Linking Records Across Data Systems, Part 4: Prospects for *Ad Hoc* Linking

A *How NCLDS Works* Brief for NCLDS Contributors and Stakeholders

Enterprise Data Office

Version 1.0 January 2026

Table of Contents

The How NCLDS Works Series

This brief is part of a series that provides details for North Carolina Longitudinal Data Service (NCLDS) users, NCLDS Data Contributors, and other stakeholders about how various technical and procedural aspects of NCLDS and the systems that contribute data to NCLDS work. The briefs focus on aspects that are not easily explained in a paragraph or two.

Each brief has been written in a way that we hope will make it accessible even to audiences without data, analysis, or technical backgrounds, but please share feedback with us about how we can make the briefs more accessible. We are also open to suggestions for other topics you would like to see covered. We can be reached at NCLDShelp@nc.gov.

Currently Available Briefs

- Linking Data: eScholar Student UID
- Linking Data: NC eLink Entity Resolution
- Linking Data: Workforce Data
- Linking Data: Prospects for *Ad Hoc* Matching (this brief)
- Making Data Requests

Planned Briefs

- Common Elements across NCLDS Data Sources
- Data Availability and Use Guide
- Using the Public Version of the NCLDS Data Dictionary
- Reviewing Data Requests
- Fulfilling Data Requests
- Reviewing Products Created by External Partners with NCLDS Data
- Cross-Sector Governance of NCLDS
- Security and Privacy
- Creating Practitioner Portals
- Cross-Sector Learning Goals

1. Purpose of this Brief

One key component of the usefulness of NCLDS is the availability across NCLDS data sources of reliable and up-to-date **record-level identifiers**. A **record** is a term used to describe information about an individual that is collected in a single row of data. A person can have many records—for example, some may contain education information, and others may contain employment information. Records often include **identifiers**, or special data points that help determine the identity of the individual whose data are included in that record. Identifiers help NCLDS connect separate records of data to each other (for example, a person's high school academic outcomes record and the same person's postsecondary course enrollment record). Without these identifiers, important data may not be included in analyses that assess the longer-term (across-time) value and impact of policies, programs, and supports.

To prepare a dataset for a Requester whose project has been approved by NCLDS's **Data Contributors** (the agencies and organizations that make data available to NCLDS Requesters), NCLDS conducts many preparation steps, such as retrieving the approved data, securely packaging it, and assessing the security of the location to which the data will be sent. This brief highlights a key step in NCLDS's preparation process: The stage during which NCLDS uses record-level identifiers to generate the links that a Requester receives as part of the approved dataset.

As described in other *How NCLDS Works* briefs (all linked on the previous page), in addition to standard personal identifiers often used for linking purposes (identifiers such as first name, last name, and date of birth), NCLDS also has access to several important numerical IDs, including: the North Carolina Department of Public Instruction's (NCDPI) eScholar Student Unique Statewide Identifier (UID), which is attached to most education records; the Government Data Analytics Center's (GDAC) NC eLink Entity Resolution Tool Identifier (eLink ID), which is attached to many of the records available to NCLDS; and, for some records, a Social Security Number (SSN), which is important for connecting records to information about wages. There also are a few other numerical IDs NCLDS may use in future iterations of its linking process.

An Important Note about Privacy

NCLDS only uses record-level identifiers during the linking stage of its data preparation process. **Once links are established, all record-level identifiers are stripped from the dataset and are not shared with Requesters.** A Requester only receives a unique ID generated by NCLDS specifically for that Requester's dataset; NCLDS-generated IDs have no meaning outside of the context of an NCLDS data request and are not retained or used in any other dataset prepared by NCLDS.

While it is very helpful to NCLDS to have access to so many different identifiers, using them effectively sometimes can be complicated, especially when there are differences in identifiers across records that belong to the same person—not only in the values of those identifiers but also in their quality, completeness, and reliability. The sections below describe how NCLDS currently manages its linking process. This brief will be updated regularly as NCLDS continues to improve this process.

2. Types of Linking Data

NCLDS refers to a set of linked records as an **entity**, and the process of linking those records to form that entity as **entity resolution**. The types of data used to link records often are referred to as **Personally Identifiable Information** (PII). Each individual piece of PII (such as First Name, Last Name, Middle Initial) is referred to as a **variable** or an **element**. These data elements are used for linking because they provide relatively stable information about an individual that is commonly collected by most agencies, organizations, and programs.

The most common elements used to link individuals across datasets are:

- Names (First Name, Last Name, Middle Initial, Middle Name, Prefix, and Suffix)
- Identification numbers (for example, Social Security Number,[1] eScholar Uniq-ID,[2] and Individual Taxpayer Identification Number)
- Addresses (for example, Street Address, City, State, County, and ZIP)
- Phone numbers and email addresses
- Linking information from other data-linking services (for example, eLink[3])

These data most often are stored in **administrative datasets**, which are datasets that an agency or organization builds and maintains for managing services, record-keeping, internal decision-making, or other program-specific purposes.

