Linking Records Across Data Systems, Part 3: Linking Workforce Data

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

Government Data Analytics Center

Version 1.1 December 2023

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: 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
- Reviewing and Approving 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 highlights one of the record-linking processes critical to understanding workforce-related outcomes—the Common Follow-Up System's (CFS[2]) approach to linking workforce data to other data available to NCLDS and its users. CFS is one of the data providers from which NCLDS can request data. It contains data dating back to the 1990s on North Carolina employees and the training and education programs in which they have participated. The system contains employment and wage information on individuals who either are working or have worked in an Unemployment Insurance-covered position[3] in North Carolina at any point over the last 25 years, making it one of the largest sources of historical wage data in the nation. CFS is maintained by the North Carolina Department of Commerce's Labor and Economic Analysis Division.[4]

2. Workforce Data Overview

Workforce Data available to NCLDS via CFS include employment status, wage information, geographic location of employment, employer information,[5] and the North American Industry Classification System (NAICS) code associated with the employer.[6] NCLDS is authorized to access up to 20 quarters (5 years) of wage data, but it can only do so when that access is part of a larger project to link wage data to any non-workforce data available to NCLDS (such as K12 and postsecondary data).

Federal regulations require that many components of workforce data be kept confidential,[7] so in most cases, an NCLDS data requester will receive only de-identified and aggregated workforce data—only authorized public officials[8] are allowed to access and work with record-level

[1] For NCLDS, a **record** typically refers to data linked to a specific individual (e.g., a student or worker)
[2] More information about CFS is available at https://tools.nccareers.org/cfs/
[3] More information about criteria for a position to be covered by North Carolina Unemployment Insurance is available at: https://www.des.nc.gov/need-help/faqs/employer-tax-faqs
[4] More information about the Labor and Economic Analysis Division is available at https://www.commerce.nc.gov/about-us/divisions-programs/labor-economic-analysis-division
[5] NC General Statute 116E-7(1): https://www.ncleg.gov/EnactedLegislation/Statutes/PDF/BySection/Chapter_116E/GS_116E-1.pdf
[6] https://www.naics.com/search
[7] Code of Federal Regulations, 20 CFR 603: https://www.ecfr.gov/current/title-20/chapter-V/part-603
[8] Per a 2016 Memorandum of Understanding between the North Carolina Department of Commerce and the Government Analytics Data Center, public officials for NCLDS include authorized staff at the Department of Commerce and in the Government Data Analytics Center.

workforce data.[9] An authorized public official will prepare for NCLDS a de-identified and aggregated dataset that includes workforce data only if the data request has been reviewed and approved by officials responsible for the use of workforce data.

3. How Workforce Data are Linked to Other Data

Most workforce data-linking for NCLDS is handled by public officials in the Department of Commerce who maintain the Common Follow-Up System. To make data linkages, CFS staff typically use a person's name, date of birth, and Social Security Number. Of note, sensitive data like these are used only for matching purposes; no sensitive data are included in final data packages prepared for and released to NCLDS data requesters. CFS has explored the use of additional, less-sensitive variables for linking, but because many of the other variables considered either tend to change over time (contact information, address) or often are incomplete or inconsistent across data sources (demographics), they are less reliable for longitudinal linking of workforce data.

CFS makes available only data that have been connected via an **exact match**, which means that CFS only confirms a link if all of the variables that are used to make the link are exactly the same in each of the data sources being linked.[10] CFS also only confirms a link when a match is **one-to-one**, which means that CFS does not link data if the workforce data is exactly matched to non-workforce data associated with two or more clearly different people.

4. How Workforce Data Linkages are Strengthened

Validation of Linking Variables

An important part of the workforce data linking process is **validation** of the data used to make the linkages—that is, determination of the degree to which those data appear to be reliable and therefore usable for matching. Perhaps the most important validation is of the Social Security Number (SSN). Workforce data provided to CFS by the Department of Employment Security includes SSN, but those data already have been validated, so most validation is of SSNs attached to the non-wage data being matched.

[9] NC General Statute 96-4{x) (https://www.ncleg.gov/EnactedLegislation/Statutes/PDF/BySection/Chapter_96/GS_96-4.pdf); NC General Statute 143B-7 (https://www.ncleg.gov/EnactedLegislation/Statutes/PDF/BySection/Chapter_143B/GS_143B-7.pdf); NC General Statute 143B-10 (https://www.ncleg.gov/EnactedLegislation/Statutes/PDF/BySection/Chapter_143B/GS_143B-10.pdf); and Federal Code of Federal Regulations, 20 CFR Part 603. Subpoart B (https://www.ecfr.gov/current/title-20/chapter-V/part-603).

[10] Other linking procedures sometimes include what is called "fuzzy matching" rules, or rules that allows for slight differences across datasets in the form or format of the linking variables (such as variations in how a name is spelled).

In the past, CFS used a federally-developed SSN algorithm[11] to help with validation, but over time the algorithm began rejecting an increasing number of viable SSNs. Part of the reason for this increase in rejections was a change in the federal approach to assigning SSNs. SSNs assigned after 2011 no longer follow many of the original SSN rules (such as a rule that associated the first three numbers of the SSN with the geographic place of its assignment).[12] As a result, the algorithm began rejecting an unbalanced number of SSNs as more and more SSNs entered the system that were no longer connected to a specific geographic location—in particular, SSNs for immigrants and international students.

Now, the validation process is more streamlined, primarily focusing on ensuring that an SSN does not have invalid numbers (such as 000, 666, or 900-999 in the leading three digits), "Woolworth Wallet" values,[13] sequences shorter or longer than 9 digits, or letters.

A Modern-Day "Woolworth Wallet" Story

Most secure data systems have validation checks to prevent someone from entering an invalid SSN, but many of those checks do not catch one specific SSN: 111-22-3333. As a result, it has become the most common "SSN" in North Carolina, entered by people who do not want to share their real SSNs.

Improving Linkages When Linking Variables Are Incomplete

Some of the key variables CFS uses to link workforce data are not always available in the systems that provide the non-workforce data to be linked. To address this challenge, CFS sometimes will temporarily enhance those non-workforce data by "backfilling" (adding on) one or more of the missing key linking variables. To "backfill," CFS first links the target non-workforce data to *other* non-workforce data that *does* include the key linking variables. One way CFS does this is by using another linking tool available to NCLDS—NC eLink[14]—to temporarily associate the missing values with the non-workforce data so that they then can be linked to workforce data. Because NC eLink uses an approach to linking that is very different from the exact-match approach used by CFS, it sometimes provides multiple "backfill" options for each variable; in those cases, CFS follows the one-to-one rule noted above and does not incorporate any of the "backfill" values suggested by eLink. Once CFS makes its matches using the enhanced non-workforce data, the variables temporarily added to the non-workforce data to improve linkages are removed from the non-workforce data.

[11] https://www.ssa.gov/employer/ssnvhighgroup.htm
[12] More information about changes in the way that SSNs are assigned is available at https://www.ssa.gov/employer/randomization.html
[13] Some wallets sold by Woolworth stores in the late 1930s and early 1940s used to include a fake Social Security card to demonstrate how the card could be stored in the wallet. This card included a ***real*** SSN. Many wallet-purchasers began using the SSN as their own, with traces of that SSN still in circulation as late as the late 1970s. The Social Security Administration has published an article about this and other interesting examples of SSN misuse: https://www.ssa.gov/history/ssn/misused.html.
[14] To learn more about NC eLink, see the *How NCLDS Works* brief, *Linking Records Across Data Systems, Part 2: NC eLink*: https://nclds.nc.gov/about-nclds/how-nclds-works-and-other-faqs