

January 2026

Summer EBT Data Wrangling: A Technical Best Practice Guide for Summer 2026 and Beyond

Introduction

Summer EBT (also known as SUN Bucks), the first new federal food assistance program in over 50 years, represents a unique opportunity to help close the summer hunger gap. The program provides at least \$120 in nutritional assistance to eligible children in participating states, territories, and Tribal Nations; this benefit helps bridge the gap in food access during the summer months when children are not in school. However, there is technical complexity in ensuring that eligible children receive benefits. Some families receive benefits automatically due to their existing enrollment in another benefit program(s) (e.g., Supplemental Nutrition Assistance Program, Temporary Assistance for Needy Families, Medicaid), while others have to apply in order to get benefits. Because of the dual pathways to enrollment and the vast array of data sources that must be checked for eligibility, program administrators face significant hurdles in identifying eligible children, minimizing the burden on families to receive benefits, and sending out Summer EBT cards to the correct home.

Over the first two years of Summer EBT implementation, Code for America provided technical assistance to a range of state partners to help ease the technical difficulties of wrangling and deduplicating datasets to enable the accurate and timely delivery of benefits. Now, after two summers of supporting agencies to navigate the new program, Code for America and No Kid Hungry, a campaign of Share Our Strength, are excited to share this updated resource.

This document defines what entity resolution is, covers its applications for Summer EBT, provides recommendations on how to approach it in your agency context, and goes into greater implementation detail with case studies from three states and one district. It is intended both for Summer EBT program administrators who want to know what they should expect out of their deduplication and matching procedures, as well as technical staff looking for implementation tips and a methodology to use. Additionally, practices laid out in this document can support states, Tribal Nations, and other jurisdictions in promoting integrity in their Summer EBT program ([per USDA guidelines](#)) as they prepare Summer 2026 plans for operations and management (POMs).

Background

What is entity resolution in the context of Summer EBT?

In the context of Summer EBT, entity resolution shows up in several crucial ways:

- To identify eligible children, program administrators look across a range of data sources that are often owned by different agencies and live in different databases.
- Across datasets, records that represent the same unique child need to be linked and then deduplicated to ensure all automatically eligible children are issued benefits and that each child receives the benefit only once.
- Additionally, Summer EBT applications must be deduplicated against records for children already streamline certified for Summer EBT to prevent overissuance of benefits.
- Finally, application data needs to be linked to National School Lunch Program (NSLP) participating school enrollment data in order to verify eligibility.

Definitions

Entity resolution, record linkage, matching, and data deduplication are often used interchangeably to describe the process of combining individual records across data sources, but each term has a distinct meaning.

Entity resolution is the process of identifying the sets of elements in one or more datasets that correspond to the same unique individual. The challenge of entity resolution is that it is not always obvious if two records represent a single individual or two different individuals with similar information. For instance, are Thomas L. Jones and Tom Jones the same person, or different people? In the case of Summer EBT, agencies must face the unique challenge of not only linking individual children's records, but often also grouping eligible children with appropriate guardians into households.

Record linkage and deduplication are entity resolution outcomes that differ in their management of corresponding sets. In record linkage, two or more datasets are searched for records that correspond to the same entity. When two records are identified as a match, the process creates a single combined entity for use in a future process. In deduplication, a single dataset is searched for records that represent the same entity, and matching records are merged into a single entity within the dataset.

When measuring the performance of an entity resolution system, two key metrics to consider are precision and recall. *Precision* is a measure of the accuracy of identified matches. A system that minimizes false positives has high precision. *Recall* is a measure of how well the system identifies true matches. A system that minimizes false negatives has high recall. Often, there is a tradeoff between these metrics, such that minimizing one causes the other to rise

Process

Designing an entity resolution process ideally involves the following steps:

1. Prepare the dataset(s) to be searched; common sources include: SNAP, TANF, Food Distribution Program on Indian Reservations (FDPIR), Medicaid, school district data on free and reduced price meals (FRP), and foster care system data.
2. Consider the harms of errors (false positives/negatives), choose one to strictly limit, and decide on the limit for that.
3. Create a validation dataset that can be used to evaluate the error rates of candidate algorithms.
4. Try to combine different algorithms that meet the error criteria, picking an overall strategy that minimizes the unconstrained error rate.
5. Use the algorithms on live data for record linkage or deduplication.

Key considerations

The choice of entity resolution algorithm depends on the consequence of its errors. There are two types of errors in linking records:

- **False positive (incorrect match):** Two records representing *different* people are mistakenly resolved to a single person.
- **False negative (missed match):** Two records representing the *same* person are mistakenly resolved to two different people.

