Data Profiling: What is it, Steps, and Tools

by James Miller - Wednesday, August 14, 2019

https://www.searchbug.com/info/what-is-data-profiling/

So, you've collected a veritable treasure chest of shiny new data. And, now you need to use it. You may be tempted to just dump your new information into your system and start grinding. However, there's a critical step between data collection and utilization: Data profiling.

Data profiling is the process of checking your information, identifying certain attributes, and determining if the information is useable.

As you may have guessed from the basic definition, profiling is a super important part of the data management process. Data profiling helps to determine whether you send a new batch of data for cleaning, or if you put it to work immediately.

We'll give you a quick rundown on how to profile your data, and some useful profiling tools. So, you can get started with data profiling, and make better decisions about using your data.

What is Data Profiling?

Data profiling is the process of examining data to collect statistics for quantifying the quality of that data or creating an informative summary of that information.

Data profiling can be used on any sort of information. It may be easiest to profile numerical data. But, you can profile other data, such as personal information.

These are the most common uses for data profiling:

1. Determine data quality or find out if data meets accuracy standards.
2. Assess the risk of using or integrating data sets.
3. Make data easier to search by implementing keyword tagging, descriptions, and categorizing.
4. Gather metadata about the information source.
5. Evaluate the accuracy of the supplied metadata about an information source.
6. Anticipate data-related problems that may delay projects.

As you may have gathered from this list, data profiling is mainly a process for checking information quality and performing quality control. Data profiling is definitely a best practice for protecting yourself from data errors and improving your business efficiency.

There are three primary types of data profiling:

- **Structure profiling.** This type of data profiling evaluates data consistency and formatting. Structural data profiling identifies how much of the data is incorrectly formatted.

For example, you'd use structure profiling to discover how many phone numbers in a list have an incorrect number of digits.

- **Content profiling.** Content profiling identifies errors in the content of the data. This type of data profiling helps you identify incorrect data entries and systemic errors in the dataset.

A good content data profiling example is checking a phone list for phone numbers that have no area code. Then, you can see what percentage of the phone numbers are missing area codes.

- **Relationship profiling. Relationship profiling shows you how datasets are related. This type of data profiling identifies common data points or references between cells or spreadsheets. Finding these relationships helps you consolidate data sources and reuse data.**

Structure profiling and content profiling are most commonly used on personal information because structural and content errors are the most problematic for businesses that rely on personal information.

To help you start improving your data quality, here's how to use data profiling on personal information.

Data Profiling Steps

We'll focus mostly on performing structural and content profiling on personal information here since that's what we deal with at Searchbug. But, you can use these steps on numerical data, as well.

This is the data profiling process:

1. **Establish data quality standards.** Data profiling doesn't help you if you don't know what's acceptable for your business. Since data profiling enables you to quantify data quality, you need to identify a data quality threshold for using data.
2. **Profile your data.** You can profile your data with the built-in tools in a spreadsheet editor like Excel or Google Sheets. Or you can use a third party data profiling tool, which we'll talk about in a moment.
3. **Gather percentage and distinct data profiling results.** You need to know both the percentage and total number of errors to perform good data quality analysis. The percentage is for your go/no go analysis. The distinct numbers are for performing a cost analysis.
4. **Identify correctable and uncorrectable errors.** Some errors can be corrected, so the entry is valid and usable. For example, invalid characters can be removed from a complete entry. If the data is incomplete or irreparable, the entry must be removed. Avoid removing entries that can be corrected, so you don't lose good data.
5. **Clean and append your data.** This last step is where you improve the data quality. Cleaning and appending your data fixes the correctable entries and completes any incomplete data. This process can turn an unusable dataset into a usable dataset.

This seems like a lot. However, the data profiling process is rather quick and easy, especially if you use data profiling tools.

Data Profiling Tools

You don't need any super-duper features to profile your data. Most open-source data profiling tools will do everything you need to profile personal information. Here are some good open source data profiling tools.

DataCleaner

Don't let the name fool you. DataCleaner is a data profiling tool that helps you profile your data with boolean analysis, completeness analysis, character set distribution, and a handful of other capabilities that help you profile your data and quickly assess data quality.

Aggregate Profiler

Aggregate Profiler is a truly open source project that offers data profiling, filtering, and governance. The software also offers metadata discovery and anomaly discovery. This one is better for numerical data profiling. But, it will work for personal contact information lists.

Talend Open Studio

Talend Open Studio gives you customizable data assessment, pattern library, and graphical analytics charts. The graphical analytics charts are handy for quickly evaluating your data quality and making a go/no go decision or presenting your data profiling results.

