**Repository Description**

The Interstate Data Breach Report Repository (IDBRR) curates, preserves, and publishes quantitative and qualitative datasets describing data breaches affecting people living in the United States. Anyone is welcome to deposit data to the IDBRR.

According to the [National Association of Attorneys General](https://www.naag.org/issues/consumer-protection/consumer-protection-101/privacy/data-breaches/) “[a]ll 50 states, the District of Columbia, Guam, Puerto Rico, and the U.S. Virgin Islands have established data breach laws to protect consumers. These laws generally require organizations to notify individuals in the case of a data breach involving certain personal identifying information.” These laws [can be found on this website](https://www.ncsl.org/technology-and-communication/security-breach-notification-laws), which links to each state’s relevant laws for private and public sector entities. While all of the states reviewed for this protocol require some disclosure from breached entities, and publish those disclosures in some way, the format and accessibility of each disclosure differs widely by state. Some states (e.g., Washington) provide relatively easy-to-access public data portals that allow users to

**Data Breach Definition**

IDBRR follows the definition of “data breach” set forth by the [National Association of Attorneys General](https://www.naag.org/issues/consumer-protection/consumer-protection-101/privacy/data-breaches/), which states that:

*A data breach can be defined as the unlawful and unauthorized acquisition of personal information that compromises the security, confidentiality, or integrity of personal information. What is considered personal information depends on state law but typically includes an individual’s first name (or initial) and last name plus one or more of the following:*

* *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*

*Additional categories may include:*

* *Medical history or health information*
* *Biometric information*
* *Email address and password*
* *Tax ID number*

**Deposit Overview**

A complete deposit must include three components, all of which are described in further detail below:

1. Data file(s)
2. Documentation file(s)
3. Dataset description (i.e., metadata)

**Data File(s)**

*Accepted Datasets*

Depositors must include at least one digital data file that includes information related to a data breach, following the above definition, that affected people currently residing in the U.S. All datasets must also follow IDBRR policies related to copyright, ethics, and preservation described in further detail later in this document.

More information on the desired format and content of deposited datasets can be found in the “Data Transformations” section of this protocol. In most cases, deposited data will take the form of tabular or text data describing a data breach affecting a single entity (e.g., company, government agency, nonprofit organization) along with some or all of the following information:

* When the breach occurred
* What states’ residents were affected by the breach
* How many people were affected
* What type(s) of information were compromised in the breach
* What type of cyberattack or other method caused the breach
* The malicious actor(s) that caused the breach, if known
* Information related to the affected entity’s response to the breach (e.g., when did the entity identify the breach, is the breach resolved, etc.)

*Datasets Not Currently Accepted*

It should go without saying that deposited datasets should not include any personally identifying information (PII), including any of the information types included in the above definition of a data breach. Accepted datasets are only those that describe such breaches, not breaches themselves. 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.

To protect the anonymity of people affected by data breaches, IDBRR currently accepts any data files describing data breaches affecting at least 100 people. Data files that describe smaller data breaches affecting fewer than 100 people, those describing data breaches affecting people not currently residing in the U.S., and/or those describing security incidents in which no individuals’ personal information was compromised (e.g., hackers broke into a company’s secure databases and installed ransomware in order to extract payment from the company) are not currently accepted. Additionally, files in an analog format (e.g., paper records) cannot currently be accepted.

*File Naming Conventions*

To ensure that IDBRR can consistently organize deposited datasets, depositors should use the following standard convention when naming their files:

*Breached Organization\_Date of Breach\_ State(s) Affected\_Number of People Affected\_Depositor Initials*

1. No spaces should be used, and “camel case” should be used where multiple words need to be written together (e.g., “MultipleWordsWrittenTogether”).
2. The *Breached Organization* should be the official legal name of the company, organization, or other entity whose data was compromised in the breach.
3. *Date of Breach* should follow the ISO-recommended YYYYMMDD format.
4. *State(s) Affected* should include the two letter state code for each state affected by the data breach. Where more than three states have been affected, write “multiple” instead.
5. Include the total *Number of People Affected* across all states in the breach.
6. Append the *Depositor Initials* to the end of the file name, indicating who has deposited the dataset
7. Underscore (\_) should be used as the transition between each element in the file name.

*Acceptable Data Formats (Adapted from the Qualitative Data Repository at Syracuse University)*

The Qualitative Data Repository describes a file format as “a specific way of structuring information for storage in a digital medium so that it is ‘understandable’ in a machine context. Formats should be readable by as many types of systems as possible (including human readable) without compromising the purpose of the data.” The submission formats IDBRR recommends help ensure that data are stored in the most appropriate ways for effective digital display, long-term archiving, and prevention of technological obsolescence.

