CIVIL JUSTICE DATA COMMONS

Data Commons Models

A report by the Civil Justice Data Commons

June 2022

Contents

About the Civil Justice Data Commons

The Civil Justice Data Commons is a joint project between the Georgetown University Law Center and Georgetown University's Massive Data Institute at the McCourt School of Public Policy that aims to create a secure, robust repository for civil legal data gathered from courts, legal services providers, and other civil law institutions. This repository enables stakeholders, researchers, and the public to better understand the civil legal system in the United States.

The Civil Justice Data Commons is led by Dr. Amy O'Hara of the Massive Data Institute and Professor Tanina Rostain of the Georgetown University Law Center.

This report was prepared by James Carey, Max Brossy, Margaret Haughney, Garrett Lance, Hannah Olsen, Anna Stone, Stephanie Straus, and Eoin Whitney.

This work is supported by a grant from the National Science Foundation (AWD 1952067).

To learn more about the Civil Justice Data Commons, please visit https://www.law.georgetown.edu/tech-institute/programs/civil-justice-data-commons/.

Introduction

This report represents foundational research that underlies the development of the Civil Justice Data Commons project. It discusses what is a data commons, what are the best practices for creating a data commons, what other example projects can a civil justice data commons learn from, and how do those other projects deal with issues such as data classification, data linkage, and text extraction. In looking at these questions, the report provides insight into the creation and architecture of the Civil Justice Data Commons. It also allows us to share what we have learned about data commons with others who wish to pursue their own projects.

What is a Data Commons?

A data commons is digital infrastructure that serves as an interoperable resource for a research community.[1] It collocates data and computing power with commonly used tools for analyzing and sharing those data. A data commons serves as a one stop shop for data research by allowing for data discovery; storage of curated data and associated metadata; linking of data with other data, publications, and citations; and easy access to tools for computing and analysis. Data commons support three functions: (1) serving as a data repository or digital library for data associated with published research, (2) storing data along with computational environments in virtual machines ("VMs") or containers so that computations supporting scientific discoveries can be reproducible, and (3) acting as a platform enabling future discoveries as more data, algorithms, and software applications are added to the commons.

There multiple stakeholders involved in a data commons.[2] First, there is the data commons service provider ("DCSP"), the entity operating the data commons. Second, there is the data contributor ("DC"), the individual or organization providing the data to the service provider. Third, there is the data user ("DU"), the individual or organization accessing the data (often academic researchers, government agencies, or journalists).

A data commons is not always the same as a "data trust," though they are related.[3] "Data trust" is often used to describe the entity that controls access to specific data, the legal framework governing that entity, or an organization operating under such an entity. A data commons involves a set of contractual obligations that govern data providers, data users, and the commons service providers. A data commons and data trust both have an overarching governance model and a common data model for interoperability across sites; however, a data trust's governance entity directly administers it, whereas the administration of a data commons might be outsourced to another operator or run by a larger institution. Nevertheless, common lessons on best practices often apply to both.

---

[1] R.L. Grossman, A. Heath, M. Murphy, M. Patterson, and W. Wells, *A Case for Data Commons: Toward Data Science as a Service*, 18 Computing in Science & Engineering 10–20 (2016) (https://papers.rgrossman.com/journal-056.pdf).

[2] *Id.*

[3] One definition, put forth by the Open Data Institute ("ODI"), is that a data trust is "a legal structure that provides independent stewardship of data." J. Hardinges, P. Wells, A. Blandford, J. Tennison, and A. Scott, *Data Trusts: Lessons from Three Pilots,* Open Data Institute (Apr. 15 2019) (https://theodi.org/article/odi-data-trusts-report). Stuart Mills says that there are subtle distinctions between data commons and trusts. Stuart Mills, *Who Owns the Future? Data Trusts, Data Commons, and the Future of Data Ownership*, (Sept. 24, 2019) (http://dx.doi.org/10.2139/ssrn.3437936). He argues that while there is no final definition of the latter term, there is some consensus that it must be a data-pooled resource that is greater than the sum of its parts. Further, Mills argues that a trust is a third-party occupying role of data steward on behalf of various stakeholders. This differs from a data commons, says Mills, as a commons is a platform that expands data access by bringing together data from more than one source. Mills notes that Grossman et al. explain that commons have interoperability, providing access and a platform for data experimentation and interaction. Mills states that a potential area of overlap between data commons and data trusts is their similar preference for controlling access and managing permissions of the data so that the original controls on the data, when it was collected, can be maintained. They differ, according to Mills, in that data commons support linking to other relevant datasets and offer relatively fewer access restrictions than trusts do. According to Kieron O'Hara, data trusts are unique because they have a trustee who administers the trust for the benefit of the user(s). Kieron O'Hara, *Data Trusts: Ethics, Architecture, and Governance for Trustworthy Data Stewardship*, Web Science Institute (Feb. 2019) (https://eprints.soton.ac.uk/428276/1/WSI_White_Paper_1.pdf).

Technical Requirements and Best Practices

Several key technical requirements and best practices govern the creation of a successful data commons. Research on data commons and data trusts suggests that there are four main areas that must be considered when determining these requirements and best practices: security, transparency, efficiency, and access.[4] Some organizations add a fifth area: trust.[5]

Security

Paprica, et al. stress the importance of having policies and processes that require data protection steps be taken and be reviewed and updated regularly. Further, they say, any prospective data users must complete a training session before being permitted to access the data, and also sign a Data Use Agreement ("DUA"), some of the provisions of which include acknowledgement that their data use will be monitored and that there will be some sort of consequences for non-compliance or other improper data usage.[6] Similarly, at the Criminal Justice Administrative Records System ("CJARS") (discussed in more detail below in Example Platforms), researchers Finlay and Mueller-Smith explain that any Personally Identifying Information ("PII") contained in their criminal-justice datasets gets removed at a very early stage.[7] Only approved CJARS staff working on linkage can see the PII, no third-party data users can.

