REMOVING PII FROM THE EDRM ENRON DATA SET

Investigating the prevalence of unsecured financial, health and personally identifiable information in corporate data

The EDRM Enron data set is an industry-standard collection of email data that the legal profession has used for many years for electronic discovery training and testing. Since this data set was published, it has been an open secret that it contained many instances of private, health and financial data. Nuix volunteered to investigate the EDRM Enron data set and remove as much of this personal information as possible before republishing a cleansed version of the data. The results of our investigation present food for thought about the prevalence of private data in all corporate data sets and the serious business risks it represents.

Under a thick blanket of privacy legislation in almost all Western countries, organizations must take extreme care to protect any personally identifiable information (PII) and personal health information (PHI) they store relating to employees or customers.

For example, the European Commission's proposed General Data Protection Regulation will impose fines of up to 2% of a company's annual global turnover for failure to protect consumers' private information.[1] This regulation, set to be adopted in 2014, will apply to businesses that operate in the EU or that hold personal information of EU citizens.

Similarly, the United States Department of Commerce's guide to protecting PII for federal government agencies details a long list of "operational safeguards, privacy-specific safeguards and security controls."[2]

Organizations that accept credit card payments must also comply with the Payment Card Industry Security Standards (commonly referred to as "PCI") imposed by the credit card companies. Under these guidelines, organizations may only store limited types of data about credit card holders and must use encryption to render this information "unreadable" wherever the organization keeps it.[3] Failure to comply with PCI standards can result in that organization losing its ability to process credit card payments.

While these laws are well known, very few organizations thoroughly comply with these regulations. Employees often make "convenience copies" and store such information without encryption in file shares and collaboration systems or send it outside the organization in emails.

E·D·R·M

nuix

Simple. Powerful. Precise.

THE SOURCE DATA AT A GLANCE

For the purposes of this exercise, Nuix used the EDRM Enron PST Data Set, which comprises:

- 1.3 million email messages and attachments from former Enron staff
- 168 Microsoft Outlook .PST files
- Almost 40 GB of data.

ORGANIZATIONS CANNOT IGNORE THE RISKS

Although inappropriately stored private, health or financial data are a serious business risk, many organizations do not take steps to address these issues based on two assumptions:

- "We don't have to worry unless our systems are hacked." Although we improperly store this information, it can't find its way outside the firewall unless there is a security breach.
- "The information is there, but no one can find it." We have masses of unstructured data and it would be virtually impossible, or too resource intensive, to trawl through millions of emails and files for privacy breaches. It would be equally hard for anyone else to find the private information stored in our systems.

Both assumptions are false.

Employees regularly take this information outside the firewall using flash memory devices, personal laptops and smartphones and cloud storage services. They also send this information to private email addresses for business-related and less legitimate purposes.

In addition, technology advances have made it much easier to index large volumes of unstructured data and locate improperly stored privacy, health or financial information within it.

THE DATA SET

Nuix volunteered to examine the prevalence of PII, PHI and PCI within the Enron PST Data Set published by EDRM and ZL Technologies, Inc. This is a worldwide standard set of test data for electronic discovery practitioners and vendors.

The EDRM Data Set Project "provides industry-standard, reference data sets of electronically stored information (ESI) and software files that can be used to test various aspects of e-discovery software and services."

This data set is sourced from the Federal Energy Regulatory Commission's investigation into collapsed energy firm Enron. The EDRM Enron PST Data Set contains approximately 1.3 million email messages which ZL Technologies distilled into 168 Microsoft Outlook .PST files. These email messages were sent and received by Enron staff in the course of day-to-day business.

THE METHODOLOGY

Nuix's Implementation Engineer Matthew Westwood-Hill and EMEA Director of Solution Consultancy Ady Cassidy analyzed the EDRM Enron PST Data Set with a series of standard investigative workflows. Nuix and EDRM are pleased to offer the legal and investigator community this methodology for identifying personal and financial data in corporate data sets.