3. NCLDS Linking Process Overview

Because most agencies and organizations use administrative data to manage their own programs only, they usually do not connect their own data records to records managed by other agencies or organizations. The types of questions that NCLDS helps to answer—for example, questions about how an experience or intervention during one period of a person's life impacted an outcome for that same person later in life—require connecting those records, so NCLDS uses the PII collected by each of its Contributors to help make those links. There are five main steps in the NCLDS linking process, summarized here.[4]

[1] For more about how NCLDS Contributors use Social Security Number for linking wage data, see the *How NCLDS Works* brief, Linking Data: Workforce Data

[2] For more about the eScholar Uniq-ID and how NCLDS Contributors generate it, see the *How NCLDS Works* brief, Linking Data: eScholar Student ID

[3] For more about the eLink entity resolution service, see the *How NCLDS Works* brief, Linking Data: eLink Entity Resolution

[4] A **Publication Appendix** includes an alternate summary version that researchers and others with technical writing responsibilities can use when describing the NCLDS linking and linking validation processes in formal publications.

Retrieving Data

After Contributors approve a data request, NCLDS acquires data relevant to the approved request from those Contributors and temporarily stores those data in a secure space where NCLDS can work with them. NCLDS separates the PII portion and the non-PII portion of each record, but, before doing so, tags each portion with a common identifier—a **source key**—so that they can be joined back together again at the end of the linking process.

Preparing Data for Linking

NCLDS then prepares the data for linking by ensuring that all instances of the same type of PII—for example, a person's last name—are in exactly the same format, and that individual data elements with obvious errors—for example, data in a "Name" field that are clearly "Address" data—are identified and, when possible, corrected. The resulting now-unified collection of PII is called a **person table**.

The Linking Process

Next, NCLDS sets the **linking rules** that will be used to link the data for the approved request. NCLDS uses these rules to compare the records to each other to determine whether they match one another. When they do match, those records are assigned a common **cluster identifier**—an identifier that lets a user know that those records all belong to the same person. This process can take a significant amount of time and computing power.

Post-Process Review and Revisions

NCLDS then examines the resulting clusters and, as needed, individual records associated with them to check for consistency and for match quality. One key part of this examination is creation of two measures: a **Cluster Variable Variability Index**, which assesses the amount of variation among variables of the same type within a given cluster of records (*e.g.*, Are all instances of First Name exactly the same, or are some slightly different?); and a **Cluster Strength Index**, which assesses how confident NCLDS can be about the records included in a given cluster. The results of this step may lead to revisiting the prior steps to address any correctable data quality issues.

Joining Records and Removing Personally Identifiable Information

Finally, NCLDS uses the source key created in the first step to join the cluster identifiers to the non-PII source data. Now that records can be linked using the cluster identifiers, the PII are no longer needed and, to ensure privacy and security, are removed from the dataset before those data are shared with the Requester.

You can find a visual flow of the entire process in the **Technical Appendix**, and we describe each of these steps in greater detail in the sections that follow.

4. Retrieving Data

Once a data request has been approved by every Data Contributor that will provide data for the request, NCLDS acquires the approved data[5] through a secure data transfer process and stores them in a secure workspace. The data can come from two primary sources:

- Data Warehouses managed by NCLDS on behalf of some NCLDS Data Contributors: These data are called **warehoused data**; NCLDS only accesses them when allowed, but they do not have to be transferred to NCLDS each time a request is approved
- Data Warehouses managed by other NCLDS Data Contributors: These data are called **federated data**, and NCLDS is only able to access them long enough to create approved datasets

NCLDS's access to and use of these data is governed by a group Memorandum of Understanding and by individual Memoranda of Agreement with each Data Contributor. These legal documents describe how NCLDS is allowed to work with the data. NCLDS's work with these data also is governed by North Carolina General Statute 116E,[6] which states the purpose of NCLDS, its compliance standards, its operating regulations, and the data-sharing guidelines by which it abides. Section 116E-2(a)(3) establishes NCLDS's role as a linker of data.

The data are brought into NCLDS's secure workspace via a process referred to as **ETL**, which stands for Extract, Transform, and Load. This process applies standard **business rules** to the data as they are transferred into the secure workspace (these rules align data formats so that they are the same across all data sources[7]). It is important to note that, although NCLDS strives to bring the most accurate data into its linking process, NCLDS does not own or directly manage any of the data; data come to NCLDS from a wide variety of sources and often are of different levels of quality—not all of which can be completely standardized by the ETL process.

Once the data are structured and situated in the dedicated workspace, they then can be prepared for linking.

5. Preparing Data for Linking

Staging

NCLDS first identifies the datasets from each Contributor that include relevant PII that can be used for linking. This time-intensive task requires a familiarity with each Contributor's data and an understanding of which elements can be selected for the linking process. To provide

[5] These include PII that Contributors allow NCLDS to use strictly for linking purposes.
[6] https://www.ncleg.gov/Laws/GeneralStatuteSections/Chapter116E
[7] For example, one ETL rule might ensure that all dates are transformed so that they all start with a two-digit representation of the month (sometimes annotated as MM), then have a two-digit representation of the day (DD), and finally end with a four-digit representation of the year (YYYY), all in a row without spaces: 01072025, 11172023, etc.

guidance for this selection process, NCLDS has developed a list of 111 total element types that can be used for linking; however, most Contributor datasets only contain about 5 to 12 of these elements.

After the linking elements are identified, they are separated from the non-PII data in the datasets and are moved into their own data tables (called “person tables,” or **PRSN tables**, for short). Before they are separated, a unique **source key** (a long string of letters and numbers) is attached to both the PII and the non-PII elements for each record, so that they can be connected again at the end of the linking process.

