Linking Records Across Data Systems, Part 2: NC eLink Entity Resolution

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

Government Data Analytics Center

Version 1.2 June 2024

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

Planned Briefs

- Linking Data: Prospects for *Ad Hoc* Matching
- Using the Public Version of the NCLDS Data Dictionary
- Making 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
- NCLDS Cross-Sector Learning Goals

1. Purpose

One key component of the usefulness of NCLDS is the availability across NCLDS data sources of reliable and up-to-date **record-level**[1] **identifiers**. Identifiers help NCLDS connect separate pieces of data to each other (for example, a person's high school academic outcomes and postsecondary course enrollment). Without these identifiers, important data may not be included in analyses that assess the value and impact of policies, programs, and supports.

This brief[2] highlights one of the identification strategies critical to understanding outcomes that span education and non-education sectors—the North Carolina Government Data Analytics Center's (GDAC) NC eLink Entity Resolution Tool. GDAC, located in the North Carolina Department of Information Technology, is the home organization of NCLDS. GDAC provides state agencies and leaders with access to the best data available to help them make program investment decisions, manage resources, and improve financial programs, budgets, and results. GDAC is legislatively mandated to manage data sharing and data integration initiatives for state agencies, institutions, and departments. GDAC identifies opportunities where data sharing and integration can generate greater efficiencies and improve service delivery.

2. NC eLink Entity Resolution Overview

The Challenge: Making the Most of State Data

North Carolina maintains many rich data sources across multiple state agencies; unfortunately, because these sources often are managed independently, they can be difficult to connect, resulting in what are sometimes referred to as "data silos." Data silos can make it difficult for a decision-maker to (for example) get a complete picture of the broad impacts of a program, or gather in one place all of the details of a problem that has wide-ranging effects. These incomplete pictures can lead to ineffective decision-making.

Connecting data across these silos can be challenging. Data from different sources—even if those data describe the same thing—typically have their own formats and structures, and there often are no identifiers common across silos that could make it easier to connect them. To build a single, complete, connected picture (or "view") of a person, place, program, or organization (an "entity"), a decision-maker often resorts to manually assembling that view from all the different sources of data—a process that not only can be time-consuming but also potentially inaccurate and unreliable.

---

[1] For NCLDS, a **record** typically refers to data linked to a specific individual (e.g., a student or worker); see **Entity Resolution Terminology** box, next page.
[2] The text for this brief has been adopted from the GDAC brief, *NC eLink: Entity Resolution in the State of North Carolina* (2023). All figures were developed by GDAC.

The NC eLink Solution: Entity Resolution

Entity Resolution is the process of objectively determining ***whether entities in different data systems are the same entity, based on how they are described or defined by the data associated with them (their data "records")***. Entity resolution systematically connects datapoints scattered across multiple systems to create a single, holistic view of the person, place, program, or organization the data are describing. Entity resolution can make those connections even when data records do not include permanent IDs or when sharing common IDs (the standard approach for connecting data) is not available as an option. Entity resolution works via a computer program that uses several key datapoints to logically group records from different sources. This program (or "algorithm") builds a complete picture of an entity by:

Entity Resolution Terminology

Element: A single piece of information (datapoint) about a larger unit (an "entity"; see below).

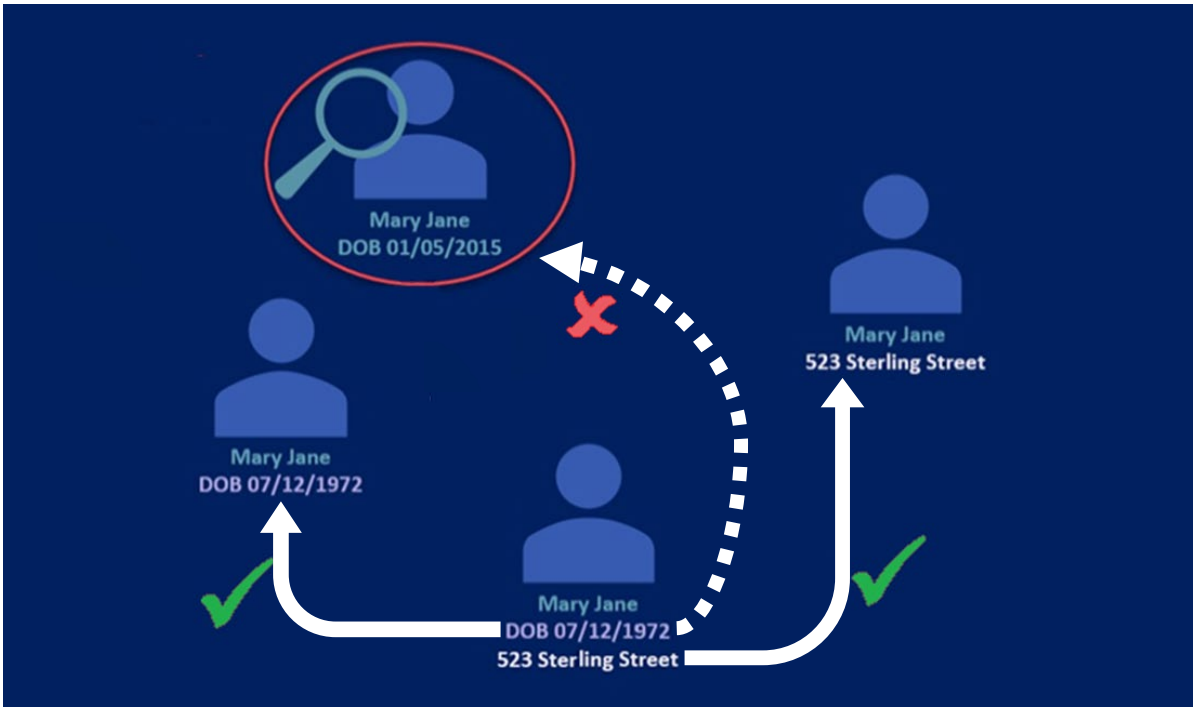
View: A prepared, single-table snapshot assembly of connected data elements from different sources.