Transparency

Paprica, et al. urge data trust operators to engage early and regularly with all affected stakeholders, including the general public.[8] Engagement discussions must be honest and open, not simply going through the motions for the sake of checking off one box on a checklist. Further, they advise that if the data trust operators have reason to believe that a particular sub-group of the population might have an outsized stake in, impacted in a greater manner by, or have extra interest in a particular data project, the data trust operator should engage in additional, appropriately tailored engagement with that sub-group. The Open Commons Consortium ("OCC") (discussed in more detail below in Example Platforms) states that sensitive data should only be released in "specialized environments," not to the public.[9] Less-sensitive data should only be released to the public after all DUA terms have been fulfilled and executed, which may include an embargo of half a year, and still in a controlled, password-protected environment. The OCC's policy for data that is not sensitive, however, is that such data should be published using open-source software available to anyone, in the interest of furthering knowledge. The OCC requires that any data they release in any capacity (open or controlled, free or proprietary) be de-identified before publication, whether the data is gathered directly from the subjects or indirectly via customer-serving technological applications, according to the privacy terms and conditions of the app in question. This includes making any software developed directly by the OCC or its partners open-

---

[4] Primarily P.A. Paprica et al, *Essential Requirements for Establishing and Operating Data Trusts: Practical Guidance Based on a Working Meeting of Fifteen Canadian Organizations and Initiatives*, International Journal of Population Data Science, 5(1) (2020) (https://doi.org/10.23889/ijpds.v5i1.1353); Stuart Mills, *Who Owns the Future? Data Trusts, Data Commons, and the Future of Data Ownership*, (Sept. 24, 2019) (http://dx.doi.org/10.2139/ssrn.3437936); R.L. Grossman, A. Heath, M. Murphy, M. Patterson, and W. Wells, *A Case for Data Commons: Toward Data Science as a Service*, 18 Computing in Science & Engineering 10–20 (2016) (https://papers.rgrossman.com/journal-056.pdf); K. Finlay and M. Mueller-Smith, *Criminal Justice Administrative Records System (CJARS)*, University of Michigan Institute for Social Research. (June 17, 2020) (https://cjars.isr.umich.edu/wp-content/uploads/CJARS_data_docs_2020_06_17_13_18.pdf).

[5] [5] J. Hardinges, P. Wells, A. Blandford, J. Tennison, and A. Scott, *Data Trusts: Lessons from Three Pilots,* Open Data Institute (Apr. 15 2019) (https://theodi.org/article/odi-data-trusts-report).

[6] P.A. Paprica et al, *Essential Requirements for Establishing and Operating Data Trusts: Practical Guidance Based on a Working Meeting of Fifteen Canadian Organizations and Initiatives*, International Journal of Population Data Science, 5(1) (2020) (https://doi.org/10.23889/ijpds.v5i1.1353).

[7] K. Finlay and M. Mueller-Smith, *Criminal Justice Administrative Records System (CJARS)*, University of Michigan Institute for Social Research. (June 17, 2020) (https://cjars.isr.umich.edu/wp-content/uploads/CJARS_data_docs_2020_06_17_13_18.pdf).

[8] P.A. Paprica et al, *Essential Requirements for Establishing and Operating Data Trusts: Practical Guidance Based on a Working Meeting of Fifteen Canadian Organizations and Initiatives*, International Journal of Population Data Science, 5(1) (2020) (https://doi.org/10.23889/ijpds.v5i1.1353).

[9] *Commons Principles*, Open Commons Consortium (https://www.occ-data.org/commons-principles).

source. An additional OCC transparency policy is to publish their findings in open-access publications or journals, as well as to deposit archival versions into similarly open-access repositories.

Efficiency

The researcher Mills suggests that data commons and trusts must be interoperable with other systems, combining access to data and experimentation.[10] Paprica, et al. advise that the policies and processes governing a trust's collection, storage, use, and disclosure of data be well-defined and standardized.[11] Hagan et al. recommend that data commons be usable by multiple parties simultaneously, scalable to include new datasets and new features over time, interoperable across multiple datasets within the commons, and explorable via multiple tools.[12]

Access

There are two prongs of access: stewardship and actual access.

Regarding stewardship, the ability to control access means the steward can dictate the flow of any potential financial or economic benefits that are gained from access.[13] Good stewardship must control access for the sake of the commons and not personal gain.

Actual access presents a logistical challenge. For security reasons, some organizations may limit access to a data commons to a secured physical space.[14] However, enabling virtual access to data commons (including through secure virtual enclaves) can allow for access at a greatly reduced monetary and logistical cost to researchers.

Trust

The Open Data Institute ("ODI") includes "trust" as a key element of a data commons. Its definition of a data trust blends both aspects of the word "trust:" the trustworthiness aspect and the legal fiduciary aspect. ODI discusses trust in the interpersonal sense of trustworthiness.[15] Trustworthiness is a fundamental requirement of data stewardship; otherwise, confidence is lost between and among the people and organizations involved in a given project. Failure to maintain trust and confidence necessarily impedes progress, success, and a project's ability to reach its full potential. In addition, under the ODI's aforementioned definition of a data trust (an independent entity that does not actually own the data in question), such a trust is legally mandated (by the binding contractual terms in their founding document) to "make decisions about its use for an agreed purpose while taking all relevant stakeholder interests into account." ODI says that this requirement will best enable the trust to "balance different – and often conflicting – views and incentives about how data should be shared and who can access it."