- Fully indexing the text and metadata. This was straightforward because the Enron data had already been converted into industry-standard PST files. In real-world situations, email might also be stored in Microsoft Exchange Server databases, archives, legacy email platforms or cloud services. Organizations might also need to examine network file shares, collaboration systems and individual computers.
- Using Nuix's "named entities" function to identify dates of birth and credit card and national identity numbers in the data set. Nuix uses regular expression pattern matching to extract intelligence from data sets during processing. Investigators can then cross-reference these intelligence items across multiple data sources.
- Searching for email messages sent to external domains of law firms known to handle personal legal matters. Nuix can group email messages by the domain name they are sent to, making these external messages easy to find.
- Searching for phrases and close groupings of keywords that could indicate personal legal or health discussions, online purchases or other private matters.
- Creating network maps and timelines of email correspondence to identify communication patterns and understand messages and documents in the context of external events.

Technology advances have made it much easier to index large volumes of unstructured data and locate improperly stored privacy, health or financial information within it.

THE METHODOLOGY CONTINUED

Having identified large numbers of suspect emails and attachments using this method, the investigators conducted further analyses on these items, including:

- Using "near duplicate" and "similar documents" functionality to find similar and related content and put together conversation threads. Nuix analyzes four- or five-word phrases, called "shingles," and compares the number of identical shingles between documents to determine the degree of similarity. A list of shingles can also provide a convenient way to narrow down searches, especially where keywords have multiple meanings or are commonly used. For example, a list of phrases containing the word "divorce" would provide a much more targeted search than a simple or proximity search containing that word.
- Using network maps to show which messages and attachments had been sent outside the company, for example to personal email addresses.

Both Nuix investigators completed their research within two days.

Because the Enron data had already been distilled into PST files, rather than including the original source material, it was not possible to forensically analyze the data set. With real-world data, it is possible to conduct deeper analysis of the complete metadata and forensic artifacts within storage, system files and email databases. For example, this could provide evidence that employees had copied sensitive files to flash drives or sent them outside the firewall.

THE RESULTS

Nuix's investigation identified more than 10,000 emails and attachments containing personal data (see table below). This included:

- 60 items containing credit card numbers, including departmental contact lists that each contained hundreds of individual credit cards
- 572 containing Social Security or other national identity numbers–thousands of individuals' identity numbers in total
- 292 containing individuals' dates of birth
- 532 containing information of a highly personal nature such as medical or legal matters.

In many cases, a single item would contain multiple instances and multiple types of information. For example, the data set included many departmental contact lists in spreadsheet form that included dates of birth, Social Security numbers, home addresses and other details of dozens of staff members. In some cases, these spreadsheets also contain the names of employees' spouses and children.

In addition, our investigations clearly showed a considerable number of these items were sent outside the company. For example, employees would forward details to their personal email addresses, presumably to work from home or while traveling.

MORE THAN 10,000 ITEMS OF PERSONALLY IDENTIFIABLE INFORMATION

Nuix's investigation of the EDRM Enron PST Data Set identified the following personally identifiable information:

TYPE OF INFORMATION	NUMBER OF ITEMS CONTAINING THIS TYPE OF INFORMATION
Credit card number	60
Date of birth	292
Highly personal information	532
National identity numbers	572
Personal contact details	6,237
Résumés containing substantial personal contact details	3,023

The cleansed data is published at www.nuix.com/enron

Nuix's investigation identified more than 10,000 emails and attachments containing personal data.

IS THE SITUATION BETTER OR WORSE TODAY?

The EDRM Enron data set is more than a decade old and organizations are much more aware today than they were in the early 2000s about the need to protect private data. In addition, since its collapse Enron has become a byword for corporate governance failings. Is it possible the Enron data set is exceptionally bad? Or do most organizations have hidden privacy data risks hidden in their information stores?

Nuix and its solution partners have conducted sweeps for private and credit card data in unstructured information stores for dozens of corporate customers. We are yet to encounter a data set that did not include some inappropriately stored personal, financial or health information.

For example, in one large insurance company, Nuix identified dozens of Microsoft Excel spreadsheets containing credit card numbers, expiry dates, CVV numbers, home addresses and dates of birth for entire departments of staff members—hundreds of employees in some cases. These convenience copies were accessible to anyone who had password access to the shared drive, making them a significant business risk.

In our experience, email and file shares also frequently contain other business risks such as inappropriate images and royalty-bearing content such as audio and video files.

In the past decade, the opportunities for private information to be stored inappropriately have multiplied. This information can be stored and taken outside the firewall using:

- Cloud email services such as Gmail and Hotmail
- Cloud storage services such as Dropbox, Box and iCloud
- Personal laptops, tablets and smartphones, often used under "bring your own device" policies
- Very high capacity USB keys and other flash memory devices.

