Criminal Justice Open Data Summit: Post-Conference Report[1]

Emily Eckart, Jacob Kaplan & Jonathan Mummolo

Executive Summary: The Criminal Justice Initiative at Princeton University hosted the Criminal Justice Open Data Summit in August 2025, convening national leaders in criminal justice data - researchers, policymakers, journalists, data scientists, and practitioners - who are defining the future of open, high-quality data in the field. The event featured a focused, one-day discussion designed to assess the state of the field, identify shared challenges, and set priorities for future progress.

The one-day program combined panels, Q&A sessions, and curated roundtables to learn from the people closest to the data, many of whom are already engaged in efforts to improve it. Discussions centered on the decentralized nature of criminal justice data in the United States and the resulting inconsistencies in quality, documentation, standardization, and accessibility. Participants examined recent efforts to address these challenges through large-scale data collection and standardization, while also confronting the risks of oversimplified dashboards, insufficient documentation, and the ethical dilemmas of releasing incomplete or misleading data. Particular attention was given to the uneven quality and release of publicly funded data, and to the limited availability of well-documented open data outside policing, especially in courts, corrections, probation, and parole.

The report that follows synthesizes these discussions and outlines five priorities for advancing open criminal justice data:

1. Investing in long-term data infrastructure
2. Shared standards, documentation, and linkable data
3. Improved academic incentives for data projects
4. Expansion of open-data efforts beyond policing
5. A long-term *moonshot* - the creation of a research-focused records management system

Together, these priorities lay the groundwork for stronger and more usable data that supports rigorous, reliable research and informed policymaking.

[1] The authors thank all participants in the Criminal Justice Open Data Summit for their insights and discussion, which informed this report. Special appreciation is extended to the presenters who shared their ongoing work and experiences with the group. The authors also gratefully acknowledge Arnold Ventures for its generous support through the Innovation Days program, which provided funding that made the Summit possible. Any errors are the authors' own. Authors listed in alphabetical order.

State of the Field: Progress and Persistent Challenges

Across four panels, five roundtables, and extensive informal discussions, participants from multiple disciplines, including political science, economics, criminology, journalism, data science, public policy, and government, shared a common goal: to improve the quality, accessibility, and usefulness of data describing the criminal-justice system. The conversations revealed both the promise and the limits of current data infrastructure. Despite the persistence of long-standing problems such as fragmentation, missingness, inaccuracies, and weak incentives for maintenance, the overall trajectory is positive. Recent initiatives show that sustained effort, collaboration, and consistent coding and data review can yield real improvements in available data.

In recent years, a wave of new projects has transformed how criminal justice data are collected, standardized, and shared. The National Archive of Criminal Justice Data (NACJD), established nearly fifty years ago, remains the cornerstone of this infrastructure, providing long-term curation, documentation, and public access to tens of thousands of datasets. More recent initiatives, including the Real-Time Crime Index (RTCI) (led by Jeff Asher), the Jail Data Initiative (JDI) (led by Anna Harvey), Civic Police Data Project (CPDP) (led by Andrew Fan), and the DATA Lab within the Philadelphia District Attorney's Office (led by Oren Gur), presented by conference speakers, demonstrate that it is now possible to produce near-real-time data on crime, jail populations, police, and prosecutorial activity. The Jail Data Initiative, in particular, represents an extraordinary technical and logistical achievement, requiring the regular scraping, cleaning, and standardization of data from hundreds of different jail websites across the country. These projects show that up-to-date, publicly accessible, and clean data on core components of the justice system are possible.

While the projects showcased at the Summit were highly technical, the discussions were accessible. Participants focused less on code or scraping methods and more on what defines *good* open data: how to balance accessibility with responsibility, how to ensure data remain reliable over time, and where the field should go next.

Discussions throughout the Summit highlighted that "open" and "useful" are not synonymous. Participants raised important questions about the responsibilities that come with making data public, and about who benefits from openness. Several presenters cautioned that poorly documented or unvalidated data, while technically publicly available, can mislead more than inform. For example, crime may appear to spike in December simply because that is when reports are entered, not when the offenses occurred. Without documentation explaining such reporting lags, users may draw incorrect conclusions.