Cleansing

After the linking data are brought into PRSN tables, NCLDS applies a **data cleansing process** before beginning the actual linking process. The cleansing process is different from the ETL process in several ways. To begin with, it is a highly configurable process, which allows NCLDS to tailor the structure of the data to a specific Requestor’s needs (without changing the information itself). Second, this process checks for and in some cases corrects common errors often found in administrative datasets. These checks include but are not limited to:

- Ensuring that a Social Security Number conforms to Social Security Administration Standards (*i.e.*, has nine digits and does not include any nine-digit sequences not used by the Social Security Administration);
- Removing common administrative data placeholders from Name elements, such as “unknown” or “claimant” or “n/a”;
- Separating common Name prefixes and suffixes from First Name and Last Name fields (*i.e.*, a First Name field with “Mr Robert” becomes a Prefix field with “Mr” and a First Name field with “Robert”); and
- Removing clearly incorrect Birth Dates, such as dates that are in the future or that are too distant in the past to be valid.

The cleansing process’s standardization of the data ensures greater accuracy in the linking process by catching ahead of time addressable anomalies in the data that might lead the process to fail to recognize links or to create inaccurate links based on data errors.

Match Codes

The now-cleansed data then are run through a process called **match code generation**. Match codes allow NCLDS to apply a concept called **probabilistic matching** (or **fuzzy matching**) during the actual linking process. Match code generation is a technical process that examines text data in a given field across all records and assigns a common code to all values in that field that have slight differences but likely should be considered the same. The field values themselves are not changed, but, to a computer, the newly-assigned common code will make them appear to be exactly the same. The process is governed by pre-established, tested rules to ensure that this standardization is both defensible and reasonable.

NCLDS currently identifies Name elements as good candidates for match code generation. The match code process helps the linking process to handle misspellings, nicknames, and name variations that may appear across datasets from different Data Contributors.

For example, if the First Name field for one record includes the name “Mciheal,” the match code generation process may (depending on the settings) identify it as being similar enough to “Michael” (in this case, because it appears to be a data entry typo) to assign both it and all instances of “Michael” the same code value, which will allow the record with “Mciheal” in the First Name field to be matched with other First Name entries spelled “Michael.”

The match code generation process NCLDS uses is an algorithm developed by the SAS Institute,[8] along with an accompanying sensitivity setting. The sensitivity setting determines how closely two or more values need to resemble each other in order to be assigned the same match code. A higher sensitivity setting will result in match codes being assigned to fewer near-match values (for example, to “Mich**ae**l” and “Mich**ea**l” but not to “M**cihea**l”); a lower sensitivity will result in match codes being assigned to a broader array of near-matches (for example, to “M**ichae**l,” “M**ichea**l,” and “M**cihea**l”).

6. The Linking Process

After the data have been staged and cleansed, and after match codes have been generated, the data now are ready to go through the actual linking process.

Linking Rulesets

To make links, NCLDS first must establish the rules the process will follow. The rules are a series of statements that define what elements the linking process should evaluate to determine whether one record belongs to the same person as another record.

Each rule can focus on a single data element or a combination of data elements. A rule that requires a combination of data elements to be the same in order to establish a link is called an “AND” statement. For the linking process to identify a link based on an “AND” rule, all elements included in the rule must match.

The NCLDS process can incorporate multiple rules, and it identifies a link whenever any one of those rules is met. To do this, each separate rule is treated as an “OR” statement, meaning that if any one of the rules is true for two or more records, the linking process will consider the records to belong to the same individual (that is, to be linked).

The current NCLDS linking rules includes two rulesets, with either or both applied and vetted, depending on the needs of the dataset:

[8] You can read more about the SAS Match Code process in the **Technical Appendix** and also here: https://documentation.sas.com/doc/en/webeditorcdc/v_058/webeditorflows/n0ftfj95fdwufen1btql44r39tp9.htm

Ruleset One

Link records if they have a:

- Matching eLink Cluster Identification Number OR
- Matching Cleaned Date of Birth AND eScholar Uniq-ID Number OR
- Matching Cleaned Date of Birth AND Cleaned First Name AND Cleaned Last Name OR
- Matching Cleaned Date of Birth AND Social Security Number [9]

Ruleset Two

Ruleset Two retains three of the rules of Ruleset One, but makes a few modifications:

- After applying the first rule of Ruleset One (matching eLink Cluster Identification Number) but before applying most of the remaining rules in Ruleset One, Ruleset Two first attempts to link records that do not have an eLink Cluster Identification Number (most often because those records were not available to eLink when it created its matches) to records that do have an eLink Cluster Identification Number, using only eLink linking rules[10]
- One Ruleset One rule is dropped: Matching Cleaned Date of Birth AND Cleaned First Name AND Cleaned Last Name
- A new rule is included: Matching eScholar Uniq-ID AND Social Security Number

The **Technical Appendix** includes additional information about these rulesets.

After the linking rulesets have been defined, all of the PRSN tables are combined to create one long PRSN table. The records are now ready to be linked.

Linking

NCLDS uses an algorithm developed by the SAS Institute called the Real-Time Entity and Network Generation Action Set to perform the heavy lifting of the linking operation.[11] The rules listed above are loaded into the algorithm and then the data are loaded. The algorithm then performs a series of steps that group the data together based on each rule. If a record matches another record based on one of the rules, those records are assigned a cluster identifier to

[9] See the **Technical Appendix** for more details about the order of these rules.
[10] NCLDS uses a subset of eLink linking rules; there are over 20 eLink linking rules, but only eight of those rules rely exclusively on the PII NCLDS currently uses to link records.
[11] You can read more about this algorithm in the **Technical Appendix** and also here: https://documentation.sas.com/doc/en/vicdc/v_027/casactrteng/titlepage.htm

group them together.[12] This process continues until all records have been compared to all other potential matching records.