---

[10] Stuart Mills, *Who Owns the Future? Data Trusts, Data Commons, and the Future of Data Ownership*, (Sept. 24, 2019) (http://dx.doi.org/10.2139/ssrn.3437936).
[11] P.A. Paprica et al, *Essential Requirements for Establishing and Operating Data Trusts: Practical Guidance Based on a Working Meeting of Fifteen Canadian Organizations and Initiatives*, International Journal of Population Data Science, 5(1) (2020) (https://doi.org/10.23889/ijpds.v5i1.1353).
[12] Margaret Hagan, Jameson Dempsey, and Jorge Gabriel Jiménez, *A Data Commons for Law*, Legal Design and Innovation (Apr. 2, 2019) (https://medium.com/legal-design-and-innovation/a-data-commons-for-law-60e4c4ad9340).
[13] Stuart Mills, *Who Owns the Future? Data Trusts, Data Commons, and the Future of Data Ownership*, (Sept. 24, 2019) (http://dx.doi.org/10.2139/ssrn.3437936).
[14] K. Finlay and M. Mueller-Smith, *Criminal Justice Administrative Records System (CJARS)*, University of Michigan Institute for Social Research. (June 17, 2020) (https://cjars.isr.umich.edu/wp-content/uploads/CJARS_data_docs_2020_06_17_13_18.pdf).
[15] J. Hardinges, P. Wells, A. Blandford, J. Tennison, and A. Scott, *Data Trusts: Lessons from Three Pilots,* Open Data Institute (Apr. 15 2019) (https://theodi.org/article/odi-data-trusts-report).

Examples of Data Commons and Similar Data Repositories

The example platforms described below include several existing data commons and similar data repositories in other fields that offer innovative solutions to the issues that a civil justice data commons will encounter. Each platform is described to highlight the lessons that can be learned from them. The fields that the platforms operate in range from criminal justice to health to data repositories that span multiple fields.

Data Repositories in Criminal Justice

Criminal Justice Administrative Records System

The Criminal Justice Administrative Records System ("CJARS") is an integrated data repository created through a partnership between the University of Michigan and the US Census Bureau. The University of Michigan collects the data through three different channels: (1) data use agreements, (2) public records requests, and (3) web scraping or bulk downloads. This data is then cleaned, harmonized and anonymized on a University of Michigan system that was built to be compliant with FBI Criminal Justice Information Services ("CJIS") standards. Personally Identifiable Information ("PII") is removed at an early stage of processing and is only available to researchers working on recording linkage. Data containing all PII is sent to the Census Bureau, where staff attempt to use all available PII to assign a Protected Identification Key ("PIK") using a probabilistic record linkage system called the Person Identification Validation System ("PVS"). Cases are linked not only by individuals but also by whether or not the records were related to the same incident. Social Security Numbers are not used in the matching. Once the PIK assignment process has occurred, the anonymized files with PIKs attached are transferred to a secure computing environment that is available at the Census Bureau headquarters and in the Federal Statistical Research Data Center ("FSRDC"). On those servers, approved data in approved projects can be linked at the person-level using the PIKs attached to each file, including the CJARS data.[16]

Measures for Justice

Measures for Justice ("MFJ") is a nonprofit based in Rochester, NY, that has been working on a criminal justice data commons.[17] MFJ aims to "collect, standardize, and publicize county-level criminal justice data from across the United States, so that policymakers, practitioners, advocates, and the general public can understand how the criminal justice system is performing, so that they can identify well-performing and low-performing jurisdictions, so that they can make changes to policy and practice, [and] so that the criminal justice system is more fair, efficient, and effective."[18] Over the past decade, MFJ has gathered court data for over 1,200 counties across at least 20 states.[19] The organization recently launched its first dashboard, displaying data for Yolo County, CA, located in the Sacramento region that also contains one of the state's University of California institutions. Already, MFJ's work has led to several states passing criminal justice data laws, and more reforms are likely in the pipeline.[20] MFJ also helped the National Center for State Courts ("NSCS") create the National Open court Data Standards ("NODS") (see Classification Systems below).[21]

---

[16] K. Finlay and M. Mueller-Smith, (2021, Mar. 22). *Criminal Justice Administrative Records System (CJARS) Data Documentation*, University of Michigan Institute for Social Research (Mar. 22, 2021) (https://cjars.isr.umich.edu/wp-content/uploads/CJARS_data_docs_2021_03_22_14_41.pdf).
[17] *About*, Measures for Justice (https://measuresforjustice.org/about).
[18] *Our Story*, Measures for Justice (https://www.measuresforjustice.org/about/story).
[19] *Our Story*, Measures for Justice (https://www.measuresforjustice.org/about/story).
[20] *About*, Measures for Justice (https://measuresforjustice.org/about).
[21] *Infrastructure*, Measures for Justice (https://measuresforjustice.org/infrastructure).