Dashboards and similar visualization tools were another point of debate. These products, such as the Real-Time Crime Index or NORC's Live Crime Tracker, make data easier to access for a wide audience, but can also strip away crucial context. When information is presented in simple, visual form, users are less likely to consult the documentation, especially for topics of high public interest such as crime and policing. Many datasets require hours of careful examination of the underlying data, code, and documentation before they can be trusted. Once released

openly, however, there is no way to ensure that users, particularly those without sufficient technical training, will take those steps. As a result, even well-intentioned users can unintentionally misuse the data or misstate its meaning. As a result, incorrect information can make its way into the public sphere, which can misinform policy and potentially waste valuable and limited resources.

A topic that came up frequently was the challenge of linking data across stages of the criminal justice process - from the initial police stop to judicial outcomes, sentencing, and release. While data on these stages exists, it is often siloed, with varying levels of availability and detail, making integration difficult and requiring significant manual effort. This lack of linkage limits research to specific parts of the system, like policing, where data is most readily available, while other stages receive less attention. One participant pointed out that even tasks that seem simple at first glance, like merging Sheriff crime data with Census data, are fraught with challenges. Although Sheriff's Offices operate at the county level, their jurisdiction is typically confined to specific cities or unincorporated areas within that county. To merge these data correctly, it's not enough to simply match by county; one must manually identify the exact areas under each Sheriff's jurisdiction. While the quicker method of merging by county may seem easier, it can lead to inaccurate results, including, as the participant noted, potentially reversing the direction of the effect being studied. Conference members agreed that improving data linkage across the criminal justice system is a critical next step for open data, and getting it right is essential for a more complete understanding of the process.

Participants also discussed the subjectivity involved in determining what constitutes "bad" or "untrustworthy" data. While many data errors are straightforward to identify, others require judgment about whether the information is correct, and different people may reach different conclusions. Some noted that withholding flawed data can sometimes be the more responsible choice, particularly when public release risks reinforcing misinformation. Refusing to release data can itself serve a transparency function, demonstrating, for example, that certain agencies are providing data of such poor quality that it cannot be responsibly shared.

There was also extensive discussion about ownership and ethics: who the data ultimately belong to: the agencies that collect and manage them, or the communities they describe. Many emphasized that open data must serve both, providing accountability and public transparency while offering tangible value to agencies and addressing concerns or topics of interest to the community. Sharing data requires staff time, technical capacity, and sometimes political risk. For that reason, sustainable open-data systems must ensure reciprocal benefit - helping agencies answer operational questions, improve internal analyses, and reduce repetitive open records requests.

Finally, participants stressed that what constitutes "useful" data depends on the audience. Researchers may prefer detailed, raw data suitable for analysis, while journalists and the public may need aggregated or visual formats. Open-data initiatives must therefore consider multiple levels of use and release, balancing transparency with responsibility and clarity.

Priorities for the Future

While there was no full consensus on every issue discussed, participants broadly agreed on the areas most critical to improving the field in the coming years. Across panels, roundtables, and informal discussions, several recurring themes emerged: the need for stronger data infrastructure, better documentation, shared standards, linkable data, incentives to encourage data work, and tackling messy data at its source. The sections that follow outline the areas of general agreement and practical steps that would strengthen open data across the criminal justice system.

1. Investing in long-term data infrastructure

Data collection and cleaning in criminal justice research are time-consuming and often unpredictable. A single website update can break an automated scraper and add hours or even days of additional work. The most valuable datasets combine long historical coverage with near-real-time updates, yet maintaining that continuity requires consistent funding and careful documentation of every change. Too many projects are funded as one-time efforts, producing static snapshots that quickly become outdated.

The presenters at the Summit demonstrated a more sustainable model: ongoing data projects in which maintaining and improving the dataset itself is the central objective. This approach treats data as a living resource that must be continually collected, cleaned, and documented. Funders can help by supporting long-term maintenance plans, and by encouraging investments that improve efficiency, such as well-tested, reproducible code that can adapt (or at least detect and warn the data team) to changes in the raw data.