The linking process is the most time-consuming and computation-heavy portion of this process. Although the logic NCLDS uses is not complex, the sheer number of records that must be compared to one another is quite large (for some requests, nearly 100 million records).

7. Post-Process Review and Revisions

Review

After the data are linked and are assigned cluster identifiers, NCLDS evaluates the output of the linking process to ensure that the linked clusters make sense. This is a critical part of the linking process, and the results of this step could lead NCLDS to return to earlier steps for re-linking. Currently, NCLDS applies three main post-process review steps.

Total Cluster Records Review

The first review step is the simplest: Counting the number of records associated with each cluster identifier, to determine how many records did not cluster with other records. Cluster Identifiers with only a single record are called “singletons,” and they can result from multiple factors, such as:

- An individual is only present in one Contributor’s data
- Data for that record are too incomplete to meet at least one of the rules criteria
- Data quality issues

A high number of singletons at the end of the linking process could indicate **underclustering**, or the identification of fewer clusters than should have been possible.

Cluster Variable Variability Index

NCLDS then generates a Cluster Variable Variability Index (CVVI) value for each cluster, which helps NCLDS assess our confidence in the *precision* of each resulting cluster, or the degree to which all instances of each variable used to create a cluster are truly *alike*. The CVVI helps identify clusters that have large internal variations. Instances of the same types of variations across clusters could point to errors that led to more records being linked than should have been linked—sometimes called **overclustering**. The **Technical Appendix** includes a more in-depth explanation of how NCLDS calculates this value.

[12] Cluster identifiers are random values that are unique to each dataset created by NCLDS in fulfillment of an approved request; cluster identifiers cannot be used across datasets, and the linking process generates entirely new cluster identifiers even if the exact same dataset is approved for a later request.

Cluster Strength Index

The Cluster Strength Index (CSI) helps NCLDS assess our confidence in the *reliability* of each resulting cluster, or the degree to which each record in a cluster matches well with each other record. Like the CVVI, the CSI also helps identify **overclustering**, but this time by assessing the overall strength of the cluster. The **Technical Appendix** also includes a more in-depth explanation of how NCLDS calculates this value.

Revisions

In both cases—underclustering and overclustering—NCLDS examines several hundred samples of singleton clusters, clusters with the highest CVVIs, and clusters with the lowest CSIs to uncover which step(s) in the linking algorithm may have led to these issues. If NCLDS is able to identify ways in which the data preparation or the underlying linking algorithm could be improved, the process begins again until either the problems are resolved or until NCLDS determines that the clustering cannot be improved.

While NCLDS looks for instances of both overclustering and underclustering, the current process prioritizes identification and resolution of overclustering, in an effort to increase overall confidence in the clusters NCLDS forms. For “high-risk” large clusters (those with a high CVVI [low precision] and a low CSI [low reliability]), NCLDS applies an additional round of revision. Affectionately referred to as the “Cluster Buster,” this revision process extracts the subset of high-risk clusters from the full set of clusters, breaks them up (unlinks all of the records in those clusters), and then performs a new round of clustering and post-processing on those records, this time with a shorter and stricter set of clustering rules. Records in the Cluster Buster subset are assigned new cluster identifiers, and the new clusters are then rejoined with the other, lower-risk clusters to form a revised dataset with clusters in which NCLDS has more confidence. The **Technical Appendix** includes a more in-depth explanation of how NCLDS conducts this process.

8. Joining Records

Recall that, during the Staging segment of the data preparation process, NCLDS added a unique source key to the PII and the non-PII components of each record that is included in the approved request. This key now allows NCLDS to rejoin the PII components used in the linking process back to the original source records.

As part of the joining stage, NCLDS also retains the CVVI and CSI values for each record, so that Requesters can see for themselves the relative confidence levels for each cluster and make decisions about whether to rely on all of the clusters identified by NCLDS or to remove some clusters from their analyses if they believe that the precision and reliability of those clusters are not high enough to meet the standards necessary for their specific data use needs. NCLDS also includes additional summary descriptive information about the dataset as a whole to further support Requesters as they make these decisions.

9. Removing Personally Identifiable Information

Since the PII component of the original record now also includes the newly-assigned cluster identifier, the PII in that component can be removed from the dataset that will be provided to the Requester, leaving just the cluster identifier in its place. The Requester can use the cluster identifier to see which records are linked, without needing access to the underlying PII that established the links. This is an important component for maintaining the privacy and security of the data with which NCLDS and approved Requesters work. At this point, NCLDS destroys all PII that were brought into the NCLDS environment for use in the linking process.

To further strengthen security and enhance privacy, the cluster IDs are unique to each approved data request and cannot be reused or applied across datasets. In other words, a cluster ID generated for one approved dataset will not be the same as a cluster ID generated for another approved dataset—even if the two datasets include exactly the same variables.

10. Future Directions for NCLDS Linking

This brief describes NCLDS's current approach to linking records, but it also describes NCLDS's *first* approach to linking records. While the current process is sound, there are several different ways in which NCLDS can grow and strengthen it as we learn more about how well our processes and rules are working, the data with which we work now, and new data that NCLDS will incorporate into its process in the future.