Data Repositories in the Health Sciences

Cancer Research Data Commons

The Cancer Research Data Commons ("CRDC") centralizes data from several other data commons and other sources, including the Genomic Data Commons, the Proteomic Data Commons, the Integrated Canine Data Commons, the Imaging Data Commons, and NCI Cloud Resources, among others. Data is made available through a cloud-based platform in which users apply for access, and if approved, are given secure credentials. The CRDC attempts to make its data as interoperable as possible, in terms of data standardization and documentation, both within the CRDC as well as ensuring interoperability with global organizations such as the Global Alliance for Genomics and Health ("GA4GH"), Digital Imaging and Communications in Medicine ("DICOM"), and Clinical Data Interchange Standards Consortium ("CDISC"). The CRDC illustrates the iterative approach of the data commons model, where different data commons can combine over time to create progressively more useful data centralization.[22]

Data Commons Pilot Phase Consortium

The Data Commons Pilot Phase Consortium ("DCPPC") is a program that allows for the development of a cloud-based platform to make three high-value datasets findable, accessible, interoperable, and reusable ("FAIR"). The Data Commons Pilot Phase Consortium is part of a broader strategy to facilitate a trans-National Institutes of Health data ecosystem planned through the Office of Data Science Strategy ("ODSS"). In doing so, the NIH intends to establish best practices in cloud infrastructure and data security, to accelerate data access and linkage, and facilitate research that has not been previously possible by linking data throughout the NIH and allowing data discovery and analysis to occur in the cloud. The Common Fund will continue to fund this work to produce future deliverables and develop best practices.[23]

Observational Health Data Sciences and Informatics

Observational Health Data Sciences and Informatics ("OHDSI", pronounced "odyssey") is a global network of researchers with health databases based out of Columbia University in New York.[24] It is a collaborative, interdisciplinary, multi-stakeholder program that aims to improve the value of health data via large-scale analytics. It provides open-source solutions. Its model is that each site is responsible for its own security, so OHDSI does not manage or advise on that from a central level. Generally, when OHDSI runs a study, they put the results directly on their public website and aggregate sites from there.[25] If a potential partner is not willing to publicly release its own data, they will not participate in the OHDSI study.[26]

Primary Care Development Corporation

The Primary Care Development Corporation ("PCDC") is a Treasury-certified Community Development Financial Institution ("CDFI") based in New York City.[27] It aims "to finance [primary care] facilities and bring culturally competent, high-quality care to underserved communities."[28] As a CDFI, it provides loans to these facilities to modernize, upgrade, and innovatively transform their operations. In line with the six CDFI tests from the Treasury Department, PCDC also engages in secondary activities beyond lending like providing expert technical assistance to healthcare providers (including its customers) and advocates for beneficial policies at all levels of government.

PCDC facilitates the collection and linkage of data from many different sources and types using technology to address inefficiencies in clinical research operations and data aggregation and analysis. Working with legal representatives from partner

---

[22] *NCI Cancer Research Data Commons*, National Cancer Institute (https://datascience.cancer.gov/data-commons).
[23] *Data Commons*, National Institutes of Health (May 19, 2019) (https://commonfund.nih.gov/commons).
[24] OHDSI – Observational Health Data Sciences and Informatics, https://www.ohdsi.org/.
[25] *OHDSI Data*, OHDSI (https://data.ohdsi.org/).
[26] *OHDSI – Observational Health Data Sciences and Informatics*, OHDSI (https://www.ohdsi.org/).
[27] *PCDC: Our Role, Primary Care Development Corporation*, PCDC (https://www.pcdc.org/about-pcdc/).
[28] *Id.*

institutions around the world, PCDC is able to streamline the process of creating international data contributor agreements to facilitate the sharing of data across international borders. They harmonize existing clinical research data and make it available to researchers to break down long-standing barriers that have kept clinical data siloed and held back advancements in research on rare diseases. PCDC is supported by many foundations, the National Institutes of Health ("NIH"), and the Department of the Interior. Other health and genomic data commons are funded by NIH, foundations, and some with pharmaceutical industry support.

Vivli

Vivli is a trusted intermediary that aggregates participant-level data from completed clinical trials to share with the international research community. "Vivli evolved from a project of The Multi-Regional Clinical Trials Center of Brigham and Women's Hospital and Harvard ("MRCT Center"), to improve access to clinical trial data and encourage data sharing and transparency."[29] The platform includes an independent data repository and search engine, and remains independent of data contributors, data users, and the broader research community.

Data Repositories Across Multiple Fields

Administrative Data Research Facility

The Administrative Data Research Facility ("ADRF") was commissioned by the US Census Bureau to inform the decision making of the Commission on Evidence Based Policy. Since 2018, it has provided services to over 180 government agency staff and researchers and hosted over 50 confidential government datasets from 12 different agencies. The ADRF is a pilot project that enables secure access to analytical tools, data storage and discovery services, and general computing resources for users, including federal, state, and local government analysts, and academic researchers.[30]

The Facility operates as a cloud-based computing environment, with federal security approvals, which currently hosts selected confidential data from the Department of Housing and Urban Development ("HUD") and the Census Bureau, as well as state, city, and county agencies, and an array of public-use data. The ADRF institutionalizes secure access to and use of confidential data. It is a secure cloud-based environment that is FedRAMP-certified[31] and has received Authority to Operate from the Census Bureau. This stamp of approval provides data owners with confidence that their data are secure. The cloud environment allows agencies within the same state or from different states to agree to share their data in a common area in the cloud for specific, approved projects. However, the ADRF does more than provide a secure environment. Data providers simply will not provide data if they cannot monitor who is using it, for what purposes, and with what results. The ADRF includes a combination of state-of-the-art technical strategies, thoughtful human oversight, and screening in order to dramatically improve privacy and usage protections. The ADRF is building a variety of standardized mechanisms for different confidentiality situations, with mechanisms for certifying the five "safes:" safe people, safe projects, safe settings, safe outputs, and safe data.[32]

If approved, staff from multiple agencies can jointly access approved areas in the cloud, so that they can work together to develop new integrated datasets, share information about coding differences or similarities, and develop common measures, without physically having to transfer data from one agency to another.

---

[29] *About Vivli: Overview*, Vivli (https://vivli.org/about/overview/).
[30] *ADRF*, Coleridge Initiative (https://coleridgeinitiative.org/adrf/).
[31] The Federal Risk and Authorization Management Program ("FedRAMP") provides a standardized approach to security authorizations for Cloud Service Offerings. *How to become FedRAMP Authorized*, GSA (https://www.fedramp.gov/).
[32] *ADRF*, Coleridge Initiative (https://coleridgeinitiative.org/adrf/).