Typically, there is a tradeoff where algorithms that try to minimize false positives do so by risking more false negatives, and vice versa. The consequences of these errors vary by context:

Context	False positive	False negative
Streamlined certification deduplication	A person is incorrectly denied benefits	A person is overissued benefits
NSLP enrollment record linkage	An applicant is incorrectly issued benefits	An applicant is incorrectly denied benefits

Importantly, the data sources available, the quality of the data in each source, and the process for linking and deduplicating records varies by state. Technological capacity, participation of vendors, and other contextual factors also vary, meaning that no two entity resolution processes for Summer EBT look the same. matches, including one completely new result, row 1 to row 3

Technical deep dive

Data formatting: Cleaning data and common issues

All entity resolution strategies depend on what data is being used. In the context of Summer EBT, there will generally be a specific set of fields to which you will have access within school data. Certain combinations of fields can do much to improve the quality of entity resolution, while others may not have a major impact. For example, approximately 87% of individuals in the U.S. can be uniquely identified simply using their gender, ZIP code, and date of birth¹. It is critical that your overall strategy does not rely too heavily on any one field. For instance, a combination of linkage algorithms that all include exact matches on the “date of birth” field ultimately will exclude someone who had accidentally input an incorrect date of birth. One minor error in one piece of data should not be the reason a child fails to receive Summer EBT benefits.

Using varied and flexible linkage strategies will identify true matches in the data while allowing for honest data entry mistakes on the parts of parents and administrators. There are many reasons why information might not be accurately recorded; a parent may mistakenly enter a child’s date of birth incorrectly, transpose their child’s last and middle names, or include nicknames rather than full names in some places and not others. To ensure children receive Summer EBT benefits, you must take these discrepancies into account when conducting the data linkage.

Detailed examples of common data discrepancies can be found in Appendix A.

Considering the context

It is impossible to measure an algorithm’s error rates (false positives/negatives) without testing it against a dataset with **known correct answers**. Creating this kind of test dataset can be time consuming, if it is even possible at all with the data sources typically used for Summer EBT.

Our approach:

To work around the limitations in Summer EBT data sources, similar datasets with reliable identifiers (like repeat SNAP applications with Social Security numbers) can be used as a stand-in for evaluating algorithms. The linking algorithm is applied to the same fields (names, date of birth, etc.) and evaluated using the reliable identifier as the truth.

Keep in mind:

This approach is most reliable if the test population is similar to the population the algorithm will ultimately be used on.

To select an algorithm, weigh the relative harm of each mistake. Minimize the risk of the lesser mistake, and constrain the risk of the greater mistake below an acceptable threshold.

¹ <https://dataprivacylab.org/projects/identifiability/paper1.pdf>

Example (streamlined certification):

The **greater harm** is typically a false positive (i.e., a child is incorrectly denied benefits).

1. Decide on an acceptable limit for the greater harm (e.g., the false positive rate cannot exceed **1%**²).
2. Test several algorithms and eliminate any that exceed the 1% limit.
3. Of the remaining algorithms, choose the one that has the lowest false negative rate (i.e., it overissues benefits the least).

Household grouping

In addition to resolving records to accurately identify individual children, it is commonly required to group children into households to simplify benefits management for guardians. This can be accomplished by examining and comparing attributes that are likely to overlap within a household, such as last names, addresses, guardian names, and/or program case numbers. As with resolving individual child records, there are tradeoffs in potential harms to consider when selecting an algorithm. A false positive (linking a child to an incorrect guardian) could result in providing access to a child's benefit information to an unauthorized person, depending on the presence of other privacy controls. A false negative (not linking a child to their correct guardian) could result in the guardian experiencing undue friction when trying to manage their child's benefits.

Entity resolution strategies

There are two main strategies to linking records in an entity resolution process: rules-based and probabilistic. Within both strategies, there are different matching approaches that are used to combine associated records. An **exact matching** algorithm requires identical information in the data fields being matched; for example, such an algorithm would likely not identify Sarah S. Smith and Sarah Smith as a match due to the missing middle initial. In contrast, **fuzzy matching** algorithms find a set of characters (sometimes referred to as "strings") that match each other approximately, rather than exactly. There are many algorithms that can be used for fuzzy matching, but most rely on some variation of string edit distance.

The **string edit distance**, also known as the *Levenshtein* distance, is the number of deletions, insertions, or substitutions between two strings, and setting an edit distance threshold determines how fuzzy your matching criteria will be. For example, an edit distance of 1 would identify "Jonas" as a possible match for "Jones" because there's only one character difference between the two. An edit distance threshold of 0 only allows for exact matches. Other fuzzy matching approaches—including variations on string edit distance and phonetic comparison—can also be used, and some common approaches are listed below:

² This is for the purpose of discussion. At the time of writing, no official maximum allowable error rates have been published by FNS. In a 2021 study, FNS estimated that certification error rates for CEP, which is similar to streamlined certification, were 2.7%.

