" lists to computer files greatly facilitated use of those lists. Although those lists were not HTML documents, they can be viewed with HTML browsers. HTML tags were added to provide formatting and highlighting. The dBase III+ program proved very well suited for the second project.

Error Checking

In the course of any bibliographic project, numerous errors creep into the data. Many times data is simply input incorrectly. Also, the sources for the data may be incorrect or inconsistent. Half the job is error checking, and a number of techniques must be used to correct the data. One technique that may prove useful is to study sorted lists of the author field, the title field, and other major fields. Sorted lists are very useful for spotting misspellings, inconsistent punctuation, inconsistent usage of abbreviations, transposition of numbers or letters, and other anomalies and inconsistencies. An example of a title list used in the course of preparing a bibliography of US Army manuals can be viewed at http://www.uidaho.edu/~mbolin/Title.html. If a field is shown in a sorted list, large errors tend to appear at the top or bottom of the list and minor errors are much less obvious. For example, if the first word of a title began with a "3" in the place of an "E," that would be obvious in a sorted title list. Transposition of "ea" within a title would be less obvious.

Of course, all errors and inconsistencies are not obvious at once. By repeated reviewing lists and making corrections, more and more errors are caught and corrected. Also, as the person preparing the bibliography becomes more and more familiar with the data, errors and inconsistencies become more apparent.

Indexing University of Georgia Bulletins

In the early 1980s, I indexed bulletins of the University of Georgia and the Georgia State College of Agriculture. Those indexes are described in Sidebar One. One purpose of the project was to try out the FAMULUS software package developed by the Pacific Southwest Forest and Range Experiment Station of the US Forest Service. FAMULUS was an innovative set of FORTRAN programs developed around 1970, which allowed individual users to create bibliographies using a mainframe computer. The user designed the basic record, created a file containing those records, and used the system to create a variety of reports.

Sorted lists were particularly useful for resolving two problems.

- One was the classic problem encountered by indexers and catalogers of distinguishing between actual titles and series names. By studying sorted lists of titles, I was able to determine what I felt the actual titles were. Although I identified several series titles, I decided that a separate series title field would not be useful.
- The second problem was that college and department names had appeared inconsistently on the actual bulletins and circulars. By studying the sorted list of "sponsoring organizations," I was able to standardize the forms used.

I had to pick up computer printouts at the computer center, which was distant from my home and my office. Picking up printouts was tedious and time consuming and disrupted my work. During the course of the project, I made scores of trips to the computer center.

Army Manuals Listed in BSIR

Recently, I prepared a list of US Army manuals, which were listed in the *Bibliography of Scientific and Industrial Reports (BSIR)* during the late 1940s. The index produced is described in Sidebar Two.

I used dBase III+ to prepare that index because a copy was available and because I am familiar with the dBase command language.1 The command language is a specialized programming language for manipulating dBase files. The features intended to facilitate printing reports can easily be used to create HTML-formatted documents. Sorted lists can be used to identify errors and inconsistencies in the master file. Rather than print lists, I used the dBase SET ALTERNATE option to "write" the lists into computer files on my PC. I viewed those files using a Web browser. With simple programs written in the dBase command language, I was able to sort the master dBase file and print fields in the records in the order I specified. I had the program add HTML tags to format and highlight the records.

Here are two examples showing the ***same*** sorted title list created using the dBase command language.

- Part of the sorted title list viewed as a text file. Note that that list is ***not*** a full-scale HTML document but rather plain text file with embedded HTML tags. Only a sample of that file viewed as text is included.

