

CECIDS Technical Memo

Understanding Race and Ethnicity Data in Early Childhood Integrated Data Systems: Standards, Crosswalks, and Guardrails

August 27, 2026

Purpose

As states build early childhood integrated data systems (ECIDS), race and ethnicity present a complex data integration challenge. These data are collected across many early childhood programs, but the categories, question structures, reporting requirements, and level of detail are not necessarily the same. They also change over time.

Federal race and ethnicity standards were substantially revised in 2024 for the first time since 1997. Federal programs are incorporating those changes on different timelines, while state programs and source systems may use different collection and reporting structures. Historical data will continue to reflect the standards in place when those data were collected.

This memo provides a framework for integrating race and ethnicity data across programs and over time. It identifies distinctions that should be preserved and the data practices needed to support changing reporting, policy, and analytic needs. The framework is intended for technical and program teams that design, map, or govern race and ethnicity data within an ECIDS.

Key Principle: An integrated data system should preserve race and ethnicity information at the most granular level available and use documented, versioned crosswalks to produce the categories needed for a particular use.

Race and Ethnicity Are Not a Single, Stable Data Element

It can be tempting to treat race and ethnicity as standard demographic fields that can be harmonized once and then used consistently across an integrated system. In practice, the same apparent data element may represent different information depending on when, where, and how it was collected. These differences can materially affect what an integrated record means and what analyses it can support.

Important source differences include:

- whether race and ethnicity are collected separately or through a combined question;
- which categories and detailed identities are available;
- whether respondents can select more than one category and how multiple responses are stored;
- how non-response, unknown values, and missing data are represented; and
- whether information is self-reported, reported by a parent or guardian, or obtained through another permitted method.

These differences are not simply inconsistencies among source systems. They may reflect the standards and program requirements under which the data were collected. For an ECIDS, the objective should therefore not be to force all incoming data into one permanent set of categories, but to preserve enough information to understand what the source data mean and translate them appropriately for different uses.

Changing Standards Illustrate the Integration Problem

The evolution of federal race and ethnicity standards demonstrates why an ECIDS needs to accommodate more than one representation of these data. The 1997 Office of Management and Budget Statistical Policy Directive No. 15 required separate questions for Hispanic or Latino ethnicity and race, established five minimum race categories, and permitted respondents to report more than one race. Under that standard, Middle Eastern and North African origins were included within the definition of White.

The 2024 revision changes both the categories and the collection structure. It establishes seven minimum race and/or ethnicity categories, including Middle Eastern or North African as a distinct category; requires a combined race and ethnicity question; allows respondents to select one or multiple categories; and requires the collection of detailed race and ethnicity information beyond the minimum categories unless an agency receives an exemption. These changes affect both how data are collected and how they can be represented downstream.

Early childhood programs are implementing the revised standards through their own reporting structures and on different timelines. An ECIDS may therefore receive data collected under legacy and revised federal standards, state-specific structures, and program-specific requirements simultaneously. Historical records add another layer: data collected correctly under an earlier standard do not become incorrect when the standard changes.

How Race and Ethnicity Standards Differ in Practice

The examples below illustrate some of the variations an integrated early childhood system may need to accommodate. The comparison is not intended to serve as a complete crosswalk. Its purpose is to show how similar categories can be embedded in different technical and reporting structures.

Overview of Variation Across Commonly Used Race and Ethnicity Standards

Standard or Program	Current Structure	Notable Features	Implication for an ECIDS
OMB SPD 15 — 1997	Separate race and Hispanic/Latino ethnicity questions	Five minimum race categories; multiple races permitted; Middle Eastern and North African origins included within White.	Historical records may not contain the information needed to map precisely to newer categories.
OMB SPD 15 — 2024	Combined race and ethnicity question	Seven minimum categories, including MENA; multiple selections; detailed categories generally required unless exempted.	The integrated model needs to support multiple selections and greater granularity while preserving historical limitations.
Head Start PIR — 2025–26	Race and ethnicity reported together using the seven revised federal categories	Families may self-report multiple identities; PIR reports a Multiracial and/or Multiethnic category and the component categories selected.	Head Start demonstrates how a program can preserve multiple identities while also producing an aggregated reporting category.