Sustained investment should also include support for documenting and modernizing historical data. Many of the most-used public datasets, including those produced by the FBI, could be significantly improved through systematic review, comprehensive documentation, and standardized formatting across years. In addition to large-scale initiatives, there is also value in smaller, targeted projects, e.g. ones funded by microgrants that help data scientists to extract, clean, and document data already available online, or through open records requests. These relatively modest investments can yield high returns by converting inaccessible or challenging data into reusable public resources.

2. Shared standards, documentation, and linkable data

Participants emphasized that many projects rely on combining multiple datasets to answer even basic research questions. Yet each dataset often requires cleaning and standardization before it can be linked. Many criminal justice datasets are produced independently, using different variable names, definitions, and file structures that make this process slower and more error-prone than it needs to be. Even small inconsistencies, such as one agency naming a column “offense,” another “offense_type,” and a third “type_of_offense”, create unnecessary work that could be avoided through shared conventions.

To make data more usable, there should be common standards for how open data are released. These include naming conventions for variables, consistent file structures, documentation standards, and release checklists that ensure datasets are accompanied by codebooks and notes on cleaning or transformation steps. Other fields have already developed lightweight solutions that criminal justice data producers could emulate, such as widely used conventions for naming variables and documenting code in a common language, e.g. R.

Having shared standards would also make the process more efficient for new contributors. Clear templates and detailed checklists provide a gentle introduction to data sharing, outlining what information to include and how to format it. This makes it easier for new users to participate while improving consistency and reducing the likelihood of errors.

Participants also called for greater attention to making data linkable. Shared identifiers, such as standardized agency codes, geographic units, and consistent date fields, allow data from different sources to be merged. Linked data enables richer analyses, helps verify accuracy across datasets, and makes it possible to include relevant variables in an analysis that may only be available in different sources. When standardized linking variables are missing, linking often requires extensive manual work. This gap can be addressed by data projects that focus on cleaning and augmenting existing datasets with shared identifiers, allowing others to merge them more efficiently.

3. Improved academic incentives for data projects

A striking feature of the Summit was that many participants doing the most sophisticated data work were not faculty members. This reflects a broader academic problem: data collection, cleaning, and sharing remain under-incentivized within universities. Faculty have little professional incentive to engage deeply in these tasks, even though their research often shapes public understanding and policy. When faculty disengage from data work, the field loses both expertise and influence.

Even when faculty want to take part in building or sharing data, the incentive structure pushes them to prioritize publications over data projects. Graduate students and early-career scholars face the same constraint, with tenure and hiring decisions tied to journal output rather than data contribution.

Graduate training often reinforces this divide by emphasizing analysis of clean, ready-to-use datasets rather than the messy, incomplete data that characterize real-world criminal justice research. Some graduate programs do not have any mandatory courses on data wrangling. However, certain universities, including Princeton, have expanded training through workshops and courses on acquiring, cleaning, and documenting data, recognizing the importance of these skills. Still, such efforts are not yet standard, and the amount of time and effort required for serious data work should be a larger focus. Data management and reproducibility should be core components of graduate curricula.

The growing use of replication archives in academic research–criminology being a notable exception–underscores this point. These archives, which require authors to post the data and code necessary to reproduce their published results, are a positive step toward greater transparency and accountability. Yet, they also raise the bar for data quality and documentation. As more fields adopt replication requirements, researchers must be prepared to share well-documented code and data that others can understand and reuse. This shift improves the trustworthiness of research but also highlights the need for stronger training and institutional support, since the added expectations come without corresponding incentives or resources.

Greater faculty participation in data work would also strengthen collaboration outside of academia. Journalists, data scientists, and private organizations now lead many of the most innovative open-data projects, but partnerships are easier and more productive when everyone understands what working with data actually entails. Faculty who lack hands-on experience may have unrealistic expectations about what is possible or how long it will take, and may struggle to engage in technical discussions. Faculty with stronger data skills can “speak the same language,” set achievable goals, and contribute meaningfully to joint projects at each step rather than only during analysis.