Title Listing
<font size = -1>
<p>
1 1/2-KVA Kohler Power Unit Model 1M21-A, TM 11-935, 44, PB 23495.
1 1/2-ton, 4x2 Truck (Ford), TM 9-806, 44, PB 3347.
1 1/2-ton, 4x4 Truck (Chevrolet), TM 9-805, 43, PB 5084.
1 1/2-ton, 6x6 Truck (Dodge T-223, Models WC-52 and WC-63), TM 9-810, 45, PB 36697.
1/4-ton, 4x4 Truck (Willys-Overland Model MB and Ford Model GPW), TM 9-803, 44, PB 3281.
10-ton Payload, 14-ton Gross, 2-Wheel Stake and Platform Semitrailer and 10-ton Converter Dolly, TM 9-892, 44, PB 3211.
10-ton, 6x4 Truck (Mack Model NR), TM 9-818, 44, PB 3203.
11-ton Payload, 15-ton Gross, 2-Wheel (2dt), Van Semitrailer (Omaha Standard Body Corp. Model F16), TM 9-894, 44, PB 3335.
12-inch Seacoast Material, 12-inch Mortar M1890MI Mounted on 12-inch Mortar Carriage M1896MI and M1896MII, TM 9-456, 42, PB 3353.
12-inch Seacoast Materiel, 12-inch Mortar M1912 Mounted on 12-inch Mortar Carriage M1896MIII, TM 9-458, 42, PB 3309.
13-ton, High-Speed Tractor M5, TM 9-786, 43, PB 36703.
155-mm Gun Materiel, M1917, M1918, and Modifications, TM 9-345, 41, PB 58403.
155-mm Gun Motor Carriage M12 and Cargo Carrier M30, TM 9-751, 44, PB 36707.
155-mm Gun Motor Carriage T83 and 8-inch Howitzer Motor Carriage T89, TM 9-747, 45, PB 62805.
155-mm Guns M1 and M1A1 and Carriage M1; 8-inch Howitzer M1 and Carriage M1; Heavy Carriage Limber M2, TM 9-1350, 43, PB 3218.
155-mm Guns, M1917, M1917A1, and M1918 MI Carriages, M1917, M1917A1, M1918, M1918A1, M2, and Limbers M1917, M1917A1, M1918, M1918A1, and M3, TM 9-1345, 42, PB 3226.
16-inch Seacoast Gun Materiel, Gun Mk. II M1, Barbette Carriage M4, TM 9-471, 42, PB 3209.
18-ton High Speed Tractor M4, TM 9-785, 43, PB 36729.

- Part of that sorted title list viewed as a hypertext document. Note the effect of the HTML tags.

Title Listing

1 1/2-KVA Kohler Power Unit Model 1M21-A, TM 11-935, 44, PB 23495.
1 1/2-ton, 4x2 Truck (Ford), TM 9-806, 44, PB 3347.
1 1/2-ton, 4x4 Truck (Chevrolet), TM 9-805, 43, PB 5084.
1 1/2-ton, 6x6 Truck (Dodge T-223, Models WC-52 and WC-63), TM 9-810, 45, PB 36697.
1/4-ton, 4x4 Truck (Willys-Overland Model MB and Ford Model GPW), TM 9-803, 44, PB 3281.
10-ton Payload, 14-ton Gross, 2-Wheel Stake and Platform Semitrailer and 10-ton Converter Dolly, TM 9-892, 44, PB 3211.
10-ton, 6x4 Truck (Mack Model NR), TM 9-818, 44, PB 3203.
11-ton Payload, 15-ton Gross, 2-Wheel (2dt), Van Semitrailer (Omaha Standard Body Corp. Model F16), TM 9-894, 44, PB 3335.
12-inch Seacoast Material, 12-inch Mortar M1890MI Mounted on 12-inch Mortar Carriage M1896MI and M1896MII, TM 9-456, 42, PB 3353.
12-inch Seacoast Materiel, 12-inch Mortar M1912 Mounted on 12-inch Mortar Carriage M1896MIII, TM 9-458, 42, PB 3309.
13-ton, High-Speed Tractor M5, TM 9-786, 43, PB 36703.
155-mm Gun Materiel, M1917, M1918, and Modifications, TM 9-345, 41, PB 58403.
155-mm Gun Motor Carriage M12 and Cargo Carrier M30, TM 9-751, 44, PB 36707.
155-mm Gun Motor Carriage T83 and 8-inch Howitzer Motor Carriage T89, TM 9-747, 45, PB 62805.
155-mm Guns M1 and M1A1 and Carriage M1; 8-inch Howitzer M1 and Carriage M1; Heavy Carriage Limber M2, TM 9-1350, 43, PB 3218.
155-mm Guns, M1917, M1917A1, and M1918 MI Carriages, M1917, M1917A1, M1918, M1918A1, M2, and Limbers M1917, M1917A1, M1918, M1918A1, and M3, TM 9-1345, 42, PB 3226.
16-inch Seacoast Gun Materiel, Gun Mk. II M1, Barbette Carriage M4, TM 9-471, 42, PB 3209.
18-ton High Speed Tractor M4, TM 9-785, 43, PB 36729.

