CRIMINAL JUSTICE RESEARCH INSTITUTE

HAWAI'I

# A REPORT ON THE CREATION OF A CENTRALIZED PRETRIAL DATA REPORTING AND COLLECTION SYSTEM, PURSUANT TO ACT 147, SLH 2023 - YEAR 2


**PREPARED BY**

Erin Harbinson, PhD and Aerielle Reynolds, MSCJA

# BACKGROUND

This report is respectfully prepared pursuant to Act 147, Session Laws of Hawai`i 2023, which appropriated funds to the Criminal Justice Research Institute (CJRI) to create a pretrial database and reporting system, and required a progress report for the Legislature over the following two years. The appropriation from Act 147 provides CJRI with the funds to fulfill a mandate outlined in Act 179 (Session Laws of Hawai`i 2019), which requires CJRI to create a "centralized statewide criminal pretrial data reporting and collection system" (HRS § 614-3). This report is the second report of two that must be submitted, which summarizes progress made approximately a year and a half since Act 147 was signed into law.

As documented by the Criminal Pretrial Task Force in 2019 [1], a major barrier to understanding how the pretrial system is operating comes from the siloed and disconnected data across agencies involved in the pretrial system. The Criminal Pretrial Task Force recommended that a Criminal Justice Research Institute be created in order to bring data across these agencies into a centralized system. This system would create capacity to report out on criminal pretrial metrics. To accomplish this mandate, the law recognized that CJRI must take several steps to develop a plan and solution to create the database. These steps acknowledged in the law include: 1) identifying databases with pretrial information, 2) determining the administrative and technological feasibility of aggregating and sharing current data, and 3) identifying gaps in pretrial data (HRS § 614-3). CJRI staff completed these steps and identified several challenges that must be addressed in the project:

- Currently, there is some data, but much of it is still stored as information and therefore information must be transformed into data for statistical analysis.

- Two branches of government house pretrial data, which means different laws and rules govern their data use, as well as differences in administrative practices and technology systems that could impact data sharing and data governance.
- Criminal justice records across criminal justice agencies are stored as different units of analysis, which makes centralizing records more complex, especially when merging, linking, and restructuring files from the Department of Corrections and Rehabilitation (DCR) (which looks at individuals who enter their facilities) and the Hawai`i Criminal Justice Data Center, Department of the Attorney General (HCJDC) (which houses arrest records for individuals) to merge with the Judiciary (which houses information on court cases that are organized by case numbers and not individuals).
- Some data can be pulled from agency databases as structured data fields. However, many are text based and even more challenging are the unstructured text fields that include long comments or court minutes.
- Extensive manual labor from several staff across pretrial agencies and CJRI are required to create a centralized source of pretrial data for research since staff must extract multiple tables from each IT database, share files across agencies, and then merge, link, and transform several fields of information into data that can be used for required pretrial statistics.

As the CJRI director interviewed several states and jurisdictions embarking on similar database projects, it became clear there was no one solution to create the pretrial database and reporting system. There are as many solutions in existence as jurisdictions embarking on these projects. While many barriers are similar across states, solutions depend on the rules, laws, technology, resources, policies, and organizational culture of these organizations. After surveying different solutions, CJRI categorized three different approaches to this work and made a recommendation to the CJRI board. The board

reviewed the costs and benefits of the different approaches, and agreed with the recommendation from the CJRI director to pursue resources for a technological solution that would extract data from all three agencies and centralize it into a data warehouse. A feasibility study was conducted in the Fall of 2022 to verify the technical requirements and estimate a budget and timeline for creating a pretrial database and reporting system. This informed the appropriation request in Act 147.

In summary, CJRI reviewed data sources across the state to develop a technical plan to create the pretrial database and reporting system. This was done in collaboration with criminal justice stakeholders which house the three main sources of pretrial data - DCR, HCJDC, and the Judiciary. The aim was to identify a solution that would address the barriers to reporting out on the pretrial system in a timely and efficient way. Act 147 provides funds and resources for CJRI to carryout this solution.

Due to the complexity of IT systems, databases, software, and criminal justice decisions involved in creating a statewide pretrial database and reporting system, the next page provides a conceptual overview of the project.

# WHAT IS HAWAI`I'S PRETRIAL DATABASE AND REPORTING SYSTEM?

Explaining the creation of the "centralized statewide criminal pretrial justice data reporting and collection system" (HRS § 614-3)

### Agencies use case and records management information systems in the pretrial system

Criminal justice agencies use several information systems to collect and store information on individuals entering the criminal justice system. These systems store information on court cases, arrest records, and more.

### Three agencies store most statewide information on the pretrial system

The CJIS system (Department of the Attorney General) contains arrest records, JIMS (Judiciary) includes information on court cases and court decisions, and OffenderTrack (Department of Corrections and Rehabilitation) includes information on people entering and exiting jails.

### Records on cases and people must be linked across the three systems

Each agency uses a unique identifier on cases or people to store information in their systems, and these identifiers must be used to link court records, arrest records, and jail records together in order to analyze system trends or evaluate impacts.

### IT tools will create data pipelines for each agency to submit data into a centralized data warehouse

The pipelines will link records and store them in one centralized data warehouse, making it possible to create one database of pretrial information without requiring additional data entry.

### Centralizing data will provide more effective and efficient data capacity for reporting

By linking records in one location, a platform will be created to build dashboards for regular reporting on metrics and will establish a central source of datasets to extract for evaluation and analysis.

