JCOIN

JUSTICE COMMUNITY OPIOID INNOVATION NETWORK

JCOIN

CURATION PROCESS

University of Michigan's Inter-university Consortium for Political and Social Research (ICPSR) has long been a global leader in scientific data stewardship. Building upon these best practices, a team of experts will add data from the Justice Community Overdose Innovation Network (JCOIN) to ICPSR's National Addiction and HIV Data Archive Program (NAHDAP). The NAHDAP/ ICPSR team will then ensure secure public access to JCOIN data through the following curation steps:

STEP 1: INGEST

The NAHDAP project manager first quality-reviews a new data submission using its deposit review checklist. During this step, the project manager may request additional information from the data contributor. ICPSR supports three distinct data curation levels, based on the type and number of curatorial activities to be performed. We expect to use levels 1 and 2 for JCOIN data.

- Level 1 is ICPSR's most basic curation. As with all ICPSR data, curators conduct disclosure risk review and create a study webpage with relevant metadata, including a description, title, investigators, and notes. Level 1 curation includes a codebook and data files for all major statistical software packages.

- Level 2 curation involves additional work to improve usability. ICPSR staff make sure variables are in appropriate formats (e.g., string, numeric, dates), missing values are documented and identified, acronyms and abbreviations are spelled out, spelling is checked and corrected, and labels are checked for completeness and readability. The data are prepared for use in the online analysis interface. These additional efforts increase researchers' ability to find and use the data in secondary research.

- The most intensive curation, Level 3, may include further customizations and additional survey question text. This level also handles non-tabular or nonnumeric data, such as GIS and qualitative data.

Once deposit is complete, project managers submit the data and curation level to ICPSR's curation unit.

STEP 2: DISCLOSURE REVIEW DETERMINATION AND MITIGATION

All curation staff receive training in disclosure risk review. Curation unit supervisors and project managers are trained in disclosure risk mediation. All staff take the University of Michigan online Program for Education & Evaluation in Responsible Research and Scholarship (PEERRS) training. Any staff who will access deposited data receive mandatory disclosure review training. The curation process includes a standard data disclosure risk review. We bring unique confidentiality issues to the ICPSR Disclosure Review Board for consultation.

(Cont)

JCOIN

CURATION PROCESS

A standard disclosure risk review is performed for each deposited file to determine if data modifications are needed to safeguard research participant confidentiality. Curators use a disclosure risk assessment tool to ensure human subject protections and to document levels of risk. Disclosure risk review evaluates the data for direct and indirect identifiers that could allow a study data record to be linked to specific research participants. Certain disclosure review steps can be minimized, such as opting to mask open-ended responses instead of reviewing all the text responses and/or selecting a higher controlled-access method.

Our assessment determines whether any data items or patterns allow reidentification of persons. NAHDAP staff addresses any such confidentiality concerns in consultation with data contributors. Such steps seek to protect research subjects while maximizing the data's analytic potential. Disclosure risk might be minimized in several ways, including collapsing numeric values into categories, top- or bottom-coding outliers, or masking variables.

STEP 3: METADATA PREPARATION AND CODEBOOK

ICPSR curates and distributes data to ensure compliance with Findability, Accessibility, Interoperability, and Reusability (FAIR) principles. All documentation is provided in PDF format and encoded in the Data Documentation Initiative (DDI) standard (see below for a full explanation of this standard). Data are curated according to the OAIS model, produced by the NASA Consultative Committee for Space Systems, and supported by the International Standards Organization (ISO). Persistent identifiers are assigned to ensure access.

ICPSR has several tools to improve speed and quality of data curation. Hermes, one of the first tools ICPSR created to assist curators, takes data inputs and produces DDI-markup and dissemination outputs, including the DDI-Codebook. Curators use the ICPSR curation unit processing plan to plan work and communicate with the NAHDAP project manager (this may also include requesting additional information from the data contributor). Data curation and dissemination begin once the curator completes their review. This detailed documentation forms the basis of the quality checks.

Metadata allows users unaffiliated with the original project to understand the data and use it properly and effectively. Full metadata adds descriptive content that enables discovery and interpretation of the data, technical content regarding physical and digital features of the data resource, and structural content describing the data configurations, such as relationships to other data or resources. Metadata is created for every NAHDAP study to make the study discoverable on the NAHDAP website on its study home page. We use a structured international metadata standard — the DDI in NAHDAP workflow.

DDI is a free international standard for describing data produced by surveys and observational methods in the social, behavioral, and health sciences.

