Data Breach Chronology

Overview

The Data Breach Chronology is a dataset of data breach notifications reported to state and federal agencies across the United States. Since 2005, Privacy Rights Clearinghouse has tracked these breaches to help understand their impact on Americans' privacy and security.

This dataset brings together breach notifications from government agencies and the detailed notification letters that organizations send to affected individuals. By analyzing both the structured agency data and the full text of notification letters, we provide a detailed picture of how breaches occur, what information is exposed, and how organizations respond.

The data includes breaches across many sectors - from small local businesses to major corporations, from K-12 schools to universities, and from local medical practices to national healthcare systems. We analyze each breach across multiple dimensions, including organization type, breach method, affected information types, geographic location, and impact scale. Through a combination of human expertise and machine learning techniques, we extract and structure information previously locked away in tens of thousands of notification letters or scattered across more than a dozen government registries, enabling researchers and analysts to explore the complex patterns of data breaches in the United States.

Table of Contents

Changelog

[2.0.0] - January 2025

Overview of Changes

Version 2.0 represents a complete overhaul of the Data Breach Chronology. The project was rebuilt from the ground up, presenting an opportunity to address longstanding limitations, improve data collection, and refine our processing methodologies. This update enhances every aspect of the database, from scraping to classification, schema design, and data access. Please note our changes below and refer to the documentation throughout this Readme for more information.

Major Updates

Scraping Infrastructure

- Developed custom scrapers for all 15 state and federal sources.
- Implemented automated monitoring to quickly adapt to changes in source websites.
- Integrated PDF downloading and archiving directly into the scraping pipeline.
- Standardized data collection across diverse formats, including APIs, dynamic JavaScript, and paginated HTML.
- Identified and captured data previously missed, effectively doubling the total number of notifications in the database.

Enhanced Processing Pipeline

- Updated AI processing models to improve classification accuracy and extraction capabilities.
- Refined AI prompts to reduce hallucinations and ensure consistent handling of nuanced details.
- Conducted a manual review of thousands of records to validate outputs and refine methodologies.
- Implemented a validation pipeline to detect inconsistencies, duplicates, and anomalies across all stages.

Improved Update Frequency

- Transitioned to a framework capable of monthly, weekly, and eventually daily updates.
- Established workflows to ensure regular and consistent updates moving forward.

Database Improvements

Schema Enhancements

- Introduced group UUIDs and related fields to consolidate related breaches across multiple reports.
- Added normalized_org_name and, group identifiers to improve tracking organizations and breach events across the database.
- Improved organization name handling with acceptable_names for aliases and DBAs, supported by detailed reasoning fields (org_name_explanation).
- Restructured the information affected fields for more precise and consistent tracking of exposed data types.

Advanced Deduplication and Grouping Logic

- Enhanced grouping algorithms to better identify related breaches
- Improved deduplication logic to reduce overcounting of individuals impacted across grouped incidents.

User-Facing Improvements

- Responding to feedback, we've transitioned to providing SQLite .db as the recommended full-database download, offering greater compatibility and usability.
- Updated documentation across the board.

[1.5.1] - May 2024

- Released a minor update to address specific data errors in breach notifications.
- Improved AI processing accuracy for particular edge cases.
- Continued full-scale development of Version 2.0 infrastructure.

[1.5.0] - January 2024

- Transitioned fully to an open data model, sourcing all information exclusively from government disclosures.
- Shifted to developing internal scraping infrastructure following the sunset of the Data Breach Archive project in December 2023.
- Included historical records scraped from government sources, eliminating dependence on third-party amalgamation.
- Marked a decisive shift in project philosophy to focus solely on government-sourced, publicly disclosed data.

[1.3.0] - January 2023

- Introduced AI-driven methods to clean and normalize data, leveraging automation to handle the growing volume of breach notifications.

- Identified third party source of breach notifications so shifted to partnership with the Data Breach Archive, at the time a comprehensive single source of scraped data breach notifications from state registries.
- Combined AI processing with our existing data to create a structured and searchable dataset.
- Highlighted flaws in merging data from multiple sources but laid the groundwork for improvements in subsequent releases.

