vertify

WHERE DOES **DIRTY DATA ORIGINATE?**

About Vertify

The Vertify customer intelligence platform (VCIP) is transforming the way revenue teams work and win. With Vertify, every revenue team has the ability to tap into the benefits of robust data integration to connect mission-critical revenue apps, and data assurance to continually clean, normalize, and enrich customer and prospect data. By sitting at the intersection of automated data connection and quality, Vertify further empowers revenue teams to harness predictive intelligence that improves marketing outcomes and accelerates growth. Learn more at www.verify.com.

What is Dirty Data?

Dirty data is any inaccurate, incomplete, outdated, or duplicated data that resides in your contact database. It's any and every typo on a form fill; it's every contact you lost touch with because they changed companies, and it's the duplicate leads that ended up in your system because of a sync issue between your Marketing Automation application and CRM. Dirty data is contact data that's not completely correct and 100% up to date.

Dirty data is the bane of every marketer's existence and kicks off a snowball effect starting with limited segmentation. Limited segmentation hinders lead routing and slows sales velocity; limited segmentation also undermines personalization. Less personalization means diluted messaging and less relevancy to the intended audience. Less relevancy kills conversion rates and causes opt-outs to spike. Lower conversion rates mean fewer leads, unhappy sales teams, and diminishing campaign ROI. Ultimately, dirty data causes misaligned sales and marketing teams and inhibits growth.

Perhaps worst of all, dirty data distorts the big picture and limits an organization's ability to reach its full potential. While that may sound embellished, consider this - data that isn't trustworthy taints analysis and makes the predictive insights you seek totally unreliable.

> "Marketers understand the problem [of dirty data], but they don't always truly respect the cost bad data imposes on the organization," said Justin Gray, CEO of LeadMD. "Predictive marketing and artificial intelligence sound sexy. But if the data is so bad, you're constantly adding new flat records that make it hard to understand who a prospective buyer is. That's where the costs start to amplify."

Data hygiene can't (and should never) be optional. Database negligence is the most effective method of marketing self-sabotage. Don't let dirty data be the reason you don't reach your goals.

Following Dirty Data Back to the Source

Now that we've established what dirty data is and the damage it causes, on to the fun stuff — where it comes from. How data is sourced is the biggest indicator of its trustworthiness and reliability.

So many B2B marketers rely on third-party data vendors to complement their existing lead-gen strategies. But when it comes to net-new contacts and contact acquisition, many third-party data vendors rely on data sourcing methods that can backfire on you down the road. Let's look at how third-party vendors source their data (and how that impacts you once you work with a vendor).

Bad List Buys

Looking to boost your marketing list in a snap? Buying those contacts from a list-purchasing company might appear too good to be true — because it is. Be mindful, that some list brokers acquire their data through questionable and sometimes illegal tactics.

SiriusDecisions reports that, on average, purchased contact lists are 15 months old at the point of sale, at best — and they continue to deteriorate at a rate of about 25% per year. While buying email lists is tempting, the long-term consequences aren't worth the short-term fix. One bad apple is all it takes to contaminate your list.

As Hubspot blogger Corey Eridon writes...

> "One customer's ill-gotten email address list can poison the deliverability of the other customers on that shared IP address."

Compiled Data

When looking to buy from a list broker, a key consideration must be where they source their data. Data that comes from the original source — i.e. self-reported information from the actual contact — is ideal. However, few data providers are able to offer data that comes straight from the source. Instead, what often happens is that the vendor will take large lists of contact data from a myriad of different sources and throw all of the lists together into their database. Accuracy and precision are ignored and the name of the game is quantity over quality.

Crowd-Sourced Data

If it's not compiled, it's likely crowd-sourced data. This means that the data informant is a secondhand source reporting information about the actual contact. In crowd-sourcing, that secondhand source declares contact records either accurate or inaccurate based on information they have that may or may not be closer to the original source (the contact). Typically, there are rewards and incentives put in place to encourage customers to correct records when they are wrong. Data accuracy is not a democracy. Just because someone says that a record is accurate does not mean it is.

Human Error

Sometimes, simple human error is the source of data quality issues. For any organization that has forms on its website where data is manually entered by a customer or prospect, there is always the risk of data being entered incorrectly. Conferences and tradeshows often provide attendee lists and the means of capturing leads onsite. While these events provide great opportunities to collect a larger volume of leads, it is not uncommon to occasionally find contact details with missing or incorrect information.