Given the difficulty of changing norms around faculty hiring and promotion, an alternative way in which universities and funders can contribute to these efforts is to support the hiring of research staff dedicated to data collection and curation. Dedicated data scientists working hand in hand with leading researchers would serve both public and academic interests. Grants can also include course releases or salary support for researchers to prepare replication archives and cleaned datasets, and provide targeted funding for documentation and dissemination. Such steps help normalize data sharing as an integral - or at least an incentivized - part of academic research.

4. Expansion of open-data efforts beyond policing

Most progress in open criminal justice data has centered on policing. Historically, police data, such as those collected by the FBI, have been the most complete, accessible, and frequently analyzed. Police are also the most visible and immediate face of the criminal justice system, which has naturally drawn public, academic, and policy attention to them. However, policing is only one part of a broader system, and understanding that system requires data on every part, including courts and corrections.

Projects such as the Philadelphia District Attorney’s Office DATA Lab and the Jail Data Initiative have begun to expand the field’s scope, demonstrating that it is possible to build and sustain open data on prosecutors and jails. The next step is to build on these examples and extend similar efforts to other justice system actors. Data on other system operators, including call takers, dispatchers, judges, prosecutors, probation and parole officers, and correctional staff remain sparse. Understanding what these professionals do and how they spend their time is essential to capturing how the system functions as a whole.

Improving data coverage beyond policing would not diminish existing efforts but complement them. Areas with limited or poor-quality data likely hold the greatest potential for research to improve practice and inform policy. A full picture of the criminal justice system depends on bringing these missing pieces into view.

5. The Moonshot: A Research-Focused Records Management System

The most ambitious idea to emerge from the Summit was the proposal for a nonprofit, research-driven records management system (RMS). This idea aims to address one of the field's most fundamental problems at its source: agency data systems themselves. Enormous amounts of researcher time and funder resources are spent cleaning and standardizing data exported from existing RMS software, systems that were never built with analysis, standardization, or transparency in mind. In some agencies, the data are so limited that even basic questions, such as identifying all officers present at an incident, cannot be reliably answered. This gap has serious implications for both research and for agencies' own early-warning systems, accountability tools, and case preparation.

A research-oriented RMS would reverse this approach. Rather than cleaning data after the fact, it would embed validation, documentation, and standardization at the point of entry. Simple features, such as limiting offense types to a standardized list or automatically flagging invalid entries, could prevent many errors that currently produce hundreds of unique spellings or inconsistent codes for the same offense. By improving quality upstream, the system would dramatically reduce the need for downstream cleaning, saving time for researchers and money for funders.

For agencies, such a system would offer a low-cost alternative to commercial RMS vendors while improving their ability to conduct the kinds of analyses they already want, such as identifying early-warning indicators, expanding CompStat, or assessing staffing and response patterns. For the research community, it would generate data that are consistent, linkable, and usable almost immediately, accelerating work that currently takes years of cleaning to complete. A nonprofit RMS built for transparency and research would raise expectations across the industry, forcing existing vendors to improve data quality and analysis tools. The competitive pressure alone could drive vendors to offer better products, lower prices, and stronger data-management capabilities, benefiting agencies even if they never switch systems.

Developing such a system would require significant investment, careful design to ensure data security, and close collaboration between researchers and practitioners. This system could potentially be operated by a nonprofit organization rather than by academics alone, allowing it to be self-funding and sustainable over time. This project would take time and sustained effort, but the potential payoff is transformative: a shift from cleaning bad data to creating good data at the source. Addressing these issues upstream would ultimately be a more efficient and durable solution for funders, agencies, and researchers alike.

Conclusion

While the field continues to face significant structural and practical challenges, the Summit made clear that progress is real and accelerating. Participants agreed that the technical ability to produce reliable, timely, and well-documented data now exists; the question is how to sustain it, coordinate it, and ensure it is used responsibly. By bringing together leaders from across the country, this conference helped clarify the state of open criminal justice data and chart a path for its continued improvement. Future work will determine whether the progress made thus far becomes a lasting foundation for transparency, accountability, and understanding in the criminal justice system.