

Jury Managers' Toolbox

Best Practices for Duplicate Removal

Overview

Because duplicate records most commonly occur as a result of combining two or more source lists, it is important to accurately remove duplicates when creating a source list. In fact, courts using source lists to compile a master jury list should employ several techniques, such as cleaning, merging, purging, and maintaining source lists, to identify and remove duplicate records. When applying these techniques, courts should recognize the possibility that two errors are possible. First, failing to identify duplicate records (a missed match) undermines the principle of random selection insofar that individuals who have more than one record on the master jury list (e.g., people who both vote and drive) have a greater probability of being selected than individuals with only one record. Second, incorrectly removing a record (a mismatch) on the mistaken belief that it duplicates an existing record disenfranchises a potentially eligible individual and decreases the inclusiveness of the master jury list.¹

Of two possible errors, the conventional belief is that disenfranchising a potentially eligible individual is worse than leaving an unrecognized duplicate on the master jury

list.² In fact, the NCSC recommends that the proportion of unrecognized duplicates not exceed 5% of the total list;³ this recommendation is made in the interest of creating an inclusive list, but also retaining the equality principle of selection.

Modeling after Indiana's statewide duplicate record identification protocol, the NCSC recommends several best practices for duplicate removal techniques outlined below in these four steps:

STEP 1 – Clean the data

Data cleaning standardizes comparable data elements within each list. This technique should be done prior to matching. For example, employ formatting rules such as using consistent abbreviations (e.g., Street (St.) Road (Rd.) or Lane (Ln.) and removing extraneous spaces or characters (e.g., commas, periods, dashes) to standardize fields for matching. This standardization of addresses will align the street numbers, street names and other directional information (e.g., First Street NW, 1st Street NW, NW 1st St., etc.), correct city spelling errors, and dictate the format. Standardization should be done, where possible, on other fields. For example, a

¹ G. Thomas Munsterman & Paula L. Hannaford-Agor, *The Promise and Challenges of Jury System Technology* 17-18 (2003).

² *Id.*

³ G. Thomas Munsterman, *Jury System Management* 4-5 (1996).

common error within the last name filed is detected for names beginning with “Mc.” The name, McHugh, may appear as, “McHugh” or “Mc Hugh,” with or without the space.

Ensuring "Clean" Data Input - Sample Formatting Rules for Database Entry			
Sample Category	Typical Input Forms	Recommended Format	Description
Date	March 10, 2009 Mar 10, 2009 10-Mar-09 3/10/2009 3/10/09	< MM-DD-YY > 03-10-09	Two numerical inputs for month, day and year. Separate with a "-".
Address	Street St. or St or Str. Road Rd. or Rd	< Xy > St	Two letter abbreviation. No punctuation.
Money	1357 1,357 \$1,357.18 \$1,357	< ##,###.## > 1,357.18	No dollar sign. Two decimal places. Comma separator. No leading zero.
Name	Mc Adam Smith-Jones Smith Jones	< Abcdef > < Abc-Def > McAdam Smith-Jones	Continuous alpha string. Use a '-' to maintain string continuity.
Phone	123.456.7890 (123) 456-7890 (123) 456 7890	< ###-###-#### > 123-456-7890	Include area code. Use '-' as a separator.

Name fields are particularly difficult to match across source lists. Adding an additional name fields will account for alternatives and facilitate the matching process. For all first names, create a new field with nicknames or alternative spellings. For example, store alternative nicknames of Jon or Jonathan for “John” to be used in the matching process. Another alternative is to use only first name initials. This will eliminate many potential alternative spellings and nicknames. Alternate last name fields should include iterations to account for hyphenated and un-hyphenated alternatives.

The cleaning process will also remove inaccurate data or records outside the court’s jurisdictional boundaries. Geocoding will correctly map county or other court jurisdiction boundaries. This process will find associated geographic coordinates using street addresses. With the identified geographic coordinates, addresses can be entered into Geographic Information Systems (GIS) to verify addresses. Typically, courts will purchase a Geocoding software package to perform this task.

Vendors licensed by the United States Postal Service (USPS) can also be used to verify the accuracy of addresses. They verify if an address exists and use the National Change of Address (NCOA) database to provide updated address records. Details on these services and the firms licensed to provide these services can be found on the USPS website.

STEP 2 – Merge the Source Lists (Identify Duplicates)

Once all individual lists have been cleaned, the next step is to match records and identify the duplicates. Matching should first be done on each list separately to identify duplicate records within a source list. Next, merge all source lists to form one master jury list.

