Chapter 7: Finding Data Everywhere

*In times like these when unemployment rates are up to 13%, income has fallen by 5% and suicide rates are climbing I get so angry that the government is wasting money on things like collection of statistics!*

Hans Rosling, quoting a caller on a radio talk show, *The Joy of Stats*

When we think of data, most people think of columns of data stored in some file like in an Excel document or comma-separated file. We think of data being *produced by something*. Data is contained in the server or application log files, it’s something we download from a website or an interface somewhere. Even throughout most of this book, we’ve focused on how we collect and prepare data that already exists. But data exists all around us and everything produces data, the majority of it is just not recorded. We want to begin this chapter by covering methods of collecting data from the environment and then discuss the challenges we face collecting data on breaches. We will talk about several methods of collecting breach data and then we will look at data analysis techniques on public breach data recording using the Vocabulary for Event Recording and Incident Sharing (VERIS) framework.

Breach data represents an extremely valuable source of information.

We want to focus this whole chapter on VERIS, Identify Verizon and the DBIR. We want to establish this framework below and then walk through what we do with breach data collection. Let’s not focus on the Data in the DBIR, but how we are able to pull information from reports.

When will want to

Four phases of a research project:

Define the goals and research questions

Establish data model and collection method

Collect and manage the data

Data analysis and presentation

*“VERIS is complex and they didn’t want to deal with it.”*

This

Planning for data collection

Collecting data from the environment typically has a higher cost than collecting data from a log file. If the benefit is great enough, we may even end up recording one data point at a time so we want to be sure we are approaching this type of data collection with a clear intent. This is where our research question (discussed in Chapter 1) will be critical. We want to begin with one or more questions that we want to answer. By jumping right into data collection with one or more research questions, the data collection ends up favoring the convenient and obvious. Take for example a call center, if we leap into data collection, we will capture what’s obvious and available: how many calls, the operator, duration, etc. But what if we want to understand how effective the call center is from the customer’s perspective? That won’t be found in a log file and all the effort in collection and preparation is wasted. We offer the following steps to follow when creating a program of data collection:

1. Create purpose and direction through a list of questions
   1. Brainstorm a list of questions that drive action or affect decisions moving forward
   2. State the questions in a way data can answer
   3. Prioritize and select one or more critical questions
      1. if we answered this, would we change? If it was affirmative or negative, anything different? (run through possible answers, possibly set a threshold)
2. Select data collection method
   1. Identify sources and methods of collecting data that address questions
   2. Establish data model and vocabulary (fields and data types - taxonomy/framework)
      1. is this field a singular value? multiple attributes?
      2. Every data point and field has a cost - go back to questions, do we care the specific value or just that it’s above or below threshold.
      3. Avoid pack-rat mentality, while it’d be awesome to have all the data, attempting to gather it is a recipe for disaster
3. Collect and manage
   1. Varying data collection methods
   2. data storage - dictated by 2b and 2c
4. Data Analysis
   1. categorical data analysis

atomic data: Boolean, integer, (float), string

collections: list or dict (assoc, object)

We touched on

Once we

* Integrating into a process (ticket system)
* document/log review (after action reports, forms)
* to answer a question, we have secondary review
* survey
* census (every member in population)
* Focus group
* Interviews
* Observation
* Survey
* Follow-up interview
* (after action report)

Issues:

* Validity and consistency Issues
* Cost to collect (resources like time, money, complexity)
* Inference / Benefit of collection

“ expected quality of the collected data, estimated costs, predicted nonresponse rates, expected level of measure errors, and length of the data collection period “

T

Type of research: confirmatory, exploratory

Research question is the reason we are doing this.

When answered, we’ll know something we didn’t before.