# TABLE OF CONTENTS

# PROVIDING PROGRESS UPDATES ON ACT 147

This report is organized with both a high level list of activities and a summary to provide updates to the Legislature on the development of the pretrial database and reporting system. Several activities related to the project are organized into thematic sections. Much of the work is sequential, therefore as tasks are reviewed further down the list, fewer have been initiated. For example, reporting cannot occur until the centralized data warehouse is created and the data pipelines are built. The list of activities are based on project deliverables from contracts, but written at an overview level to capture the main components of the system. If there is no progress update such as “in progress” or “ongoing,” then the task has not been initiated. A written summary is provided to capture context or to explain future activities for the project.

This project cannot be accomplished without the full cooperation from each of the agencies CJRI is collecting data from for the pretrial database and reporting system. This project was planned in consultation with operations, IT, research, and other staff across the three agencies to ensure it was a feasible approach. CJRI is grateful for their continued support of this project, including their testimony in support of House Bill 68 during the 2023 legislative session, which became Act 147. CJRI staff would like to acknowledge the continued involvement of DCR, HCJDC, and the Judiciary to make this project a priority. This includes meeting with IT vendors, sharing datasets, answering questions about pretrial operations, and other activities requested of them. CJRI is grateful for this continued support.

# PROGRESS UPDATE

| Administrative and Initial Development Activities | In Progress | Ongoing | Completed |
|---|---|---|---|
| Executed contract | | | ● |
| Met with IT partners to establish project plans and timelines | | | ● |
| Met with three primary agencies to introduce data transfer protocols and expectations | | | ● |
| Developed draft data governance agreement | | | ● |
| Meet with criminal justice peers in other states with similar centralized data and reporting systems | | ● | |

**Summary of Progress**

Several administrative steps occurred in 2023 to create the pretrial database and reporting system. A contract was executed in September 2023 and the project is now in the second fiscal year of the contract. At this time, the work is expected to be finished within the contract deadline and budget. The data warehouse has taken longer to complete due to the complexity of the Judiciary's case management system. The data pipelines were expected to be completed by this fall, but now they are expected early next calendar year. As a result, some aspects of the data pipelines and datasets have been scaled back to stay on budget and on time. This will be covered in more detail throughout the report.

The project approach has shifted some, but the most important component - the data warehouse approach - remains the same. CJRI is using an Extract, Transfer, Load (ETL) tool to map data pipelines into a centralized data source, a data warehouse. The ETL streamlines the several steps needed to prepare data for analysis, such as restructuring files to link by case level or transforming key concepts from many fields into one column of data. Instead of staff spending months requesting data and cleaning data for analysis, the tool automates some of this work. The data warehouse will store two different unified datasets based on research

(case and person level). With unified datasets, CJRI can download centralized pretrial datasets for research and respond to questions more efficiently. Additional datasets were curated for a subset of metrics that could be restructured feasibly within the contract timeframe.

| Identification of Data and Data Transfer Process | In Progress | Ongoing | Completed |
|---|---|---|---|
| Obtain sample dataset extracts from all three agencies | | | ● |
| Create data map demonstrating linking of data records | | | ● |
| Develop calculations for key performance indicators | | | ● |
| Create data codebook | | ● | |