In addition, organizations expend a great deal of effort to protect data, making multiple copies and backups, often across several data centers. This multiplies the number of places private, health and financial information is stored, once it is in the system.

WHAT IS PERSONALLY IDENTIFIABLE INFORMATION (PII)?

Personally identifiable information includes any combination of data that could be combined to identify an individual. This may not include the person's name or address—for example, one researcher found the combination of gender, zip code and date of birth was enough to uniquely identify 87% of the population of the United States.[4]

TAKE STEPS TO ADDRESS A SUBSTANTIAL BUSINESS RISK

It is hard to avoid the conclusion that most organizations have PII, PHI and PCI data stored inappropriately, and that this information does not stay within the firewall. The increasing burden of privacy and data breach regulations, combined with the strict requirements of credit card companies, make this an unacceptable risk.

Using the tools and methodology outlined in this paper, organizations can identify inappropriately stored PII and PCI data and take immediate steps to mitigate the risks involved.

[1] European Commission, "Proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)," 2012, http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf.

[2] E. McCallister, T. Grance and K. Scarfone, "Guide to Protecting the Confidentiality of Personally Identifiable Information (PII): Recommendations of the National Institute of Standards and Technology," National Institute of Standards and Technology, 2010, http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf.

[3] See for example PCI Security Standards Council, "PCI DSS Quick Reference Guide: Understanding the Payment Card Industry Data Security Standard version 2.0," 2010, https://www.pcisecuritystandards.org/documents/PCI%20SSC%20Quick%20Reference%20Guide.pdf.

[4] L. Sweeney, "Computational Disclosure Control: A Primer on Data Privacy Protection," Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science, 2001, http://hdl.handle.net/1721.1/8589.

We are yet to encounter a data set that did not include some inappropriately stored personal, financial or health information.

Ady Cassidy joined Nuix in 2011, as Director of Solution Consultancy based in London, UK. Ady is a computer forensic investigator and eDiscovery consultant who has worked for more than 20 years as a computer forensic analyst with the Suffolk Police High Tech Crime Unit. Before joining Nuix, Ady was Managing Consultant with 7Safe, where he was responsible for managing the London based eDiscovery team deploying end-to-end forensic and eDiscovery services.

Ady has had a number of white papers published covering aspects of metadata within the EDRM and eDiscovery practices. Ady has managed large onsite eDiscovery data collections for global partners and has undertaken work at the highest security levels.

Matthew Westwood-Hill joined Nuix in 2013 as an Implementation Engineer advising clients on best-practice ways to use our software and developing solutions that meets their needs. Matthew is an expert in computer forensics, computer investigations and enterprise-wide electronic discovery with more than 15 years experience in the IT industry.

Prior to joining Nuix, Matthew worked for one of the top law firms in Australia. He also ran a computer forensic and electronic discovery company that delivered a full spectrum of digital forensic investigation services including warrants, recovering and analyzing a wide range of electronic devices, reporting on these findings and acting as an expert witness in court. He also has extensive experience recovering deleted and encrypted data.

NUIX INVESTIGATOR

Nuix Investigator software enables corporate law enforcement and regulatory investigators to search and correlate across vast amounts of data quickly and efficiently. With Nuix, investigators can gather all available data in a single location and use advanced investigative techniques to understand the content and context of digital evidence. Nuix offers a range of products for different case sizes, with unmatched capabilities to handle the largest data sets and the finest forensic details.

ABOUT NUIX

Nuix enables people to make fact-based decisions from unstructured data. The patented Nuix Engine makes small work of large and complex human-generated data sets. Organizations around the world turn to Nuix software when they need fast, accurate answers for digital investigation, cybersecurity, eDiscovery, information governance, email migration, privacy and more.

ABOUT EDRM

EDRM (www.edrm.net) creates practical resources to improve eDiscovery and Information Governance. Launched in May 2005, EDRM was created to address the lack of standards and guidelines in the eDiscovery market. EDRM published the Electronic Discovery Reference Model in January 2006, followed by additional resources such as IGRM, CARRM and the Talent Task Matrix. Since its launch, EDRM has comprised more than 260 organizations, including 170 service and software providers, 63 law firms, three industry groups and 23 corporations involved with eDiscovery.