[1.2.0] - 2018-2019

- Began experimental automation efforts, including early scraper development, to process and clean data. Encountered challenges in scaling due to the volume of breaches and the varied quality of source data.
- Froze the historical database at the end of the year, creating the historical archive as a snapshot of all breach data through 2019.

[1.1.0] - 2010

- Developed the first formal taxonomy for classifying breaches, informed by consultations with data scientists.
- Transitioned from a managed text list to a structured database, introducing a taxonomy for breach type and organization type.

[1.0.0] - 2005

- Launched the Data Breach Chronology as a manually maintained text list on the Privacy Rights Clearinghouse website.
- Focused on manually curated entries documenting breaches reported by U.S. organizations, government agencies, and media outlets.
- Operated without a formal taxonomy or structured database, relying on narrative descriptions.

Data Sources

The Data Breach Chronology draws from fifteen U.S. government agencies that maintain public records of data breach notifications. Each agency has its own reporting requirements and methods of sharing information.

Federal Sources

- U.S. Department of Health and Human Services

State Sources

- California Office of the Attorney General
- Delaware Department of Justice
- Indiana Office of the Attorney General
- Iowa Attorney General's Office
- Maine Office of the Attorney General
- Maryland Office of the Attorney General
- Massachusetts Office of Consumer Affairs and Business Regulation
- Montana Department of Justice
- New Hampshire Department of Justice
- Oregon Department of Justice
- Texas Office of the Attorney General
- Vermont Office of the Attorney General
- Washington State Office of the Attorney General
- Wisconsin Department of Justice

Each state has unique reporting thresholds and requirements, leading to variations in what information is available. For example, some states require reporting of any breach affecting state residents, while others set minimum thresholds. Some states make notification letters public, while others provide only summary data.

When a breach affects residents of multiple states, it may be reported to several agencies. We track these related reports and group them together to provide a more complete picture of each incident.

Data Fields

The Data Breach Chronology database contains the following fields, organized by their function:

Core Identification Fields

Field Name	Description	Example

Field Name	Description	Example
id	Unique identifier for each record	`"550e8400-e29b-41d4-a716-446655440000"`
source	The government agency that reported the breach	`"California Office of the Attorney General"`
org_name	Primary name of the breached organization	`"Global Tech Solutions, Inc."`
acceptable_names	Alternative names, including DBAs and common variants	`"GTS, Global Tech, Global Tech Solutions"`
org_name_explanation	Details about how the organization name was determined	`"The California Attorney General reported the organization as 'Global Tech Solutions Inc'. The notification letter refers to the company as both 'GTS' and 'Global Tech Solutions, Inc.' The latter appears to be the full legal name based on its use in the formal letterhead."`
organization_type	Classification of organization (BSF, BSO, BSR, etc.)	`"BSO"`
organization_type_explanation	Reasoning behind the organization classification	`"While the organization handles healthcare data, it is explicitly described as 'a healthcare software and technology services provider,' making it a technology company (BSO) rather than a healthcare provider."`

Incident Information

Field Name	Description	Example
incident_details	Summary of the breach event	`"On April 2, 2024, ABC Healthcare reported to the Maryland Office of the Attorney General that it experienced a`

Field Name	Description	Example
		data breach potentially exposing patient information. The breach, discovered on March 20, 2024, involved unauthorized access to an employee email account."
breach_type	Method or nature of the breach (HACK, CARD, PHYS, etc.)	"HACK"
breach_type_explanation	Reasoning for the breach classification	"Classification as HACK is based on explicit description of 'unauthorized access to our network through a malware attack.' The technical details provided clearly indicate external cyber compromise rather than accidental disclosure or physical theft."
information_affected	Structured data about types of information compromised	[Complex JSON structure detailing affected information types]
information_affected_explanation	Details about how affected information was determined	"The notification letter specifically lists exposed data including names, Social Security numbers, and medical records. The data was confirmed to be unencrypted based on explicit statements in the letter."

