Handling of Criminal History Export from Department of Public Safety

This document describes the methods used in association with the export of criminal history data from the CT Department of Public Safety, including ingest and normalization of data, algorithms used to associate records with CSSD clients, and automation of data retrieval for analysis.

1. Background

The Department of Public Safety (DPS) stores arrest data for all clients in two databases.

The first database contains CCH records, in which arrests are matched to known individuals via fingerprints. Persons in this database are given a unique identifier known as an SPBI number, and all subsequent arrests are tied to this number by a fingerprint match.

The second database contains Suspense records, which are either non-fingerprint identified arrests or recent fingerprinted arrests that have yet to be matched to a known individual with an SPBI number. This database acts both as a queue for fingerprint matching and a permanent location for storage of non-fingerprinted arrest data.

CSSD has a need to periodically retrieve records from the DPS databases for use in recidivism studies. A process was devised to export the entire contents of the CCH and Suspense databases for this purpose.

2. Data Export

DPS produces text file extracts of all data from the CCH and Suspense databases. Suspense records are sent as one file, named “sus.txt”. Due to the sheer number of records held in the CCH database, these records are broken up into numerous files, typically containing two years’ worth of data, to avoid loss of performance in the DPS system.

These files are transmitted from DPS to CSSD through a secure FTP site, where they are archived after being loaded (next step).

3. Data Ingest

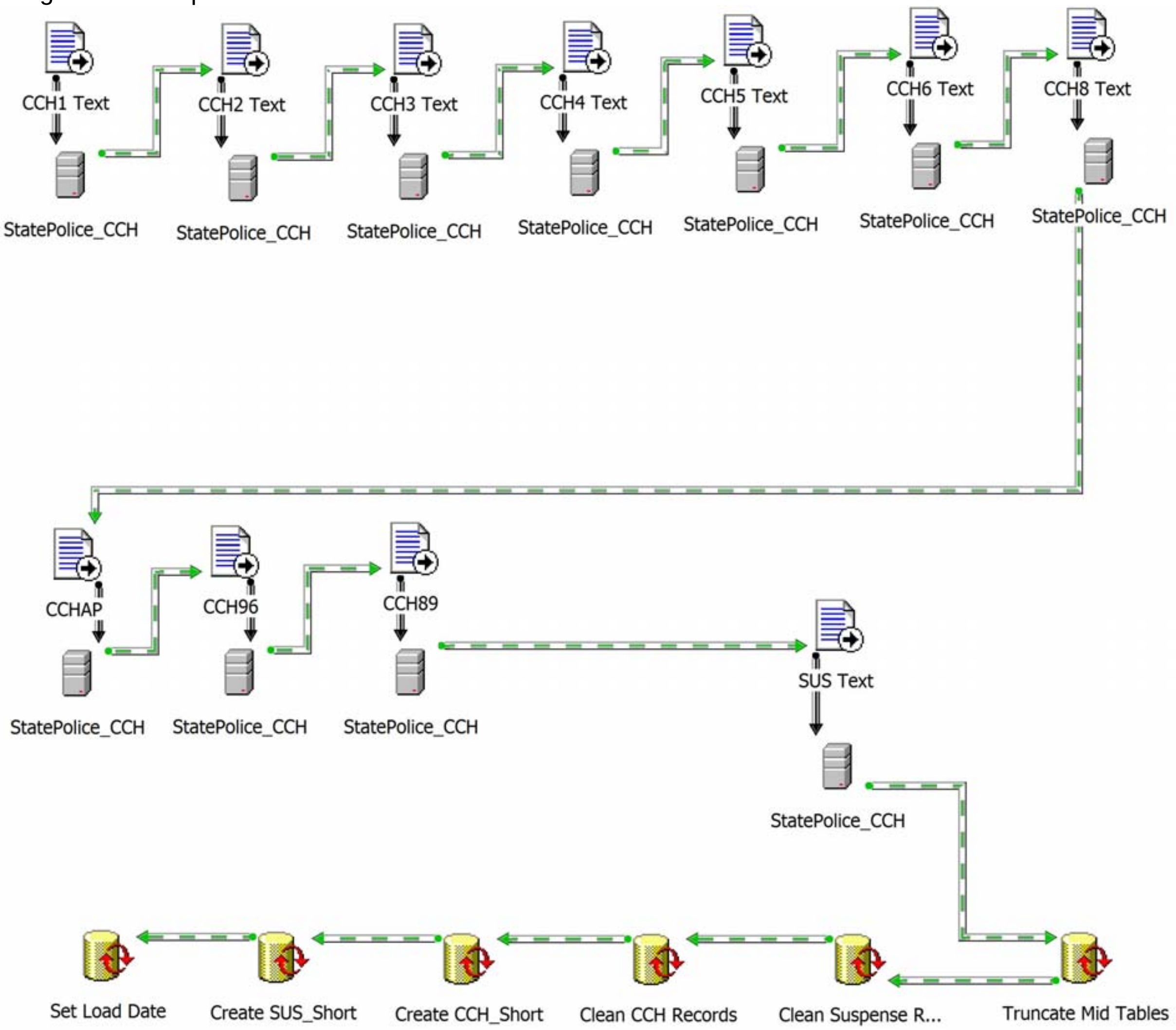
An automated load procedure has been developed to ingest the data files from DPS into searchable tables in a local database. A data transformation services (DTS) package steps through the following procedures:

- CCH files are loaded into a temp table
- Suspense records are loaded into a second temporary table.
- Suspense records are loaded into a permanent table. During this step, the records are cleaned and normalized to the following specification:
    - Extra space removed from client name
    - Extra space removed from SPBI number
    - SSN normalized to a nine-character string, missing values set to ‘000000000’
    - Statutes, charge descriptions, verdict data, and sentencing data are trimmed of extra space; missing values set to null
- CCH records are loaded into a permanent table. During this step, the records are cleaned and normalized to the following specification:
    - Extra space removed from client name
    - Extra space removed from SPBI number
    - SSN normalized to a nine-character string, missing values set to ‘000000000’
    - Statutes, charge descriptions, verdict data, and sentencing data are trimmed of extra space; missing values set to null
- Shortened versions of both the CCH and Suspense tables are created containing only identifying information and portions of the client name. These tables are used to increase performance during the matching process.

The records are then examined to determine the most recent sample date that can be guaranteed across both tables. Each table contains two dates (arrest date and verdict date). The most recent values in each of these four columns are gathered, and the earliest of the four is recorded as the sample date. In the initial run, this procedure plays out as follows:

CCH Arrest Date	11/19/2006
CCH Verdict Date	11/20/2006
Suspense Arrest Date	10/12/2006
Suspense Verdict Date	10/13/2006
Earliest Data Integrity:	**10/12/2006**

Diagram of load process:



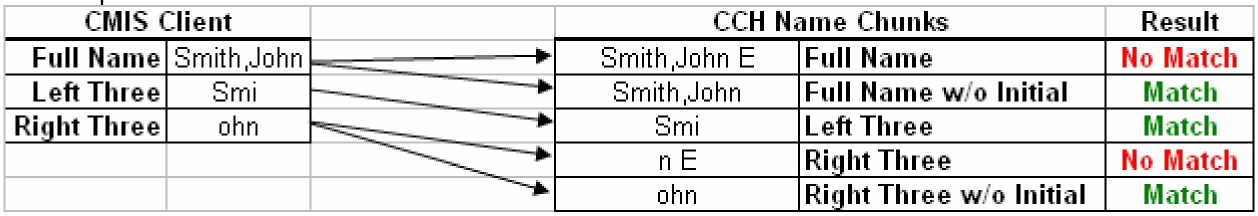
4. Associating DPS Records to CMIS Clients

Since records in the Suspense and CCH files do not always have a client identifier in common with CMIS client records, and other identifying information is inconsistent between the two systems, an algorithm was developed to make confident matches based on several fields.

CCH fields used to compare the data are below:

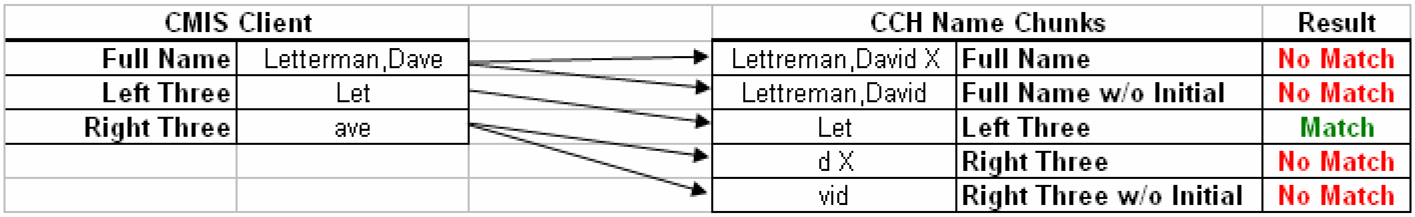
SPBI	Full Name	Left Three	Right Three	Right Three w/o Initial	Full Name w/o Initial	Gender	Date of Birth	SSN
1234567	Smith,John E	Smi	n E	ohn	Smith,John	M	7/4/1965	041-99-9999

Including the client's name as match criteria can become difficult if not handled carefully. To ensure a high match confidence, numerous parts of the name are compared in conjunction with other data elements. An example of the comparison:



CMIS Client			CCH Name Chunks		Result
Full Name	Smith,John		Smith,John E	**Full Name**	**No Match**
Left Three	Smi		Smith,John	**Full Name w/o Initial**	**Match**
Right Three	ohn		Smi	**Left Three**	**Match**
			n E	**Right Three**	**No Match**
			ohn	**Right Three w/o Initial**	**Match**

Using this method, matches can be found despite data entry errors, missing initials, and most name changes:



CMIS Client			CCH Name Chunks		Result
Full Name	Letterman,Dave		Lettreman,David X	**Full Name**	**No Match**
Left Three	Let		Lettreman,David	**Full Name w/o Initial**	**No Match**
Right Three	ave		Let	**Left Three**	**Match**
			d X	**Right Three**	**No Match**
			vid	**Right Three w/o Initial**	**No Match**

CCH records are matched to CMIS records in one of two ways:

- If the CMIS record has an SPBI number, CCH records are considered a match based on SPBI, Gender, and at least one other identifier. The third identifier can be date of birth, SSN, or either the left or right three-character comparison of the name.

