**Repository Architecture**

The Interstate Data Breach Report Repository (IDBRR) uses the Comprehensive Knowledge Archive Network (CKAN) repository, which is an open-source data management system used across the world to power data hubs and data portals, including many civic and open government data systems. CKAN offers a wide range of features and extensions, which allow users to upload many types and formats of data, automatically apply common metadata elements and/or create their own, view and visualize data in different formats, ensure the security of their data from deposit through download, and more. CKAN is open-source and widely used, making it a fairly sustainable option for our repository architecture.

It appears that CKAN provides some virus scanning capabilities, but we would want to ensure these and apply additional safeguards if they are insufficient to ensure the safety of deposited datasets.

**Data Deposits and Transfers**

Depositors to IDBRR will be asked to create a simple account with the repository, including their preferred name, contact information, and any institutional affiliations or personal identifiers (e.g., ORCID number) they wish to disclose). This account setup will provide initial authentication of depositors through CKAN’s features. After completing sign-up, prospective depositors will have the option to submit a dataset for upload using a structured form embedded in the IDBRR repository, which will ask for some basic metadata about the dataset to aid in constructing a metadata record for the deposit, as well as further authenticating the data prior to its publication in the repository. The deposit form will ask for the following information:

* Depositor contact information and institutional affiliation, which can be pulled from their profile automatically
* A description of each dataset submitted, to include (more information in the “Metadata Application Profile” section):
  + The dataset’s creator
  + The date of the dataset’s submission (populated automatically)
  + A title for the dataset
  + Brief description or abstract for the dataset
  + The type of entity affected by the breach (e.g., government, private industry)
  + The type of breach(es) described in the dataset
  + The state(s) covered by the breach
  + The time/timeline of the breach
  + The size of the breach, using the following “select one” categories:
    - 100 – 499 people affected (small breach)
    - 500 – 1,999 people affected (medium breach)
    - 2,000 – 9,999 people affected (large breach)
    - 10,000+ people affected (very large breach
  + The dataset’s copyright license, if any
  + The version of the dataset, including any cleaning or other edits made to the data prior to submitting it for publication
* Declarations allowing the depositor to deposit the dataset and IDBRR to publish the dataset, including confirmation that:
  + The depositor has the legal authority to share the dataset
  + The dataset does not include any confidential or otherwise personally identifying information (PII)

The IDBRR repository system will then create unique internal identifiers for each deposited dataset to support internal communication, and a curator will review each dataset to confirm that it contains what the deposit form claims it contains. After initial review and confirmation, the curator will contact the depositor via email to share this update, along with each dataset’s internal identifier so the depositor can inquire and/or track the progress of their submissions.

As stated in the “Collection Policies” section, curators will review all deposits for personally identifying information, and will require depositors of datasets that include PII of any kind, or that raise any sensitivity concerns will be asked to remove, redact, anonymize, or otherwise recode the relevant data, and resubmit it along with documentation of the modifications made.

Because IDBRR only publishes datasets that describe data breaches affecting groups of at least 100 people, and does not ask for demographic or otherwise potentially identifying information about those people (see more in the “Data Transformations” section, below, as well as the “Metadata Application Profile” section), we hope that the risk of deposited datasets including PII is minimal. However, to be safe, IDBRR curators review all submitted datasets for PII, using the [Department of Labor](https://www.dol.gov/general/ppii) definition of PII, in conjunction with the [National Association of Attorneys General](https://www.naag.org/issues/consumer-protection/consumer-protection-101/privacy/data-breaches/) definition of PII that defines a data breach, to determine what information cannot be included in a submitted dataset. Information that cannot be included in a dataset published by IDBRR includes:

* An individual’s first or last name in combination with one or more of the following pieces of information:
  + Social Security Number
  + Driver’s license number or state-issued ID card number
  + Account number, credit or debit card number, combined with any security code, access code, PIN or password needed to access an account
  + Medical history or health information
  + Biometric information
  + Email address and password
  + Tax ID number
* Contact information of any kind
* Demographic information that could allow for the identification of a relatively small group of people who could be individually identified (see examples in the next bullet)
* Any information that allows a user of the dataset to directly or indirectly infer the identity of an individual affected by a breach (e.g., a small breach affecting a specific organization with known members or attendees, such as a local religious institution or school with a published list of students, faculty, and/or staff)

Once datasets are confirmed to not include any PII, an IDBRR curator will obtain a persistent, universally unique identifier for each dataset from the [DataCite DOI service](https://datacite.org/create-dois/). Additional metadata will be created and added to the information submitted by the depositor using the automatic form, to align with the requirements set forth in the “Metadata Application Profile” section.

Finally, prior to publication, the IDBRR curator will contact the depositor, asking them to:

* Affirm the accuracy of the information provided during dataset submission
* Review the deposited (and potentially transformed, see more below) dataset and confirm its accuracy
* Grant permission to publish the dataset under the relevant license (see the “Data Licensing” section for more)
* Answer any questions the IDBRR team might have about the data prior to its publication

**Data Transformations**

*File / Format*

In order to ensure the preservation of published datasets and their widest possibly use by end users, datasets accepted of the previously listed types (see the “Data Collection Policies” section) will be converted into open, machine-readable formats to be stored in the IDBRR repository as well as distributed to end users.