Federal Statistical Research Data Centers

Federal Statistical Research Data Centers ("FSRDCs") are partnerships between federal statistical agencies and research institutions that provide authorized individuals with secure facilities to access restricted-use microdata. This data comes from statistical agencies, such as the National Center for Health Statistics, the Bureau of Labor Statistics, and the Bureau of Economic Analysis, through coordination with the Census Bureau, and allows researchers to examine various social and economic issues by linking their existing data to FSRDC data. The 31 FSRDCs nationwide have application processes including project review and approval procedures that align with permitted uses of each data source.[33] Approved users must access the microdata through the FSRDC's physical secure enclave on site, or, recently with the COVID-19 pandemic, a virtual enclave. Users must work in a specific set of software on FSRDC's secure computing network. If users would like to use additional software, they need approval from FSRDC security staff or managers.[34]

Iowa State: DataShare

DataShare is an open access repository for sharing and publishing research data created by Iowa State University scholars and students. The University Library reviews applications, which describe the data contents, the project, and the license for the data. The review considers such aspects as file format and extent of data documentation. If approved, the data is stored on the DataShare site and is publicly available for download. Data use is restricted to personal, non-commercial educational purposes. It is currently free of charge and not intended for big data. As data storage grows, DataShare will consider whether to charge a fee.[35]

National Opinion Research Center

The National Opinion Research Center ("NORC"), a "nonpartisan and objective research organization at the University of Chicago," operates the NORC Data Enclave.[36] The NORC Data Enclave stores and provides remote access to confidential microdata in healthcare, education, and social research. Part of their Advanced Data Solutions Center, the Data Enclave offers standard data services, such as secure remote data access and statistical disclosure control; research data services, such as data harmonization, de-identification, and linking; and platform-as-a-service, such as large-scale data warehousing, custom data extract, transform, and load ("ETL"), and database management.[37] Authorized users–from government agencies to research institutes to universities–utilize the secure storage, access, and analysis services through contracts or grants with NORC. The enclave utilizes a computing cloud-based environment that allows for projects such as all-payer claims databases and the sharing of Centers for Medicare & Medicaid Services data.

Open Commons Consortium

The Open Commons Consortium ("OCC") "manages and operates cloud computing infrastructure, data commons, and data ecosystems to advance scientific, medical, health care, and environmental research for human and societal impact."[38] The OCC requires its members to sign membership agreements so that the terms, conditions, and obligations of all parties are clear and

---

[33] *Federal Statistical Research Data Centers: Locations*, US Census Bureau (Nov. 18, 2021) (https://www.census.gov/about/adrm/fsrdc/locations.html). J. Shen and L. Vilhuber, *Physically Protecting Sensitive Data,* in: Cole, Dhaliwal, Sautmann, and Vilhuber (eds.), *Handbook on Using Administrative Data for Research and Evidence-based Policy* (2020) (https://admindatahandbook.mit.edu/book/v1.0-rc5/security.html#).

[34] *Federal Statistical Research Data Centers*, US Census Bureau (Dec. 16, 2021) (https://www.census.gov/about/adrm/fsrdc.html).

[35] *2021 Datashare Resource Fair One-Pager Final*, Iowa State University Library (2021) (https://drive.google.com/file/d/1LjTYmSYX9p5ADgmzZ6QGyEI5BbcWA2k0/view).

[36] *NORC Data Enclave*, NORC (https://www.norc.org/Research/Capabilities/Pages/data-enclave.aspx.

[37] *Advanced Data Solutions Center,* NORC (https://www.norc.org/About/Departments/Pages/advanced-data-solutions-center.aspx).

[38] R.L. Grossman, A. Heath, M. Murphy, M. Patterson, and W. Wells, *A Case for Data Commons: Toward Data Science as a Service*, 18 Computing in Science & Engineering 10–20 (2016) (https://papers.rgrossman.com/journal-056.pdf). *What We Do*, Open Commons Consortium (https://www.occ-data.org/what-we-do).

standardized.[39] The OCC has a dedicated administrative director who can devote all their time and energy to leading the operation. Further, the OCC's activities are organized into working groups, tasked with the following roles and scopes defined in the Consortium's charter: managing governance, compliance, and security; establishing a data model for its ecosystem and formats for submitted data; determining how to integrate, harmonize, and analyze data submitted to the data ecosystem; establishing infrastructure and standards for patient-partnered data; covering legal agreements; selecting research projects, and providing computing resources.

The OCC's core operating principles are that they are open source, standards-based, use open APIs, and their data are available without restrictions. They are an ecosystem of other systems, data sources, computational resources, software services, and applications. Whether the primary source is open or proprietary the data must satisfy their legal and ethical guidelines and regulatory requirements.

Research Improving People's Lives

Research Improving People's Lives ("RIPL") is a Rhode Island-based tech nonprofit that works with federal and state government agencies to "use data, science, and technology to improve policy and lives."[40] In line with its aquatic nomenclature theme, RIPL hosts Research Data Lakes ("RDLs"), which are high-security cloud data platforms that link siloed agency data for relevant policymakers.[41] RIPL says its software can automatically link and match cross-agency and even external (third-party) data before "permanently" anonymizing it, in order to provide policymakers the ability to design or redesign programs in a holistic manner. The provision of adequate, wraparound public services to constituents, based on the insights provided by the Data Lakes should "alleviate poverty and increase economic opportunity."[42]

This is a sample list of data commons and similar repositories to accelerate collaboration and research. Many other data commons and similar repositories are being developed in the social, life, and environmental sciences.