- **Damerau-Levenshtein distance** is an extension of the *Levenshtein* distance that allows for transpositions, matching “Davis” with “Davi.”
- **Jaro-Winkler** is a string similarity metric that measures the likeness of two strings on a scale of 0 to 1, with 1 indicating identical strings and 0 indicating no similarity. It is a modification of the *Jaro distance metric* that weights strings with common prefixes as more similar. This weighting can make it especially helpful for matching names.
- **Metaphone** is a set of phonetic algorithms used for encoding the pronunciation of words, allowing for fuzzy matching of strings that sound similar when spoken. The original version of the algorithm is focused on common English pronunciation, while later variants (Double Metaphone and Metaphone 3) improve performance for variant pronunciations, words derived from other languages, and names commonly found in the United States.

When developing an entity resolution strategy, it helps to begin with a test dataset to check exact matches before moving on to fuzzy matching algorithms, evaluating the quality of the match at each step in terms of false positives and false negatives. By relying on approximations instead of exact matches, fuzzy matching will allow more positive links overall, both false and true. Using multiple algorithms in an iterative fashion allows one to evaluate new matches as you go so that you don’t have to sift through previously-identified matches at each step. Monitoring edge cases is also important; for example, with middle names, it is possible to match against an initial, which could lead to a high number of false positives. In some cases, you may want to constrain these matches to only work for names that are at least three or four characters long. This will help ensure your command is not incorrectly linking a very short first name, such as “Bo,” with a longer different name, like “Bobrow.”

Blocking and computational complexity

Probabilistic entity resolution and fuzzy matching approaches can be computationally costly, especially over large datasets. For states with large populations, examining all possible pairs of records using computationally costly comparison functions may end up taking a prohibitive amount of time or computing resources. One common strategy is to “block” the comparisons by first requiring an exact match on a simplified element, and only looking for links within those blocks. For example, blocking on the first letter of the last name would only look for matches in the pairs of children whose last name begins with “A,” then “B,” then “C,” and so forth. It is possible to further reduce the run time by blocking on multiple elements simultaneously.

Linkage over time

Examining the dates of records can help you determine whether they refer to the same entity. For example, by paying attention to the start and end attendance dates for school records, one can reasonably conclude that—despite the names on them being the same—records that show simultaneous attendance at two different schools likely refer to two different people.

As long as a record exists only in a particular data set for a single time period, calculating the overlap in time between datasets is relatively straightforward. If a student leaves and comes back, or if records are already merged, comparing multiple time intervals for overlap can be somewhat tedious. A simpler approach is to keep these records separate and simply note the link between them.

Probabilistic record linkage

Agencies that are already familiar with the basic record linkage strategies outlined above can implement probabilistic record linkage using open-source packages available for popular programming languages such as [reclin2](#) in R or [splink](#) in Python. Rather than manually specifying a sequence of strategies that depend on distance thresholds between fields, probabilistic record linkage automatically learns how to combine similarity measures between fields into a single numeric measure, which can be used to decide on potential matches. Such software packages also work with more advanced comparison functions between fields, such as preprocessing names into phonetic representations.

Completing entity resolution

When designing an entity resolution process, you should consider how to represent data once the process is complete. There are multiple approaches with different tradeoffs to weigh depending on the context:

- **Approach 1:** Determine which attributes from the source data to retain, create a single representation of the entity in the dataset, and discard other attributes from the original source. While this may seem straightforward, there can be undesirable consequences. For example, future data that needs to match a deduplicated record set won't be able to target the data that have been discarded and valuable fallback information might be lost in the case of an unreachable client.
- **Approach 2:** Alternatively, after selecting attributes to keep and creating a uniform representation of an entity, links to the source records for each entity can be preserved. New data can be resolved against known entities by comparing to both the final and source records for each entity, using rules such as based on whether any two records match or the average distance between pairs.

Regardless of whether all of the source data is kept, there must be a strategy for deciding which data should be representative of an individual. In particular, it is important to establish a preferential order for mailing addresses. Simply preferring one source over another isn't always ideal; for example, a SNAP address that was valid for just the first two months of a new school year is not inherently better

than an FRP certification made by application at a new address in the spring. Ultimately, the goal is to establish a manner of ordering all addresses. For example:

- An address that was supplied more recently comes before one that was supplied earlier
- In the event of ties, addresses are potentially resolved in order of: SNAP, TANF, Medicaid, FDPIR, categorical eligibility program, and FRP certification.