These tools will help you quickly assess your data quality and generate a quantitative data quality rating. However, they're not data providers or resellers. To clean and complete your data, you need different tools.

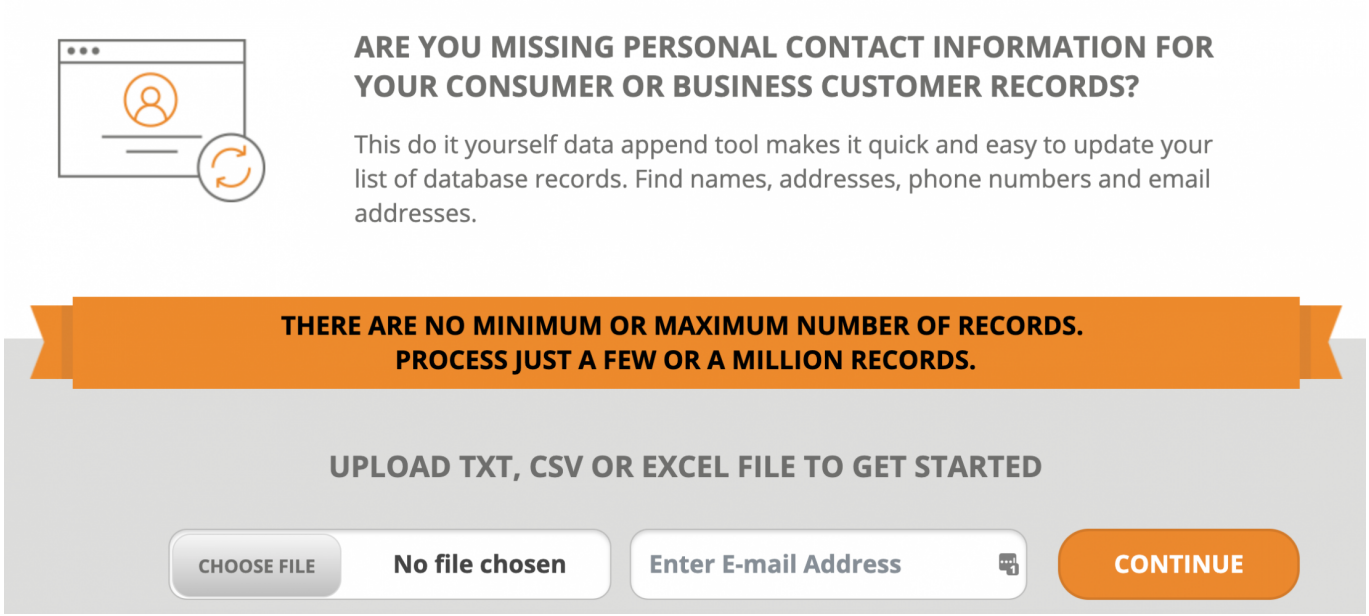
Once you've profiled your data, here's how to clean and append your personal contact information.

Batch Processing

Batch processing is a simple way to clean and append personal information without coding or any real technical expertise.

All you need is an Excel, .csv, or .txt file that's been profiled and had the uncorrectable data entries removed.

Then, simply upload your data file to your data processor. If you use a good data reseller to get contact information, your results will usually come back in minutes. However, very large files may take a bit longer.



That's it. The returned file will have the most recent addresses, phone numbers, and email addresses for each entry. A quality data provider will achieve match rates between 71% and 85%, depending on the type of search you purchase.

Once you get your results from your data partner, your data is ready to use.

Data Integration

Data integration takes a little more time and technical expertise. However, it can almost completely automate the data cleaning process.

You need to manage your data with some sort of software, like a CRM system, to use data integration. This is because you use an API to connect your data management software to your data partner's system.

The API usually integrates using a custom URL. So, you don't need an API key. And, the integration is pretty simple. However, you may need your tech team to setup the API.

But, once you have the API connected, you can setup your system so that the data cleaning process triggers automatically. That way, your data is automatically sent out for cleaning and appending before you use it. Data integration can save you a ton of time, in addition to saving you money and improving efficiency.

Data cleaning is *technically* the last step in the data profiling process. But, data profiling is more like determining whether or not your data needs to be cleaned and preparing the data for cleaning.

Even so, data profiling is still an extremely valuable practice that will help you improve your data quality and get more return from your data.

Ready to clean your profiled data? Get the most dependable data with Searchbug data cleaning tools.

Searchbug.com