ADEP WORKING PAPER SERIES

Coverage of Prisons and Detention Facilities in the 2020 Census

M. Michaellyn Garcia
U.S. Census Bureau

Keith Finlay
U.S. Census Bureau

Clare E. Speer
U.S. Census Bureau

Elizabeth Willhide
U.S. Census Bureau

Kayla N. Patti
U.S. Census Bureau

Tracy A. Loveless
U.S. Census Bureau

ADEP-Working Paper 2024-08
November 2024

Associate Directorate for Economic Programs
U.S. Census Bureau
Washington DC 20233

Coverage of Prisons and Detention Facilities in the 2020 Census

M. Michaellyn Garcia, U.S. Census Bureau
Keith Finlay, U.S. Census Bureau
Clare E. Speer, U.S. Census Bureau
Elizabeth Willhide, U.S. Census Bureau
Kayla N. Patti, U.S. Census Bureau
Tracy A. Loveless, U.S. Census Bureau

ADEP-WP-2024-08
November 2024

Abstract

During every decennial census, the Census Bureau enumerates people living in prisons and detention facilities as part of its group quarters (GQ) operations. The Census Bureau uses the Master Address File (MAF) to define the universe of all living quarters including GQs. Among the challenges of enumerating GQs, maintaining the currency of the MAF remains a critical step for conducting both the decennial census and intercensal surveys. This project evaluated the benefits of leveraging censuses of facilities from the Department of Justice (DOJ) as alternative sources of information to help maintain the list of justice facilities in the MAF. The DOJ data were matched to the MAF using probabilistic linkage methods. This integrated dataset enabled the identification of potential gaps in coverage or classification errors in all contributing data frames. For example, more than 99.5 percent of the prisons and detention facilities in the DOJ frames matched to records in the MAF. Of these matched records, 86.8 percent were enumerated, 4.2 percent were listed as vacant, and 9.1 percent were not enumerated in the 2020 Census. The DOJ data included population counts that served as benchmarks for evaluating the decennial census counts of the corresponding populations. Of the matched facilities, 97.4 percent of the DOJ-measured population resided in facilities linked to GQs or housing units that were enumerated in the 2020 Census. These results suggested that enumeration occurred in the facilities that housed nearly all the DOJ-measured population. However, 9.1 percent of the matched records not enumerated represented a segment of this universe for which alternative data sources and further research could be leveraged to improve coverage. This study demonstrates achievable and accurate data reconciliation which has the potential to reduce costs and respondent burden in future surveys. These results bolster the case for continued development of an integrated database as a long-term solution to facilitate the linkage and validation of further alternative data sources and to increase maintenance frequency.

Keywords: group quarters, prisons and detention facilities, 2020 Census
JEL Classification Codes: K49, Y10

Introduction

During every decennial census, the Census Bureau enumerates people living in prisons and detention facilities as part of its group quarters (GQ) operations. GQs are places where people live or stay in a group living arrangement, which are owned or managed by an organization providing housing or services for residents. The Census Bureau uses the Master Address File (MAF) to define the universe of all living quarters including GQs. Among the challenges of enumerating GQs, maintaining the currency of the MAF remains a critical step for conducting both the decennial census and intercensal surveys.

This project evaluated the benefits of leveraging censuses of facilities from the Department of Justice (DOJ) as alternative sources of information to help maintain the list of justice facilities in the MAF. The DOJ data used in this project came from censuses of justice facilities conducted by the Bureau of Justice Statistics (BJS) and the Office of Juvenile Justice and Delinquency Prevention (OJJDP), referred to throughout the report collectively as the "DOJ frames". The DOJ frames included lists of facilities along with their addresses, characteristics, and population counts.

Probabilistic linkage methods were used to match facility records from the DOJ frames to GQ records from the MAF. The MAF lists all potential living quarters and includes attributes such as addresses, address type classifications, enumerations status, and Census geographies. Subject-matter experts validated the probabilistically linked records using facility characteristics in both the DOJ frames and the MAF. This created a high-quality integrated dataset that offered the ability to identify potential gaps in coverage or classification errors in all contributing data frames.

Analysis of the integrated frame showed that more than 99.5 percent of the prisons and detention facilities in the DOJ frames matched to records in the MAF. Of these matched records, 86.8 percent were enumerated, 4.2 percent were listed as vacant, and 9.1 percent were not enumerated in the 2020 Census. Non-enumeration was related to each record's classification code. Units coded in the following ways were less likely to be enumerated: as non-existent; as duplicate; as non-residential structures prior to 2020 operations; or as permanently closed before the 2020 Census. The COVID-19 pandemic also caused facilities to temporarily close during the 2020 Census which contributed to an increase in facilities recorded as vacant.

The DOJ data included population counts that served as benchmarks for evaluating the decennial census count of the corresponding populations. Of the matched facilities, 97.4 percent of the DOJ-measured population was linked to MAF GQs or housing units enumerated in the 2020 Census. The remaining DOJ-measured population (2.6 percent) was linked to GQs listed as vacant or not enumerated. These results suggested that enumeration occurred in the facilities that housed most of the DOJ-measured population. However, 9.1 percent of matched facilities were not enumerated. These units contained 1.6 percent of the DOJ-measured population and represented a segment of this universe for which alternative data sources and further research could be leveraged to improve coverage. Overall, the population-level coverage estimate (97.4 percent) was greater than the facility-level coverage estimates (86.8 percent), which suggested a greater likelihood for enumeration in more populous facilities.

Juvenile justice and community corrections facilities both tend to be smaller, and both had lower 2020 Census coverage rates than other types of prisons and detention facilities.

Summary of Findings and Recommendations

This study demonstrates that data can be reconciled accurately across databases, which can potentially reduce costs and respondent burden by identifying coverage issues prior to enumeration. We propose continued development of an integrated database as a long-term solution to facilitate the linkage and validation of further alternative data sources and to increase maintenance frequency. Analysis of the integrated database identified additional similarities and differences between the data sets. Key findings and recommendations are:

1. The DOJ frames are high-quality, facility-based data sets that, when integrated with the MAF GQ frame, improve the ability for the Census Bureau to conduct verification, updates, and quality assurance. While the MAF and DOJ frames ultimately rely on updates from the same federal, state, and local law enforcement agencies, observed differences between the 2019 DOJ data and the 2020 Census data identify a need for continuous updates. The Census Bureau should streamline these efforts by continued development of the integrated frame. This will increase MAF update frequency, streamline processing, and reduce in-office verification, field verification, de-duplication costs, and respondent burden.

2. Integration of the data sets revealed new information such as facility coverage gaps and ways of improving both Census Bureau and DOJ frames. This enables identification of alternative data sources to fill gaps and increase record quality by use of targeted research and web scraping. The Census Bureau should continue development and inclusion of new data sources, source evaluation criteria, and coordinated updates in collaboration with internal Census Bureau and other external partners. An integrated relational database designed to meet the needs of internal and external customers, increases the likelihood for discovery of additional gaps at both facility and GQ levels. This facility database will allow rapid evaluation and integration of alternative data sources about facilities. Increased collaboration benefits all agencies by improving data currency and coverage for both public and protected data while assuring compliance with data stewardship and security requirements.

3. DOJ records primarily list facilities, while MAF records primarily list GQs that represent buildings or units within facilities; thus, frame linkage results in a combination of one-to-one, one-to-many, and many-to-many relationships. The Census Bureau should continue development of procedures for a data integration strategy that allows for easy identification of parent-child relationships. The parent-child relationships within this database will enable new methods of analysis including benefit-cost analysis of collection strategies and the integration of person-level data sources at either the facility or GQ levels, allowing for the reduction of respondent burden.

Challenges of Covering Prisons and Detention Facilities in Enumerations

The enumeration of people living in prisons and detention facilities for the 2020 Census presented unique challenges which included identifying all facilities, confirming the types of GQs located within

these facilities, and finally contacting the right officials to gain access and collect information (GAO 2021). These challenges increased when facilities were inside commercial buildings or single-family homes not easily identifiable as GQs; when services and administrations varied across authorities and communities; when facilities housed people under the authority of multiple responding agencies; and when facilities changed status or moved.

The Census Bureau created the MAF as a reference database of the universe of living quarters for the United States, Puerto Rico, and the Island Areas. While the MAF receives updates throughout each decade, comprehensive validations only occur during the decennial census. The Census Bureau is increasingly using linkages to other comprehensive, often publicly available, local, state, and federal administrative records to increase the accuracy of the MAF.

