Data Cleansing: What It Is (and Why You Need It)

by James Miller - Monday, September 03, 2018

https://www.searchbug.com/info/2018/09/data-cleansing-why-you-need-it/

In the digital age, data is almost as good as currency.

This trend is largely powered by the advancement of AI and automation. Machine learning requires massive amounts of data. Not only that, but if you're going to create more efficient processes, you need more material to work with.

Having a stockpile of accurate data helps on the analog side, too. This includes good, old fashioned snail mail sales letters.

Better data makes customer outreach more reliable and personalized, which empowers sales teams and marketing departments.

However, more efficient data usage also amplifies the effect of bad data.

Thousands of emails might be sent or hundreds of prospects called before errors are identified and corrected.

Entire batches of data analytics results can be ruined by bad data.

That's why good data management policies are critical.

A vital part of your data management ecosystem is data cleansing.

What is data cleansing?

Data cleansing is the process of combing through data, and correcting errors or completing missing information.

No matter how thorough your data collection systems are, it's impossible to prevent errors from getting into your databases.

People make mistakes on data entry forms. People intentionally omit or enter invalid information. There's just no way to account for every possibility.

Moreover, there are massive benefits to having huge stores of prospect and client data. New data collection methods are being created all the time to help exploit these benefits.

Then, the expansion of digital resources and ways that customers can engage with companies means that error-free data collection is probably impossible. If it is possible, it won't happen for a long time.

That means data cleansing is one of the most important mechanisms in data management.

Obviously, you should still be vigilant in catching errors during data collection, but regularly cleansing your data is really the only way to prevent errors from causing problems in many parts of your organization.

The best way to keep your data error free is to employ efficient data cleansing tools.

How to cleanse your data

Even with data cleansing tools, cleaning your data is still a process.

Since you need to regularly clean your data, creating an easily repeatable data cleansing process will make cleansing new data and refreshing old data far more efficient and saves money in the long run.

Identify key data fields

The most important pieces of information will vary from organization to organization. It depends on your customers, product, marketing strategy, and even your employees.

In any case, the first step in creating a solid data cleansing process is identifying what information is most valuable to your company.

This step also helps you create data validation guidelines, which improve your data collection process. With your key data identified, you can minimize the amount of unusable data entries that appear in your databases.

This step alone improves the quality of your data.

Analyze your data

With your key data in hand, you can go to your data stockpile and identify the gaps.

The analysis phase of cleaning your data is also an opportunity to organize it and remove any data fields that you don’t really need.

If you’re using Excel or Google Sheets, you can create scripts and workflows to streamline and automate a surprising amount of this process.

Here are a few key things to do as you’re analyzing and organizing your data:

Remove duplicate rows.

Duplicate rows can cause problems if you need to import your data. So cleaning these out can save you

headache later on.

Remove spaces and nonprinting characters.

Extra spaces (unicode character set values 32 and 160) and weird characters (unicode character values 0 to 31, 127, 129, 141, 143, 144, and 157) can cause issues for sorting, filtering, and searching.

Getting rid of these makes your data management life much easier.

Merge and split columns.

If you've imported data, make sure that it's divided up the way you need it to be. For example, you may want to split a single name column into a first and last name column.

This will help you identify missing information later on.

Having your data properly organized will also increase the internal efficiency of your company, since people will be able to more easily search and find what they need in your databases.

Append missing data

This is where the first two steps really start to pay off.

If you've got your data organized and you know which information is most important to you, you can pretty easily process your data in batches.

You have a couple options here.

- **You can use a batch data appending service. Simply upload a CSV, TXT, or Excel file. After that, the service will append the missing data that you need to your list.**
- **You can use an API to create an integrated data append process so that your data processing is more internal to your organization.**

The cool thing about using an API, is that you can also embed the data completion capabilities into the customer facing side of your website or app. With some creativity, this offers opportunities to streamline and automate your data validation at the point of collection.

Automate

Using the same data append services that you used to complete your data, you can automate continued maintenance and cleaning of your databases.

This is especially easy if you use an API to integrate data appending into your internal tools. Simply create a script that runs the data append function at regular intervals.

One mistake companies make is running through the first three steps every once in a while, like a spring cleaning for their data.

The trouble with this method is that these companies end up working with bad data for a month or two or more until they go through their databases again.

Creating automation that will continually refresh your data will keep you working with accurate information all the time.

Manage

Even with a good system for cleaning your data in place, you still need to consistently manage your databases to maintain data sets that your automated tools can work with, while remaining relevant to your needs as your business evolves.

Here’s what you need to do to keep your data in good shape:

Data backup

This one goes without saying. Backup your data, or you could find yourself having to recreate or buy entire silos of data all over again.

Input validation

Input validation is your first line of defense against bad data. As we mentioned before, perfect input validation probably isn’t possible.

But crafting your opt in processes, creating scripts that disallow incorrectly formatted data to be entered, and establishing procedures for manual data entry can help minimize the amount of data cleansing you’ll need in the long run.

Data validation

This is actually kind of a subset of input validation.

Data validation pairs with input validation to create a sort of two-step data entry process. Essentially, what you want is a mechanism that checks newly inputted data before it actually gets put into use.

A great way to do this is to build a script that simply removes incomplete or invalid data from a database and creates a separate container of data that will eventually become the file that gets run through the data cleanser.

Cleansing

Yep. Data cleansing is a key element of good data management. Refer to steps one through four in this article for more details.

Aggregating

Data aggregation can be achieved using built-in functions in your data management software like Excel or Google Sheets.

Aggregating your data essentially prepares it for use based on what elements of the data set are important for your company and objectives.

Another benefit of aggregating your data is that it provides one more layer of protection against bad data, since missing or invalid data can cause errors in the aggregation process.

So, aggregating your data can tell you if you've got any more tidying up to do before you send the data over to the teams who are going to use it.

Filtering

This part will be pretty easy for data sets that you've constructed with your own collection methods, since you probably only collect the data that you need.

However, if you're using imported data that you acquired from a secondary source, there's a good chance that there's a bunch of stuff in there that's not particularly useful to you.

In this case you'll need to filter the data and toss anything that's not useful. This makes it much easier for your teams to work with the data and get what they need from it.

If you frequently work with imported data, it's a good idea to create a way to automate this process or at least have a workflow to streamline it.

Merging

This one is pretty straightforward:

If you can, you should merge multiples of the same type of data into a single container, so that it's easier to search and process.

Appending

We covered data append when we talked about cleansing. Append any missing data to your lists to create complete data sets.

Deduping

Deduping is just a super technical sounding term for removing duplicates of any data you have. Unless you have a specific need for duplicate data, it's best not to have more than one of each (with the exception of your backups).

Transforming

Data transformation is converting data from one structure or format into another. Typically, this is most useful if you use imported data. Sometimes you'll get data in a format that your systems can't work with. So you'll need to transform it to make it workable for you.

Often, the program you use to work with your data, like Excel, won't have good data transformation capabilities.

Fortunately, there's a wide range of data transformation tools for getting this done.

Standardization makes this process simple. Avoid working with multiple data formats whenever possible. It just complicates things.

So that's data cleansing (and data management).

While data cleansing alone isn't enough to keep your data in shape, it will prevent headaches and can save your business a lot of money.

Feel free to check out our data cleansing services. Then leave a comment and let us know how you keep your data in order!

Searchbug.com