The NCSC recommends that the court use relatively few commonly occurring data elements when both matching within a single source list and matching between multiple source lists. For example, matching on only a few key elements (e.g., last name, first initial, the last four digits of a Social Security Number, and date of birth) will improve your confidence in the decision to remove a duplicate. The fewer elements used for matching will reduce the probability of discovering an unknown match due to missing records. For example, if the social security number is frequently missing in the records on a source list, matching on this field will result in a higher number of unresolved duplicates.

As part of this step, the NCSC recommends that the court creates a juror identification number (ID) as a way to track records across lists when a duplicate is identified. The juror ID number should be a unique number maintained by the court which is linked to other ID numbers on the source lists. This will enable the court to trace the original source of the record and facilitate updates to records maintained in the master source list.

Storage of a juror ID assumes the court maintains previous year lists and periodically corrects or updates data. All data updates should be funneled from the local jury manager's offices to the person or agency maintaining the master jury list. For instance, Indiana uses a state-wide master jury list, so the updated juror information is funneled to the state through a computer system by the local jury managers.

Each update is assigned an effective date, similar to a date stamp, which records the date associated with juror records. The effective date is tremendously helpful in the situation for which two *different* addresses are provided from two *different* lists, for the *same* individual. It is optimal to select the record from the list that is known to contain the most recent and accurate information. However, without an effective date associated with the individual record (or the source list's creation date), it proves difficult to correctly identify which record should be removed. Use of an effective date allows the court, upon identification of a duplicate, to retain the record with the most recent information and purge the older, more likely incorrect, record.

STEP 3 – Purge the Records

Once the lists are merged together to form one master jury list, the master jury list should be evaluated for ineligible records to be purged. Before purging records, your court should develop accepted protocol for records of an individual who is under 18 years of age, possesses an out of state address, or has become inactive on the source list (e.g., has not renewed a driver's

license in a period of years, or is not a U.S. citizen). Such individuals may become jury-eligible in the future, so courts should retain these records in a separate file. Temporarily ineligible records should not be subsumed into the permanent suppression file, but retained as inactive for the current master jury list.

The master jury list records should be compared against a suppression file.⁴ Suppression files contain records of individuals who have been permanently excluded from jury service (e.g. deceased, or permanently disabled).

STEP 4 – Maintain Updated Records

The master jury list should be maintained periodically as new or updated information becomes available. The NCSC recommends that a new master jury list is created at least once a year. When the new list is created, the list will be verified against the previous year's updated list. In other words, if a master jury list is created in January of 2009 and updates are stored within the list throughout the year, in January of 2010, when the new master jury list is created, the final step is to compare the 2010 master jury list to the 2009 list. This comparison assumes that the 2009 list contains periodic updates or corrections that do not appear on the 2010 list. Recall the value of assigning an effectiveness date to records, which enables the court to select the most

⁴ Courts should be cautious of using suppression files. For a discussion of the associated hazards, see: Paula Hannaford-Agor, *Jury News: Suppression Files – Valuable Tools or Traps for the Unwary*, in 23(3) Ct. Mgr. 75 (2008).

up to date and accurate record for inclusion in the master jury list.

Conclusion

The NCSC has studied the effectiveness of duplicate removal techniques for courts using combined registered voter/licensed driver lists and advocates that duplicate rates not exceed 5%. When the matching criteria are exact matches on the last name, first name, middle initial, birth month and day, and street number or post office box number, the probability of a duplicate record being missed is approximately 6% and the probability of mistakenly removing a unique record is less than 1%.⁵ Data elements containing missing information as well as the existence of extraneous spaces, punctuation, or non-standardized formatting in any of the fields used for matching can result in an unrecognized duplicate being left on the master jury list while the use of fewer matching criteria (e.g., surname, first initial, and date of birth only) will result in fewer unrecognized duplicates. Commercial jury automation software generally employs more sophisticated (trademark protected) matching criteria, which typically results in 2% to 3% unrecognized duplicates left on the master jury list.

The NCSC Center for Jury Studies examined Indiana's state-wide master jury list system to serve as a model for this research. Indiana, per the Indiana Supreme Court *Order Approving the Master List for Jury Pool Assembly and Jury Reporting Requirements*, No. 94S00-0501-MS-19 (Nov. 1, 2005) uses two source lists (the Bureau of Motor Vehicles list and the Department of Records tax return list) to compile the master jury list. The NCSC sincerely thanks Dave Remondini and the Indiana AOC staff for sharing their duplicate record processes for the benefit of this tool.

Disclaimer: The guidelines discussed in this document have been prepared by the National Center for State Courts and are intended to reflect the best practices used by courts to identify and remove duplicate records from the master jury list.

⁵ *Munsterman, supra*, note 1, at 18-20.