[39] *Commons Governance*, Open Commons Consortium (https://www.occ-data.org/commons-governance).
[40] *Homepage*, RIPL (2022) (https://www.ripl.org/).
[41] *Research Data Lakes*, RIPL (2022) (https://www.ripl.org/research-data-lakes/).
[42] *Id.*

Classification Systems

There are several classification systems created by organizations that aim to increase efficiency among disparate organizations and streamline data sharing in the justice system. The most prevalent system among Legal Services Providers (LSPs) is the National Subject Matter Index ("NSMI"). The National Center for State Courts ("NCSC") created a competing classification system called the National Open Court Data Standards ("NODS") that captures a wider variety of data.

National Subject Matter Index

The National Subject Matter Index ("NSMI") is a centralized, comprehensive taxonomy of Americans' legal issues that is promulgated by the Legal Services Corporation so that LSPs can index relevant topics.[43] Each category and subcategory has a specific number attached to it (e.g. "Financial Identity Theft 1090400"). One of the larger categories in the NSMI taxonomy is "consumer." "Consumer" itself has multiple subcategories, including "collection / repossession," which has multiple sub-subcategories of its own. NSMI does not provide a taxonomy for data related to outcomes, hearing, pleadings, demographics, or any other data point besides case type. While NSMI allows for granular data collection regarding case types, it does not collect data on anything else, thus severely limiting its usefulness for researchers.

National Open Court Data Standards

The National Open Court Data Standards ("NODS") seeks to develop business and technical court data standards to support the creation, sharing, and integration of court data by ensuring a clear understanding of what court data represent and how court data can be shared in a user-friendly format. NODS is a creation of NCSC, an independent, nonprofit court improvement organization. The NODS taxonomy allows for the collection of granular data on demographics, participants, case type, and judicial proceedings.[44]

Each data type has myriad subtypes. NODS allows for more comprehensive collection of data because it calls for the collection of data beyond just case types. NODS does not have a specific debt case type, but rather has it subsumed under contract, along with whether the contract is small enough to be under small claims.

Researchers find NODS more useful because it enables the collection of multiple types of data. If all LSPs used NODS and shared anonymized data from it, researchers would be able to assess countless questions. They would be able to link different data types to see what effect any input variable has on outcomes in the aggregate. As NODS is the creation of a non-governmental actor, NODS adoption is not mandatory for LSPs and courts.

National Information Exchange Model

The main classification system for criminal justice data in the United States is the National Information Exchange Model ("NIEM").[45] NIEM was created to develop reliable, reusable content to meet community needs. It provides for a common vocabulary and a standardized exchange development process to streamline efforts and promote consistency while lowering operating costs.[46] Fundamentally, NIEM is a classification system for the criminal justice system, but, crucially, it allows different systems such as local, state, tribal, and national jurisdictions to communicate even if they are using different programming languages or operating systems. NIEM also includes easily-accessible historical data.[47] The NIEM criminal

[43] *The NSMI Database*, NSMI (https://nsmi.lsntap.org/).
[44] NCSC, *National Open Court Data Standards (NODS)*, (Apr. 11, 2022) (https://www.ncsc.org/services-and-experts/areas-of-expertise/court-statistics/national-open-court-data-standards-nods).
[45] *NIEM's History*, National Information Exchange Model (https://www.niem.gov/about-niem/history).
[46] *NIEM Model*, National Information Exchange Model (https://www.niem.gov/about-niem/niem-model).
[47] *Global Justice XML (Archive)*, Bureau of Justice Assistance (https://bja.ojp.gov/program/it/national-initiatives/gjxdm).

justice standards has five main categories: subject, person, assessment, hearing, and parole. Each category has hundreds of subcategories.

NIEM is a boon to researchers. With a government-backed taxonomy of easily-accessible criminal justice data, researchers can link data in any way necessary. Hopefully, the US Government will try to create something similar for the civil justice system, but there are currently no signs that such work is underway.

Standards Advancement for the Legal Industry

The Standards Advancement for the Legal Industry ("SALI") Alliance, an independent nonprofit, aims to produce "standards for the legal industry to accelerate innovation, and improve efficiency."[48] Its members include stakeholders from across the entire legal industry.[49] SALI's main offering is the Legal Matter Specification Standard ("LMSS"), which is made up of more than 10 sets of codes "that define the 'common language' for describing different aspects of legal matters" and is structured as a database "that defines how the codes, descriptions, and values relate to each other."[50] The goal is for any member or stakeholder to use the LMSS, in part or in whole, as their metadata framework, taxonomy, and "experience database" for a variety of organizational tasks. SALI claims to be unique in the legal industry because of its independence: no member or type of stakeholder exerts control over the Alliance's operations; thus, its neutrality enables all stakeholders to have a voice at the table. It also claims to be helpful to members because of its flexibility, which allows tailored solutions to very specific organizational needs.

Learned Hands

Learned Hands is a joint project between Suffolk Law School's Legal Innovation and Technology Lab, lead by David Colarusso, and Stanford Law School's Legal Design Lab, lead by Margaret Hagan, that aims to use machine learning to create a taxonomy of legal help issues.[51] Users are presented with real-world legal issues from online sources and asked to categorize them, and the results are fed into a machine learning algorithm.[52] The project represents one of the innovative ways new technologies can be used to aid in legal data classification.

International Classification Systems

Australia

Under the *National Legal Assistance Partnership 2020-25*, legal aid commissions, community legal centers, and indigenous legal services groups across the continent must collect legal assistance data.[53] This is by far the most comprehensive civil justice data collection standards collected among these three nations. The program's manual outlines the types of data that must be collected, which goes beyond any of the recommended data collection standards provided by NODS, NSMI, and even the UK report (detailed below).[54] It includes detailed demographic data, granular legal aid service information, specific type of case, and type of court. Its extensive requirements allow researchers to stratify data at a very fine level.