* Tabular data will be stored as flat CSV files
* Text data will be stored as flat TXT files
* Geospatial data, where accepted, will be stored in a file determined on a case-by-case basis, but likely as GeoJSON

Tabular data stored in text files, PDF files, or other formats will be extracted and saved as CSV files, where possible, using open source tools such as Tabula. Where submitted data cannot be converted into one of the above types without significant loss of information or functionality (e.g., underlying formulas in an Excel document, special characters or images in a PDF, charts that cannot be converted into tabular format), the original file will be published in the IDBRR repository. As indicated in the “Data Collection Policies” section, IDBRR will accept data submitted in a proprietary format on a case-by-case basis, but in most cases strongly discourages submissions of this type. Where proprietary data formats are published, a “README” file will accompany that file, submitted by the depositor, documenting the name, version, and manufacturer of the software required to access the proprietary file, as well as suggested analyses or means to view the data and draw inferences from it.

In all cases, originally submitted files will accompany converted, preferred files in published datasets, to help users understand how a dataset was transformed by IDBRR, and to allow users maximum flexibility in how they engage with the data in the repository. For example, if data were submitted as an Excel file and converted into a CSV, IDBRR will publish both files, but will indicate a preference for the CSV file as the curated file which the repository has invested resources in preserving. Additionally, dataset metadata will include information about the data’s version history, indicating when the file was last updated. If necessary, a small free text field will allow an IDBRR curator to provide additional explanation about what changes were made to the data during the transformation phase.

*Data Values*

IDBRR curators take a number of steps to “tidy” and normalize tabular data, relying on the open source [OpenRefine tool](https://openrefine.org/) to complete transformations and track changes for publication along with the curated data (see the “Curated Datasets” section for examples). [The University of Illinois Library](https://guides.library.illinois.edu/openrefine/commontransform) suggests a number of common transformations of “messy” data at the cell level that curators can perform using OpenRefine ([as does OpenRefine itself](https://openrefine.org/docs/manual/cellediting)), which IDBRR curators follow for all datasets submitted to the repository. These include:

* Trim leading and trailing whitespace
* Collapse consecutive whitespace
* Unescape HTML characters (e.g., “&” vs. “&amp;amp;”
* Remove special characters
* Replace smart quotes with ASCII
* Perform necessary case transformations (e.g., upper, lower, title) as stipulated by the data dictionary, below
* Remove blank rows and columns
* Split multi-value cells into multiple columns
* Standardize missing values as blank cells
* Standardize dates, addresses, and geospatial coordinates, following the guidelines in the data dictionary, below
* Transforming tabular data into “tidy data,” following those principles (i.e., each variable is a column, each observation is a row)

Because CSV and TXT files do not store information related to data types (e.g., Boolean, string, number, etc.), IDBRR curators will not perform data-type transformations.

Additionally, to promote the integration and analysis of data about data breaches across states and across datasets, IDBRR endeavors to normalize the variable names and standardize values included in all datasets. IDBRR has created the below dictionary of preferred data values, which serve as a baseline and benchmark for published datasets (though depositors may expand upon the available elements as needed). All datasets should endeavor to include the following values, where possible and applicable, and IDBRR curators will ensure that such values are standardized to their preferred form. Submitted datasets that do not include all of the below elements will be curated to include as many of the below elements as possible.

IDBRR reviewed a number of states’ data breach disclosures and datasets (where available), and developed the below data dictionary based on this review. Washington State offered the most extensive, robust, and detailed disclosure data available in its open data hub, and IDBRR has largely followed that state’s reporting standards to design the data dictionary ([see example, here](https://data.wa.gov/Consumer-Protection/Data-Breach-Notifications-Affecting-Washington-Res/sb4j-ca4h/about_data)).

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name (Column Title)** | **Description** | **Data Type** | **Encoding Scheme** |
| DateAware | The date the breached entity became aware that users’ data had been compromised | Date | Date format: YYYY-MM-DD |
| DateSubmitted | The date the breached entity notified a relevant authority (e.g., Attorney General’s office) | Date | Date format: YYYY-MM-DD |
| BreachCause | The cause of the data breach | String; controlled vocabulary | Possible entries:   * Cyberattack: when a third party aggressively attempts to access private data |
| BreachedEntity | The owner of the data that was breached, whether or not this was the entity that was breached | String | N/A |
| BreachedEntityLongitude | The longitudinal value of the breached entity, where applicable | Number | Longitude value |
| BreachedEntityLatitude | The latitudinal value of the breached entity, where applicable | Number | Latitudinal value |
| BreachedEntityStreet | The street address of the breached entity | String | Street address |
| BreachedEntityCity | The city of the breached entity | String | City |
| BreachedEntityState | The two string state code [according to the FAA](https://www.faa.gov/air_traffic/publications/atpubs/cnt_html/appendix_a.html) | String | State code |
| BreachedEntityZip | The zip code of the breached entity | Number | Five-digit zip code |
| CyberattackType | The type of cyberattack that caused the data breach | Controlled Vocabulary | * Malware * Ransomware * Phishing * Skimmers * Other * Unknown |
| StateAffected | The state in which the people whose data was breached reside | String | Two-letter state codes ([see here](https://www.faa.gov/air_traffic/publications/atpubs/cnt_html/appendix_a.html)) |
| ResidentsAffected | The number of state residents affected | Number | N/A |
| IndustryType | The industry type that the breached entity belongs to | String | * Business * Education * Finance * Government * Health * Nonprofit or Charity |
| DaysToContain | The number of days it took the breached entity to end the exposure of consumer data | Number | N/A |