Using Windows Effectively

The sorts of data checking and correction, which took days during the earlier project, took only seconds. I could run the web browser and dBase at the same time using Windows. It was easy to switch between those two programs. I reviewed a list "printed" to a computer file using a web browser. If I spotted an error, I simply toggled into the dBase program, make the correction, and toggled back to the web browser. Using computer files for working copies allowed me to work rapidly and effectively without having to break my concentration to get printouts from a printer.

Basic Sorted Lists

As with the earlier project, establishing correct titles was difficult. The army's changing conventions for listing titles and series titles on the covers and title pages of manuals lead to confusion. Also, the indexers who had prepared the *Bibliography of Scientific and Technical Reports* were inconsistent and occasionally sloppy. Sometimes, they confused series names with titles and occasionally paraphrased titles. A list sorted by title helped me to distinguish between titles and series names and to spot cases where titles were paraphrased. A separate series name field was not necessary in this bibliography because the manual numbering system is coded to show the general subjects of the manuals.

Exploiting a Pattern in the Data

I assumed that PB numbers—the order numbers given to documents listed in *BSIR*—were assigned using a mechanical number-stamping machine as the documents were

unboxed. I expected many of the PB numbers to be in sets of sequential numbers. I reasoned that:

- Gaps in the numbering sequence might represent items that I had missed while searching through the *Bibliography of Scientific and Industrial Reports* for army manuals and
- Duplicate PB numbers must represent transcription errors from when I entered the data.

I wrote a program that created a list in PB-number order showing the PB number, manual number, and part of the title on a single line. A blank line indicated a gap in the PB-number sequence, and a line printed in blue showed a duplicate PB number. Here is a sample of that PB-number list.