The MAF includes all types of living quarters, including housing units which typically house families, group quarters which includes prisons and detention facilities but also dormitories and nursing homes, and finally transitory locations which include campgrounds and motels. The classification of each type of living quarters type is important for determining how the Census Bureau will enumerate the unit and how the data will be used after enumeration. These classifications were important for determining which GQ records were in-scope for this project.

The MAF is not the only federal source for the universe of prisons and detention facilities. The Bureau of Justice Statistics and the Office of Juvenile Justice and Delinquency Prevention, in the Department of Justice, also conduct censuses of these facilities. The Census Bureau works with these facility censuses, or frames, as part of collections it conducts on behalf of those agencies. The DOJ frames differ in how facilities are classified and surveyed. This project is the first to assess whether these data could be integrated with MAF records despite those differences, and whether that integration could support continual MAF updates.

During the 2020 Census, the COVID-19 pandemic worsened the known challenges associated with enumerating GQs. Jails and prisons went on lock-down to prevent contagion. Early release of some individuals living in prisons and detention facilities was granted to stop disease spread or to isolate high-risk individuals. All of this may have affected data quality and required the Census Bureau to conduct nonresponse follow-up operations and data imputation for incomplete or partial responses (Census Bureau 2024).

Given the challenges of the 2020 Census and recommendations from the Government Accountability Office (GAO), the Census Bureau has been conducting additional coverage research to compare the current GQ enumeration frame with alternative data sources. Likewise, other areas of the Census Bureau have been exploring data collection strategies to improve responses from facilities (GAO 2021, 28; Census Bureau 2024).

This project focused on assessing automated linkage methods and by matching GQs where people live under the supervision of the justice system, including prisons, jails, community corrections housing, and juvenile justice facilities, to administrative lists of justice facilities available from the DOJ. The linked data helped identify potential gaps in coverage, classification, or enumeration in all contributing data frames.

These results demonstrate the benefits of continued development of an integrated database as a long-term solution to support the incorporation of additional administrative sources and increased maintenance frequency.

Project Scope

Census Bureau Master Address File and 2020 Census GQ Types

To identify links between DOJ frames and the MAF, a subset of records from the Master Address File Extract (MAFX) were used for automated matching, initial manual matching, and research for records within GQ classification types that mapped well to the universes of the DOJ frames.[1] After all potential MAFX matches were exhausted, additional resources such as the MAF Browser and the Master Address File/Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) System were used to identify records that existed within the MAF but were not included in the MAFX.

The GQ types listed below were used to select records from the MAFX to match to the DOJ frames. See "Frame Linkage Results and Coverage Assessments" below for summary statistics on each of the GQ types from the 2020 Census.

a. Federal Detention Centers (101)
   GQ type 101s are stand alone, multi-level, federally run correctional facilities that provide short-term confinement or custody of adults pending adjudication or sentencing. These facilities may hold pretrial detainees, holdovers, sentenced offenders, and Immigration and Customs Enforcement (ICE) inmates, formerly called Immigration and Naturalization Service (INS) inmates. These facilities include:
   i. Metropolitan Correctional Centers,
   ii. Metropolitan Detention Centers,
   iii. Federal Detention Centers,
   iv. Bureau of Indian Affairs Detention Centers,
   v. ICE Service Processing Centers, and
   vi. ICE contract detention facilities.

a. Federal and State Prisons (102 and 103)
   GQ type 102 and 103s are adult correctional facilities where people convicted of crimes serve their sentences. Common names include prison, penitentiary, correctional institution, federal or state correctional facility, and conservation camp. The prisons are classified by two types of control: (1) "federal" (operated by or for the Bureau of Prisons of the Department of Justice) and (2) "state." This category may include privately operated correctional facilities.

b. Local Jails and Other Municipal Confinement Facilities (104)

[1] The MAF Extract (MAFX) is a product generated from the MAF to provide customers with a consolidated list of MAF Units and associated addresses, geospatial data, and other attributes. It is used for a variety of purposes within the Census Bureau, including census and survey data collection and tabulation. This study used the 2020, 2022, and 2023 vintages.

GQ type 104s are correctional facilities operated by or for counties, cities, and American Indian and Alaskan Native tribal governments. These facilities hold adults detained pending adjudication and/or people committed after adjudication. This category also includes work farms and camps used to hold people awaiting trial or serving time on short sentences.

c. Correctional Residential Facilities (105)
GQ type 105s are community-based facilities operated for correctional purposes. The facility residents may be allowed extensive contact with the community, such as for employment or attending school, but are obligated to occupy the premises at night. Examples are halfway houses, restitution centers, and prerelease, work release, and study centers.

d. Military Disciplinary Barracks and Jails (106)
GQ type 106s are correctional facilities managed by the military to hold those awaiting trial or convicted of crimes.

e. Correctional Facilities Intended for Juveniles (203)
GQ type 203s include specialized facilities that provide strict confinement for residents and detain juveniles awaiting adjudication, commitment, or placement, and/or those being held for diagnosis or classification. Also included are correctional facilities where residents are permitted contact with the community, for purposes such as attending school or holding a job. Examples are residential training schools and farms, reception and diagnostic centers, group homes operated by or for correctional authorities, detention centers, and boot camps for juvenile delinquents.

DOJ Frames and Their Universes

This project used four frames from the DOJ that covered prisons or detention facilities. The 2019 vintages of the BJS Census of State and Federal Adults Correctional Facilities (CCF), Census of Jails (COJ), and Annual Survey of Jails in Indian Country (SJIC) were acquired from the public National Archive of Criminal Justice Data. The 2020 vintage of the Juvenile Residential Facility Census and the facility frame used for the 2019 vintage of the Census of Juveniles in Residential Placement were used with permission from OJJDP. The 2019 vintages for the CCF and COJ were the most current censuses available at the time of the study. These vintages also reflected the most accurate pre-pandemic facility counts and removed the potential for nonresponse because of COVID-19.

a. Census of State and Federal Adults Correctional Facilities (CCF)
Conducted approximately every five to seven years, the Census of State and Federal Adult Correctional Facilities (CCF) collects facility-level data on the operations of facilities and the conditions of confinement, including facility capacity and crowding, court orders, safety and security within prisons, security-staff workload, overall facility function, programming, work assignments, and special housing. The CCF furnishes the sampling frame for the nationwide Survey of Prison Inmates.

The 2019 CCF covered adult correctional facilities operated by state departments of corrections, the Federal Bureau of Prisons (BOP), and private contractors in all 50 states, including the combined jail and prison systems in Alaska, Connecticut, Delaware, Hawaii, Rhode Island, and

Vermont (Bureau of Justice Statistics 2022a). Facilities that were included in the CCF housed prisoners primarily for state or BOP authorities; were operational on the day of the census; and were physically, functionally, and administratively separate from other facilities.

b. Census of Jails (COJ)

The Census of Jails (COJ) is part of a series of data collections that studies the nation's local jails and the 12 Federal Bureau of Prisons (BOP) detention facilities that function as jails in the United States. Within this series, the COJ is the only collection that enumerates these locations and provides inmate counts at the jail facility level. The COJ also provides the sampling frame for the Survey of Inmates in Local Jails (SILJ) and the Annual Survey of Jails (ASJ).

The 2019 COJ collected data necessary for producing estimates on local jail populations, including one-day custody counts by sex, race and Hispanic origin, conviction status, and severity of offense (felony and misdemeanor); counts of non-U.S. citizens by conviction status; juvenile counts; holds for state and federal authorities; admissions and releases; and average daily population by sex (Bureau of Justice Statistics 2022b). It also collected data on facility information, including rated and design capacity, staffing, and opioid testing, and treatment programs.

c. Annual Survey of Jails in Indian Country (SJIC)

The Bureau of Justice Statistics' Annual Survey of Jails in Indian Country (SJIC) includes all Indian country correctional facilities operated by tribal authorities or the U.S. Department of the Interior's Bureau of Indian Affairs (BIA). The survey is designed to collect detailed information on confinement facilities, detention centers, jails, and other facilities operated by tribal authorities or the BIA. Information is gathered on inmate counts, movements, facility operations, and staff.

The 2019 SJIC includes reservations, pueblos, rancherias, and other appropriate areas as specified in 18 U.S.C. § 1151 (Bureau of Justice Statistics 2022c).

d. Juvenile Residential Facility Census (JRFC) & Census of Juveniles in Residential Placement (CJRP)

The Juvenile Residential Facility Census (JRFC) was administered for the first time in 2000 by the Census Bureau through an interagency agreement with the OJJDP. The scope of JRFC is designed to collect information about specific facilities that hold juveniles. Not included in the JRFC are adult prison and jails or facilities used exclusively for mental health, substance abuse, or for abused or neglected children.