---

[48] *SALI Alliance—Welcome*, SALI Alliance (https://www.sali.org/).
[49] *SALI Alliance—FAQs*, SALI Alliance (https://www.sali.org/Frequently-Asked-Questions).
[50] *Id.*
[51] *Learned Hands*, Stanford Law School (https://learnedhands.law.stanford.edu/).
[52] Bob Ambrogi, *Stanford and Suffolk Create Game to Help Drive Access to Justice*, LawSites (Oct. 16, 2018) (https://www.lawnext.com/2018/10/stanford-suffolk-create-game-help-drive-access-justice.html).
[53] *National Legal Assistance Data*, Australian Government, Attorney-General's Department (https://www.ag.gov.au/legal-system/legal-assistance-services/national-legal-assistance-data).
[54] *National Legal Assistance Data Standards Manual*, Australian Government, Attorney-General's Department (Mar. 2020) (https://www.ag.gov.au/sites/default/files/2020-03/National-Legal-Assistance-Data-Standards-Manual.pdf).

Canada

Canada collects a large amount of data from its civil justice courts. Its federal statistical website allows researchers to sort data by case type, court type, and jurisdiction.[55] All data is collected annually and published online. It currently has eight civil justice categories.[56] The government provides a downloadable, sortable spreadsheet with all civil cases.[57] Canada also provides a yearly assessment of legal aid and related data sorted by jurisdiction,[58] as well as a yearly qualitative report on the state of access to administrative justice.[59]

Although Canada has some of the most accessible data in the world, the data it collects is not as granular as NODS or the UK's 2010 report. For example, they do not collect any demographic data, a critical gap for understanding how to improve access to civil justice in Canada.

United Kingdom

The UK collects and reports data on civil justice statistics, such as all defended cases, specified and unspecified money cases, mortgage and landlord cases, total numbers of claims, and privacy injunctions.[60] They do not collect more granular data on a large scale. Like many American states, it seems that the data collected is mostly to see court expenditures and aggregate trends in total case numbers.

The UK did produce a very comprehensive, granular report on Civil Justice in England and Wales in 2010.[61] This report collected data on 11 distinct civil justice categories and types of plaintiff representation. The details within provide ample opportunity for researchers to link the data to see what effect something like race might have on outcomes. The report also collected qualitative data about people's perceptions of the civil justice system.

The judicial system in England and Wales has a 1 billion pound sterling court reform and modernization program encompassing multiple elements aiming to create a court and tribunal system that is just, proportionate, and accessible to all.[62] Since 2016, the government has pledged to maintain and improve access to justice and to measure such progress through empirical evidence while regularly reevaluating success as the data gets updated. The judiciary is working with researchers and academics to test their evaluation approach. The Legal Education Foundation has compiled a report that they hope will enable the judiciary to design inclusive services and strengthen public trust and confidence in the justice system.

United Nationals Economic Commission for Europe Classification

According to the US National Academies, starting around 2007, the United Nations Economic Commission for Europe ("UNECE") partnered with the UN Office on Drugs and Crime ("UNODC"), the European Commission, and the UN Statistical Division ("UNSD") to develop a new Europe-specific standard for the classification of crime for statistical

---

[55] *Civil Courts*, Government of Canada, Statistics Canada. (https://www150.statcan.gc.ca/n1/en/subjects/crime_and_justice/courts/civil_courts).
[56] *General Civil Court Cases, by Type of Action, Canada and Selected Provinces and Territories*, Government of Canada, Statistics Canada (Mar. 21, 2013) (https://www150.statcan.gc.ca/t1/tbl1/en/tv.action)
[57] *Civil Court Cases, by Level of Court and Type of Case, Canada and Selected Provinces and Territories—Open Government Portal*, Secretariat, T. B. of C. (https://open.canada.ca/data/dataset/5641ad22-190a-4486-8c5d-3884328a51a5)
[58] *Legal Aid in Canada: 2018-2019*, Department of Justice Canada (2020) (https://www.justice.gc.ca/eng/rp-pr/jr/aid-aide/1819/1819.pdf).
[59] S. McDonald, *Development of An Access to Justice Index for Federal Administrative Bodies*, Department of Justice Canada (2017) (https://www.justice.gc.ca/eng/rp-pr/jr/fab-eaf/fab-eaf.pdf).
[60] *Civil Justice Statistics Quarterly, England and Wales, April to June 2019 (Provisional)*, UK Ministry of Justice (Sept. 5, 2019) (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/832991/civil-justice-statistics-quarterly-Apr-Jun_FINAL_new.pdf).
[61] P. Pleasence et al, *Civil Justice in England and Wales: Report of Wave 1 of the English and Welsh Civil and Social Justice Panel Survey*, UK Legal Services Commission (2011) (http://doc.ukdataservice.ac.uk/doc/7643/mrdoc/pdf/7643_csjps_wave_one_report.pdf).
[62] N. Byrom, *Digital Justice: HMCTS Data Strategy and Delivering Access to Justice. Legal Education Foundation* (Oct. 2019) (https://research.thelegaleducationfoundation.org/wp-content/uploads/2019/09/DigitalJusticeFINAL.pdf).