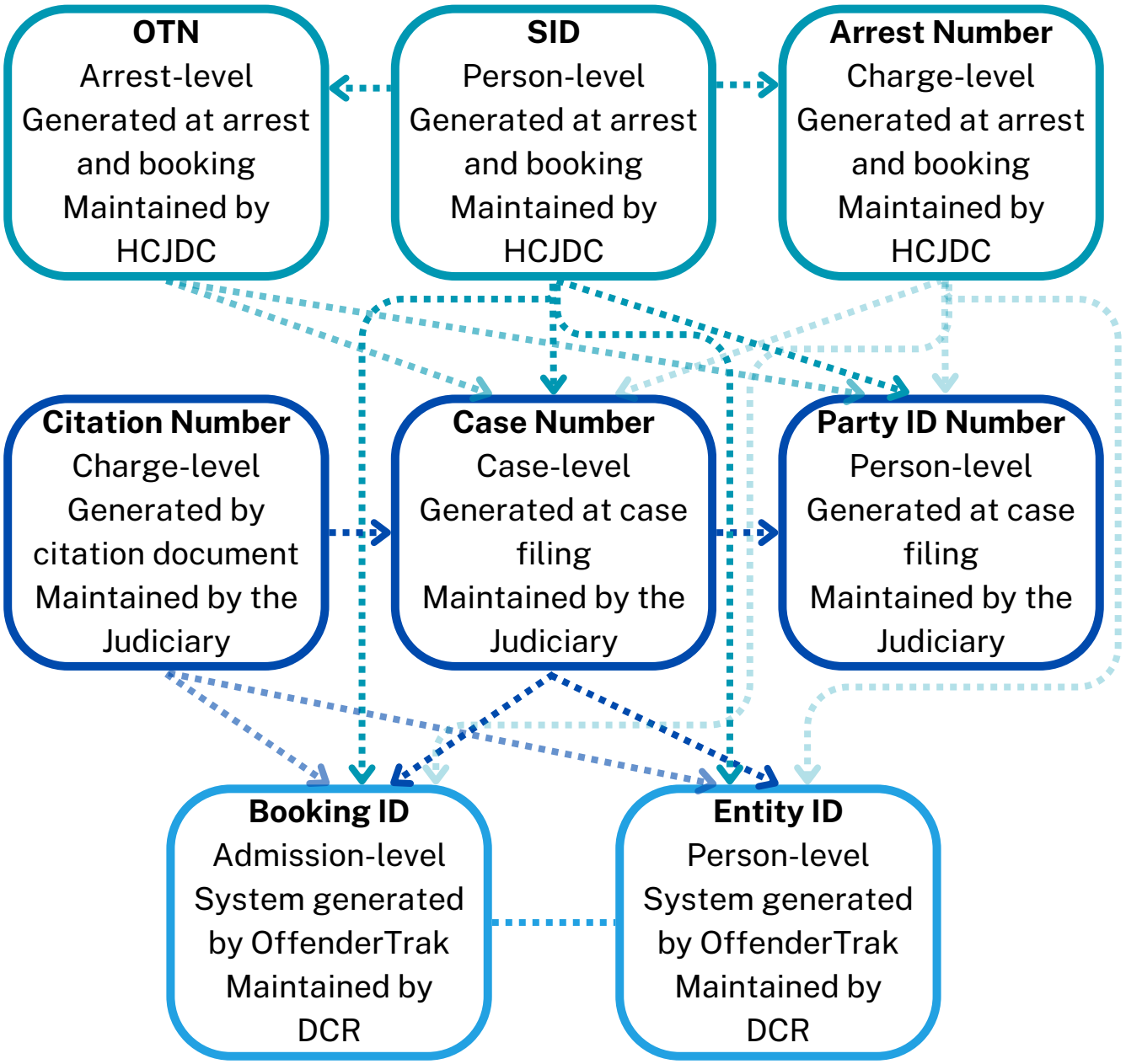
**Summary of Progress**

Over the course of the past year, CJRI and its technology and software partners established a process to ingest data into the centralized data warehouse, which required a significant amount of background work. While each agency had shared sample data at the onset of this work, in some cases it took several iterations of data extractions to assist CJRI staff in identifying necessary data elements for producing pretrial metrics and develop data transformation rules to normalize and clean data in preparation for analysis. The sample data allowed CJRI to determine key identifiers, which are needed to link records effectively across agency datasets. This mapping of records and identifier fields was necessary to ensure that data pipelines were created accurately to reflect these maps, and allow CJRI research staff to follow an individual's trajectory throughout the criminal pretrial system. Additionally, sample data was used to assist CJRI staff in developing the logic to calculate key metrics. CJRI also worked with agency research staff to create data codebooks, which will benefit all researchers in the state.

CJRI is grateful for the continued collaboration and support of staff from DCR, HCJDC, and the Judiciary on this project. Their responsiveness to data requests, time taken to answer data related questions, and insight into operations being captured within pretrial data has been invaluable.

**Figure 1. Mapping Identifier Connections to Merge Records Across Agencies**

CJRI staff reviewed various identifiers maintained by criminal justice agencies for their utility in merging records across agencies. With siloed data, unique identifiers for people, cases, and charges are essential in linking records at the case level. Some identifiers are only used within one agency, and thus they are only useful for restructuring data specific to that agency. Other identifiers are used across agencies, and can be leveraged to merge records. The graphic below is a simplified mapping of criminal justice records by unique identifiers to create a centralized source of data.



| Creation of the Centralized Data Warehouse | In Progress | Ongoing | Completed |
|---|---|---|---|
| Hold meetings with each agency's IT department to understand technological capabilities | | | ● |
| Install ETL software | | | ● |
| Establish data sharing connections for each agency to share data | | | ● |
| Develop data pipeline strategy | | | ● |
| Receive training on ETL | | ● | |
| Receive training on data warehouse | | ● | |
| Finalize ETL architecture to promote most efficient method of data sharing long-term | | ● | |
| Finalize security policies and procedures | | | ● |