Entity: A single person, place, program, or organization that various data elements (sometimes from different sources) collectively describe.

Record: A collection of data elements that describe an entity; an entity can be described by a single record or by multiple records.

- Recognizing when two or more data records are related to the same identity, even if they have been described *differently*; and
- Recognizing when two records do *not* relate to the same identity, even when they have been described *similarly* (Figure 2.1)

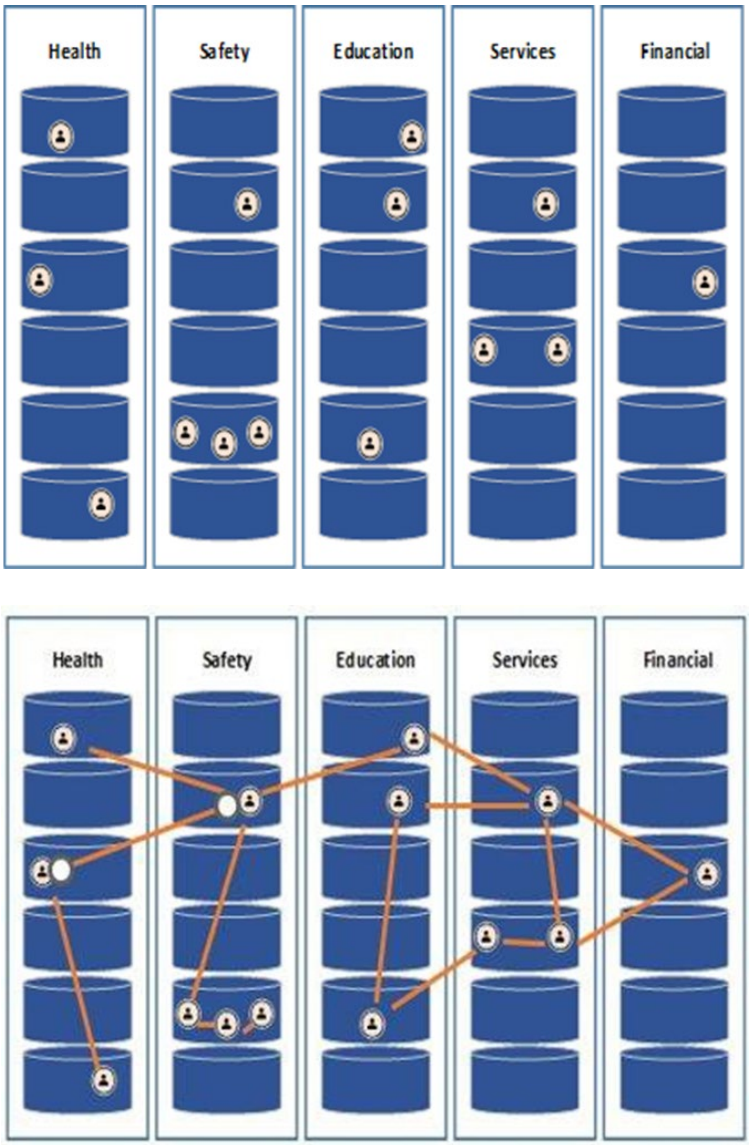
2.1 The Entity Resolution Concept



3. NC eLink

To provide this kind of entity resolution service, GDAC has developed an “enterprise” (universal) utility called **NC eLink**. NC eLink reduces data silos by integrating information from a variety of data sources provided by multiple data partners via a single data-matching process. (Figure 3.1). Because it is housed in GDAC, NCLDS has access to this tool when fulfilling requests that include connecting data from agencies and organizations that participate in NC eLink.[3]

3.1 Unconnected Data for a Single Entity (Top), Connected via NC eLink (Bottom)



[3] Currently (2024), these NCLDS data contributors either already participate in NC eLink or are scheduled to do so soon: the Early Childhood Integrated Data System (ECIDS), the North Carolina Department of Public Instruction, the North Carolina Community College System, and the North Carolina Department of Commerce.