The biennial JRFC is one part of OJJDP's annual efforts to describe both the youth placed in residential facilities and the environments of these facilities. The companion data collection to JRFC, the Census of Juveniles in Residential Placement (CJRP), collects information in alternating years on the demographics and legal attributes of youth in placement from the same facilities that meet JRFC inclusion criteria. Since the facility universe of the two collections are the same, updates occurred annually.

The JRFC used was the 2020 vintage. The 2019 CJRP vintage was also included to minimize the potential impact of COVID-19 on 2020 facility records.

Mapping Census 2020 GQ Types to DOJ Frame Universes

The Census Bureau, BJS, and OJJDP have different operational requirements for classifying prisons and detention facilities. For example, BJS and OJJDP tend to categorize a facility according to its role in the function of the criminal justice system, while the Census Bureau is more likely to categorize a group quarters unit in a way that reflects how the GQ's operator submitted data or the physical structure of the facility containing the GQ. As a result, the universe of each DOJ frame does not correspond perfectly to a Census 2020 GQ type.

The DOJ frames used in this project also do not include all types of justice facilities that the Census Bureau must enumerate and therefore maintain in the MAF. The following units are out-of-scope for the DOJ frames: some facilities operated by or for U.S. Immigration and Customs Enforcement (ICE) or the U.S. Marshals Service (USMS) (under GQ type 101), some community corrections facilities that may be operated by local governments (under GQ type 105), all military disciplinary barracks and jails (under GQ type 106), and facilities in Puerto Rico and the Island Areas (across multiple GQ types). In the future, frame development will include web scraping to identify and collect new data sources that contain these additional facilities such as facility web sites and audit reports.

Table 1 maps Census 2020 GQ types for prisons and detention facilities to the universes of each DOJ frame.

Table 1: Census 2020 GQ Type to DOJ Frame Crosswalk

Census 2020 GQ Type	DOJ Frame or Alternative Data Source	Relationship	Notable Differences
101: Federal Detention Facilities	Census of Jails	COJ included 12 Federal Bureau of Prisons (BOP) detention facilities.	
	Annual Survey of Jails in Indian Country	SJIC included confinement facilities, detention centers, jails, and other facilities run by tribal authorities or the Bureau of Indian Affairs (BIA).	Census records classified SJIC facilities as federal if managed by a federal agency. SJIC also included state prisons and local jails.
	Alternative data sources	The DOJ Frames did not include all dedicated ICE and USMS facilities.	Federal agency websites that included ICE and USMS facilities were identified as alternative data sources.
102 & 103: Federal & State Prisons	Census of State and Federal Adults Correctional Facilities	The CCF covered adult correctional facilities run by state departments of corrections, the BOP, and private contractors in all fifty states, including the combined jail and prison systems in Alaska, Connecticut, Delaware, Hawaii, Rhode Island, and Vermont.	CCF listings were facility level where as the MAF listed prisons with a single ID and other prisons with multiple IDs (an ID for each GQ within the larger facility). It was difficult to know if the MAF listing(s) captured all the potential GQs and population within the facility.

	Annual Survey of Jails in Indian Country	SJIC included confinement facilities, detention centers, jails, and other facilities run by tribal authorities or the BIA.	Census records classified SJIC facilities as federal if managed by a federal agency. SJIC also included state prisons and local jails.
	Alternative data sources	Prison facilities located in Puerto Rico (PR) and the Island Areas.	Based on comparisons to the MAFX, additional research and sources are needed to verify coverage.
104: Local Jails and Other Municipal Confinement Facilities	Census of Jails	COJ included data from the nation's local jails.	
	Census of State and Federal Adults Correctional Facilities	The CCF included the jails run by combined jail and prison systems in Alaska, Connecticut, Delaware, Hawaii, Rhode Island, and Vermont.	
	Annual Survey of Jails in Indian Country	SJIC included confinement facilities, detention centers, jails, and other facilities run by tribal authorities or the BIA.	
105: Correction Residential Facilities	Census of State and Federal Adults Correctional Facilities	CCF included community corrections facilities, such as halfway houses, run by state departments of corrections, the Bureau of Prisons, or private contractors.	
	Alternative data sources	Community corrections housing run by state courts, local courts, or local governments.	GQ type 105 included smaller community corrections facilities run at a more localized level. These were not captured in the DOJ data. Additional research is needed to identify alternative data sources.
106: Military Disciplinary Barracks and Jails	Alternative data sources	All military disciplinary barracks and jails.	Web scraping may be able to identify sources for future Census Bureau GQ frame development.
203: Correctional Facilities Intended for Juveniles	Juvenile Residential Facility Census, Census of Juveniles in Residential Placement	JRFC included long-term secure facilities, reception or diagnostic centers and detention centers.	JRFC listings were for individual GQs within larger facilities. The MAF listings had a mix of GQs, sometimes listing each individual GQ, and sometimes listing just the larger facility. JRFC listings also included juvenile GQs housed within an adult facility. MAF records did not always list the juvenile record within the adult facility.

	Alternative data sources	Facilities located in Puerto Rico and the Island Areas.	Based on comparisons to MAFX, additional research and sources are needed to verify coverage.

Source: Authors' analysis of collection methodology documentation and the integrated frame that was created for this report.

Matching Methodology

Because the DOJ frames and the MAF did not share a common identifier, probabilistic record linkage methods were used to make prospective matches based on similarity in facility names and addresses. Subject-matter experts validated all unmatched cases and all but the closest matched cases. The product was a high-quality crosswalk between the MAF and the DOJ facility censuses. The crosswalk was then used to assess the coverage of prisons and detention facilities in the 2020 Census, and to identify where and how coverage might be improved. See Appendix A for additional details on the methodology.

Probabilistic Matching Techniques

The first step of the matching methodology was to convert all the names and addresses to term frequency-inverse document frequency (TF-IDF) matrices, using 3-character n-grams as the unit of analysis. The nearest neighbor algorithm was then used to compare the matrices and identify the highest probable matches based on defined thresholds.

Table 2 shows a simplified example of the record linkage output where the addresses were an exact match and there was variation in the name fields in the MAF. GQ records in the MAF had names for the facility or for the component GQs, which might correspond to buildings or wings of a larger facility. The DOJ facility name was compared to both the MAF facility name and the MAF GQ name since the facility name can appear in either name field. The distance score (Score 1, Score 2, Score 3) was the nearest neighbor score for the subsequent two columns. An exact match received a score of 0.0.

Table 2: Example of record linkage output

Score 1	DOJ Facility Name	MAF Facility Name	Score 2	DOJ Facility Name	MAF GQ Name	Score 3	DOJ Address	MAF Address
0.4	Main Correctional Facility	Main Correctional	1.0	Main Correctional Facility	Unit 1	0.0	123 Main St.	123 Main St.

Source: Fictional example of facility names and addresses created by the authors and not based on 13 U.S.C. § 9a-protected data.

Often a facility with one record in a DOJ frame had multiple plausible matches in the MAF. The most common cases involved MAF records that represented a different dorm, building, or unit at the same facility, or when the MAF had additional records if the facility housed people under contract with another agency, such as ICE or the USMS. The record linkage output was separated into three groups: DOJ records with one potential match in the MAF, DOJ records with multiple matches in the MAF, and DOJ records that did not match to any MAF record. After cleanup of the data files and the addition of other unit features from the MAF, the singular and multiple match files were validated.

Validation Process

The record linkage results went through a combination of automated and manual validation. If the DOJ record had only one potential match or had multiple matches that fit into one of a series of deterministic patterns explained below, the match was considered validated. The remaining potential matches were reviewed manually. See Appendix A for a more detailed explanation on the matching methods and thresholds.

Automated Validation

Automated validation was achieved by pattern identification for DOJ facilities that matched to more than one MAF record. For example:

a. A DOJ record (Main County Jail) matched to three MAF records such as Main County Jail, Main County Jail USMS, and Main County Jail ICE, where the USMS and ICE records signify that the facility had contracted capacity to USMS and ICE and were enumerated directly from those agencies.
b. A DOJ record matched to five records in the MAF. All five MAF records matched to the same DOJ Facility Name and Address, and their MAF GQ names were Dorm A, Dorm B, Dorm C, Dorm D, and Dorm E.
c. Two different DOJ records with similar names matched to two different MAF records with similar names. The example below shows cross-matched records. The highlighted rows were non-matches.