Dates and Timeline:

Field Name	Description	Example
reported_date	When the breach was reported to the agency	"2024-03-15"
breach_date	When the breach occurred or was discovered	"2024-01-15"
end_breach_date	When the breach	"2024-01-31"

Field Name	Description	Example
	ended (if applicable)	
date_info_explanation	Details about date determinations	"The California Attorney General reported this breach on 2024-02-15. The notification letter details an unauthorized access period that began January 1, 2024 and continued until January 31, 2024, when the intrusion was detected and contained."

Impact Metrics

Field Name	Description	Example
total_affected	Total number of individuals impacted	"15000"
residents_affected	Number of state residents affected	"5000"
impact_info_explanation	Details about how numbers were determined	"The Maine Attorney General reported that 5,000 Maine residents were affected. The notification letter states that 15,000 individuals were impacted across all states."

Location Information

Field Name	Description	Example
breach_location_street	Street address where breach occurred	"5550 Peachtree Parkway, Suite 500"
breach_location_city	City where breach occurred	"Peachtree Corners"
breach_location_state	State where breach occurred	"GA"
breach_location_zip	ZIP code where breach occurred	"30092"
breach_location_country	Country where breach occurred	"United States"
breach_location_explanation	Details about location determination	"The breach location was explicitly stated in the provided breach notification letter."

Group/Relationship Fields

Field Name	Description	Example
group_uuid	Identifier for related breaches	`"a123e4567-e89b-12d3-a456-426614174000"`
normalized_org_name	Normalized organization name for the group	`"Global Healthcare Systems"`
normalized_org_name_explanation	Explanation of group name standardization	`"Multiple variations of the organization name appear across state reports. This standardized name represents the most complete legal name found in notification letters."`
group_org_breach_type	Primary breach type for the group	`"HACK"`
group_org_breach_type_explanation	Explanation of group breach classification	`"All related notifications describe the same ransomware incident."`
group_org_type	Primary organization type for the group	`"MED"`
group_org_type_explanation	Explanation of group organization classification	`"Consistently identified as a healthcare provider across all notifications."`

Source Documentation

Field Name	Description	Example
source_url	URL of the government agency's report	`"https://oag.ca.gov/privacy/databreach/list"`
notification_url_original	Original URL of the breach notification	`"https://example.com/breach-notice-2024-001.pdf"`
pdf_contents_cleaned	Processed text from notification	[Full text of cleaned notification letter]

Field Name	Description	Example
	letter	

Metadata

Field Name	Description	Example
updated_at	Last modification timestamp	`"2024-03-16 09:15:00"`

Understanding Classifications

Organization Types

Organizations in the database are classified using the information available in a particular breach notification into specific categories based on their primary function:

- **BSF** (Financial Services Business): Banks, credit unions, investment firms, insurance carriers (excluding health insurance)
- **BSO** (Other Business): Technology companies, manufacturers, utilities, professional services
- **BSR** (Retail Business): Physical and online retail merchants
- **EDU** (Educational Institutions): Schools, universities, educational services
- **GOV** (Government and Military): Public administration, government agencies
- **MED** (Healthcare Providers): Hospitals, clinics, HIPAA-covered entities
- **NGO** (Nonprofits): Charities, advocacy groups, religious organizations
- **UNKN**: Used when insufficient information is available

Breach Types

Similarly, each incident is classified based on the primary method of breach, again based solely on the information available in the breach notification:

- **CARD**: Physical payment card compromises (skimming devices, POS tampering)
- **HACK**: External cyber attacks (malware, ransomware, network intrusions)
- **INSD**: Internal threats from authorized users
- **PHYS**: Physical document theft or loss
- **PORT**: Portable device breaches (laptops, phones, tablets)
- **STAT**: Stationary device breaches (desktops, servers)
- **DISC**: Unintended disclosures (misconfiguration, accidents)
- **UNKN**: Used when the breach method cannot be determined

Understanding Breach Groups and Normalized Organization Name

The Data Breach Chronology includes two separate grouping systems to help track both organizations and individual breach events across the database.

The “normalized_org_name” field is a human readable, single organization identifier that can be used to track an organization across all breaches. Because there is a small risk that this normalization is incorrect, we also include the `org_name` and `acceptable_names` fields as well, with explanations for how those were determined.

When an incident affects multiple states, organizations often report to several state agencies. Additionally, organizations may file updated reports as more information becomes available.