Monitoring outcomes

Tracking the success of various algorithmic approaches and fine tuning your methods is an important part of improving entity resolution processes over time. Given the nature of record linkage and deduplication, technical work is constantly evolving. However, it is also critical to monitor the actual outcomes of your overall Summer EBT program. The following metrics can help you identify potential issues with your entity resolution processes or algorithmic selections and inform key stakeholders about how implementation of the program is going. Some common metrics to consider tracking include:

- Overissuance rate
- Underissuance rate
- Time between eligibility data collection and Summer EBT benefits issuance
- Percentage of records requiring manual human review

Case studies

These case studies are derived from in-depth interviews conducted with key technical and program administration stakeholders in Washington, DC, Ohio, Wisconsin, and New Jersey. Each interview lasted for 30 to 60 minutes, with questions covering three broad categories:

- Data sources and cleaning
- Technologies, algorithms, and matching methods
- Monitoring, outcomes, and lessons learned

By diving into these topics with our partners, we were able to gain a deeper understanding of how Summer EBT implementation is progressing two years in, sample a comparison of technical choices across the Summer EBT landscape, and learn about each state's key challenges and successes.

While the insights gained from these case studies may not provide definitive conclusions about ideal technical choices, the states interviewed were selected because they have made notable improvements over the first two years of implementation. We want to share the valuable learnings they've generated for other states seeking to improve their Summer EBT entity resolution processes.

Case study #1: Washington, DC

Preserving source data relationships and simplifying manual review

Introduction

To deliver Summer EBT (S-EBT) benefits to eligible children across the district, Washington, DC incorporates data from a variety of agencies and sources using a record linkage solution custom built by the data team at the Department of Human Services (DHS). This agency operates and provides data for the SNAP and TANF programs. Additional eligibility data for Medicaid is provided by the Department of Healthcare Finance (DHCF); data for Free and Reduced Price Meals (FARM) and NSLP-participating school enrollment is provided by the Office of the State Superintendent of Education (OSSE). Data used to validate and standardize DC mailing addresses is provided as a web service by the Office of the Chief Technology Officer (OCTO). DHS and OSSE are the two primary public agencies that collaborate to provide S-EBT benefits, with DHS taking the lead on program administration for the summer 2025 benefits period. Two years into the program, DC's data team has built a sophisticated rules-based entity resolution system and is continuing to refine it to increase data integrity and reduce manual review.

Record linkage and deduplication

Data sources, quality issues, and cleaning

DHS and OCTO provide the technical infrastructure for the administration of S-EBT. Data from multiple sources is linked to determine eligibility and create a complete data record for each eligible child:

- SNAP/TANF enrollment
- Medicaid enrollment
- FARM enrollment
- Categorical eligibility sources
- S-EBT applications

Common data quality issues include potential duplicate entries from historical SNAP/TANF records, out-of-date addresses, inconsistencies in formatting, and misunderstanding of manually entered data resulting in issues like month/day transposition in birthdates. Data cleaning is automated and includes the following processes:

- Removing diacritic marks to normalize alphabetic characters, and removing hyphens and extra whitespace
- Normalizing all alphabetic characters to lowercase
- Splitting multi-part names into tokens on space/hyphen delimiters and storing them for later comparison

- Computing double metaphone codes on all name parts for phonetic matching
- Standardizing and correcting addresses using OCTO data and a commercial address service

Data sharing takes place across the following agencies and vendors: DHS (SNAP/TANF), DHCF (Medicaid), OSSE (FARM, NSLP enrollment), and Alchemer (vendor for public S-EBT application).

Matching and deduplication process

Several of DC's source datasets contain unique record identifiers (SNAP/TANF person record number, student ID, etc.), but there is not a reliable unique identifier across the data sources. The DC data team introduced an S-EBT unique identifier (UID) that is assigned during the entity resolution process and linked back to source records to improve system performance and increase data integrity in subsequent record linkage and deduplication runs. To perform initial linkage and deduplication of records without an assigned UID, several primary data attributes are used for matching, with secondary attributes used to guide resolution of possible matches:

- **Primary:** Core PII (first and last name, date of birth, SSN, local school ID)
- **Secondary:** Address, email, phone number, gender

Additionally, as a performance optimization, they perform blocking using several attributes (including name metaphone codes) to reduce the size of datasets being compared for possible matching.

Matching is performed with T-SQL scripts by comparing exact and fuzzy matches (using edit distance and phonetic comparisons) and computing a match confidence score using a tunable, weighted algorithm. The algorithm's output is a 0-100 match confidence, with 100 representing an exact match. A confidence score below 80 is considered a non-match, 95-100 is a match, and 80-95 a near match requiring manual review.