Table 3: Example of many-to-many matches due to slight name variations

DOJ Facility Name	MAF Facility Name	DOJ Address	MAF Address
Washington County Jail - North	Washington County Jail North	123 Main Street	123 Main St
Washington County Jail - North	Washington County Jail South	123 Main Street	98 St Hwy 63
Washington County Jail - South	Washington County Jail North	98 State Highway 63	123 Main St
Washington County Jail - South	Washington County Jail South	98 State Highway 63	98 St Hwy 63

Source: Fictional example of facility names and addresses created by the authors and not based on 13 U.S.C. § 9a-protected data.

d. There could be differences in naming practices between the MAF and a DOJ frame, but if there were standard naming conventions across frames, reconciliation would be a smaller burden. In this example, the DOJ used “Satellite Prison Camp” (SCP) while the MAF used “Camp”.

Table 4: Example of one-to-many relationship (from DOJ to MAF)

DOJ Facility Name	MAF Facility Name	DOJ Address	MAF Address
Washington - USP and SCP	Washington - Camp BOP_D	456 Main Street	456 Main Street
Washington - USP and SCP	Washington USP BOP_D	456 Main Street	456 Main Street

Source: Fictional example of facility names and addresses created by the authors and not based on 13 U.S.C. § 9a-protected data.

Manual Validation

Records with singular matches and above-threshold distance scores, and records with multiple matches that could not be automatically validated, were validated manually. This process relied on a combination of supplemental data from the MAF and the DOJ frames (e.g., population, capacity, or facility functions); internet mapping services; state and local prison and jail websites; and subject-matter expertise. Internal Census Bureau resources were also used to research potential matches using facility addresses and names in the MAF for records that may not have been available in the MAFX.

Residual Units in the DOJ and MAFX GQ Frames

DOJ records that had no initial match or had potential matches that were determined to be incorrect were processed again on a different subset of MAF records. See Figure 1 in Appendix A for the details of this process. Results from each subsequent linkage went through the same process of validation as described above. After all iterations of record linkage and validation were complete, there was a small set of remaining records from each of the DOJ frames and the MAFX for which there was no corresponding matching record.

Manual linkage of the residual DOJ records using internal Census Bureau resources was the last step. Manual linkage included matches to non-GQ housing units or non-residential structures, such as government or commercial properties.

There was an additional set of remaining records in the MAFX that were in scope (prisons and detention facility GQ types that were enumerated in 2020) and for which there were no matching records identified in any of the DOJ frames. Some of these records were used to identify gaps in the DOJ frames, including some community corrections facilities that were operated by local governments (under GQ type 105), all military disciplinary barracks and jails (under GQ type 106), and facilities in Puerto Rico and the Island Areas (across multiple GQ types).

Frame Linkage Results and Coverage Assessments

Probabilistic record linkage combined with validation by subject-matter experts resulted in a high-quality integrated frame of justice facilities. This frame mapped unique identifiers for records in the DOJ frames to MAF Identifiers (MAFIDs) that index GQs. Linked MAFIDs were used to extract additional unit features from the MAF which were used to assess overall coverage differences between the DOJ frames and the MAF.

As discussed above, the DOJ frames were not a perfect reference point for the MAF frame used to enumerate prisons and detention facilities in the 2020 Census. Not all GQ types were covered by the DOJ frames and there was some temporal misalignment between the 2019 vintage DOJ frames and the 2020 MAF frame, during which there were known facility function or location changes due to either natural change or Covid-19.

Prisons and Detention Facilities Records by Data Source and Facility Type

The tables below show the number of records in each of the DOJ frames and in the MAFX by GQ type. Table 5 shows the number of records in each of the DOJ frames. There were 6,535 units across the four frames, and almost 45 percent of the records were from the Census of Jails.

Table 5: Prison and Detention Facilities in All DOJ Frames, by Frame

DOJ Frame	Number of DOJ Facilities	Percent of All DOJ Facilities
Census of State and Federal Adult Correctional Facilities (CCF)	1,677	25.7%
Census of Jails (COJ)	2,936	44.9%
Juvenile Residential Facility Census (JRFC)	1,842	28.2%
Annual Survey of Jails in Indian Country (SJIC)	80	1.2%
All DOJ Frames	6,535	100.0%

Source: Authors' calculations from BJS and OJJDP frames using no 13 U.S.C. § 9a-protected data.

Using records sourced from the MAFX, Table 6 shows the number of prison and detention facility GQs that were enumerated in the 2020 Census, by GQ type. Military Disciplinary Barracks and Jails (GQ type 106) were excluded here and in further analysis because there are few records and no overlap with the four DOJ frames. The most prominent difference between Tables 5 and 6 is that there were almost three times as many in-scope GQs in the MAFX as there were facilities in the DOJ frames. GQs listed in the MAF included individual buildings or wings, while records in DOJ frames represented facilities that might include multiple GQs. In the MAFX, state prisons (GQ type 103) constituted more than half of the records. There are more than five times as many state prison GQs as there are facility records in the CCF, which includes federal prisons and many community corrections facilities as well. In comparison, the ratio of juvenile justice GQs (GQ type 203) to facility records in the JRFC is much closer to one.

Table 6: Prison and Detention Facility Group Quarters from the 2020 Census, by Group Quarters Type

Census 2020 Group Quarters Type	Number of Group Quarters	Percent of All In-Scope Prison and Detention Facility GQs
101 Federal Detention Centers	2,300	12.2%
102 Federal Prisons	250	1.4%
103 State Prisons	9,600	50.6%
104 Local Jails	3,800	19.8%
105 Correctional Residential Facilities	1,100	6.0%
203 Correctional Facilities for Juveniles	1,900	9.9%
All In-Scope Prison and Detention Facility Group Quarters from the 2020 Census	19,000	100.0%

Source: Authors' calculations from the Master Address File Extract (Project number 7510276, approval CBDRB-FY24-020).
Note: This table includes records from the MAFX that were enumerated or found vacant in the 2020 Census. All in-scope GQ types were included except for Military Disciplinary Barracks and Jails (GQ type 106) because that group includes few GQs and no overlap with the DOJ frames. All estimates rounded according to Census Bureau DRB rounding rules. Because of rounding, cells may not sum to column totals.

Match Rates for Prison and Detention Facility Group Quarters from the 2020 Census

Records from the DOJ frames were first matched to group quarters records from the MAFX that were included in the 2020 Census. Linkage was restricted to the GQ types that covered prisons and detention facilities and to records that were vacant or enumerated during the 2020 Census. Table 7 shows the linkage rates between this set of MAFX records and the DOJ frames. In total, 86 percent of these MAFX records matched to at least one DOJ record.

About 2,700 MAFX records, or 14 percent of in-scope MAFX records, were not matched to a DOJ record. Qualitative analysis of unmatched records showed that they tended to be coded as vacant in 2020, which could reflect closures or pandemic releases. Some of the unmatched records were enumerated or found vacant but were outside the scope of the DOJ frames such as some ICE or USMS facilities; facilities located in Puerto Rico and the Island Areas; and community corrections facilities run by local governments. Unmatched records tended to be small GQs associated with small county jails, police holding cells, or halfway houses.

Analysis of the in-scope MAFX records that could not be linked to the DOJ frames helps identify data sources and areas of future research that could help accomplish comprehensive coverage in the

integrated frame. This research would include spatial analysis of GQ building coordinates relative to DOJ facility address locations. Potential data sources include parcel data sets; collections of geographic prison boundary footprints, such as data from the Department of Homeland Security Homeland Infrastructure Foundation-Level Data (HIFLD); and web scraping.