purposes.[63] This project, which took several to complete, was housed in the UNSD's Conference of European Statisticians.[64] According to the UNECE, the goal was not to impose a new system upon local and national law enforcement agencies across the continent, which would have been virtually impossible to coordinate, but to create a separate and parallel system for recording data in a manner not limited by any jurisdiction's specific legal definitions.[65] With input from many nations around the world (including the United States, via USDOJ's Bureau of Justice Statistics), the task force created a framework for standardizing and classifying crimes. First, the primary unit of analysis was the event or incident itself.[66] Each incident could then be narrowed down via at least a half dozen subcategory tags, which would allow researchers to stratify datasets at any level with great detail. A hypothetical example described by the UN task force would describe a given event (a shooting) with tags recording the perpetrator (an organized crime figure), the victim (a female), the weapon (a firearm), the apparent motivation (intentional homicide), and the outcome (attempted, but failed, because the shot(s) missed the target). While this one hypothetical would be tagged as such, researchers could sort continent-wide data for all sub-tags for any given input (e.g., all events involving female victims, or firearms, or organized crime figures, or intentional homicides, or attempted murders. Presumably, more demographic and geographic information is gathered as well. After receiving international feedback and incorporating other revisions, the framework was scaled up globally and adopted by the full UNODC in 2015.[67]

[63] National Academies of Sciences, Engineering, and Medicine, *Modernizing Crime Statistics: Report 1: Defining and Classifying Crime,* National Academies Press (2016) (https://doi.org/10.17226/23492). UNODC*, International Classification of Crime for Statistical Purposes, Version 1.0.* (Mar. 2015) (https://www.unodc.org/documents/data-and-analysis/statistics/crime/ICCS/ICCS_English_2016_web.pdf).
[64] National Academies of Sciences, Engineering, and Medicine, *Modernizing Crime Statistics: Report 1: Defining and Classifying Crime,* National Academies Press (2016) (https://doi.org/10.17226/23492). *Principles and Framework for an International Classification of Crimes for Statistical Purposes: Report of the UNODC/UNECE Task Force on Crime Classification: Report to the Conference of European Statisticians*, UNODC/UNECE (Sept. 2011) (https://www.unodc.org/documents/data-and-analysis/statistics/crime/ICCS/9_UNODC-UNECE_taskforce_report.pdf).
[65] National Academies of Sciences, Engineering, and Medicine, *Modernizing Crime Statistics: Report 1: Defining and Classifying Crime,* National Academies Press (2016) (https://doi.org/10.17226/23492).
[66] *Principles and Framework for an International Classification of Crimes for Statistical Purposes: Report of the UNODC/UNECE Task Force on Crime Classification to the Conference of European Statisticians*, UNODC/UNECE (June 2012) (https://www.unodc.org/documents/data-and-analysis/statistics/crime/Report_crime_classification_2012.pdf).
[67] National Academies of Sciences, Engineering, and Medicine, *Modernizing Crime Statistics: Report 1: Defining and Classifying Crime,* National Academies Press (2016) (https://doi.org/10.17226/23492).

Linkage Approaches

Court records containing party name and address can be matched to other datasets, particularly those containing race and Hispanic origin data. Some datasets have self-reported race and Hispanic origin, like the decennial Census. In other datasets, like commercial data, demographic characteristics may be derived or modeled.

Linking court records will employ probabilistic linkage methods. Person-level linkages work when the same person is found in List A and in List B. For example, consider a civil docket is List A and the 2010 decennial Census is List B. For that to happen, we need complete, accurate identifying information for the individual to be present in both lists. If one list is incomplete, or has flawed information, the match will fail. This could occur if people in the court docket did not fill out the 2010 Census (or did not provide their full names), or if the clerk mistyped a person's last name.

For the linkage to be successful, we need to assume that court records capture complete, accurate information for case parties. Especially for defendants, we need complete first names and last names (middle names or middle initials help, too). We also need a complete residential address to inform the match. Matching on name alone does not produce the best results – too many possible matches are found for most names. Searching for a specific name residing at a specific place (at a specific point in time) produces better match rates.

Matching within the FSRDC could seek a person-place match between the docket and the Census, but that would only work for calendar year 2010. The Census Bureau has government records containing name and address data for the population for the other nine years of the decade. Through the FSRDC, the person-place data from the court records (List A) can be matched to a composite of government records (List C) that includes name and date of birth information from the Social Security Administration for every Social Security Number ever issued, along with current-year address data from sources like the Internal Revenue Service, Medicaid, and Medicare, which are used for linkages by the Census Bureau. This intermediate step allows us to add a linkage key, a unique identifier for each individual in the records, facilitating linkage to other government files.

The process in our hypothetical example above is currently used by the CJARS project to conduct linkages with the decennial Census and survey data, and with other administrative data sources for FSRDC research.

Text Extraction

Through our work, we have realized that the civil justice field has a growing need for data science applications such as text extraction and classification. With text extraction, we can pull important information from different types of documents based on our desired parameters. Text can be extracted from structured data sources such as digital text files or XML files, but they can also be extracted from unstructured data including images and PDFs via optical character recognition (OCR). In the civil justice field, we can use OCR to more easily extract data like attorney names and claims amounts from scanned court case documents. We can even mine thousands of court records to learn the mitigating factors why defendants may have been withholding rent, such as inadequate heating or illegally holding a security deposit. This allows us to see trends in the real-world context surrounding rent disputes. With text classification tools, different terms can be sorted into predefined categories – the tool will analyze the extracted text and sort based on its understanding. Extracted text that includes the words "lack/withholding of heat/electricity/water" may be sorted into a category we title "Utilities" while text including "security deposit" would sort into a "Money" category.

Conclusion

While still at an early stage of development, data commons and similar data repositories, classification schemes to collect and analyze data, data linkages, and text mining hold the promise of significantly accelerating research on the justice system. These platforms and tools can yield knowledge about the operations of courts and the value of legal services. They can also shed light on the antecedents and longer-term consequences of civil justice involvement.