PB 982 [BSIR 1:79] - **TB SIG E1** German Radio Sets, Torn. Fu. bl. and Torn. Fu. f.
PB 983 [BSIR 1:79] - **TB SIG E2** German Radio Set, Torn. Fu. d2
PB 984 [BSIR 1:79] - **TB SIG E3** German Radio Set, Torn. E. b.
PB 985 [BSIR 1:79] - **TB SIG E4** German Radio Set Receiver, Spez. 445 b Be.
PB 986 [BSIR 1:79] - **TB SIG E5** German Radio Transmitter, 10 W.S.c and 10 W.S.h an
PB 987 [BSIR 1:79] - **TB SIG E6** German Radio Transmitter, 5 W.S./24b-104
PB 988 [BSIR 1:78] - **TB SIG E7** German Radio Set, Fusprech. a.
PB 989 [BSIR 1:79] - **TB SIG E8** German Radio Transmitters, 20 W.S.c and 20 W.S.d
PB 990 [BSIR 1:79] - **TB SIG E9** German Radio Transmitter, 100 W.S.a.
PB 991 [BSIR 1:79] - **TB SIG E10** German Radio Transmitter, 80 W.S.a.
PB 992 [BSIR 1:79] - **TB SIG E11** German Radio Transmitter, 30 W.S.a.
PB 993 [BSIR 1:78] - **TB SIG E12** German Radio Sets, Feldfu. b. and Feldfu. c.
PB 994 [BSIR 1:78] - **TB SIG E13** German Radio Set, SE469 A.
PB 995 [BSIR 1:78] - **TB SIG E14** German Radio Receiver, Kw.E.a
PB 996 [BSIR 1:76] - **TB SIG E-15** Description and Use of Captured Enemy Field Wire a
PB 997 [BSIR 1:79] - **TB SIG E16** German Radio Set, Torn Fu. g.
PB 998 [BSIR 1:80] - **TB SIG E17** Japanese Radio Set, Model 94 Mark 6 Wireless Set,
PB 999 [BSIR 1:81] - **TB SIG E18** Japanese Radio Set, Model 97 Light Wireless Set
PB 1000 [BSIR 1:80] - **TB SIG E19** Japanese Radio Set, Model 94 Mark 5 Wireless Set,
PB 1001 [BSIR 1:80] - **TB SIG E20** Power Supplies for German Radio Sets
PB 1002 [BSIR 1:80] - **TB SIG E21** Japanese Radio Set, Mobile Wireless Set C Mark 1 M
PB 1003 [BSIR 1:81] - **TB SIG E22** Japanese Radio Set, Model 95, Mark 4 Short Wave Tr
PB 1004 [BSIR 1:80] - **TB SIG E23** Japanese Radio Set, Model 94 Mark 2B Wireless Set
PB 1005 [BSIR 1:78] - **TB SIG E24** German Field Telephone, Model 33
PB 1006 [BSIR 1:79] - **TB SIG E25** German Radio Transmitter, 30 W.S./24b-120
PB 1007 [BSIR 1:80] - **TB SIG E26** Japanese Radio Set, Model 94 Ground-Air Mark 2 Wir
PB 1008 [BSIR 1:78] - **TB SIG E27** German Office Connection, Model 33
PB 1009 [BSIR 1:354] - **TB SIG E28** Japanese Radio Set, Mobile Wireless Set B
PB 1010 [BSIR 1:80] - **TB SIG E29** German Small Switchboard for 10 lines

I checked that list against the *Numerical Index to the Bibliography of Scientific and Industrial Reports* (*BSIR),* v. 1-10, 1946 - 1948 which listed PB numbers with the volume and page numbers of the corresponding entry in *BSIR*.[2] I found:

- Many cases where I had simply missed citations to manuals in *BSIR* that should have been included.

- Many cases where a particular PB number had not been used. That is, those PB numbers did not appear in *BSIR*.
- A number of cases where I had garbled PB numbers during data entry.
- A case where several pages were left out of the copy of *BSIR* that I was consulting. Because of that printer's error, I had missed several citations to Army manuals.
- A few cases where the manual number had been left out of the citation in *BSIR*. I was able to supply those numbers by consulting other sources.
- A few cases where the PB numbers printed in *BSIR* were clearly wrong.

Using a Spelling Checker

The dBase program does not contain a spell checker. I wrote a simple program to write the manual numbers and titles from each citation to a plain-text computer file. Then I used the spell checker in Microsoft Word to check the titles. I toggled between the dBase master file and Word making corrections and then checking further. When I used the spell checker, I had already checked and rechecked all the entries. I was amazed at how many misspelled words the spell checker found. Of course, the spell checker also found many peculiar Army words and spellings which had been correctly transcribed.

Printing the Finished Bibliography

I wrote a dBase command language program to produce the finished bibliography in the form of a full-scale, finished HTML document and to "write" it to a file. The dBase TEXT...ENDTEXT command allows text which is to be printed verbatim to be included in the program. I simply incorporate the HTML to be included in the final document into the program using TEXT...ENDTEXT. That has made editing my bibliography simple.

- To correct the text of the bibliography, I edit the dBase command language program to change the HTML-formatted text which is to appear in the final document.
- To correct the data, I corrected records in the master dBase file. When I had made the corrections, I ran the program to "write" an updated copy of the finished bibliography.