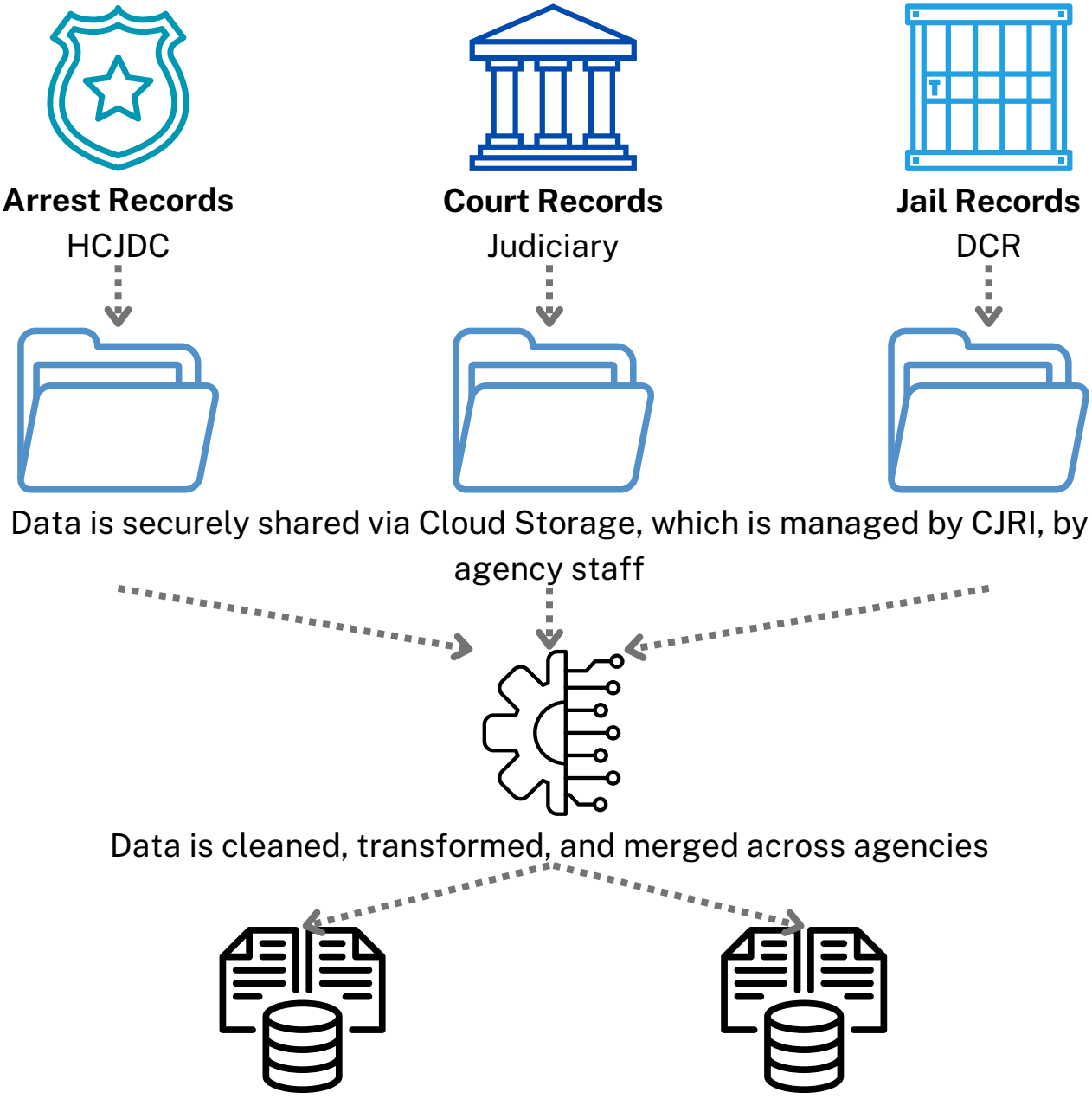
**Summary of Progress**

The ETL provides several ways for each agency to share data with the data warehouse. Currently, data is shared through cloud storage instead of encrypted emails and flash drives. Data sharing processes were developed based on feedback regarding the technology and needs of each agency. While the cloud data sharing connections have been used thus far, the ETL can also be set up with a direct connection that would create an automatic process for data sharing with CJRI in the future. This would reduce the workload of staff and create more sustainability in data sharing. This will be explored after the unified datasets and data warehouse are finalized.

In addition, a strategy was developed for the data pipelines for the two unified datasets. These pipelines enable the transformation and normalization of raw data to prepare it for use in research and analysis. Staff have received training on the ETL tool and the data warehouse, yet some trainings will occur after the unified datasets are finalized. Some aspects of data governance may need finalization, but several security practices are in place with the software. The database has been developed to make data sharing efficient for source agencies, while respecting agency rules and research principles governing sensitive data.

**Figure 2. Pretrial Database and Reporting System Workflow**

CJRI's pretrial database and reporting system uses an Extract, Transform, Load (ETL) tool to ingest data into a centralized location. Criminal pretrial data from HCJDC, the Judiciary, and DCR, are shared to a cloud-based data warehouse managed by CJRI. This data is automatically uploaded to CJRI's pretrial database. Then, the data pipelines in the ETL automatically clean, transform, and merge files to produce curated datasets that are used for analysis by research staff.



| Development of Reporting Process | In Progress | Ongoing | Completed |
|---|---|---|---|
| Author data visualization guide for dashboard and reporting methodology and design | | ● | |
| Receive training on software for reporting and dashboards | | ● | |
| Develop draft dashboards with software for subset of metrics | | ● | |
| Provide demonstration of priority metrics for feedback | | | |
| Conduct stakeholder engagement to receive feedback on reporting | | | |

**Summary of Progress**

CJRI initiated preliminary work related to the development of the reporting process this year. Before reporting can begin, data must be ingested into the centralized data warehouse. Some reporting relies on more efficient data wrangling, which comes from the unified datasets. The unified datasets will include data from all three sources and therefore allow research staff to calculate a range of metrics more effectively. Other reporting will come from dashboards, which will report out on certain pretrial metrics regularly. Ideally, there would be several dashboards for all of the pretrial metrics, yet this was not feasible at this time. Datasets have to be curated for dashboards specifically, and it was not possible to curate the number of datasets necessary for all the pretrial metrics within the project's timeframe. Instead, dashboards will target a few metrics that were possible to create with the data and other reporting will occur through reports. Under HRS § 614-3, it was expected that research staff would use a centralized source of data to report out annually on pretrial metrics and this will occur for those metrics that cannot be produced in dashboards. Once the data warehouse is finished, CJRI will develop a plan to build more curated datasets to expand the dashboards since these are a more digestible and timely reporting mechanism.

In tandem with work related to the development of the data warehouse, CJRI staff have undertaken several activities related to the reporting

process. A data visualization guide was drafted to standardize reporting and outline a research methodology for pretrial metrics. Over the past year, a dataset was tested with the dashboard software to ensure it would integrate with the data warehouse. Additionally, staff received training on customizing the dashboard. This data tested successfully, though dashboards are not available until the data warehouse is finalized since they are built on the datasets.

Prior to public reporting, CJRI plans to share draft dashboards and reporting templates with key stakeholders to ensure that information is presented in a clear and digestible manner. This review process will encompass a wide range of stakeholders, including legislators, criminal justice agency staff, and members of the public. CJRI will provide updates on their website about opportunities to provide feedback on the new system as they become available.

| Finalizing the Pretrial Database and Reporting System | In Progress | Ongoing | Completed |
|---|---|---|---|
| Execute final data governance agreement | | | |
| Finalize data sharing MOUs | | | |
| Revise dashboards based on feedback from stakeholder engagement | | | |
| Make dashboards public on CJRI website | | | |
| Create new section of annual report summarizing pretrial metrics and reporting | | | |

**Summary of Progress**

The concluding list of tasks is related to finalizing the policies and technical features for the pretrial database and reporting system. The data pipelines, the data warehouse, and the dashboards are dependent on the unique nature of all three data sources. As the system is finalized, CJRI and its technology and software partners will validate the unified datasets to ensure that data was ingested into the system and merged properly. Additionally, policies will be adjusted to reflect the final database architecture ranging from data structure, sharing, security, and other features. Near the end of the project, CJRI will seek feedback from

different stakeholders to ensure that reporting (i.e., dashboards and reports) is informative and accessible for a variety of users.