Manual review

The data team at DHS has worked to tune the entity resolution system to reduce the number of near-matches needing manual review. For the summer 2025 benefit period, about 1,500 records required review. The team created a Python-based tool to facilitate manual review by staff, presenting the details of each source record side by side for easy comparison. Ordinarily, a record takes 30-60 seconds to review. Once the database is initially populated and data updates begin flowing into the system, there are typically only five to ten manual reviews required daily. The DC team is also considering future adoption of AI-based tools to handle some near match cases, enabling them to tune for even greater accuracy without adding undue manual processing burden.

What worked, what's next

Two years into the program, and only one year into DHS leading its administration, DC has built an effective system for managing their S-EBT streamlined certification and deduplication processes. Key enablers of success have been the introduction of a record UID linked back to source data primary keys, a sophisticated rules-based matching algorithm, and a custom tool to streamline the manual review process for near matches. A continuing challenge is the quality/validity of upstream data, which could be alleviated by improved validation of data as it's input into source systems.

Case Study #2: Ohio

Revamping data cleaning and matching criteria to minimize overissuances

Introduction

Ohio has a number of data sharing agreements across key agencies to facilitate entity resolution that identifies eligible children for S-EBT. Currently, data is shared by the Department of Education and Workforce (DEW), the Department of Jobs and Family Services (JFS), and Medicaid. In 2023, the Ohio legislature created a new Department of Children and Youth (DCY), which is another key stakeholder. The Management Council, a nonprofit entity that manages school data alongside Ohio's Information Technology Centers, leads the data standardization and entity resolution processes for Summer EBT. Once the program has concluded for the year, the technical implementation team receives a file with all benefit issuances, which they use to check for duplicates. This internal review process has led to identification of important challenges and solutions to reduce error.

Record linkage and deduplication

Data sources, quality issues, and cleaning

The agencies listed above share data with the Management Council. The data they provide covers key federal programs such as SNAP, TANF, and Medicaid. The online S-EBT application is another source of data along with data on eligibility for free and reduced meals provided by school districts. When SNAP data is available for a given record, it's the primary source against which other data is checked, as it usually contains the most current information due to how frequently families update their address through the program.

In 2024, data formatting was a major source of error, with differing punctuation in names, variations in address formats, as well as human error with misspellings or transposition of numbers. Ohio prioritized addressing these issues to streamline their data standardization process; now, they have a largely automated, scripted approach that removes all punctuation (except periods) and white spaces from each data source prior to matching. Gender is reported differently across data sources, so that has presented a challenge as well.

Matching and deduplication processes

Ohio uses SQL with custom scripts to conduct rules-based entity resolution. They run multiple scripts to identify and handle duplicate records, with some requiring manual review based on matching criteria. They also have scripts to identify edge cases such as twins. The overall matching process is:

- Exact matches are processed automatically, using the first 6 characters of first and last name, date of birth, gender, the first 8 characters of the mailing address, and ZIP code.
- If the addresses and ZIP code do not match, records are considered unique.

The team has improved their match criteria over time to address various challenges with different sources truncating data using different rules, as well as varying data quality across sources. Ohio maintains a database of known duplicates that is used as an additional check in deduplication.

Manual review

After the data is standardized and loaded, one staff member preprocesses the data. They run a script on all remaining partial matches (around 40,000 records) that helps them sort through different types of potential duplicates. A team of three people review this list and refine it using additional scripts over several days. At the end of this process, they are left with several hundred records that are manually reviewed in detail.

What worked, what's next

Ohio has implemented an iterative improvement process to increase their match rates and reduce overissuances. After the first year of implementation, they found that overissuances (where children were receiving double benefits) were common. They refined their data standardization and match criteria to reduce this issue, and after the second year of S-EBT, only had a handful of overissuances. Ongoing areas of growth include securing additional data sharing agreements to bring in more records that can help increase the match rate as well as improving data quality for addresses. Ohio is currently evaluating a vendor solution to help with address standardization.

Case Study #3: Wisconsin

Automating data cleaning and building on an integrated eligibility system

Introduction

In Wisconsin, a range of stakeholders work together to deliver S-EBT benefits to all eligible children statewide. An external vendor, Deloitte, is the primary technical implementer. Deloitte has a long-running relationship with the Department of Health Services (DHS), which administers FoodShare (SNAP), and the Wisconsin Department of Public Instruction (DPI), which runs school nutrition programs. DHS and DPI are the two primary public agencies that collaborate to deliver S-EBT benefits. Because of this long-running

relationship and the state's prior experience with delivering Pandemic EBT, they've been able to build on existing infrastructure to help with record linkage and deduplication as well as streamlined certification. Two years in, they're continuing to refine a rules-based approach to entity resolution and are focused on reducing the number of manual reviews. Wisconsin was one of the first states to provide a self-service portal for S-EBT, which allows families to share information with the state for eligibility determination purposes.