Standard or Program	Current Structure	Notable Features	Implication for an ECIDS
CCDF / ACF-801	Current case-level reporting is transitioning to a revised combined race and/or ethnicity structure	Current guidance distinguishes certain types of non-response; beginning with the October 2027 submission, revised reporting incorporates MENA and a combined construct.	An ECIDS may need to interpret different ACF-801 structures depending on the reporting period.
Illinois PFA / ISBE SIS	SIS Student Demographics uses one mandatory coded Race field.	Current codes include Hispanic or Latino, six race categories including MENA, and Two or More Races.	A state system may use categories similar to the revised federal standard while representing them in a different technical structure.

The examples make an important point: differences among standards involve more than category labels. The current Head Start PIR allows families to select multiple race and/or ethnicity categories and requires programs to report the component categories for people reported as Multiracial and/or Multiethnic. By contrast, the current Illinois SIS Student Demographics specification stores race in a single mandatory two-character field, includes Two or More Races as one allowable value, and now includes Middle Eastern or North African as a separate value.

Both systems can describe race and ethnicity, but the information available for subsequent analysis is not equivalent. A technical team can standardize these data for a defined use without assuming that the underlying records carry the same level of detail or meaning. That distinction should remain visible in the integrated data model and its documentation.

Key Question: *What differences in source standards must an ECIDS preserve in order to interpret race and ethnicity data accurately across programs and over time?*

Preserve Granularity Before Applying Reporting Categories

Where a source provides detailed race and ethnicity information, an ECIDS should preserve that information rather than immediately collapsing it into broader reporting categories. For example, the 2024 federal standard includes Asian as a minimum category and specifies detailed categories such as Chinese, Asian Indian, Filipino, Vietnamese, Korean, and Japanese. A detailed value can be aggregated to Asian when needed; if it is replaced with Asian during integration, the original information cannot be recovered later.

Key Principle: **More granular information can generally be aggregated into broader categories. Broader categories generally cannot be disaggregated after collection.**

The integrated representation should therefore preserve the most granular information available from participating source systems, even when a current reporting requirement uses broader categories. This does not mean every downstream product should use the most detailed data. It means the integration layer should avoid discarding details that may be needed for future reporting, research, or policy uses.

Preserve Source, Integrated, and Derived Representations

Race and ethnicity data become easier to govern when the system distinguishes among what the source supplied, how the ECIDS represents the information for integration, and what a particular output requires. These representations are related, but they should not be treated as interchangeable. Keeping them distinct allows reporting standards to change without requiring the historical source record to be rewritten.

Different Representations of Race and Ethnicity Data Relevant to Data Integration and Reporting

Representation	What It Captures	Why It Matters
Source	The original value or values and, where available, the source data dictionary, collection structure, and reporting period.	Preserves what the originating program or system actually provided.
Integrated	A consistent representation that accommodates granular identities, multiple selections, and source limitations.	Supports linkage and analysis without reducing all records to the least detailed source.
Derived	Categories produced for a particular report, analysis, or public use through documented transformation rules.	Allows the same underlying data to support different uses without replacing source information.

Multiple Identities Must Be Preserved Where Available

Race and ethnicity should not be modeled on the assumption that a person has a single value. Source systems may represent multiple identities as separate fields, multiple records, a multi-select response, or an aggregated value such as Two or More Races or Multiracial and/or Multiethnic. Where the source provides the individual selections, an ECIDS should preserve them rather than retaining only the aggregate category.

The current Head Start PIR illustrates the value of doing so. A child reported as both Black and Hispanic or Latino is counted as Multiracial and/or Multiethnic, while the underlying Black and Hispanic or Latino selections are also reported. Where another source provides only Two or More Races, however, the ECIDS cannot reconstruct the races originally selected and should retain that limitation rather than infer detail that is not present.