Research and experience from Cornell’s eCommons repository have shown that the likelihood of successful long-term preservation of data is much higher when file formats possess the following characteristics (as do IDBRR’s recommended formats):

* Complete and open documentation
* Platform-independence
* Non-proprietary (vendor-independent)
* No "lossy" or proprietary compression (i.e., no processes enacted on the data that remove or degrade aspects of the data critical to its understanding and/or use)
* No embedded files, programs or scripts
* No full or partial encryption
* No password protection

If you possess data in a unique, proprietary, or otherwise non-recommended format not described below, please reach out to IDBRR to determine if we can accept your deposit.

|  |  |
| --- | --- |
| Data Type | Recommended Formats (adapted from the Qualitative Data Repository) |
| Tabular Data | * For data with rich metadata: Stata, SPSS, and R data files (.dta, .sav, .RData) * For data with little metadata: Comma or tab separated values (.csv or .tab), newer versions of Microsoft Excel files (i.e., .xlsx, 2003 or later) |
| Text | * PDF/A, Rich Text Format (.rtf) * Plain text/html (.txt, .html) * Newer versions of Microsoft Word files (i.e., .docx, 2003 or later) |
| Geospatial Data | Data depositors may wish to include data visualizing/mapping the physical location of data breaches, though IDBRR prefers that geographic data be included in text or tabular formats. Depositors wishing to include geospatial data may use the following formats:   * ESRI Shapefiles (essential - .shp, .shx, .dbf, optional - .prj, .sbx, .sbn), Geo-referenced TIFF (GeoTIFF, .tif) |

As part of IDBRR’s mission to ensure sustainable access to its deposits, repository staff will also periodically monitor submitted datasets for possible obsolescence of formats. Obsolescence will be evaluated by comparing submitted data formats with common software available for free to the public, as well as by format-related support requests to the site. If IDBRR decides to change the format of a file, the original depositor – if available – will be consulted. However, as part of the original deposit process all depositors grant IDBRR the right to modify their data files in accordance with this policy.

*Data Deposit Size Limits (Adapted from the Cornell eCommons Repository)*

Individual files deposited in IDBRR should not exceed 5 GB in size. Note that most web browsers have a file size upload limit, which may be less than 5GB. Growth in the total size of a collection of files associated with a particular data breach and deposited into IDBRR should not exceed 50 GB per year. If you have questions about these size limitations, please contact the IDBRR administrator.

**Documentation File(s) (Adapted from ICPSR guidelines)**

Documentation files are integral to reuse of a data collection and should thoroughly explain the data collection. Examples of documentation files include codebooks, copies of data collection instruments, user guides, README files, summary statistics, and bibliographies of publications using the data. Documentation can be submitted as Microsoft Word, ASCII, or DDI XML files, among other formats. User guides and README files should include, at a minimum, a manifest that describes what files are in the submission, an MD5 checksum value, information about how the data were collected, an explanation of the variables and/or elements included in the dataset, and an explanation of non-obvious values in the dataset (if necessary).

Documentation files should follow the file naming conventions described above, with the addition of “\_DOCUMENTATION” at the end of the filename.

**Dataset Description (i.e., Metadata) (Adapted from ICPSR guidelines)**

IDBRR users are able to discover, understand, and analyze data because every new or revised IDBRR dataset includes descriptive information (metadata) such as the name of the breached organization, the date or timeframe of the breach, the type of attack that caused the breach, etc.

Please see the “Metadata Application Profile” section of this protocol for more information on creating robust and effective metadata for your dataset deposit.

**Data Ethics (Adapted from Cornell eCommons and the QDR)**

Submitting work to IDBRR requires depositors to attest that the work contains no confidential or proprietary information. Confidential information includes data that can uniquely identify someone, such as names, addresses, or a Social Security number, credit card number, or driver's license number. Proprietary information is information, such as patentable information, that is owned, copyrighted, or may be owned or copyrighted, by someone else.

All depositors are required to adhere to ethical and legal constraints when sharing data. If you have collected data that are confidential and/or sensitive, extra attention will be necessary to ensure that they are shared ethically and legally. IDBRR offers detailed guidance on managing and sharing sensitive data that applies to various stages of the data collection and deposit process:

* During your data collection, make sure to manage and handle data responsibly and keep all data, and especially sensitive data, secure.
* As you prepare your data for sharing, make sure you follow best practices in de-identifying data.
* As you deposit your data, talk with IDBRR about options for setting access controls that limit who can view and download some or all of the data.

**Copyright (Adapted from QDR and the Cornell University eCommons Repository)**

The copyright status of one or more elements of a data deposit may pose challenges to sharing them via IDBRR. Because several digital copies might have to be produced for preservation and administration purposes, and because dissemination to a larger group of users is often intended, typical copyright protections for research uses might not apply to IDBRR deposits.

IDBRR’s curators can help depositors to find legal ways to provide access to data that are under copyright. For instance, we may be able to aid you in obtaining necessary licenses and permissions to publish the materials in question. Alternatively, you might provide a detailed data listing (e.g., full bibliographical information) for data files that cannot be shared.

In any case, the author/owner of data deposited in IDBRR must be willing and able to grant IDBRR the right to preserve and distribute the work via our repository.

**Data Preservation Guarantees (Adapted from the Cornell University eCommons Repository)**

IDBRR is committed to responsible and sustainable management of data deposited in our repository, and to ensuring long-term access to those data. All data deposited in IDBRR will be assigned a persistent identifier in the form of a DOI obtained by [DataCite](https://datacite.org/), as well as a persistent Web address (URL). While prudent efforts will be taken to preserve work in any digital format submitted to IDBRR, submission in a recommended file format (described earlier) is strongly encouraged in order to facilitate long-term preservation.

At this time, IDBRR is committed to preserving the binary form of the digital object. Further practical measures to preserve as much functionality (i.e., "look and feel") of the original content as possible will be taken as resources permit.

Digital preservation is an evolving field. Current long-term preservation strategies and technologies employed by IDBRR are shaped by the Open Archival Information System (OAIS) reference model (ISO 14721:2012) and informed by relevant international standards and emerging best practices. IDBRR preservation activities and policies will be reviewed regularly to ensure that they remain current as technology and institutional practices evolve.