Benefits of NC eLink

While entity resolution algorithms are not a new idea, not all algorithms are alike. Some key characteristics of NC eLink relevant to NCLDS data requests include its ability to:

- Connect entities across education and non-education data sources;
- Standardize how those connections are indicated, regardless of data sources;
- Conduct the linking process in a centralized location behind state privacy and security curtains for uniform security compliance that meets all state and federal standards; and
- Reduce the need to share personally identifiable and sensitive information for one-off matching tasks.

4. The NC eLink Process

Figure 4.1 shows a high-level flow of steps in NC eLink's entity resolution process. The sections that follow the figure include summary and detailed explanations of each of those steps.

4.1 NC eLink Entity Resolution Process

Data Input is the process of onboarding data from a participating agency or organization into the NC eLink environment.

Data Preparation includes an exploratory analysis to identify basic characteristics of and patterns in the data, as well as completion of data quality checks that identify and resolve issues with the data (such as transforming a datapoint found in multiple data sources so that its format is standardized).

Entity Resolution includes three steps:

- Data Linkage—Making initial, linear, datapoint-to-datapoint linkages.
- Clustering—Using advanced linking rules to group together multiple pairs of linked data.
- Entity Validity Index (EVI) Scoring—Assigning a numerical indicator of the confidence NC eLink has in each cluster.

Entity Management is the process of assigning and managing a unique ID for each entity cluster identified by NC eLink.

NC eLink Services are ways in which the NC eLink process can be applied to specific requests.

Step 1: Data Input (Onboarding)

As will be explained in greater detail in Step 4, below, NC eLink works by regularly updating its assessment of connections among data from *all* participating agencies and organizations simultaneously. As a result, agencies and organizations update the data they share with NC eLink on a pre-determined, regular schedule (for instance, twice a year or monthly)—*not* every time the NC eLink service is used. Data from a given agency may not be part of a specific data request, but data from that agency—in conjunction with data from all the other agencies and organizations that participate in NC eLink—help *inform* the connections made for *every* data request.

Agencies and organizations do not provide NC eLink with access to every datapoint they hold—only datapoints most useful for linking. These key datapoints include names, demographic characteristics, personal identification numbers, and addresses (Figure 4.2). All data brought into NC eLink—including data that might be classified as "personally identifiable information" (PII)—are used only to generate links across systems and are not automatically released from NC eLink to someone who requests data via NCLDS (a "Requester").[4] Only the linking information those data help to generate (described in later steps) is shared directly with a Requester.

4.2 Data Element Groups and Preparation Operations for Each Group

		Operation				
		Parsing	Cleaning	Standardization	Normalization	Match Code
Data Element Group	Name	●	●			●
	Demographic			●		
	Personal ID		●	●		
	Date		●	●	●	
	Address	●	●	●	●	●
	Contact Info	●	●	●		

Step 2: Data Preparation

Exploratory data analyses and data quality checks are performed to identify any issues with the data that might make it difficult for NC eLink to use them. These include operations such as cleaning (i.e., ensuring that all values for a given datapoint are valid) and standardization (i.e.,

[4] A Requester may receive PII as part of a data request for which NC eLink provided linking services, but decisions about what data can be shared with a Requester are managed by each data owner on a request-by-request basis following owner, state, and federal protocols—not by NC eLink.

ensuring that a datapoint follows the same format as similar datapoints already in NC eLink). Figure 4.2 (above) outlines the different combinations of preparations applied to the main categories of data (or "Data Element Groups") used by NC eLink. For example, as part of the preparation process, NC eLink might:

- Remove titles from names;
- Remove special characters and punctuation;
- Validate and standardize personal identification numbers to ensure that they follow a consistent format;
- Organize individual data elements so that they follow the same order across all datapoints; and/or
- Format dates, addresses, and contact information for consistency.

Step 3: Entity Resolution

The next step in the process is entity resolution, which is the process of logically grouping data records by comparing and associating individual data elements across those records. Data records are not physically merged during this process; instead, they are linked only via the assignment of a shared unique ID.