Record linkage and deduplication

Data sources, quality issues, and cleaning

Deloitte provides the technical infrastructure that supports the rollout of S-EBT, which is built on the state's existing integrated eligibility system, CARES. There are three levels of data consulted to identify eligible students:

- **Primary:** Integrated eligibility system (known as CARES and includes SNAP, TANF, kinship care and foster care system data, direct certification, certain categories of healthcare)
- **Secondary:** School data, including NSLP participation, provided by DPI
- **Tertiary:** Online S-EBT application data

Common data quality issues include duplicate entries that result from misspelled names or nicknames being used, missing addresses for parents/head of household, as well as incomplete information for children who enter only through the S-EBT application and are not already in one of the other administrative data sources such as the integrated eligibility system or school data provided by DPI. Data cleaning is largely automated, particularly for any data in the integrated eligibility system, but these cleaning processes are fine tuned each year based on prior year performance.

Data sharing takes place across the following agencies and vendors: DHS (SNAP/EBT), DPI (school nutrition), DCF (direct certification/child welfare), Deloitte (primary vendor), Nutri-Link (public application host). Key stakeholders have prioritized cross-agency collaboration to enable the successful uptake of Summer EBT.

Matching and deduplication process

Wisconsin's integrated eligibility system contains a unique identifier, called a master customer index (MCI), which is a non-PII field used to identify children across the various datasets. Deloitte has developed a rules-based, hierarchical algorithm for matching records. First, they rely on data from the integrated eligibility system and use the MCI to match records across data sources. If an MCI is not available, the algorithm ranks using first and last name as well as date of birth, based on data source. If a student appears in multiple sources, the higher priority data source "wins." Additionally, there are special case rules for students in foster/kinship/out-of-home care: benefits go to current placement, overriding other matches in the eligibility system. These rules are tested, reviewed, and fine tuned on an ongoing basis.

Matching is done in batches, with the vast majority of records processed in one load (approximately 90-

95% of records). Subsequent data that comes through is matched in a daily or monthly run, though the volume is low following the initial processing. Proportionately, most records that are flagged for manual review come out of the first batch run; last year there were approximately 3,000 records that required manual review during the initial run.

Manual review

The rate of non-matching has decreased each year since implementation of Summer EBT began, but typically there are a few thousand cases (out of approximately 450,000) that require manual review by DHS staff. Deloitte recently added a backend feature into the state's integrated eligibility system that allows staff to easily review records and determine whether they are a match. Deloitte conducts load testing on production data to help anticipate the volume of discrepancies that will require manual review by agency staff.

What worked, what's next

While Wisconsin continues to refine their processes and algorithms for entity resolution, two years in they have a well-functioning system that enables them to serve hundreds of thousands of students each summer. They've focused on improving how they share data across agencies to decrease the wait time for benefits and minimize manual intervention to facilitate the program. Extensive partnership with other agencies has enabled a streamlined system that processes the majority of cases in an automated fashion and makes it easy for agency staff to complete manual review when necessary. The Summer EBT self-service portal for families, paired with the broader integrated eligibility system's mobile app and desktop application, have reduced call center volume because recipients can easily provide and access information about their case. Currently, they're tackling missing or uncertain address information, which is a common challenge across states we spoke to.

Case Study #4: New Jersey

Synthetic data and an iterative, probabilistic strategy

Introduction

New Jersey's Summer EBT program is led by its Department of Agriculture (NJDA), in collaboration with the Department of Human Services (DHS) and Department of Education (DOE). The NJDA additionally coordinates with the Department of Children and Families (DCF) for foster care data and local education agencies (LEAs) across the state for school enrollment data. The New Jersey State Office of Innovation (NJ OOI) also provides significant technical support and the Governor's Office has assisted with administrative coordination across agencies for operating the program effectively. Two years into S-EBT implementation, New Jersey has taken a sophisticated approach to entity resolution, using probabilistic linkage methods and a test-driven, iterative approach to continuously improve their delivery of summer nutrition benefits to over 500,000 eligible children.