Table 7: Prison and Detention Facility Group Quarters from the 2020 Census, by Group Quarters Type and Whether Records Matched to DOJ Frames

Census 2020 Group Quarters Type	MAFX records matched to DOJ records	Percent of MAFX records matched to DOJ records	MAFX records not matched to DOJ records	Percent of MAFX records not matched to DOJ records
101 Federal Detention Centers	1,400	62.6%	850	37.4%
102 Federal Prisons	250	89.2%	30	10.8%
103 State Prisons	9,100	94.8%	500	5.2%
104 Local Jails	3,300	88.4%	450	11.6%
105 Correctional Residential Facilities	650	58.7%	450	41.3%
203 Correctional Facilities for Juveniles	1,500	80.5%	350	19.5%
Total	16,000	86.0%	2,700	14.0%

Source: Authors' calculations from integrated frame based on links between the Master Address File Extract, Census of State and Federal Adult Correctional Facilities, Census of Jails, Annual Survey of Jails in Indian Country, Juvenile Residential Facility Census, and Census of Juveniles in Residential Placement (Project number 7510276, approval CBDRB-FY24-020).
Note: This table includes records from the MAFX that were enumerated or found vacant in the 2020 Census. All in-scope GQ types were included except for Military Disciplinary Barracks and Jails (GQ type 106) because that group includes few GQs and no overlap with the DOJ frames. All estimates rounded according to Census Bureau DRB rounding rules. Because of rounding, cells may not sum to column totals. Rows in this table are components of the rows shown in Table 6, but the rows here may not sum to those in Table 6 because of rounding.

Matched Records in the Integrated Justice Facility Frame

After linking records from the DOJ frames to GQs that were enumerated or found vacant in the 2020 Census, not all records from the DOJ frames could be linked so other records from the MAF had to be considered. Once all possible candidate records were considered, more than 99.5 percent of records from the DOJ frames were linked to MAF records. In this section, we describe some of the features of that integrated frame.

Table 8 describes matched records from the integrated frame by GQ type and by DOJ frame. The table includes both matches to MAFX records as provided in Table 7 and matched MAF records that resulted from manual validation and research of MAF records not enumerated or found vacant in the 2020 Census.

Most DOJ records matched to MAF records with the anticipated GQ types. For example, 88.3 percent of CCF records matched to 103 State Prisons and 66.0 percent of COJ records matched to 104 Local Jails.

Facilities in the JRFC/CJRP data were most likely to match to out-of-scope GQ types at 37.1 percent. Many of these JRFC/CJRP records matched to non-correctional juvenile group quarters (201 Group Homes for Juveniles and 202 Residential Treatment Centers).

Table 8: Matched Records from Integrated Justice Facility Frame, by DOJ Frame and Census 2020 Group Quarters Type

	DOJ Frame				
Census 2020 Group Quarters Type	CCF	COJ	JRFC/ CJRP	SJIC	All DOJ Frames
101 Federal Detention Centers	1.1%	25.5%	0.9%	51.4%	8.0%
102 Federal Prisons	2.2%	0.2%	0.0%	D	1.3%
103 State Prisons	88.3%	1.5%	0.5%	D	50.0%
104 Local Jails	0.3%	66.0%	1.7%	28.8%	18.6%
105 Correctional Residential Facilities	5.5%	1.7%	0.9%	D	3.7%
203 Correctional Facilities for Juveniles	0.7%	0.3%	51.8%	11.7%	8.9%
All other GQ types	0.8%	0.2%	37.1%	0.0%	6.5%
Matched to address without GQ type	1.0%	4.7%	7.1%	D	3%
Total, percent	100%	100%	100%	100%	100%
Total, count	10,500	5,000	3,000	100	18,500

Source: Authors' calculations from integrated frame based on links between the Master Address File Extract, Master Address File, Census of State and Federal Adult Correctional Facilities, Census of Jails, Annual Survey of Jails in Indian Country, Juvenile Residential Facility Census, and Census of Juveniles in Residential Placement (Project number 7510276, approval CBDRB-FY24-020).
Note: This table includes all matched records from the MAF and the DOJ frames. After matches were made to MAFX records with in-scope GQ types, nonmatched records were linked to MAF records not included in the MAFX including records not in-scope for the 2020 Census. "All other GQ Types" included links to GQs in the following categories: Juvenile Group Homes / Residential Treatment Centers (201, 202), Skilled Nursing Facilities (301), Psychiatric / Hospice Facilities / Residential Schools for People with Disabilities (401, 403, 405), Emergency and Transitional Shelters (701), Adult Group Homes & Treatment (801, 802), Workers/Religious/Natural Disaster GQs (901, 902, 903). "Matched to address without GQ Type" included records not associated with GQs in the MAF. All estimates rounded according to Census Bureau DRB rounding rules. Because of rounding, cells may not sum to column totals. D: Statistic not disclosed because of small cell size.

There is anecdotal and recorded evidence that the MAF frame sometimes classifies facilities that contain more than one GQ as a single entity and sometimes as multiple entities. More standardized coding of parent (facility)-child (GQ) relationships could enhance the Census Bureau's ability to identify the correct contacts or respondents (GAO 2021). The DOJ frames had some variation in facility classification, but overall were more likely to represent entire facilities as a single entity. This study confirmed that the parent-child or facility-GQ relationships were often represented differently in the DOJ frames versus in the MAF frame. These relationships can be difficult to identify due to variation in methodology as to how and why the data were collected, recorded, and updated. As discussed in "Matching Methodology" above, facility-GQ relationships were classified as follows:

a. One DOJ record matched to one MAF record;

b. One DOJ record matched to multiple MAF records or one MAF record matched to multiple DOJ records; or
c. Groups of many DOJ records matched to groups of MAF records.

Table 9 shows the match relationships overall and by frame. Most matches (54.9 percent) involved one DOJ record matching to one MAF record followed by one DOJ record matching to many MAF records (36.9 percent).

CCF records (63.5 percent) had the highest proportion of DOJ records coding to multiple MAF records while COJ and JRFC/CIRP had higher rates of one-to-one record relationships. SJIC records were distributed among the relationship categories with a higher proportion of records in the one MAF to many DOJ and many-to-many groups than the other DOJ Frames.

Table 9: Matched Records from Integrated Justice Facility Frame, by DOJ Frame and Match Relationship

Facility/Group Quarters Match Relationship	Percent of matched CCF records	Percent of matched COJ records	Percent of matched JRFC/ CJRP records	Percent of matched SJIC records	Percent of all matched records from DOJ Frames
One DOJ record : one MAF record	31.9%	59.3%	69.3%	50.0%	54.9%
One DOJ record : many MAF records	63.5%	37.0%	13.0%	20.0%	36.9%
Many DOJ records : one MAF record	1.4%	1.7%	10.0%	10.0%	4.1%
Many DOJ records : many MAF records	3.3%	2.0%	7.6%	20.0%	4.1%
Total, percent	100%	100%	100%	100%	100%
Total, count	1,700	2,900	1,800	80	6,500

Source: Authors' calculations from integrated frame based on links between the Master Address File Extract, Census of State and Federal Adult Correctional Facilities, Census of Jails, Annual Survey of Jails in Indian Country, Juvenile Residential Facility Census, and Census of Juveniles in Residential Placement (Project number 7510276, approval CBDRB-FY24-020).
Note: This table includes all matched records from the MAF and the DOJ frames. All estimates rounded according to Census Bureau DRB rounding rules. Because of rounding, columns may not sum to 100%.

Measuring Census 2020 Coverage with the Integrated Frame

Facility records from the 2019 DOJ frames represent an alternative reference point for assessing coverage of prisons and detention facilities in the 2020 Census. The integrated frame of MAF and DOJ records enables the identification of which facilities in the DOJ frames were included in the 2020 Census. Facility population counts in the DOJ frames allow similar coverage analysis at the resident population level. In this section, coverage of prisons and detention facilities is assessed at the facility-level and the resident population-level using the integrated frame.

To operationalize the coverage analysis using the integrated frame, business rules were required for matched groups that did not involve one DOJ facility matched to one MAF GQ. To simplify analysis, the relatively rare many-to-many matches were excluded from the coverage assessment. For one DOJ record to many MAF record matches and many DOJ to one MAF record matches, a hierarchical rule shown in Table 10 was used to assign a single Census 2020 enumeration status code to the matched group of records. If at least one linked MAFID was enumerated as a GQ, the attributes from this MAFID were used in the summary of matched facilities. If no enumerated GQ was found, then the priority rank continued in the following order: enumerated as an HU, non-enumerated GQs, non-enumerated HU, then finally matched MAF records that did not have any GQ, HU, or enumerated attributes.

Once a Census 2020 enumeration status was assigned to each match group, the associated DOJ-measured population was assigned to the group. If multiple MAF records were in the match group, their Census 2020 populations were summed to create a single match group population. Where a population count was available for both DOJ and MAF records in a match group, the DOJ population was found to be a good proxy for the Census 2020 population measure.[2]

Table 10: One-to-Many Priority Ranking Business Rules

	Priority Ranking When One DOJ Record Matched to Many MAF Records
1.	Enumerated GQ
2.	Enumerated housing unit
3.	Not enumerated GQ – vacant
4.	Not enumerated GQ – duplicate
5.	Not enumerated GQ – non-residential
6.	Not enumerated housing unit – not in Census
7.	Not enumerated GQ – non-existent
8.	Not enumerated GQ – unknown
9.	No enumeration status or code

Source: Authors' creation for analysis purposes only.