Conclusions

During the course of a bibliographic project sorted lists can be very useful for error checking and standardizing entries. Printing those files to computer files and viewing them using a web browser can greatly speed and simplify the work. Also, using HTML allows you to highlight and format the entries for viewing.

References

1. Although dBase III+ is an antique, it is flexible and powerful program. It can easily be used to write full-fledged HTML-formatted documents. Since it is obsolete, thousands of licensed copies are available for the asking on campuses around the country. Also, many people are familiar with dBase programming.

2. *Numerical Index to the Bibliography of Scientific and Industrial Reports*, Volume 1-10, 1946-1948. (New York: Special Libraries Association, 1949). (Available from University Microfilm International, PB2-OP04032).

Sidebar Number 1

Bulletins and Circulars of the University of Georgia

Bulletins of the University of Georgia are a window on the history of the school. Over the years a great variety of publications were published in the bulletin series because they could be mailed cheaply. A sample page from the index can be viewed at http://www.uidaho.edu/~mbolin/uga.gif and shows the diversity of material printed in the Bulletin.

The publications of the semiautonomous Georgia State College of Agriculture are very interesting because the history of that college is a classic study in academic empire building. The rapid expansion of the college and its programs and activities is clearly reflected in its publications.

I used the FAMULUS program to produce separate indexes to university Bulletin series and to the College of Agriculture publication series and a merged list: *A FAMULUS Index to The Bulletin of the University of Georgia. (Athens, Georgia: Political Science Department, University of Georgia, 1984.)*

A FAMULUS Index to Georgia State College of Agriculture Bulletins and Circulars. (Athens, Georgia: Political Science Department, University of Georgia, 1984.)

A FAMULUS Index to Bulletins and Circulars of the University of Georgia.
(Athens, Georgia: Political Science Department, University of Georgia, 1984.) Each of those indexes contained sequential lists of publications; title lists, and lists arranged by the sponsoring college, department, or organization. Several copies of those indexes were given to libraries at the University of Georgia holding the publications indexed. The index to the bulletins and circulars of the College of Agriculture was also given to the National Agricultural Library.

This project had a useful byproduct. The bulletins and circulars of the Georgia State College of Agriculture had been microfilmed by the Southeastern Land Grant College Library Microreproduction Project. However, publications of the college which had been issued in the University Bulletin before the College of Agriculture started its own publications series had been missed. With a small grant from the College of Agriculture, I organized those early agriculture publications and had them microfilmed as: *University of Georgia Agricultural Extension Service Bulletin, Supplement. (Athens, Georgia: University of Georgia College of Agriculture, 1984.)* The master copy of the film was sent to the National Agricultural Library.

Sidebar Number 2

Army Manuals listed in *BSIR*

Shortly after World War II a large variety of publications were released for distribution by the Office of Technical Services which was a precursor of the National Technical Information Service. They were listed in the *Bibliography of Scientific and Industrial Reports*. Those reports are now available from the Photoduplication Service of the Library of Congress. However, they must be ordered using the PB number they were assigned by the Office of Technical Services.

I used the dBase III+ program to create a master database containing records describing U.S. Army manuals listed in *BSIR*.[1] Then I used the dBase command language to write an HTML-formatted document for "publication" on the World Wide Web. A list in manual-number order is the key component of the bibliography

since users probably will know a manual number and be looking for the PB number with which the manual could be ordered. Title and PB number listings are also provided. Color-coding is used in the manual-number listing to distinguish between technical manuals and technical bulletins which are interfiled. Color coding is used in the title and PB-number listings to distinguish between the various manual series.

The list of ***Army Manuals Listed in the Bibliography of Scientific and Technical Reports, 1946-1949*** **is available on the web at:**

http://unllib.unl.edu/Bolin_resources/bsir/Army%20Manuals%20Listed%20in%20BSIR,%201946-49.htm