Known Limitations of the Current Process

- NCLDS currently has access to only a limited number of data sources to support entity resolution.
- In the data to which NCLDS does have access, only a limited amount of entity resolution information is available.
- The current rulesets include only Name, ID, and Date of Birth; as noted in Section 2, there are several other variables that could be used to enhance linking, such as Address, Phone, and Email.
- NCLDS has defined only two matching rulesets so far, but more are possible.

Possible Enhancements for Later Versions of NCLDS Entity Resolution

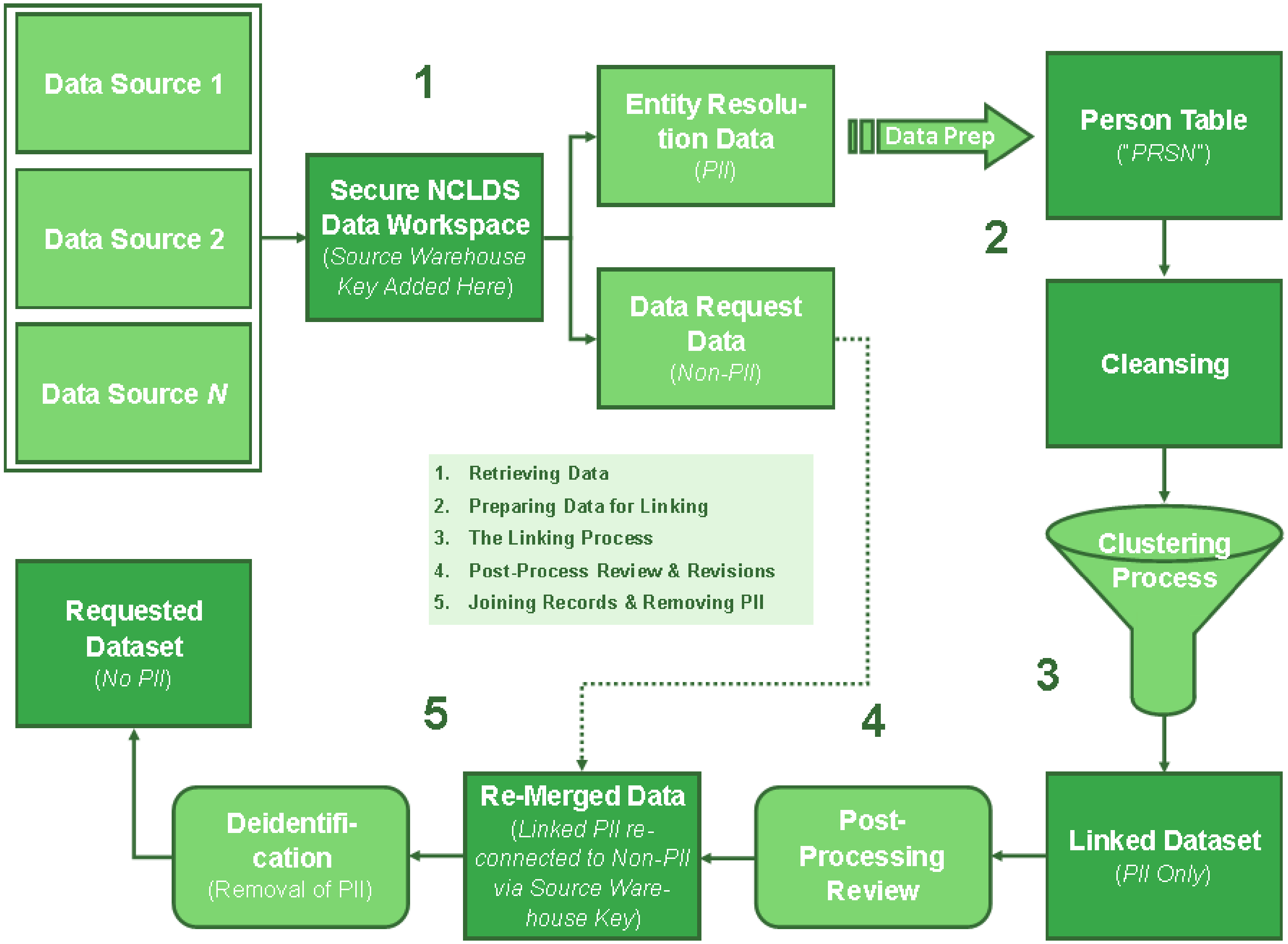
- NCLDS can increase the number of variables available for linking by introducing a preparation step called **Backfill and Validate**, which will allow NCLDS to accurately expand the number of variables in each record that are available for linking.

  Backfill and Validate also can enhance the completeness of final datasets. For example, because wage data currently are linkable only via SSN, a Backfill and Validate process for SSN can help NCLDS increase the number of records that have an associated SSN (and therefore can be matched to wage data).

- There are several ways in which NCLDS can expand and enhance its cleansing rules—including the addition of data source-specific cleansing rules and incorporation of information about the date ranges for which certain changeable variables (such as address) are valid.
- NCLDS can expand its current limited approach to probabilistic (“fuzzy”) matching.
- Match rules currently all have the same “weight” or significance; NCLDS can apply weights to reflect the relative value of rules.
- As noted above, NCLDS can test the value of adding more match variables and match rules.
- The current process results in a single approach to assigning cluster IDs; NCLDS can apply multiple linking processes to the same dataset and then compare the results across processes to determine the best process for a request on a request-by-request basis.
- Similarly, NCLDS can create a User Interface that allows a data-savvy Requester to provide input about the match rules NCLDS applies to that Requester’s dataset.
- Like the current set of match rules, the current set of tools NCLDS uses to evaluate cluster strength can evolve; for example, NCLDS can explore expansion of the CSI to include information about how many rules a cluster meets.
- Finally and most importantly, NCLDS can establish feedback loops to ensure that NCLDS’s stakeholders can contribute more directly to continuous improvement of the NCLDS linking process.