## Addressing Challenges

There are several challenges to modernizing data in any setting, but there are some barriers unique to criminal justice and public sector settings. Because CJRI anticipated this, a feasibility study was conducted in 2022 to inform the appropriation request made in H.B. 68 (2023). As the project is now in its second year, an array of challenges are summarized in order to provide context on the current state of the project and share lessons learned for other government agencies taking on data modernization projects. Some of these challenges have generated a delay in creating unified datasets in the data warehouse, which is the foundation of the pretrial reporting.

*Coordinating with several subject matter experts*: One overarching challenge in a data modernization project is the reliance on several subject matter experts (SMEs) to carry out this work. Staff at CJRI have experience with criminal justice issues and research, but must rely on people in different areas of agency operations and technology for input. Implementing an ETL and creating a data warehouse is commonplace for IT departments, but not for research agencies. It can take several conversations to "translate" a request or concept to different people. This means that tasks can take longer as more people are brought in, and many key concepts have to get translated to different terminology depending on the SME. For example, a social science researcher and a software designer might not mean the same thing when they say "data restructuring." The involvement of several SMEs has resulted in a longer timeline for project development. Yet, a few strategies have been developed to address this challenge. Regular meetings with a broad range of subject matter experts has helped ensure that people with varying perspectives hear the same information and can help catch things others may not. This is particularly important so that SMEs voice concerns before projects progress too far off course. In addition, it is helpful to continue to define key concepts and follow-up with IT partners to ensure everyone is clear on the requests and deliverables, even when it feels like a question or concept is elementary.

*Acquiring and understanding pilot data:* As previously mentioned, pilot data from all three agencies was an essential component of database development. This pilot data allowed CJRI research staff to determine how agencies were storing data elements within their systems – some data elements are stored as coded variables, while others are in date fields or string variables. Following the execution of contracts in Fall 2023, it took several months and several iterations of datasets being curated by agency staff to ensure that CJRI had all necessary data elements to be able to calculate pretrial metrics. In addition to acquiring pilot data, it took considerable time for CJRI research staff to understand all the pilot data that was shared by each agency. This involved determining which data elements would serve as the key identifiers, which are needed to link and merge records across agencies, as well as determining which data elements suffer from quality issues. Much of the criminal pretrial data across the state is inputted into agency data systems by a number of operational staff, and as such, the consistency and accuracy of data entry can be impacted. Data entry practices can also change over time to be responsive to changes in policy or operations, and it is important that CJRI research staff understood how such changes affected data quality. Staff from all three agencies continue to provide feedback to CJRI to ensure that data elements are being utilized in a way that is reflective of pretrial operations across the state.

*Judiciary data:* The Judiciary's data proved to be very complex to understand and work with. While data from DCR and HCJDC were shared as one file per agency, the Judiciary's data must be extracted as several dozen tables that must be linked before they can be merged with data from the other agencies. JIMS data is stored in several tables and fields depending on case type (e.g., circuit, district, traffic), which requires mapping all possible fields required to constitute a single data element, such as defendant release status. In other instances, several fields were required to construct a new data element. This was necessary for even basic concepts like the *date pretrial ends*, which is a critical data point to calculate many performance metrics. JIMS staff provided substantial support to this project to ensure that all necessary data elements were identified and properly mapped to one another across case types, and this information was built into the data warehouse and unified datasets. While

it was expected that JIMS data would be complex, it was not clear how the technical complexity would impact the development of data pipelines and the curated datasets for dashboards until later on in the project. To address this, CJRI had to adapt in a few ways. Some of the transformations for the unified datasets were scaled back, which means the unified datasets automate some "data cleaning" but, research staff will still have to do more manual cleaning than expected. Also, there were fewer datasets built for the dashboards. Scaling back the data pipeline work was necessary because it would require more time to map and in some cases, it would require the addition of a data architect from the software partner.

*Gaps in critical identifiers*: Key identifiers, including case numbers, arrest number, state identification number (SID), and offense tracking number (OTN), are some of the most essential data elements to the development of the pretrial database and reporting system. These data elements help to identify unique court cases, charges, individuals, and arrests, and allow for the linking and merging of data across agencies. While several of these identifiers are present in the data of more than one agency, CJRI and its technology and software partners encountered gaps and data quality issues within these identifiers which needed to be overcome. Many of these issues are related to manual data entry, with identifier fields not being inputted (e.g. SID not being entered when a case is filed) or because identifiers are truncated (e.g. an identifier missing a hyphen). Fortunately, this challenge was easier to address than others. A set of rules was developed to normalize identifiers across agencies, as well as logic to use a combination of identifiers to ensure that records were properly linked across agencies. These were all documented and have been incorporated into the data pipelines, which means they will be applied automatically when data is ingested into the data warehouse.

*Preponderance of critical information in text-based documents*: While CJRI will be able to analyze many of the metrics identified in HRS § 614-3 with readily available data across the datasets provided by DCR, HCJDC, and the Judiciary, several pieces of information critical to analyzing specific pretrial outcomes are most consistently found not in agency datasets, but rather in text-based documents (i.e., pdf files). For instance, a defendant's failure to appear, while tracked in several other data elements within JIMS,

is most consistently found within bench warrants, while the pretrial release recommendation made by Intake Service Centers staff is found within pretrial bail reports. An Intelligent Document Processing tool (explained in greater detail in Appendix A of this report) could be used by CJRI to extract these data elements from text-based documents into a dataset. Then, this information would need to be linked to CJRI's unified datasets for use in analysis. However, relying on this could end up as a costly approach to transforming this data since it charges per page. Alternatively, agencies may need to consider entering this data differently or CJRI could partner with someone using other data science techniques for text (i.e., natural language processing).