Preserve Non-Response Distinctions Where They Exist

Not every absence of race or ethnicity data means the same thing. Depending on the source, the record may reflect one of several conditions that have different implications for reporting and data quality:

- the respondent was asked but did not provide the information;
- the information is unknown;
- the source program did not collect the information; or
- an expected value is technically missing from the transmitted record.

An ECIDS should preserve these distinctions where they exist in the source data and avoid collapsing them further during integration. It should not, however, create distinctions that the source did not collect. For example, the 2025–26 Head Start PIR defines Unspecified to include both cases in which race or ethnicity is unknown and cases in which parents declined to identify it, so those conditions cannot be reliably separated downstream from the PIR value alone.

Current ACF-801 technical guidance, by contrast, distinguishes a family that was asked but did not provide race or ethnicity information from a situation in which the lead agency did not collect the information. Preserving the distinctions available from each source supports accurate reporting and makes data quality limitations easier to diagnose. The integrated representation should therefore be no less precise about missingness than the source, but no more precise than the source can support.

Source Provenance Matters

A standardized race or ethnicity value alone may not provide enough information to interpret the record over time. Where feasible, an ECIDS should retain the metadata needed to trace the value back to its originating context. The technical record should identify:

- the originating program and source system;
- the original code or response;
- the applicable source data dictionary or standard;
- the date or reporting period;
- the collection method, where available; and
- the crosswalk or transformation used to produce a standardized value.

Provenance becomes particularly important when definitions change. Under the 1997 federal standard, White included origins in Europe, the Middle East, and North Africa; under the 2024 standard, Middle Eastern or North African is a separate minimum category, and White refers to European origins. A historical record containing only White therefore cannot reliably be divided retrospectively into White and Middle Eastern or North African because the information needed to make that distinction may never have been collected.

Key Principle: An integrated system should make source limitations visible rather than create precision that does not exist in the underlying data.

Crosswalks Are Core Integration Infrastructure

Crosswalks provide the translation layer between information received from source systems and the categories needed for reporting or analysis. An ECIDS may need mappings among historical and current OMB standards, Office of Head Start PIR requirements, CCDF/ACF-801 requirements, state education and early childhood program requirements, source-system-specific categories, and approved analytic or public reporting categories. These mappings should be treated as governed data assets rather than embedded only within transformation code.

At a minimum, a crosswalk should identify the source standard and version, the source value, the target standard and version, the mapping rule, the effective period, and any known information loss. It should also document where a valid mapping cannot be made. This is particularly important during periods when multiple source standards are active simultaneously.

The purpose of a crosswalk is not to make unlike data appear equivalent. It is to make explicit what can be translated, what information is lost in translation, and where comparability is limited. Versioning the crosswalk separately from the source data allows the ECIDS to revise mappings as standards change while retaining a reproducible record of prior transformations.

The Use-Case Question

At its core, this framework responds to a broader question facing an ECIDS:

Key Question: *How can an integrated system maintain usable and interpretable race and ethnicity data when contributing programs collect and report the information differently and the standards themselves change over time?*

The system does not need one permanent categorization. It needs to preserve what was collected, understand the standard and structure under which it was collected, and generate appropriate representations for different use cases. This approach allows standards and reporting requirements to change without repeatedly redefining historical source data or redesigning the underlying data model.

Use-Case Implications

Different use cases may require different representations of the same race and ethnicity data. Federal program reporting may require one set of categories, state reporting another, and research may benefit from more granular identities. Public reporting may require additional aggregation because of privacy, confidentiality, or small-cell concerns.

The category required for a single output should therefore not determine what information the integrated system retains. For each use, teams should identify the governing standard, target categories, the necessary level of detail, the treatment of multiple identities and non-response, and the applicable disclosure rules. Maintaining granular source information separately from derived reporting categories gives an ECIDS the flexibility to support current and future uses while applying appropriate rules to each. No output should imply distinctions, comparability, or precision that the source data and collection context cannot support.