We’ve created the various “group_*” fields to help identify singular breach events that were subsequently reported to multiple registries or had updating information provided. We identify potential relationships between breaches based on:

- Matching Normalized_org_name, and
- Similar reporting dates (within a 90-day window)

Please note that this grouping is based on available information and may not capture all related breaches. Some incidents that are actually related may not be grouped, while others may be grouped despite being separate incidents. The grouping information is provided to help users understand potential relationships between breach reports, but should not be considered definitive.

Working with the Data

The Data Breach Chronology is available in two formats, each designed to serve different research needs:

SQLite Database

The complete dataset is provided as a single SQLite database file, preserving every detail of our data breach records. This format is ideal for researchers performing complex analysis or working with the full text of breach notifications.

SQLite’s universal format means you can explore the data in whatever way works best for you: - Browse and query visually using DB Browser for SQLite, a free and open-source tool - Analyze programmatically using languages like Python or R - Import into other database systems

Here’s a simple example using Python:

```
import sqlite3
import pandas as pd

with sqlite3.connect('databreach_chronology.db') as conn:
  # Find breaches affecting more than 100,000 individuals
  large_breaches = pd.read_sql_query('''
```

```
    SELECT
      org_name,
      reported_date,
      breach_type,
      total_affected
    FROM final_breach_notifications
    WHERE CAST(total_affected AS INTEGER) > 100000
    ORDER BY reported_date DESC
''', conn)
```

Excel Spreadsheet and CSV export

For immediate access to breach records, we also provide an Excel version of the dataset as well as a CSV file.

Important: To prevent issues when importing into Excel, **the pdf_contents_cleaned field is omitted** from both the Excel and CSV exports. Despite our best efforts—such as truncating long text fields or splitting data across multiple sheets—this field was frequently causing problems. Excel has strict limits on the number of characters allowed per cell (32,767 characters) and a maximum number of URLs per sheet; including the full text often led to workbook corruption or import errors.

The CSV export now uses a pipe (|) as the column separator and double quotes (") as text qualifiers, ensuring that commas within text do not break the column alignment.

To open the CSV in Excel:

1. Go to **Data > From Text/CSV**.
2. Choose the file and set the delimiter to the pipe character (|).
3. Ensure that text qualifiers are set to double quotes (").

We strongly recommend working with the SQLite (.db) export if you need access to the full text of breach notifications. The Excel and CSV formats have been optimized for quick access and ease of use, but they omit some fields to ensure data integrity and compatibility with Excel.

Data Quality and Accuracy

Data is sourced from public, government-maintained sources, and the database should not be considered a complete record of all data breaches in the United States. It reflects reported breaches made publicly available.

The Data Breach Chronology is provided "as is." While we employ multiple validation steps and integrity checks, the complex nature of breach notifications and our processing methods means there may be inaccuracies in the data. We actively work to minimize these through automated and manual review processes, but they cannot be entirely eliminated.

If you identify any inaccuracies in the database, please contact us at databreachcorrections@privacyrights.org. Your feedback helps us maintain and improve the quality of this resource.

Acknowledgments

The Data Breach Chronology began in 2005 under the leadership of Beth Givens, Privacy Rights Clearinghouse's founder and former Executive Director. The current version was developed and is maintained by Emory Roane, Associate Director of Policy at Privacy Rights Clearinghouse.

We are also thankful for the contributions of The Rose Foundation for Communities and the Environment, Consumer Federation of America, Coleman Research Lab, Ahmed Eissa, Ava Watson,, and everyone else who has supported the project in its various forms over the years.

Licensing and Usage Terms

The Data Breach Chronology Database and the DataBreachChronology_Archive are subject to specific licensing terms outlined in the accompanying TermsofService.pdf file. Users are advised to review these terms to understand any restrictions or permissions regarding the use of this data. The licensing terms are designed to protect the integrity of the data and ensure its appropriate use in research, policymaking, and advocacy.

For any licensing-related queries or permissions, please contact `databreachchronology@privacyrights.org`.

Contact Information

For questions about the data or to report corrections: `databreachcorrections@privacyrights.org` For licensing and permissions: `databreachchronology@privacyrights.org`