*Balancing varying approaches to database development:* The way an ETL is leveraged to prepare data for analytics can vary by the way data pipelines are built in part because they are based on business needs, but also on the software. The initial approach undertaken in the first nine months was shifted to a second approach to ensure the project met the intent of the law and stayed within the contract timeframe. First, CJRI research staff developed several layers of logic to calculate pretrial performance metrics identified in HRS § 614-3, which were going to be the source for several curated datasets specific to a single dataset per metric. While this approach was beneficial for metrics in reports and dashboards, these datasets could not be used for more complex analysis. For example, a dataset would calculate the average length of detainment, but it could not be used to evaluate what factors were predicting length of detainment. Because the court data is stored in a multi-dimensional, transactional database and pretrial decisions are multi-relational, it is less effective to create datasets that correlate data through metrics (i.e., the first approach). Instead a second approach was undertaken to develop unified datasets at specific units of analysis (e.g., case-level or person-level). This approach is more appropriate for a database and reporting system for research purposes. It cleans up foundational data merging and linking that needs to occur before calculating most metrics and it establishes datasets that include an array of pretrial data. This resulted in less dashboard datasets, but these can be created in the future with the IT specialist. More importantly, the project prioritizes the creation of research datasets that streamline data merging and linking of millions of pretrial records.

## Additional Considerations

*Staff support:* As part of Act 147 (2023), funds were allocated to create the new system, while also providing staff support to make the project successful. The law created two temporary IT positions - one for DCR and one for the Judiciary. Both have been established and were filled in 2024. These positions have been critical to providing their agencies with the capacity to front-load a significant amount for this project. The new law also provided CJRI with a new permanent IT project specialist position, which will support the database long-term. This position was created with feedback from technical experts, in order to limit the need for external assistance in the future, and recruitment for this position is ongoing.

*Database security:* A substantial amount of pretrial-related data can be found in public records, such as arrest reports or court records. However, many of these systems also include confidential information associated with these records. As often as possible, CJRI is extracting data that is public, while omitting and/or not utilizing confidential information, as well as personally identifiable information (PII), such as social security numbers and home addresses. The ETL and cloud storage have policies and procedures on data security, and will be set-up to limit the storage of PII. For example, if it is needed for matching records, the ETL can be set-up to match records with PII but remove or mask the PII before it is stored in the cloud-based database. Additionally, the data warehouse that CJRI selected is used in many states and has security protocols appropriate for confidential data, including the CJIS data housed by HCJDC.

Even if records are public, social science researchers strive to keep people anonymous and to protect them from additional scrutiny, in accordance with research principles of ethical data stewardship. CJRI is working with their IT partners to make it difficult for a specific individual to be identified even if the information is public. The system is designed for reporting at the aggregate and is not a system to query individuals to piece together their entire criminal justice record. CJRI data sharing MOUs address these concerns, and are modeled after research data sharing agreements that cover sensitive information.

*Long-term planning*: As CJRI staff have assisted with other research requests, they encounter the same barriers to criminal justice research

regardless of the topic. Criminal justice data is disconnected, and many operational decisions are made in silos. Because of this, it is difficult for agencies to turn around information on other policy questions in a timely manner. If the Legislature wanted to learn more about probation or parole, similar barriers exist as those in the current pretrial data landscape. The new pretrial database can serve as a model to help the state strategize about solutions to collect criminal justice data beyond the pretrial system.

Finally, it is not enough to rely on metrics from a pretrial database and reporting system to change policy. The Legislature, agencies, and the public must embrace data driven decision-making. CJRI research staff have partnered with agencies and learned about their operations to ensure data is accurate and represents the system fairly. This includes working with agencies to ensure CJRI research staff know the strengths and limitations of the data, and developing a reporting process that is objective with documented methods that are transparent and clear. CJRI staff is coordinating with other statewide efforts to modernize data and build data capacity for the state. For example, this includes meeting with the Chief Data Officer, Office of Enterprise Technology Services, attending conferences on local data initiatives in the public sector, or collaborating with local organizations such as the Hawai`i Data Collaborative to bring more expertise to this work. The pretrial database and reporting system is a unique criminal justice project that benefits from the support and feedback of several members of the community engaged in innovative data strategies.

# PLANNING FOR FY 2025 & FY 2026

The CJRI appropriation in Act 147 included one-time development funds and support for annual maintenance to establish a "centralized statewide criminal pretrial justice data reporting and collection system" (HRS § 614-3). By centralizing and linking data across agencies in the criminal pretrial system, CJRI will have capacity to report out on pretrial metrics to assess how the pretrial system is performing and analyze data to make recommendations for policy. This report has summarized progress in achieving this work through the appropriation request. Summarized below are the features of the system that require ongoing maintenance, which CJRI's is requesting in their operating budget for the next two fiscal years.

## Data Warehouse

The data warehouse serves as the **centralized source of pretrial data**. It is a cloud-based platform that charges monthly based on the amount of data stored. The CJRI operating budget includes an estimate of ingesting and storing data from three criminal justice agencies on a monthly basis. Data records have been collected as far back as January 2011 for all three agencies and will continue to update with new records each month. All criminal JIMS data must be re-ingested each month due to the complexity of the JIMS system.

---

## ETL Subscription

The ETL is a subscription based service. However, with Act 147 it included one time development work to create the data pipelines and provide training. Moving forward, an annual subscription covers the ETL software including the **preservation of the data pipelines that are critical to the ingesting, merging, and transforming of siloed data** into the data warehouse.

