

A Checklist for Evaluating Record Linkage Software

Charles Day, National Agricultural Statistics Service

From the 1950's through the early 1980's, researchers and organizations undertaking a large record linkage project had little choice but to develop their own software. They often faced the choice of using less accurate methods or expending dozens of staff years to create proprietary systems. For example, in the late 1970's, the U.S. National Agricultural Statistics Service spent what is conservatively estimated as 50 staff years to develop a state-of-the-art system. Happily, today's record linkage practitioners no longer need do this any more than they need to write their own word processing software, nor should they attempt it. Powerful, flexible, relatively inexpensive software that implements all but the most sophisticated methods is available in the form of generalized packages that can stand alone or software components that can be integrated into a surrounding application. There is no longer any reason for anyone but researchers into the theory of record linkage to attempt to write record linkage software from scratch.

The Record Linkage Workshop and Exposition featured six vendor representatives who exhibited their software on site. This checklist is provided as an aid in evaluating the record linkage software they sell, along with other products that may enter the market. While the authors have endeavored to make the checklist as complete as possible, there may still be important characteristics for your application that the checklist does not cover. There is no substitute for a thorough analysis of your individual needs. Comments on the checklist are welcome. Please email them to cday@nass.usda.gov.

General

- 1.1 Is the software a generalized system or specific to a given application?
- 1.2 Is the software a:
 - Complete system, ready to perform linkages "out of the box?"
 - Set of components, requiring that a system be built around them? If so, how complete are the components?
 - Part of a larger system for performing integrated mailing list functions?
- 1.3 What types of linkages does the software support?
 - Unduplication (one file linked to itself)?
 - Linking two files?
 - Simultaneously linking multiple files?
 - Linking one or more files to a reference file (e.g., geographic coding)?

- 1.4 Can the software be used on the following computers:
- Mainframes?
 - Mini-computer?
 - Workstation?
 - IBM-compatible microcomputer?
 - Macintosh?
- 1.5 Can the software run under the following operating systems:
- MS/PC DOS?
 - OS/2?
 - Windows 3.1/95?
 - Windows NT?
 - UNIX?
 - VMS?
 - Mac OS?
 - Novell NetWare?
 - Mainframe OS (e.g., IBM MVS)?
- 1.6 For PC based systems, what level of processor is required? How much memory? How much hard drive space?
- 1.7 Can the system perform linkages interactively (in real time)? Can it operate in batch mode?
- 1.8 How fast is the software on the user's hardware and files the size of the user's files? If the software is interactive, is its performance adequate?
- 1.9 If the software is to be used as part of a statistical estimation system, are the methods used in the software statistically defensible?
- 1.10 Is the vendor reliable? Can the vendor provide adequate technical support? Will they continue to exist for the projected life of the software? If this is in question, is a software escrow available? Is the user prepared to support the software him/herself?
- 1.11 How well is the software documented? Can a new user reasonably be expected to sit down with the manual and begin using the software, or will training be necessary? Does the vendor provide training? At what cost?
- 1.12 What features does the vendor plan to add in the near future (e.g., in the next version)?
- 1.13 Is there a user group? Who else is using the software? What features would they like to see added? Have they developed any custom solutions (e.g., front ends, comparison functions) they would be willing to share?
- 1.14 Is other software, such as database packages or editors, needed to use the system?
- 1.15 Does the system provide security and data integrity protection features?
- 1.16 How many and what type of staff personnel will be required to develop a system from the software? To run the system? What type of training will they need and will the vendor provide that training?

Linkage Methodology

- 2.1 What record linkage method is the software based on?
 - Fellegi-Sunter?
 - Information-Theoretic methods?
 - 2.2 How much control does the user have over the linkage process? Is the system a "black box," or can the user set parameters to control the linkage process?
 - 2.3 Does the software require any parameter files? If so, is there a utility provided for generating these files? How effectively does it automate the process? Can the utility be customized?
 - 2.4 Does the user specify the linking variables and types of comparisons?
 - 2.5 What kinds of comparison functions are available for different types of variables? Do the methods give proportional weights (that is, allow degrees of agreement)?
 - Character-for-character?
 - Phonetic code comparison (Soundex or NYSIIS variant)?
 - Information theoretic string comparison function?
 - Specialized numeric comparisons?
 - Distance comparisons?
 - Time/Date comparisons?
 - Ad hoc methods (e.g., allowing one or more characters different between strings)?
 - User-defined comparisons?
 - Conditional comparisons?
 - 2.6 Can the user specify critical variables that must agree for a link to take place?
 - 2.7 How does the system handle missing values for linkage variables?
 - Computes a weight like any other value?
 - Uses a median between agreement and disagreement weights?
 - Uses a zero weight?
 - Allows user the option to specify treatment?
 - 2.8 Does the system allow array-valued variables (e.g., multiple values for phone number)? How do array-valued comparisons work? What is the maximum number of values in an array?
 - 2.9 What is the maximum number of linking variables?
 - 2.10 How does the software block records? Do users set blocking variables? Can a pass be blocked on more than one variable?
 - 2.11 Does the software support multiple linkage passes with different blocking and different linkage variables?
 - 2.12 Does the software contain or support routines for estimating linkage errors?
-