[2] To understand the comparability of population counts in the MAF and in the DOJ frames, the correlation between facility population counts in the MAF from the 2020 enumeration and the facility population counts from the linked facility records in the DOJ frames were calculated. The correlation coefficients were calculated separately for each DOJ frame, using the linked facility records within a particular DOJ frame. Correlation coefficients represented facilities with one-to-one or one-to-many relationships in our linked frame, where at least one DOJ record and at least one MAF record had non-missing population. For facilities with one-to-many relationships, populations within a facility were summed to calculate a single total facility population value. GQs not enumerated or found to be vacant in 2020 were coded as having zero population. Once these aggregations were made to records in the linked frames, correlation coefficients were representative of the units in each DOJ frame with non-missing populations. The correlation coefficients for population were 0.97 for the CCF, 0.90 for the COJ, 0.58 for the JRFC, 0.64 for the SJIC, and 0.96 across the combined frame. A higher correlation for the COJ reflected the relative homogeneity of jails, while the low correlation for the JRFC reflected how that collection was different and the juvenile residential facilities may be found within larger GQ units. (Source: Authors' calculations from the integrated frame based on links between the Master Address File Extract, Census of State and Federal Adult Correctional Facilities, Census of Jails, Annual Survey of Jails in Indian Country, Juvenile Residential Facility Census, and Census of Juveniles in Residential Placement [Project number 7510276, approval CBDRB-FY24-020].)

Table 11 shows that 86.8 percent of matched DOJ records were enumerated in the 2020 Census, with 84.9 percent enumerated as GQs and 1.9 percent enumerated as traditional housing units (not as GQs). Around 4.2 percent of facilities were coded as vacant. Of the matched facilities, 9.1 percent were not included in the 2020 Census. These units were coded as non-residential or nonexistent prior to Census 2020 operations.

The population coverage estimates in Table 11 are based on population counts in the DOJ frames (i.e., no 2020 Census population data were used). The population in the matched DOJ justice facilities represented an estimated 98.4 percent of the total count from the DOJ-measured population from matched facilities. For facilities not enumerated or found vacant in the 2020 Census, the missed population, as measured by the DOJ populations, was a smaller fraction of the total than the missed share of facilities (1.6 percent non-enumerated DOJ-measured population compared to 9.1 percent non-enumerated facilities). These results confirmed that while facilities that housed most of the DOJ population were enumerated, additional facilities were identified that may have not been represented correctly in either the MAF or the DOJ frames because of change over time.

Incorrect classifications potentially impacted overall coverage. The remaining non-enumerated matched records (9.1 percent) or misclassified units (1.9 percent) represent areas where there would be immediate benefits of integrated data, which would help target research and data corrections.

During the 2020 Census, some facilities were closed to reduce the spread of COVID-19. Of matched facilities, 4.2 percent were found to be vacant. The share of the DOJ-measured population count attributed to those vacant facilities was comparatively lower at 1.1 percent. These results indicate that COVID-19 pandemic facility closures had a greater impact on facility coverage than on the population count.

Table 11: Facility-Level and Resident Population-Level Coverage Estimates Using the Integrated Justice Facility Frame

Census 2020 Enumeration Status of Matched Records		Percent of Matched DOJ Facilities	Percent of DOJ Population Residing in the Matched MAF Facility
Enumerated or found vacant in 2020 Census	Enumerated as group quarters	84.9%	97.1%
	Enumerated as housing unit	1.9%	0.3%
	Vacant group quarters	4.2%	1.1%
	Subtotal	90.9%	98.4%
Not enumerated nor found vacant in the 2020 Census	Residential	5.1%	1.0%
	Nonresidential	4.0%	0.5%
	Subtotal	9.1%	1.6%
Total		100%	100%

Source: Authors' calculations from integrated frame based on links between the Master Address File Extract, Census of State and Federal Adult Correctional Facilities, Census of Jails, Annual Survey of Jails in Indian Country, Juvenile Residential Facility Census, and Census of Juveniles in Residential Placement (Project number 7510276, approval CBDRB-FY24-020).

Note: The DOJ frames are from 2019 and so there may be some expected differences in the universe of facilities and the populations residing in those facilities during the 2020 Census. This table includes all matched records from the MAF and the DOJ frames except for group of many DOJ records matched with many MAF matches. For single DOJ records that matched to more than one MAFID, the hierarchy shown in Table 10 was used to classify each DOJ record with priority going to MAF records with GQ codes, followed by housing unit codes, vacant codes, and finally not-in-census codes. Populations used for coverage estimates came from the DOJ frames, not from the 2020 Census. Coverage estimates by DOJ frame are available in Appendix B. All estimates rounded according to Census Bureau DRB rounding rules. Because of rounding, columns may not sum to 100%.

Because of the process of aggregating GQs to facilities in a match group, the enumeration results shown in Table 11 assess whether any portion of the facility was included in enumeration. This only tells one part of the coverage story. A disadvantage of this simplified classification is that it may overrepresent coverage of larger DOJ facilities because a single MAF record may only be attributed to a single GQ, whereas the DOJ record represents an entire facility containing multiple GQs.

Additional analysis is recommended to assess coverage in facilities with one-to-many and many-to-many parent-child type relationships to assess internal GQ coverage. These relationships will need to be understood to develop a comprehensive integrated frame and to accurately associate administrative data with specific justice facilities.

Observations and Recommendations from Frame Integration

The linkage process and analysis of the integrated frames were used to develop a set of observations about prisons and detention facilities in the MAF and recommendations for improving data quality and enumeration coverage.

a. DOJ frames matched to MAF records at a rate around 99.5 percent. Of these, 90.9 percent were enumerated as GQs, enumerated as housing units, or found to be vacant. The population in the DOJ justice facilities that matched to enumerated and vacant MAF records represented an estimated 98.4 percent of the total population count from the DOJ-measured population. The remaining non-enumerated matched records (9.1 percent) or misclassified units (1.9 percent) demonstrated additional areas of achievable coverage improvement.

    Recommendation 1: Most of the population associated with DOJ records reside within facilities that match to the MAF. This report confirms that DOJ frames are high-quality sources that the Census Bureau should use for updating GQ and address records, validating census data collection, and assessing MAF coverage.

    Recommendation 2: Observed differences between the 2019 DOJ data and the 2020 Census data required some manual matching, indicating a need for continuous updating as GQ attributes can change within a short amount of time. The Census Bureau should continue the development of an integrated MAF/DOJ database that will increase MAF update frequency and streamline processing, and thereby reduce in-office and/or field verification costs and deduplication efforts across frames. Continued maintenance and

research can support both the decennial census and reimbursable survey needs by building on existing Census Bureau infrastructure.

Recommendation 3: Use the integrated database to identify gaps and deficiencies in the data sources. Alternative data sources should then be identified by targeted research or web scraping and, when publicly available, shared with BJS/OJJDP. This increased collaboration will benefit all agencies by improving data currency and coverage for both public and protected data while assuring compliance with data stewardship and security requirements.

b. Although the MAF contains MAFIDs that represent individual GQs within larger facilities, the MAF does not have easily identifiable attributes that establish facility-to-GQ relationships. These missing data structures could potentially lead to overcounts (double counting a child GQ), undercounts (missing a child GQ), or misclassifications of the population within a child GQ or attributing it to a different GQ type. Identifying these relationships is key to reducing future risks, incorporating administrative datasets, and improving coverage.

   Recommendation 4: Support continued development of the integrated database for capturing GQs located within larger facilities and differentiating between a facility name and address and associated GQ names and addresses.

   Recommendation 5: Explore relational data configurations that allow for easy identification of parent-child relationships to simplify linkages, enable administrative updates, and enable extraction specific to survey needs. Tools include improved MAF classification codes, new parent-child codes, crosswalk / bridge tables spatial tools. Parent-child links will also enable new person-level data sources to be used for enumeration.

   Recommendation 6: Use the integrated database to discover data gaps at both the facility and GQ level. Web scraping and compliance form parsing will be important for a comprehensive frame update system.

c. Probabilistic automated address matching produces reliable links between facilities and addresses. Review of matching outcomes found variations that impacted automated matching. Common variations include differing yet related facility and GQ names (e.g., “Youth Detention” versus “Juvenile Corrections”); differing acronyms (e.g., “SCP” versus “CAMP”); differing addresses or zip codes; and incorrect cross-matching of facility names and addresses.