---

## Dashboards

The dashboards will host a handful of **pretrial metrics for regular reporting**. Dashboards must be built off of a dataset curated specifically for metric calculations. In the begining, CJRI piloted dashboards with a user-friendly software tool. It was found to be less efficient than a service that is part of the data warehouse, which CJRI learned about later on. This will be implemented instead, which is significantly cheaper and charges monthly based on data use (i.e., amount of data processed through dashboard).

# OUR ORGANIZATION

## CJRI STAFF

**Erin E. Harbinson, PhD**
Director

**Aerielle Reynolds, MSCJA**
Research Analyst

**Samuel Choi, PhD**
Research Analyst

**Mariah A. McCaskill**
Secretary

**Pamela Oda**
Undergraduate Research Intern

**I ka nānā no a ʻike.**

By observing, one learns.
-ʻŌlelo noʻeau

Through observing, or researching, Hawaiʻi's criminal justice system, CJRI is dedicated to helping stakeholders, lawmakers, and the public learn more about Hawaiʻi's criminal justice system.

## BOARD MEMBERS

**Judge Matthew J. Viola**
*CJRI Board Chair*
Judge, First Circuit Criminal Division
*Judiciary*

**Nicole C. Fernandez**
Offender Services Section
Administrator, WCCC
Department of Corrections and Rehabilitation
*Governor's Office*

**Rep. Gregg Takayama**
District 34
*Hawai`i House of Representatives*
(Term starting July 31, 2024)

**Peter Wolff**
Federal Public Defender (Retired)
*Hawai`i Senate*

**Francis Young**
Acting Corrections Program Services Division Administrator
*Department of Corrections and Rehabilitation*

**Rep. Scot Z. Matayoshi**
District 49
*Hawai`i House of Representatives*
(Term ending July 30, 2024)

The staff at CJRI could not accomplish their work successfully without the expertise of the board. Each of the board members brings valuable knowledge from their respective roles and experience across the criminal justice system and the policymaking realm. The criminal justice system is wide-ranging, and the board is essential in helping prioritize projects and providing feedback on ways to communicate research. Their collective experience has improved the work of CJRI in several ways. The CJRI staff thank the board members for their ongoing work and support.

## ENDNOTES

1. *Hawai'i Criminal Pretrial Reform: Recommendations of the Criminal Pretrial Task Force to the Thirtieth Legislature of the State of Hawai'i* (2018): https://www.courts.state.hi.us/wp-content/uploads/2018/12/POST_12-14-18_HCR134TF_REPORT.pdf

## ACKNOWLEDGEMENTS

Special thanks to: Representative Scot Matayoshi, who sponsored House Bill 68, which later became Act 147 (2023) and provided the appropriation to establish the pretrial database and reporting system. We are grateful for the input from many individuals across the pretrial system, and would like to recognize staff that continue to dedicate their time and expertise in creating the pretrial database and reporting system. A special thanks for:

*Department of Corrections and Rehabilitation*: George King, Judy Yamada, Frank Young

*Hawai`i Criminal Justice Data Center*: Wendy Char, Philip Higdon, Susan Yonemura

*Judiciary*: Irene Mae Abut, Adam Cohen, Velma Kam, Ariel Maeda, Dana Nakasato, Mai NguyenVan, Sudarat Pindavanija, Sanghoon Yun

# CONTACT INFORMATION

Criminal Justice Research Institute
The Judiciary - State of Hawai`i
417 South King Street
Honolulu, HI
96813-2943

808 - 539 - 4881

CJRI@courts.hawaii.gov

https://cjrihawaii.com/

https://www.courts.state.hi.us/criminal-justice-research-institute-cjri

# INTELLIGENT DOCUMENT PROCESSING

## What is intelligent document processing?

Intelligent document processing (IDP) is a tool that automates the extraction of data from paper-based documents or document images (e.g., PDFs) through a combination of generative artificial intelligence (GEN-AI), natural language processing (NLP), and/or machine learning (ML) [1]. IDP tools extract the unstructured information contained within documents and transform it into structured data, which can then be used in research and analysis. Built into IDP tools is the ability to verify the data they extract and collect, by applying predetermined rules to check for errors and/or cross referencing existing databases. Like other AI powered tools, IDP systems continually learn and improve over time, adapting to changes in document formats and learning from previous errors to improve the accuracy of the data they collect.

## What are the potential benefits of IDP for the Judiciary?

The implementation of an IDP tool could be beneficial to the Judiciary. Throughout the courts, a myriad of documents are filed, many of which contain information that could be used to inform decisions, increase efficiency, and advocate for resources across the judicial system. However, this information is not readily accessible for research and analysis purposes, currently requiring staff to manually read these documents to extract and log relevant information into a spreadsheet or document. The courts have limited capacity for this type of work due to staffing constraints. Additionally, manual document processing has the potential to introduce human error into the data collection and entry process, in which data is entered incorrectly into a dataset, and is inconsistent with what is reflected in the source document. Moreover, when multiple individuals are involved in manual document processing, they may differ in the information they are collecting, also creating inconsistencies between the data collected and the source documents. An IDP tool would allow the Judiciary to collect information from a myriad of documents at scale, and expand the court's capacity for data collection

from within its own case and records management systems. Additionally, an IDP tool would promote the collection of more consistent information collected from documents, through the elimination of additional human data entry, and through the inclusion of protocols that promote accuracy and reliability.