JCOIN

CURATION PROCESS

It can be used to document and manage data across the research life cycle, from conceptualization, collection, and processing to distribution, discovery, and archiving. DDI facilitates users' understanding of data and the harvesting of data by software systems and computer networks at libraries and other data repositories worldwide, as well as strengthening the interoperability of these entities. As an XML markup standard, DDI stores codebook information in a human-readable and machine-actionable form, allowing for tagging that embeds "intelligence" in the metadata. This is ideal for long-term preservation, permitting flexibility in rendering the information for display and allowing ICPSR to build tools that take advantage of detailed, variable-level information. All study and codebook metadata are prepared in DDI-XML and are used for online searching and to create study information (descriptions and related citations) and codebooks as PDF files. The advanced search capabilities of ICPSR's Social Science Variables Database (like the variable comparison tool) are based on DDI metadata.

Each JCOIN dataset will be accompanied by a codebook (produced by ICPSR or the original investigator) with variable names, variable and value labels, missing data declarations, full question text, and univariate statistics; data collection instruments; any other documentation (such as a user guide or manual with code used to create and interpret score and other derived or recoded variables), linking of datasets, construction and use of weight variables, and any other information needed to understand and analyze the data appropriately.

Relevant hard-copy documents will be converted to searchable PDF files.

STEP 4: USE OF PERSISTENT IDENTIFIERS

After release, the release system automatically assigns a Digital Object Identifier (DOI)—a persistent identifier—to that release version of the data collection. Updates to the data collection have a version number extension on the DOI. For example, the DOI for the Research on Pathways to Desistance Subject Measures data is doi:10.3886/ICPSR29961.v2. "ICPSR 29961" is the data collection's study number, and version 2 added the follow-up data to the previously released baseline data.

STEP 5: QUALITY CONTROL

NAHPAP curators use a multi-step quality control process. First, a peer or senior curator conducts a "first QC" of files using a structured form. The form prompts the curator to verify that the data were curated according to the processing plan approved by the curation supervisor and that internal curation processes were recorded. A knowledgeable senior curator then conducts a "second QC." This big-picture review focuses on data usability. Any issues uncovered during QC must be resolved before data release.

STEP 6: SELECTING AN ACCESS LEVEL

Using disclosure review reports and input from data contributors, NAHDAP will

JCOIN

CURATION PROCESS

Idetermine the release option appropriate for each study. Data release options include the following:

- Public release with agreement to the ICPSR Terms of Use
  - Download from NAHDAP website and the Helping to End Addiction Long-term Initiative, (HEAL) Data Platform
  - Online analysis using Survey Documentation and Analysis (SDA)
- Restricted release with an approved Data Use Agreement
  - Secure transfer of encrypted data with an approved data security plan to the HEAL Data Platform
  - Virtual data enclave (VDE) remote, secure access with approved DSP
  - Physical data enclave (PDE) access
- Delayed dissemination

Full public release is only warranted when there is little risk of reidentification or data have been transformed to substantially reduce that risk. Still, access to public release data requires data users to agree to ICPSR Terms of Use attesting they will not use the data for investigation of specific research subjects, will make no use of a subject's identity if discovered inadvertently, and will advise ICPSR of such discovery. Dissemination delays may allow enough time to pass for public data release. Otherwise, if the data undergoes restricted release, data users must submit requests for the data and receive approval from NAHDAP staff. NAHDAP currently uses two standard restricted data use agreements, one for data disseminated via secure download and one for data disseminated via the VDE.

Both agreements were vetted by staff with the University of Michigan Office of Research and Sponsored Projects. The restricted data use agreement for confidential data through VDE differs from the first agreement in its data security plan, under which data users do not take physical possession of the data; also, it specifies that data users receive access to their output files only after NAHDAP staff or a designee conducts a disclosure review on the output. We will develop a data security plan that includes one option for secure access to the data using the enclave that is part of the HEAL Data Platform.

STEP 7: DATA PRESERVATION

ICPSR has over six decades of experience in preserving data and adapting to rapid technological change. A key practice is ensuring the redundancy of stored data, which ICPSR achieves by employing multiple and varied methods and locations to back up its holdings. ICPSR participates in several community initiatives that focus on preserving data for coming generations. ICPSR is a CoreTrustSeal-certified data repository, demonstrating its commitment to sustainability and trustworthy data infrastructures. Certification requires rigorous peer review of compliance with digital preservation best practices, including organizational infrastructure, continuity of access, preservation, and security. CoreTrustSeal is the most widely recognized certification for data repositories. ICPSR's director of metadata and preservation is a member of the CoreTrustSeal Standards and Certification Board. ICPSR also makes use of

JCOIN

CURATION PROCESS

the DuraCloud service and Fedora, a community-maintained, open-source repository system that supports durable access to digital objects.

One product of curation workflow is the preservation of all files (e.g., original files deposited, processing history, and all resulting curated files for dissemination). ICPSR maintains five copies of its data (data, documentation, metadata, and associated materials) and requires that any off-site backup data be encrypted. Some of these copies are held locally or nearby, while others are stored in commercial cloud resources.