   Recommendation 7: Develop facility record linkage code so that new frame lists can be integrated with higher match rates in the future. This code would improve the import of enumeration data from agencies with differently structured data, such as rosters from federal and state agencies that contract detention services from other agencies. These linkage tools will also need to differentiate between facility name and address and the associated GQ names and addresses to avoid false matches.

Recommendation 8: Research additional methods for using the DOJ frames data as a tool for quality assurance such as population or capacity counts and location information.

Recommendation 9: Use linked and residual non-matched results to identify records that were coded incorrectly or may be GQs that are difficult to locate. For example, units currently coded as non-residential and can be verified as current GQs, correcting incomplete delete/move actions conducted during address canvassing, verifying updates from the Local Update of Census Addresses (LUCA) program, identifying missed GQs, and correcting duplicate GQs.

Recommendation 10: Research spatial relationships between parent-child relationships to identify potentially missed relationships and gain understanding of how facilities and associated GQs are spatially related.

Recommendation 11: Categorize facilities that were not covered by the DOJ frames, determine whether they were out-of-scope or inexplicably missed, and either address their absence from the DOJ frames or find other frames to integrate that cover these facilities.

Next Steps for Implementing Recommendations

To move the linked frame forward as an ongoing maintenance tool and reconcile the records identified for coverage improvement, a database schema must be developed that can support customers in many areas including those that support reimbursable survey, Census of Governments, and decennial census programs. This includes:

1. Creation of an integrated prisons and detention facilities frame that identifies parent-child relationships and includes development of relational tools to enable continued maintenance. A database specification will be developed to account for differences in how facilities and units within facilities are classified differently by the Census Bureau and the DOJ. The data structures should also allow for export based on different survey needs. For example, the possibility that facility data is collected from an operating agency like a county sheriff or a contracted group or agency such as ICE.

2. Research on alternative data sources such as websites and regulatory compliance documents. Develop a web scraping prototype, tools for parsing websites and compliance documents, and modules for incorporating new data into the integrated database.

3. Development of procedures to incorporate maintenance of the integrated frame into reimbursable survey updates. Agreements will be pursued with potential partner agencies to integrate their operational frames into the project. Planning will begin on infrastructure to support frame maintenance during survey operations.

Glossary

Acronym	Definition
BIA	Bureau of Indian Affairs
BJS	Bureau of Justice Statistics
BOP	Federal Bureau of Prisons
CCF	Census of State and Federal Adults Correctional Facilities
CJRP	Census of Juveniles in Residential Placement
COJ	Census of Jails
DOJ	Department of Justice
ERD	Economic Reimbursable Surveys Division
GAO	Government Accountability Office
GQ	Group quarters
ICE	U.S. Immigration and Customs Enforcement
JRFC	Juvenile Residential Facility Census
LUCA	Local Update of Census Addresses
MAF	Master Address File
MAFID	Master Address File Identifier
MAFX	Master Address File Extract
OJJDP	Office of Juvenile Justice and Delinquency Prevention
PR	Puerto Rico
SCP	Satellite Prison Camp
SJIC	Bureau of Justice Statistics Annual Survey of Jails in Indian Country
TD-IDF	Term frequency-inverse document frequency
USMS	U.S. Marshals Service

References

Bureau of Justice Statistics. 2022a. Census of State and Federal Adult Correctional Facilities, 2019. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR38325.v2

Bureau of Justice Statistics. 2022b. Census of Jails, 2019. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR38323.v1

Bureau of Justice Statistics. 2022c. Annual Survey of Jails in Indian Country, 2019. Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR38111.v2

Census Bureau. 2024. 2020 Census Topic Report: Potential Quality Impact of the COVID-19 Pandemic. https://www2.census.gov/programs-surveys/decennial/2020/program-management/evaluate-docs/EAE-2020-quality-impact-covid19.pdf

Government Accountability Office [GAO]. 2021. Decennial Census: Bureau Should Assess Significant Data Collection Challenges as It Undertakes Planning for 2030. Report Number GAO-21-365. https://www.gao.gov/assets/gao-21-365.pdf.

Government Accountability Office [GAO]. 2023. A More Complete Lessons Learned Process for Cost and Schedule Would Help the Next Decennial. Report Number GAO-23-105819. https://www.gao.gov/assets/d23105819.pdf

Appendices

Appendix A: Matching Methodology

Details regarding the probabilistic record linkage are provided below.

Probabilistic Matching Techniques

The first step in the record linkage algorithm was to convert the facility name and addresses into TF-IDF matrices. A TF-IDF score represented the importance of a particular word within a record. This score reflected how often the word appeared in the record (the more the word occurred, the more important it was) as well as how often the word appeared in the full set of records (the more the word occurred across records, the less important it was). Once these records had been converted to TF-IDF matrices, the Nearest Neighbors algorithm compared these matrices to find which records in the MAF matched to the records in the DOJ frame. Thresholds were set here to exclude neighbors with high distance values.

The output file from the record linkage contained one row for every potential match between a DOJ record and a MAF record. Each row contained distance values (between 0 and 1) to quantify the similarity between the DOJ record and the MAF record attribute. A distance value of 0 meant the fields were an exact match. The fields used from the MAF were MAF Facility Names, MAF GQ Name and MAF Address. These were compared with the facility name and address fields from the DOJ frames (except for the Survey of Jails in Indian Country, for which addresses were not available). This process was done independently for each of the four DOJ frames with slight modifications since the frames varied by available attributes.

Table 12 below contains an example of a subset of columns from one row of the record linkage output. Columns 1, 2 and 3 compared the DOJ Facility and the MAF Facility Name. They were similar, but slightly different, so the distance value in the Match 1 field was 0.4. Columns 4, 5 and 6 compared the DOJ Facility and the MAF GQ Name. The MAF GQ Name identified a GQ unit within the facility. This name was entirely different from DOJ Facility Name; therefore, these two attributes did not match. As a result, the Match 2 field was left empty. Columns 7, 8, and 9 compared the DOJ Facility Address with the MAF address. They were the same, so the distance value equaled zero.

Table 12: Example of Record Linkage Output

1	2	3	4	5	6	7	8	9
DOJ Facility Name	**MAF Facility Name**	**Match 1**	**DOJ Facility Name**	**MAF GQ Name**	**Match 2**	**DOJ Address**	**MAF Address**	**Match 3**
Central Correctional Facility	Central Correctional	0.4	Central Correctional Facility	Unit 1		123 Main Street	123 Main Street	0.0

Source: Fictional example of facility names and addresses created by the authors and not based on 13 U.S.C. § 9a-protected data.

Validation process

The record linkage results went through a combination of automated and manual validation. For DOJ

records with one exact potential match in the MAF, if the address and names had distance values below a certain threshold, they were automatically validated. Thresholds were determined by manually validating a random sample of the linkage output and identifying the distance values beneath which linkages were always verified as valid. Thresholds varied by frame and were chosen to be low enough that the chance of validating an incorrect match was low. Remaining potential matches which did not meet the threshold were reviewed manually.

For DOJ records with multiple potential matches in the MAF, if all potential matches for a specific DOJ record followed a certain pattern and were below set matching thresholds (see "Automated Validation" section below), they were automatically validated. All remaining potential matches were manually validated. In addition to determining that certain potential matches were valid, relationships between MAF records were also identified in cases where multiple MAF records matched to one DOJ record. If all the MAF records matched to one DOJ record had a different MAF GQ Name specifying a unit within the facility, the relationship was considered a "GQ unit". If there was a generic MAF record matching a DOJ facility but there were other matching MAF records that indicated a contract with another agency, the unique records were marked as "child" records whereas the generic record was marked as the "parent".

DOJ records that either had no initial match, or their potential matches were determined to be incorrect after validation, were run through the linkage process again on a different subset of MAF records. See Figure 1 below for the details of this process, including the linkage criteria for each subset of MAF records. Results from each subsequent linkage went through the same process of validation as described above. After all iterations of algorithmic linkage, facilities from the DOJ with no match were searched for manually on the MAF browser. DOJ and MAF records that still had no valid match were then determined to exist only in their respective frames.