Record linkage and deduplication

Data sources, quality issues, and cleaning

NJDA and NJ OOI provide the technical platforms for the administration of S-EBT. Data from multiple sources is linked to determine eligibility and create a complete data record for each child:

- SNAP/TANF enrollment
- Medicaid enrollment
- Foster enrollment
- FRP school meals enrollment
- Categorical eligibility sources from LEAs
- Combined FRP School Meals and Summer EBT applications from LEAs

Common data quality issues include transpositions in birthdates, inverted first and last names, nicknames and culture-specific issues (e.g. Hebrew names in one data source but not another), invalid and non-standardized addresses, and general errors arising from manual data entry. Data cleaning and preparation is automated and includes the following processes:

- Normalizing characters with diacritic marks (accents), removing symbols and extra whitespace, standardizing case
- Handling titles (e.g. Mr. or Ms.) encountered in name fields
- Ensuring semantic alignment between attributes from different data sources
- Data sharing takes place across the following agencies: NJDA, DHS (SNAP/TANF, Medicaid), DOE, DCF (foster), and LEAs across the state.

Matching and deduplication process

New Jersey uses a hybrid strategy incorporating both rules-based and probabilistic approaches to entity resolution. The first layer of the process is rules-based, implemented using SQL database procedures, searching for records that are clear exact or near matches (i.e. string edit distance ≤ 1) on core attributes (e.g., names, date of birth, address). The remaining records are processed probabilistically using the open-source [splink](#) library. The model's column weights and decision cutoff points were tuned using a combination of manually-labeled real data and synthetic data representative of real-world conditions, including simulated errors. The team designed a suite of test cases based on known real-world, challenging scenarios (e.g. families with similarly-named twins) which was used for initially tuning the model and guarding against regressions as new data and scenarios are added over time. Amazon SageMaker and Jupyter notebooks were key tools that enabled rapid building and iteration of the probabilistic system.

Manual review

The New Jersey team initially used significant amounts of manual review as they built their probabilistic model, allowing them to both manually verify records that were possible matches and label those records for use in subsequent model testing. However, the manual review workload has declined significantly with the adoption of probabilistic models.

What worked, what's next

Two years into the program, New Jersey has built a sophisticated system for entity resolution in S-EBT, with accuracy verifiable via robust test suites, manual review minimized, and a process in place for continuous improvement. Data science enabled the automatic enrollment of approximately 100,000 additional children each year since implementation began. Arriving at this point has not been without obstacles; the team cited cross-agency coordination, frequent address changes for clients, and inconsistent validation in source data systems as central challenges. They highlighted several key enablers of their successes: deep knowledge of their data domain, executive buy-in and support, and a skilled project manager to coordinate work across the many stakeholders involved in their program.

Common themes

Upstream data quality: a universal challenge

Source data quality remains a challenge for all of the states we spoke with, although many have made notable progress in reducing errors due to data quality issues over the first two years of implementation. Each state uses some form of automated cleaning, though the approaches differ. Up-to-date addresses are a key pain point when it comes to entity resolution. Most states have a perspective on which data source contains the most accurate address information, and have a hierarchical system for identifying the right record. While there are commonalities in this hierarchy, the exact method varies from state to state.

Pathways to matching

Three of the states we spoke with—Ohio, Washington, DC, and Wisconsin—use a rules-based approach to link records. All three rely on a blend of exact and fuzzy matching algorithms and update their matching processes each year. Typically, they start with exact matches and then progress through fuzzy matching, followed by manual review for those records that are still unresolved. DC has built a Python-based tool to assist in speedy review; in Wisconsin, Deloitte built a tool for manual review that is housed inside the integrated eligibility system. New Jersey deploys more advanced methods, drawing on both rules-based and probabilistic methods.

Barriers to measuring performance

One noticeable area for growth is for states to begin measuring and monitoring the performance of the algorithms they use. Generally, states are checking match rates, and in Ohio they have focused on reducing overissuances, but more sophisticated measures of performance, such as precision and recall, are not currently deployed. To help improve performance of entity resolution processes in 2026, we believe additional education and technical assistance could help states develop and implement performance measurement tactics in alignment with USDA guidance around promoting [integrity in Summer EBT programs](#). This type of work is particularly in sync with respect to developing and maintaining an accurate master issuance file, as well as efforts to detect and prevent dual participation in other states/Indian Tribal Organizations.

Advice from the field

When asked what has been most crucial to improvements over time, states emphasized the importance of data quality and integrity; understanding the broader data landscape across sources and agencies; knowing the unique characteristics of the data each state works with; and ultimately, partnerships. Having buy-in and alignment in goals and purpose across stakeholders has been central to each state's ability to refine their approaches over time.