Technical Appendix

Process Flow Diagram



1. Retrieving Data
2. Preparing Data for Linking
3. The Linking Process
4. Post-Process Review & Revisions
5. Joining Records & Removing PII

Match Code Generation

Match codes, created with the SAS DQMATCH function, are a way of helping computers recognize two pieces of information that might mean the same thing, even if they are written differently. For example, people often spell the name Katherine in many ways—like Catherine and Katharine. A match code translates each of those spellings into the same standardized code so that the computer knows they are likely to refer to the same person. A DQMATCH user also can control how strict or flexible the match should be by adjusting the sensitivity. For instance, a sensitivity setting of 50% groups together a wider range of similar spellings, while a sensitivity setting of 85% is stricter and only allows closer matches. The function also includes a toggle for locale (such as "ENUSA" for US English), which tells the function which language rules and cultural spelling patterns to apply for data from a given region or context.

Real-Time Entity and Network Generation Action Set (RTENG)

The SAS RTENG action set first ingests unlinked records, then applies configurable matching rules (exact match, fuzzy match, or hybrid) to compare record pairs, and finally applies clustering algorithms to cluster related records under a single cluster ID that can be used to represent an individual.

Rationale for Linking Rule Order

As noted above, NCLDS constantly is reviewing and revising existing linking rules, as well as testing new rules. The current process includes two distinct rulesets (Table TA.1).

Table TA.1 NCLDS Linking Rulesets

RULESET ONE	RULESET TWO
RULE_EER_CLUSTER_ID	RULE_EER_CLUSTER_ID
	"NC eLink Light" Rules: RULE_FN85_LN85_DOB_SSN RULE_FN_LN_SSN RULE_FN_LN_DOB_SSN4 RULE_FN90_SSN_DOB RULE_FN_MN_LN_DOB RULE_FN_MI_LN_DOB RULE_ESCHOLAR_ID_FN_LN_DOB RULE_ESCHOLAR_ID_FN_LN_SSN
RULE_ESCHOLAR_ID_DOB	RULE_ESCHOLAR_ID_DOB
—	RULE_ESCHOLAR_ID_SSN
RULE_FN_LN_DOB	—
RULE_SSN_DOB	RULE_SSN_DOB

Ruleset One was used to test the end-to-end entity resolution process. It includes four exact match conditions:

- NC eLink Cluster ID (which allows NCLDS to take advantage of data pre-clustered by the NC eLink entity resolution algorithm)
- eScholar ID and Date of Birth (which allows NCLDS to take advantage of data pre-clustered by the eScholar entity resolution algorithm; it also allows NCLDS to loosen the NC eLink rules)
- SSN and Date of Birth (which further loosens the NC eLink rules. NC eLink has 14 rules that include SSN but of those rules include 3 conditions; this rule allows for matching on only two conditions)
- First Name, Last Name, and Date of Birth (which also loosens NC eLink rules; NC eLink does not allow for a match on First Name, Last Name, and Date of Birth unless at least a fourth condition (such as Middle Initial, SSN, or Zip Code) also is met

Analysis of the application of Ruleset One alone revealed several large clusters and evidence of overclustering due, in part, to the relatively loose First Name/Last Name/Date of Birth rule, as well as to data quality issues associated with placeholder birthdates (*e.g.*, partially masked birthdates created by including a person's real month and year of birth but a common or random date of birth) and common first and last names (*e.g.*, John Smith).

To address the overclustering in Ruleset One, Ruleset Two expands the number of match rules from four to 12 by including eight additional rules used by NC eLink. Similar to Ruleset One, NC eLink Cluster ID is used as the first match rule to take advantage of data pre-clustered by NC eLink (true for 67% of the records to which NCLDS currently has access). All records then are subjected to another round of matching using the eight eLink match rules. This additional step enables matching of records previously matched by NC eLink to some of the 33% of records that have not been processed previously by NC eLink. The final three rules applied are looser variations of NC eLink match rules: eScholar ID and Date of Birth, eScholar ID and SSN, and SSN and Date of Birth. An exact match on First Name, Last Name, and Date of Birth is not used as a match rule in Ruleset Two.

Post-Process Review Technical Details

Calculation of the Cluster Variable Variability Index

First, NCLDS identifies each type of PII in a given cluster. For example, a cluster may include one or more instances of First Name, Last Name, Date of Birth, and eScholar Uniq-ID Number. Next, NCLDS counts each instance of each of those individual data elements within the cluster that has a unique value. For example, a cluster of five records with First Name values of "Michael," "Michael," "Michael," Michael," and "Mich**eal**" has two unique values for First Name ("Michael" [four times] and "Micheal"). Then a sum is taken across all the distinct counts where the count is greater than one. If a cluster contains two different eScholar Uniq-ID values and three different First Name values but the same value across all records in the cluster for Last

Name and Date of Birth, this cluster will receive a CVVI of 5 (2 eScholar Uniq-ID values + 3 First Name values + 0 variations in Last Name + 0 variations in Date of Birth; Table TA.2). Note that a value is not considered a variation unless there is more than one unique value for a given variable. Thus, a variable with only one value does not result in a variation count of 1, but a variable with 2 values results in a variation count of 2 (with the presence of a second value making the first value also a variation). If the cluster identifier has only one distinct value across all rule-related data elements, the resulting CVVI value will be 0.