- 2.13 Does the matching algorithm use techniques that take advantage of dependence between variables?

Fellegi-Sunter Systems

- 3.1 How does the system determine m- and u-probabilities? Can the user set m- and u-probabilities? Does the software provide utilities to set m- and u-probabilities.
- 3.2 How does the system determine weight cutoffs? Are they set by the user? Does the software provide any utilities for determining weight cutoffs?
- 3.3 Does the software allow linkage weights to be fixed by the user? What about weights for missing values?

Data Management

- 4.1 In what file formats can the software use data?
Flat file?
SAS Dataset?
Database? If yes, what kind of database?
 Dbase?
 Fox Pro?
 Xbase?
 Informix?
 Sybase?
 ORACLE?
 Other database package?
- 4.2 What is the maximum file size (number of records) that the software can handle?
- 4.3 How does the software manage records? Does it use temporary data files or sorted files? Does it use pointers?
- 4.4 Can the user specify subsets of the data files to be linked?
- 4.5 Does the software provide for "test matches," of a few hundred records to test the specifications?
- 4.6 Does the software provide a utility for viewing and manipulating data records?

Post-linkage Functions

- 5.1 Does the software provide a utility for review of possible links? If so, what kind of functionality is provided for? What kind of interface does the utility use, character-based or GUI? Does the utility allow for review between passes, or only at the end of the process? Can more than one person work on the record review simultaneously? Can records be "put aside" for later review? Is there any provision for adding comments to the reviewed record pairs in the form of hypertext? Can pairs of groups of records be updated? Can the user "back up" or restore the possible links before committing to decisions? Can a "master" record be created which combines values from two or more records for different fields?

- 5.2 Does the software provide for results of earlier linkages (particularly reviews of possible links) to be applied to the current linkage process?
- 5.3 Does the software provide a utility for generating reports on the linked, unlinked, duplicate, and possible link records? Can the report format be customized? Is the report viewed in character mode, or is the report review done in a graphical environment? Can the report be printed? If so, what kind of printer is required?
- 5.4 Does the software provide a utility for extracting files of linked and unlinked records? Can the user specify the format of such extracts?
- 5.5 Does the software generate statistics for evaluating the linkage process? Can the user customize the statistics generated by the system?

Standardization

- 6.1 Does the software provide a means of standardizing (parsing out the pieces of) name and address fields?
- 6.2 Does the software allow for partitioning of variables to maximize the use of the information contained in these variables (for example, partitioning a phone number into area code, exchange, and the last four random digits)?
- 6.3 Can name and address standardization be customized? Can different processes be used on different files?
- 6.4 Does address standardization meet U.S. Postal Service standards?
- 6.5 Does standardization change the original data fields, or does it append standardized fields to the original data record?
- 6.6 How well do the standardization routines work on the types of names the user wishes to link?
- 6.7 How well do the standardization routines work on the addresses the user will encounter? (E.g., how well does it handle rural addresses? Foreign addresses?)

Costs

- 7.1 What are the purchase and maintenance costs of the software itself, along with any needed additional software (e.g., database packages), and new or upgraded hardware.
- 7.2 What will be the cost of training personnel to use the system.
- 7.3 What are the projected personnel and (in the case of mainframe systems) computer-time costs associated with running the system.
- 7.4 Is the cost of developing a system for the intended purposes using the software within the available budget?

Empirical Testing

- 8.1 What levels of false match and false nonmatch can be expected with the system? Are these levels acceptable?
- 8.2 How much manual intervention (e.g., possible match review) will the system require.
- 8.3 How rapidly can typical match projects be completed using the system?