Key Guardrails for Integrating Race and Ethnicity Data

- Preserve the most granular information available from the source.
- Support multiple race and ethnicity selections where the source provides them.
- Preserve source values and sufficient provenance to interpret how the information was collected.
- Preserve distinctions among non-response states where those distinctions exist in the source.
- Do not infer distinctions or details that the source did not collect.
- Treat reporting categories as derived representations rather than replacements for source information.
- Maintain crosswalks as documented, versioned data assets.
- Identify the applicable standard and version when producing standardized values.
- Reassess mappings when authoritative standards or program requirements change.
- Apply disclosure and privacy rules before releasing granular race and ethnicity information.

Ideal-State Data Principle

In an ideal data environment, an ECIDS would preserve the most granular race and ethnicity information available from each source, including multiple selections and relevant provenance, while maintaining versioned crosswalks that produce the categories needed for different reporting and analytic purposes. The integrated model would accommodate multiple valid source structures simultaneously rather than requiring every participating system to use a single collection format.

This approach provides greater flexibility than designing the integrated system around a single reporting standard. It allows the same underlying data to support program reporting, longitudinal analysis, research, and future standards without discarding information during integration or redefining historical records as standards evolve. Where source data are less detailed, the system should preserve that limitation rather than attempt to reconstruct detail that was never collected.

Policy and Data Note

Race and ethnicity standards and program reporting requirements will continue to evolve. An ECIDS should establish a formal process to review authoritative guidance and update its data dictionaries and crosswalks. That process should identify authoritative sources, assign responsibility for monitoring changes, document effective dates and the rationale for revisions, assess impacts on current and historical data, test revised mappings, and retain prior versions so historical transformations remain reproducible.

The transition from the 1997 to the 2024 OMB standards—and the differing implementation approaches already visible across Head Start, CCDF, and Illinois education data—demonstrates why this function is necessary. Change control should be treated as ongoing data stewardship rather than a one-time implementation activity. Periodic review with program and data owners can also identify changes in field use that are not obvious from the source schema alone.

Related CECIDS Resource: *Understanding Source Data Through Structured Data Dialogues* describes the ongoing practice of validating the meaning of sources and changes with program and data owners.

What This Framework Does Not Attempt to Do

This framework does not establish a single preferred taxonomy of race and ethnicity for all early childhood programs, require source systems to collect identical information, or assume that historical data can be perfectly converted to current standards. It also does not determine the level of race and ethnicity detail appropriate for public release. Those decisions depend on the applicable program requirements, the analytic purpose, the governance rules, and the disclosure standards.

Instead, the framework is designed to support accurate interpretation and responsible reuse of race and ethnicity data across programs and over time. It emphasizes preserving source information, documenting transformations, and designing the integration layer to accommodate changing standards. These practices allow an ECIDS to standardize data for defined uses without treating unlike source records as more equivalent or precise than they are.

About This Memo

This memo combines authoritative federal and Illinois program guidance with generalized design recommendations for early childhood integrated data systems. The recommendations are informed by CECIDS experience integrating administrative data across programs and are intended to be transferable beyond Chicago. Program requirements and source-system practices should be confirmed with the responsible data owners before implementation.

Authoritative Regulatory References

1. U.S. Office of Management and Budget. *Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity*. 2024.
2. U.S. Office of Management and Budget. *1997 Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity*.
3. U.S. Department of Health and Human Services, Administration for Children and Families, Office of Head Start. *2025–2026 Head Start Program Information Report, Version 1.0*.
4. U.S. Department of Health and Human Services, Administration for Children and Families, Office of Child Care. *ACF-801 Case Level Data Standards, Technical Bulletin #3r-V9*.
5. U.S. Department of Health and Human Services, Administration for Children and Families, Office of Child Care. *Child Care and Development Fund ACF-801 Form and Instructions*; approved revisions effective with October 2027 reporting.
6. Illinois State Board of Education. *Student Information System: Student Demographics Data Elements*.
7. Illinois State Board of Education. *Early Childhood Block Grant / Preschool for All program and reporting materials*.