Table TA.2 Cluster Variability Index Example

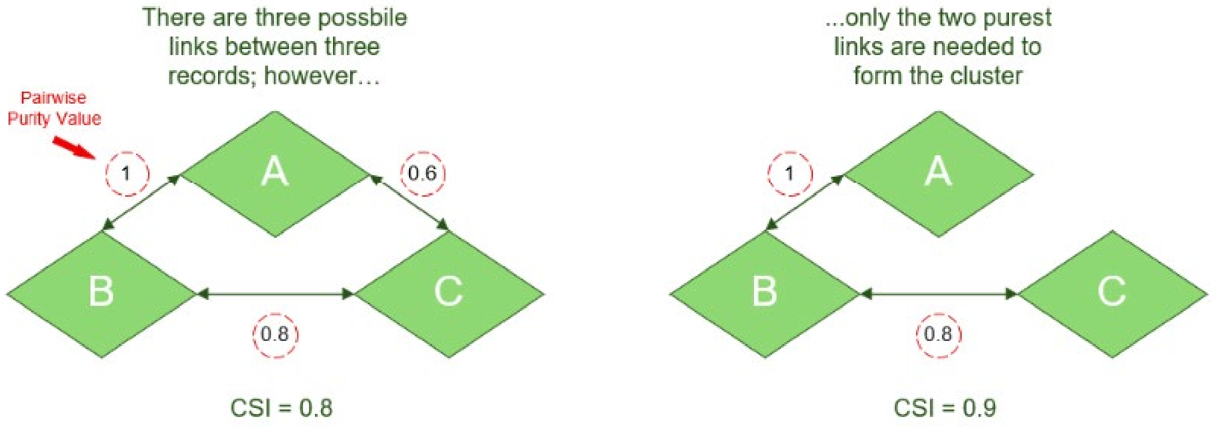
eSCHOLAR ID		DOB	FIRST NAME		LAST NAME
1234	*1*	07-22-1991	*John*	*1*	Doe
1234		07-22-1991	*John*		Doe
1234		07-22-1991	**Jon**	**2**	Doe
1234		07-22-1991	**Jon**		Doe
1243	*2*	07-22-1991	*Jack*	*3*	Doe
1243		07-22-1991	*Jack*		Doe
2 Variations	**+**	***0 Variations*** **+**	***3 Variations***	**+**	***0 Variations*** **=**
			CVVI: 5		

Calculation of the Cluster Strength Index

To calculate CSI, NCLDS compares every possible pair of records in a cluster. For instance, in a cluster made up of Record A, Record B, and Record C, Record A is compared to Record B, then to Record C, and Record B is compared to Record C. Similar to the CVVI, the comparison looks at differences in each variable used to match records, but this time the comparison is at the level of each pair. In other words, the CVVI may have identified variations in all of the values of First Name, but when looking only at Record A and Record B during the CSI process, there may be no differences in how First Name appears in those two records.

The CSI process gives each comparison a similarity score between 0 and 1, then the process combines these results into a single score for the whole cluster. The formula that combines the scores includes four factors: the *average similarity* (how consistent the cluster is overall), the *minimum similarity* (the weakest two-record link in the cluster), the *density* (how many record pairs are above a predetermined similarity threshold value), and the *purity*. To measure purity, the CSI process tallies the similarity scores of each record-to-record connection, then retains only the highest of those scores for the *minimum number of connections needed to create the cluster*. Using our example of Record A, Record B, and Record C, there are a total of 3 possible connections across all of those records (A to B, A to C, and B to C), but only 2 of those connections are needed to make the cluster (for example, A to C and B to C, or A to B and B to C; Figure TA.1, following page), and the purity measure is based on the stronger of those two sets of connections, based on similarity scores.

Figure TA.1 Generation of Cluster Strength Index Value


There are three possbile links between three records; however…
Pairwise Purity Value
1
A
0.6
B
C
0.8
CSI = 0.8
...only the two purest links are needed to form the cluster
1
A
B
C
0.8
CSI = 0.9

Management of the Highest-Risk Clusters

NCLDS uses CVVI and CSI values to place clusters into “risk” categories that indicate whether a given cluster may need to be broken up. The first step is to group clusters into quadrants using CVVI and CSI values. A CVVI value lower than the number of clustering rules (*i.e*., for Ruleset 2, less than 12) is characterized as a “Low CVVI”; a CVVI value between the number of clustering rules used to just under twice the number of rules (for Ruleset Two, between 12 and 23) is categorized as a “Medium CVVI”; a CVVI value of twice the number of rules or higher is categorized as a “High CVVI”. A CSI of less than 0.6 is categorized as a “Low CSI”; a CSI of less than 0.85 but greater than 0.6 is categorized as a “Medium CSI”; a CSI of 0.85 or greater is categorized as a “High CSI.”

Using these values, NCLDS groups each cluster into one of four quadrants. NCLDS separates out clusters in the “high risk” quadrant (those with both a Medium or High CVVI [indicating low precision] and a Medium or Low CSI [indicating low reliability]), un-clusters the records in those clusters, and then performs a new round of clustering and post-processing; however, for these records, the clustering cycle uses a more restricted set of rules:

- The RULE_EER_CLUSTER_ID rule is retained to leverage NC eLink’s conservative clustering methodology; and
- A new rule—RULE_FN_LN_DOB_SSN—is added (in place of a similar rule in Ruleset Two that allowed for fuzzy matching on FN and/or LN) to make only highest-confidence additions to existing eLink clusters.