Automated Validation

Validation was automated by identifying patterns across cases where one DOJ record matched to multiple MAF records. These patterns included:

a. A DOJ record (Washington County Jail) matched to three MAF records (Washington County Jail, Washington County Jail USMS, and Washington County Jail ICE)
b. A DOJ record matched to five records in the MAF. All five MAF records had the same MAF Facility Name and the same address, and their MAF GQ Names were Dorm A, Dorm B, Dorm C, Dorm D, and Dorm E.
c. Two different DOJ records with similar names matched to two different MAF records with the same similar names.
d. There could be differences in naming practices between the MAF and a DOJ frame, but if there were standard naming conventions across frames, reconciliation would be a smaller burden. In this example, the DOJ used "SCP" whereas the MAF used "Camp".

A python script was written to identify cases that fit into one of the patterns and then automatically categorize the potential matches as correct (if below the determined thresholds) or incorrect. Remaining potential matches that were unable to be validated through the automated process were then manually validated.

Manual Validation

Potential matches from the algorithmic record linkage needed further investigation to determine validity. Below is a list of the scenarios encountered for potential matches and the methods used to resolve the records:

a. The DOJ record and the MAF record had the same or similar names but different addresses.

   i. This difference may be related to the name of a cross street leading to the same building. The difference could also be the result of a street having two different names (verified by searching on a web mapping service). Additionally, the MAF Browser may also list additional addresses for this record and one of the additional addresses might match the DOJ address. If any of those cases were true, this was a match.

   ii. If the addresses lead to separate places, there were scenarios which still qualified as a match.

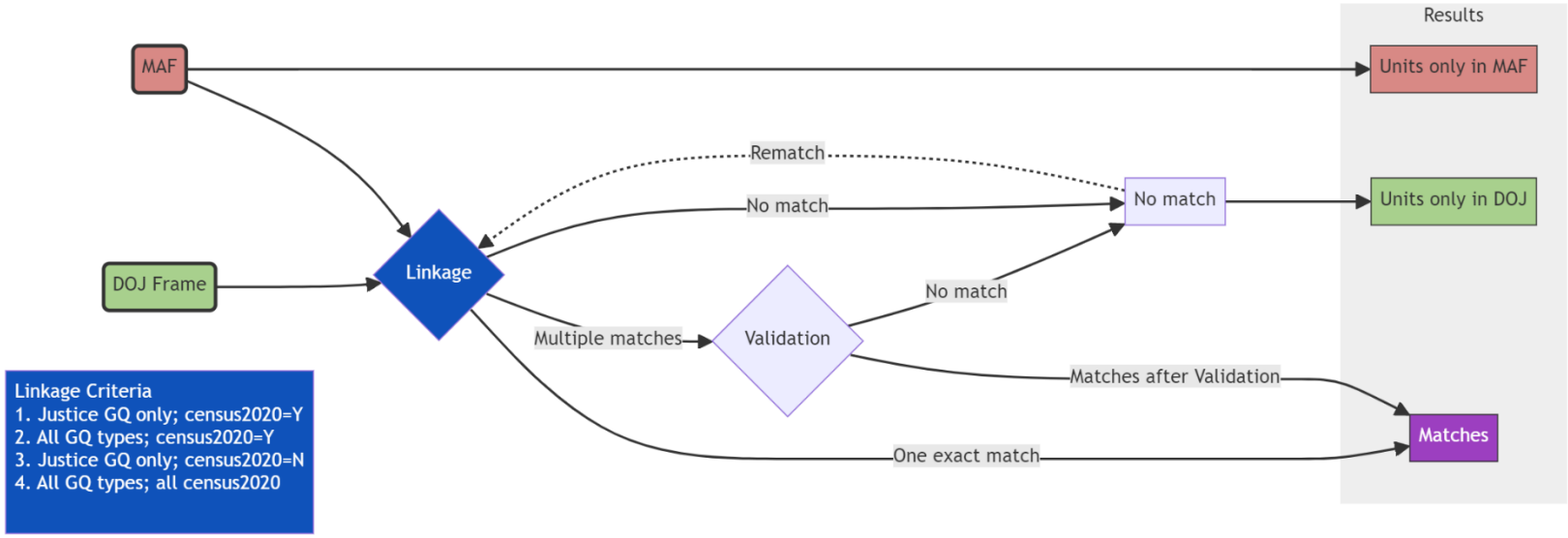
      a) The MAF address went to the county courthouse or sheriff's office, whereas the DOJ address went to the county jail. This would be coded as "A" (administrative) because it was unclear whether the MAF was referencing population at the sheriff's office or whether the record referenced the population held at the county jail.
      b) Sometimes facilities moved and records still had the old address. A web search of the name of the facility along with the word "moving" helped identify sources that verified the move. If it was determined that the MAF address was old, the record was marked as a match, along with a comment that this was an old address.
      c) The DOJ address was a mailing address rather than the actual physical location of the facility.

b. The DOJ record and the MAF record had the same address but different names.

   i. Internet searches verified if there were two different facilities at the same location or if a singular facility went by different names. Multiple names for a single facility were more common when a facility shared a name with an individual, for example Washington County Jail was the same facility as the John Doe Correctional Facility.
   ii. The DOJ record might match to multiple MAF records within the same facility. Looking at the population of the DOJ record and comparing it to the populations of the MAF records helped determined if the DOJ record covered all the internal GQ MAF records.
   iii. The DOJ record (for one of the non-juvenile frames) matched to a MAF record for a juvenile facility or a work release facility with the same address. If the DOJ facility name does not specify that it was a juvenile or work release facility, the function variables can determine if juveniles or people in work release programs were represented in the DOJ record.

Figure 1: Matching Process Flow Chart

Flow chart below maps out the iterative approach for linking the DOJ Frame to the MAF.



Appendix B: Coverage by DOJ Frame

Table 13 breaks out the facility-level and population-level coverage estimates by DOJ frame. In the 2020 Census, most facilities associated with the DOJ frames (CCF, COJ, JRFC/CJRP and SJIC) were enumerated or determined to be vacant at a high rate (95.7, 90.5, 87.0, and 90.0 percent of facilities respectively). Likewise, the total percent of the population associated with DOJ records that were enumerated was higher than the facility-level enumeration rate for the CCF, COJ, and JRFC/CJRP frames. In 2020, SJIC facilities had the highest rate of vacancy—around 30.0 percent of facilities.

Table 13: Facility-Level and Resident Population-Level Coverage Estimates Using the Integrated Justice Facility Frame, by DOJ Frame

	Census 2020 Enumeration Status of Matched Records	Percent of Matched CCF Facilities	Percent of CCF Population Residing in the Matched Records	Percent of Matched COJ Facilities	Percent of COJ Population Residing in the Matched Records	Percent of Matched JRFC/CJRP Facilities	Percent of JRFC/CJRP Population Residing in the Matched Records	Percent of Matched SJIC Facilities	Percent of SJIC Population Residing in the Matched Records
Enumerated or found vacant in 2020 Census	Enumerated as GQ	91.2%	98.5%	85.5%	95.3%	78.5%	83.7%	60.0%	62.0%
	Enumerated as Housing Unit	2.7%	0.4%	0.5%	0.1%	3.5%	1.7%	0.0%	0.0%
	Vacant GQ	1.8%	0.6%	4.5%	1.7%	5.0%	3.7%	30.0%	30.2%
	Subtotal	95.7%	99.5%	90.5%	97.1%	87.0%	89.1%	90.0%	92.1%
Not enumerated or found vacant in the 2020 Census	*Subtotal*	4.3%	0.5%	9.5%	2.9%	13.0%	10.9%	10.0%	7.9%
Total		100%	100%	100%	100%	100%	100%	100%	100%

Source: Authors' calculations from integrated frame based on links between the Master Address File Extract, Census of State and Federal Adult Correctional Facilities, Census of Jails, Annual Survey of Jails in Indian Country, Juvenile Residential Facility Census, and Census of Juveniles in Residential Placement (Project number 7510276, approval CBDRB-FY24-020).

Note: The DOJ frames are from 2019 and so there may be some expected differences in the universe of facilities and the populations residing in those facilities during the 2020 Census. This table includes all matched records from the MAF and the DOJ frames except for group of many DOJ records matched with many MAF matches. For single DOJ records that matched to more than one MAFID, the hierarchy shown in Table 10 was used to classify each DOJ record with

priority going to MAF records with GQ codes, followed by housing unit codes, vacant codes, and finally not-in-census codes. Populations used for coverage estimates came from the DOJ frames, not from the 2020 Census. Coverage estimates by DOJ frame are available in Appendix B. All estimates rounded according to Census Bureau DRB rounding rules. Because of rounding, columns may not sum to 100%.