Data Linkage

The first part of entity resolution—data linkage—is a process during which pairs of data records are determined to be related enough to each other to be identified as belonging to the same real-world entity. NC eLink has access to several different rules for comparing datapoints, with the rule used for any given pair of datapoints dependent on the specific data available for each member of the pair. For example, one rule is applied when two records have information about first name, last name, date of birth, and address. Currently, there are 23 separate rules available for linking data points.

Clustering and Confidence Level of Matches

Next, these two-way linkages are assembled into larger and broader clusters of records that share similar characteristics. Similar to the initial linking process, the clustering process looks for different valid combinations of data records, but it is during this clustering process that the algorithmic part of NC eLink takes center stage: While clustering is, like linking, rules-based, these rules can include special statistical (e.g., deterministic and probabilistic) methods, in addition to exact-match rules.

After creating a cluster, NC eLink then calculates an *Entity Validity Index (EVI)* value, which indicates how much "confidence" NC eLink has in the overall set of matches in a cluster. The EVI confidence value is higher when information such as name, date of birth, and personal identification number exist for at least some parts of the cluster; it is higher still when that information is present consistently across records within a cluster.

Step 4: Entity Management

As noted above, one important difference between the NC eLink process and other ID assignment processes is that NC eLink connects data *across* data sources (i.e., not only across datasets *within* a participating agency or organization, but also across datasets from *different* agencies). Another important difference is that the cluster IDs NC eLink generates can be updated when new data are introduced that helps the algorithm better understand how all the data shared with NC eLink are related. Most of the connections NC eLink makes across datasets are not likely to change over time (for example, when the data that originally connected Mary Jane's information in Figure 2.1, above, are reliable and consistent, the linkages and the ID assigned to those linkages will not change), but additional new data may help NC eLink improve and tweak connections it made earlier when the information available to it was incomplete or inaccurate. The Entity Management step is where those improvements are made.

After initial clustering is complete, an incremental update process manages the inclusion of any new records that have been received since the last round of clustering, as well as any re-clustering triggered by data in these new records. When new data are brought into NC eLink, those new data go through the entity resolution process described above. During this new round of entity resolution, NC eLink determines whether each new datapoint connects to an existing cluster. If a new datapoint links to an existing cluster, it will be assigned the same ID as the cluster to which it matched. However, NC eLink may instead determine that the new datapoint either provides enough information to suggest that NC eLink should break apart an existing cluster (that is, the new datapoint helped NC eLink to see that Mary Jane really was two people after all), or that the new datapoint does not connect to any existing cluster. In the first case, two new IDs are created—one for the new cluster formed by the new datapoint and some of the datapoints from the original cluster, and one for the datapoints that remain in the original cluster. In the second case (when a new datapoint does not appear to link to any existing clusters), a new ID will be created once more data are connected to that datapoint in later updates, forming an entirely new cluster.

As a result, NC eLink IDs are not always permanent, sometimes changing as new data arrive that may lead to a cluster split or cluster merge. This evolutionary framework helps NC eLink improve linkages and clusters over time, but NC eLink does not lose track of past linkages; instead, it tracks and maintains a complete audit history of all entity ID changes. This history is important for services like NCLDS, which on occasion may need to create datasets with linkages that mirror those in datasets constructed earlier using older versions of NC eLink clusters (for example, when a before-and-after comparison is needed, or when an analysis needs to be repeated to validate the original results).

Step 5: NC eLink Services (Entity Resolution Output)

In addition to creating, maintaining, and updating linkages and clusters, NC eLink also currently offers two Core Services, both of which can contribute to the fulfillment of an NCLDS data request:

- **Core Data Services**—NC eLink can provide cross-reference information (crosswalks) to Requesters that includes cluster ID information and other data keys so that a Requester can make additional matches across their data if the need arises during analysis.
- **Bulk Match Requests**—In certain cases, Requesters with approval from NC eLink data contributors can submit their own record-level data to NC eLink and receive in return their original records plus additional context data from the NC eLink clusters to which their records matched.

NC eLink also plans to make available a third Core Service of particular importance to NCLDS:

- **Clustering as a Service**—This standalone process will apply NC eLink's data preparation and clustering approach to data not currently available to the main eLink service. Because these data are not formally associated with NC eLink, this service will not add the Requester's data to the NC eLink warehouse, and the resulting standalone clusters will not impact existing NC eLink clusters, but the information returned by this service can be applied within in the confines of a specific project. NCLDS has requested an opportunity to work with the NC eLink Team to create a standardized Cluster-as-a-Service option for NCLDS requests that involve data from agencies and organizations that do not formally participate in NC eLink, as well as Requester-provided data that otherwise are not available through NCLDS.[5]

[5] Because such data would be incorporated into analyses using data provided by NCLDS data contributors, fulfillment of a Requester's submission of external data for linking to data provided via the NCLDS process first will require data contributor approval as part of the standard NCLDS Request Review Process.