Most of the clusters submitted for “Cluster Busting” comprise two or more smaller and more reliable clusters that then were incorrectly linked to each other by a single rule; subjecting the records in these super-clusters to a stricter ruleset typically restores the smaller and more reliable clusters that were within the super-cluster but does not then re-form the original high-risk super-cluster.

After the highest-risk clusters have undergone this more conservative linking process and have once again gone through post-processing, they are assigned new cluster identifiers that do not conflict with identifiers already in use among the lower-risk clusters (the original cluster identifiers for the highest-risk clusters are retained for post-process comparison purposes) and then are rejoined with the lower-risk clusters to create an overall lower-risk clustered dataset.

Publication Appendix

This appendix includes language researchers and other data users with technical writing responsibilities can use in formal or peer-reviewed publications to describe the NCLDS linking process and the NCLDS linking validation process.

Description of the NCLDS Linking Process

Preparation of Data for Linking

For each request, NCLDS assigns a unique identifier (a source key) to every record that NCLDS Data Contributors have approved for the requested dataset. Next, NCLDS separates the Personally Identifying Information (PII) portion and the non-PII portion of each record in the approved dataset. Both the PII and the non-PII portions of records approved for the request retain the source key value, so that they can be rejoined after the linking process is complete.

PII from the records in the approved dataset are then supplemented with PII from other records available to NCLDS that are not part of the approved dataset but that NCLDS is allowed to use for linking purposes. Using all available records during the linking process allows the process to identify transitive links, or otherwise-missable links between two records in an approved dataset (Record A and Record C) that only are linkable by way of their mutual linkage to an intermediary record (Record B) that is not in the approved dataset.

NCLDS then prepares the PII portion of each record available for linking by ensuring that all instances of the same type of PII—*i.e.*, a person's last name—are in exactly the same format, and that individual data elements with obvious errors—for example, data in a "Name" field that are clearly "Address" data—are identified and, when possible, corrected. Finally, NCLDS uses the SAS DQMATCH function[13] to generate "fuzzy matching" match code versions of each First Name (sensitivities: 85 and 90) and Last Name (sensitivity: 85) element. The resulting prepared collection of PII comprises an NCLDS "person table," which is the table that NCLDS subjects to the linking process

The Linking Process

NCLDS uses the underlying Action Set of a SAS Viya 4 macro—Real-Time Entity and Network Generation (RTENG)[14]—to apply its linking rules. Records that are linked to each other via RTENG are assigned a common cluster identifier. NCLDS does not currently apply different weights to the rules, so RTENG adds a record to a cluster at the first instance of any rule match.

The current standard NCLDS ruleset is (in order):

- eLink Cluster ID

[13] https://documentation.sas.com/doc/en/webeditorcdc/v_058/webeditorflows/n0ftfj95fdwufen1btql44r39tp9.htm
[14] https://documentation.sas.com/doc/en/vicdc/v_027/casactrteng/titlepage.htm

- Applicable[15] eLink Rules:
  - First Name (fuzzy match; sensitivity: 85) + Last Name (fuzzy match; sensitivity: 85) + Date of Birth + SSN
  - First Name + Last Name + SSN
  - First Name + Last Name + Date of Birth + Last Four Numbers of SSN
  - First Name (fuzzy match; sensitivity: 90) + Date of Birth + SSN
  - First Name + Middle Name + Last Name + Date of Birth
  - First Name + Middle Initial + Last Name + Date of Birth
  - eScholar ID + First Name + Last Name + Date of Birth
  - eScholar ID + First Name + Last Name + SSN
- eScholar ID + Date of Birth
- eScholar ID + SSN
- Date of Birth + SSN

[*If applicable*:] For this project, we requested the following customizations to the base NCLDS linking ruleset: [*Provide details about the ruleset adjustments made, and the rationale*].

Description of the NCLDS Validation Process

NCLDS examines the records associated with the resulting clusters to check for consistency and for match quality. One key part of this examination is creation of two measures: a Cluster Variable Variability Index (CVVI), which assesses the *precision* of a cluster by calculating the amount of variation among variables of the same type within a given cluster of records (*e.g.*, Are all instances of First Name exactly the same, or are some different?); and a Cluster Strength Index (CSI), which measures NCLDS's confidence in the *reliability* of a cluster's contents via pairwise comparisons of each record in the cluster to every other record in the cluster. The results of this step may lead to revisiting prior steps to address any correctable data quality issues, or to breaking up clear instances of overclustering by subjecting records associated with problematic clusters (those with high CVVI values and low CSI values) to a truncated and stricter set of clustering rules that results in more defensible albeit smaller clusters for those records.

At the end of the validation process, the PII portion of records approved for a requested dataset are rejoined to the non-PII portions of those records via the source key generated at the beginning of the process to form the final dataset. Because only records approved for sharing with a Requester are included in the final dataset, any data from records that were used for linking but that are not a part of the approved dataset are removed after the linking stage.

[15] *Note*: There are over 20 eLink rules, but only eight apply to the PII available to NCLDS. NCLDS applies those rules at this stage to link records not available to eLink to records already clustered by eLink.