Appendix A

Examples of common discrepancies

Name

- First name and middle name combined
- Last name hyphenated vs. combined into one name
- Middle name included in one place and not another
- Jr. or Sr. included or not included
- Nicknames used in some places, full names in another
- Special or accented characters used in some places and not others
- Names transposed, e.g. first and middle or middle and last names swapped
- Middle names different in different datasets
- Sometimes name fields have character limits, so a very long name might not be included in full in one dataset vs. another

Date of Birth

- Dates transposed, e.g. month and date swapped
- Siblings' birth dates swapped
- Wrong year listed due to a simple typo

Address

- Address could differ across datasets because a child has moved and the data in one set is old or hasn't been updated
- St. vs. Street, etc.
- Apartment number included vs. not included

School Names

- The name of a school can be written many ways, and could differ across datasets
- People tend to refer to schools colloquially, rather than by their formal names; e.g. "central high" or "Small Town elementary"
- Because of this, if possible it is ideal to match using a unique school id rather than the name of the school

Parent or Guardian Name

- Parent or guardian name fields present the same matching challenges as student name fields (see above)
- Sometimes a different parent or guardian is listed in different datasets, e.g. in school data vs. SNAP system data

Appendix B

Detailed example

Here's an example of how we at Code for America evaluated Pandemic-EBT (P-EBT) application matching criteria on a test dataset derived from repeat SNAP applications.

Table 1: True and false positive rates as fuzzy matching criteria expand

Allowed distance		First and last name included		Last name excluded	
Name fields	DOB, ZIP	False positive rate	False negative rate	False positive rate	False negative rate
0	0	0.019%	24.7%	0.229%	23.2%
0	1	0.032%	18.6%	0.243%	17.0%
0	2	0.104%	12.0%	0.314%	10.5%
1	0	0.021%	23.3%	0.278%	22.5%
1	1	0.035%	17.2%	0.291%	16.4%
1	2	0.106%	10.6%	0.363%	9.8%
2	0	0.039%	23.1%	0.588%	22.4%
2	1	0.052%	16.9%	0.601%	16.2%
2	2	0.124%	10.4%	0.673%	9.7%
3	0	0.164%	23.0%	2.455%	22.3%
3	1	0.177%	16.8%	2.468%	16.1%
3	2	0.249%	10.2%	2.540%	9.6%

In this table, “Allowed distance” is the edit distance—the number of single character changes one would have to make to turn one string into another. A larger allowable distance implies a fuzzier match.

This example illustrates the importance of trying different strategies. In the first pair of results columns, matches are made on first name, last name, date of birth, and ZIP code. When using these fields, even after allowing for a substantial number of errors to be made, it's impossible to lower the frequency of failing to make matches below 10%. The second pair of results columns is obtained by excluding last name entirely from analysis (note: this is purely for example). In this set, there are now multiple possible approaches which bring the false negative rate below 10%.

If it has been decided that the maximum allowable false positive rate that meets program goals is 1%, we would choose the matching strategy that excluded last names with edit distances of two for the name field and date of birth / ZIP code, as that set had the lowest false negative rate while respecting the constraint. Even if no specific threshold is included in regulations, tradeoffs like this allow you to be explicit about how your linkage strategy will make various kinds of errors and can be helpful when reporting to stakeholders.

Complete example

This is an example of a linkage strategy that Code for America ran in a state to match student records in one dataset against another to verify their eligibility for P-EBT benefits. The table below details each step, the number of matches captured with each step, and the percentage of the total number of identified matches that each step contributed.

This strategy included 14 individual steps, with the comparison getting fuzzier each time. Most steps capture fewer possible matches than the one before, but all contribute to the total number of matches. Had we stopped at exact matches on first name, last name, date of birth, and gender, we would have missed out on 16.1% of the matches identified.

Step / Strategy	Step description	Matches Captured	Percent of Total
1	Exact match on FN, LN, DOB, gender	103,515	83.9%
2	Exact on DOB & Exact on LN & FN or MN contained	6,070	4.9%
3	Exact on DOB & LN contained & FN or MN contained	4,891	4.0%
4	Exact on School ID & LN contained & FN exact & gender	4,030	3.3%
5	Exact on DOB & LN contained & FN LevDist =< 2	1,345	1.1%
6	Exact on DOB & LN LevDist =< 2 & FN or MN contained	881	0.7%
7	Exact on School ID & LN LevDist =< 2 & FN or MN contained	560	0.5%
8	Exact on DOB & LN LevDist =< 2 & FN LevDist =< 1	445	0.4%
9	Exact on School ID & DOB LevDist =< 2 & LN contained & FN or MN contained	202	0.2%
10	Exact on School ID & DOB LevDist =< 2 & LN LevDist =< 1 & FN Dist =< 2 & gender	160	0.1%
11	Exact on DOB & FN contains LN & LN contains FN (either direction is a match)	95	0.1%
12	Fuzzy name & DOB	1,010	0.8%
13	Fuzzy name & DOB & school ID	132	0.1%
14	Fuzzy name & school ID	5	0.0%
	Total matches	123,341	100.0%