- If the CMIS record in question does not have an SPBI number, we can still find a match. Depending on which identifiers are intact, a match is considered confident on either of the following conditions:

    - Complete match on Gender, Birth Date and SSN, combined with a match on either the left or right three-character comparison of the name.
    - Match on Gender, match on either SSN or Birth Date, and a match on the full-value comparison of the name.

  When a match is found on a CCH record using the second method, all records with that SPBI number are then considered a match, as the DPS system has already made the association.

Suspense records rely on the same algorithm as the CCH records without CMIS SPBI, however every record considered a match must meet the criteria since the Suspense file does not contain a unifying record.

A docket-level matching module is currently being developed that will add another association tier to the process by searching for CCH and Suspense records based on a CMIS docket list for the client.

5. Automation of Record Association

Once a method of associating CCH records to CMIS clients has been established, a goal was set to create a process by which criminal histories could be easily gathered for a list of clients. Another DTS package was assembled to make the entire process accessible to CSSD staff.

- The package first gathers identifying information from CMIS based on a list of client identifiers (known as resource IDs) that have been placed in a table by research staff.

- A list of non-CMIS-SPBI CCH matches is assembled using shortened versions of the CCH table.

- The first round of matches is produced for CMIS records that have a valid SPBI value.

- The second round of matches pulls all records for SPBI numbers matched to non-CMIS-SPBI records using shortened tables.

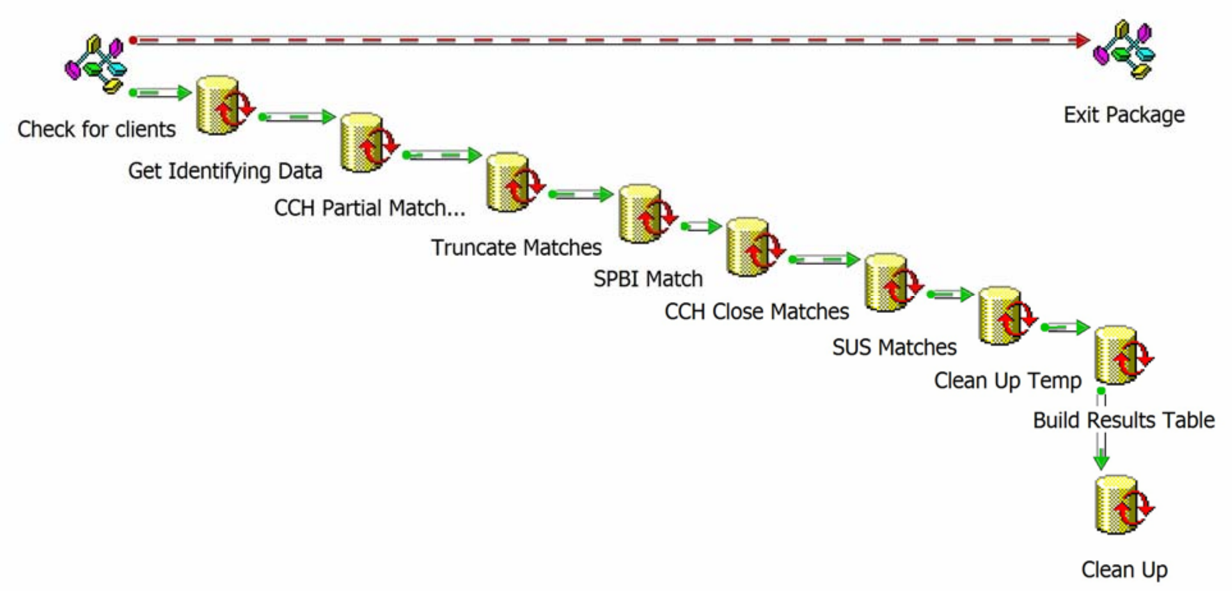
- Third round matches CMIS clients to Suspense records.

- The package creates a table containing all arrest data found for the clients specified in the original table.

The data is normalized for easy analysis during this procedure. Date fields are converted from eight-character strings to a date/time format, verbose descriptions of verdict and consecutive/concurrency codes are added and all sentencing data is converted to days for easy comparison.

This process can be utilized by research staff in one of two ways. For small client lists (under 1,000 names) the process can be triggered manually via a stored procedure on the SQL server, taking approximately one hour to complete. The package also checks every night for the existence of a client list, so large jobs can be handled during the evening hours.

Diagram of automated matching process:


Check for clients
Exit Package
Get Identifying Data
CCH Partial Match...
Truncate Matches
SPBI Match
CCH Close Matches
SUS Matches
Clean Up Temp
Build Results Table
Clean Up