### What are recommendations for adopting an IDP tool for the Judiciary?

There are many documents that are filed with the courts that are already amenable to processing with an IDP tool in their current form. However, many others are not. IDP tools are limited in their ability to process and extract information from handwritten documents, such as requests for an Order for Protection, which are often filled out by hand by petitioners. By adapting similar forms to be fillable PDFs, which allow users to enter information into specific fields, instead of being printed out, filled out by hand, and scanned as PDFs for filing, would make them more amendable for processing by an IDP tool. These forms would still be printable should a court user want or need to print them out and fill them in by hand, and can be flagged by the IDP tool for manual review prior to being used for analysis.

### What are the considerations and limitation of IDP use for the Judiciary?

In addition to the benefits of an IDP tool, there are also several considerations and limitations to keep in mind when deciding if the courts should adopt an IDP tool. IDP systems can be expensive, related to data storage and per use (per document read costs), especially when deployed across a wide variety of documents. IDP tools are only as reliable as the information that is captured within these documents and forms, which may be filed by court staff, attorneys, or members of the public. This may require training, policy, or guidance on entering information so that the information contained within documents and forms are amendable to the IPD tool. Moreover, IDP tools are relatively new, and a quality assurance process should be implemented to validate the accuracy of the information being captured by the tool before relying on it for research and analytics. Finally, subject matter experts within the Judiciary should be identified,

and these individuals should assist with setting up keywords to code and extract information from documents and forms.

### What are potential use cases for IDP within the Judiciary?

- Identifying how often bench warrants are issued related to failure to appear.
- Examining criminal complaint documents related to specific offenses.
- Exploring judgments of conviction for more comprehensive information on plea deals, sentencing, and probation conditions.
- Tracking HRS § 704 related filings and outcomes.
- Evaluating orders for protection and temporary restraining orders.

**Example Use Case - Criminal Complaint Documents and Marijuana Charges**

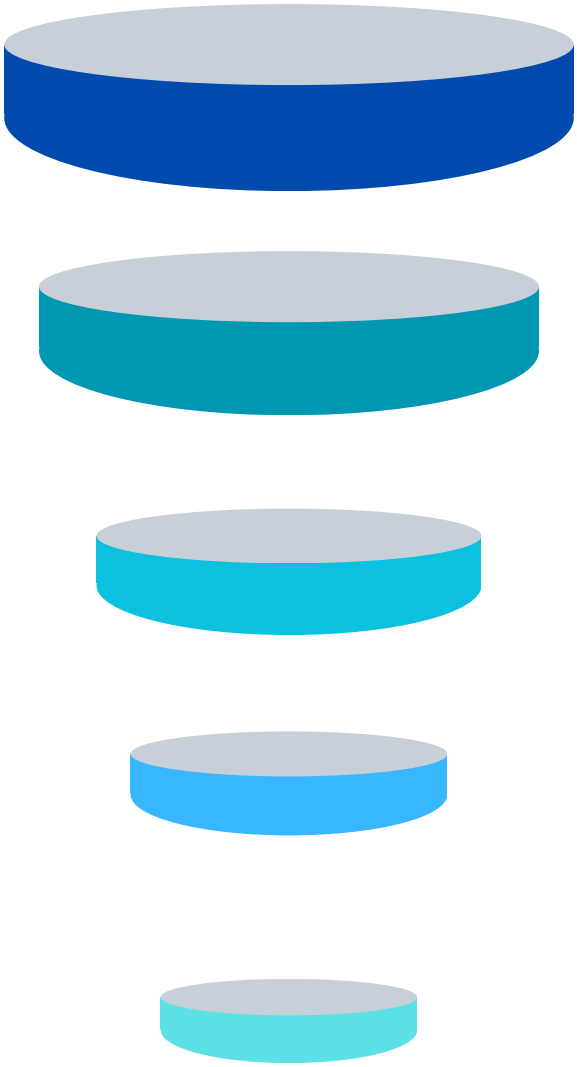
During the 2023 legislative session, there was notable interest in determining how often individuals are charged specifically for marijuana, as opposed to other drugs. The Hawai`i Revised Statues does not have offenses specific to marijuana only; rather, marijuana is one of several drugs included (e.g., HRS § 712-1249 lists marijuana and Schedule V drugs). In the fall, CJRI piloted the use of an IDP tool to identify individuals whose charges were related to marijuana, through the processing of their charging documents. CJRI staff developed logic, which was deployed by the IDP tool to extract specific data points from the charging documents, including the offense date, the type of drug involved in the offense, and the amount of drugs specified in the complaint.

### References

1. *What is Intelligent Document Processing? - IDP Explained - AWS.* (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/what-is/intelligent-document-processing/#:~:text=Intelligent%20document%20processing%20(IDP)%20is,when%20stock%20levels%20are%20low.

**Figure 1. Criminal Justice Funnel for Pretrial Data**

*Why is data mapping important for a centralized data warehouse?* As someone progresses through the criminal justice system, decisions points filter people and cases out like a funnel. The graphic below demonstrates this.



**Arrests**
Arrests occur with local police, and arrest data are aggregated in CJIS.

**Charges**
Prosecutors must file charges to initiate a case, which is tracked in Judiciary JIMS data. Not all arrests are charged.

**Jail bookings**
Not all individuals charged will be brought into a DCR facility, whether for booking or pretrial detainment.

**Adjudications**
Pretrial ends when a case is adjudicated, either dismissed or sentenced.

**Convictions**
Not all adjudications are convictions, an individual must plea or be convicted by jury. Some individuals may have their cases dismissed or be found innocent.

As records are matched across data from HCJDC, DCR, and the Judiciary, identifiers are used to link these records. In some cases it is a unique identifier for a person, other times it is a unique identifier for an arrest report or a court case. Not all records will have matches as cases filter out of the system. And for some metrics, it requires all three data sources to calculate metrics for pretrial.