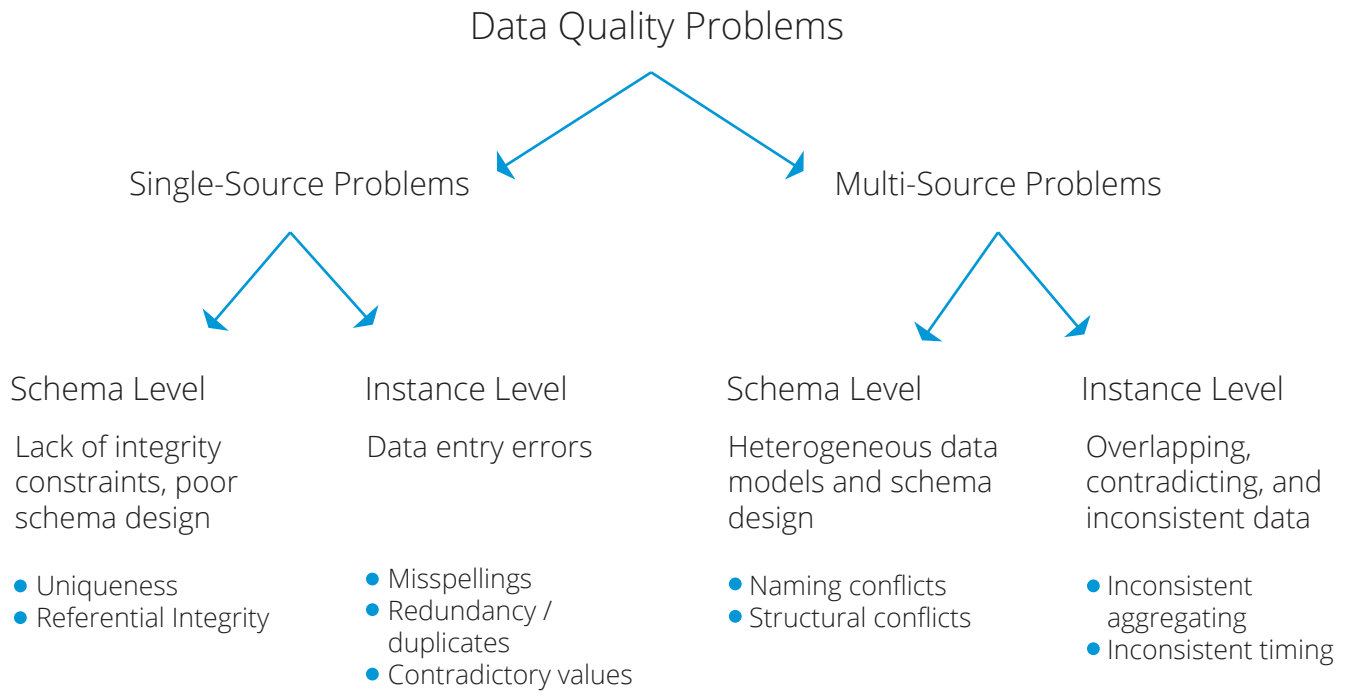
Natural Decay

Today's workforce is more transient than ever. Between people changing jobs/companies and organizational changes at the company level, contact data is a fluid asset rather than a fixed one. Contact data natural decays and expires as time passes, to the tune of about 32% per year (according to SiriusDecisions).

That may seem like a high percentage, but it's worth considering that, according to Dun & Bradstreet, in the next hour alone, 271 businesses will move, 1274 telephone numbers will be changed or disconnected, 673 new businesses will open, and eight companies will change their name.

A Deep Dive into Dirty Data: Single-Source vs. Multi-Source Problems

Dirty data can occur within a single set of records or between multiple sets of data that rely on each other and need to be merged. These two major data quality problems are referred to as single-source problems and multi-source problems and can be restored by cleaning the data.



Single-Source Problems

Schema Level Problems: Problems that occur within the theory or standardization of database organization. Schema level problems will be reflected in instance-level problems. Typically these problems occur when a database is not properly engineered or has poor integrity constraints.

Examples:

Illegal values - values are outside of the domain range

Uniqueness violations - duplication of unique fields, like a Social Security number Referential integrity issue - the referenced field is not defined

Violated attribute dependencies - one field depending on another field incorrectly, like city and zipcode

Instance Level Problems: Errors, inconsistencies, and inaccuracies in the actual contents of each record. These problems are not visible at the schema level. Errors that occur at the instance level encompass a wider range of inconsistencies that do not reflect the database construction but feature conflicting data points.

Examples:

Misspellings - typos or phonetic errors

Missing values - unavailable values during data entry

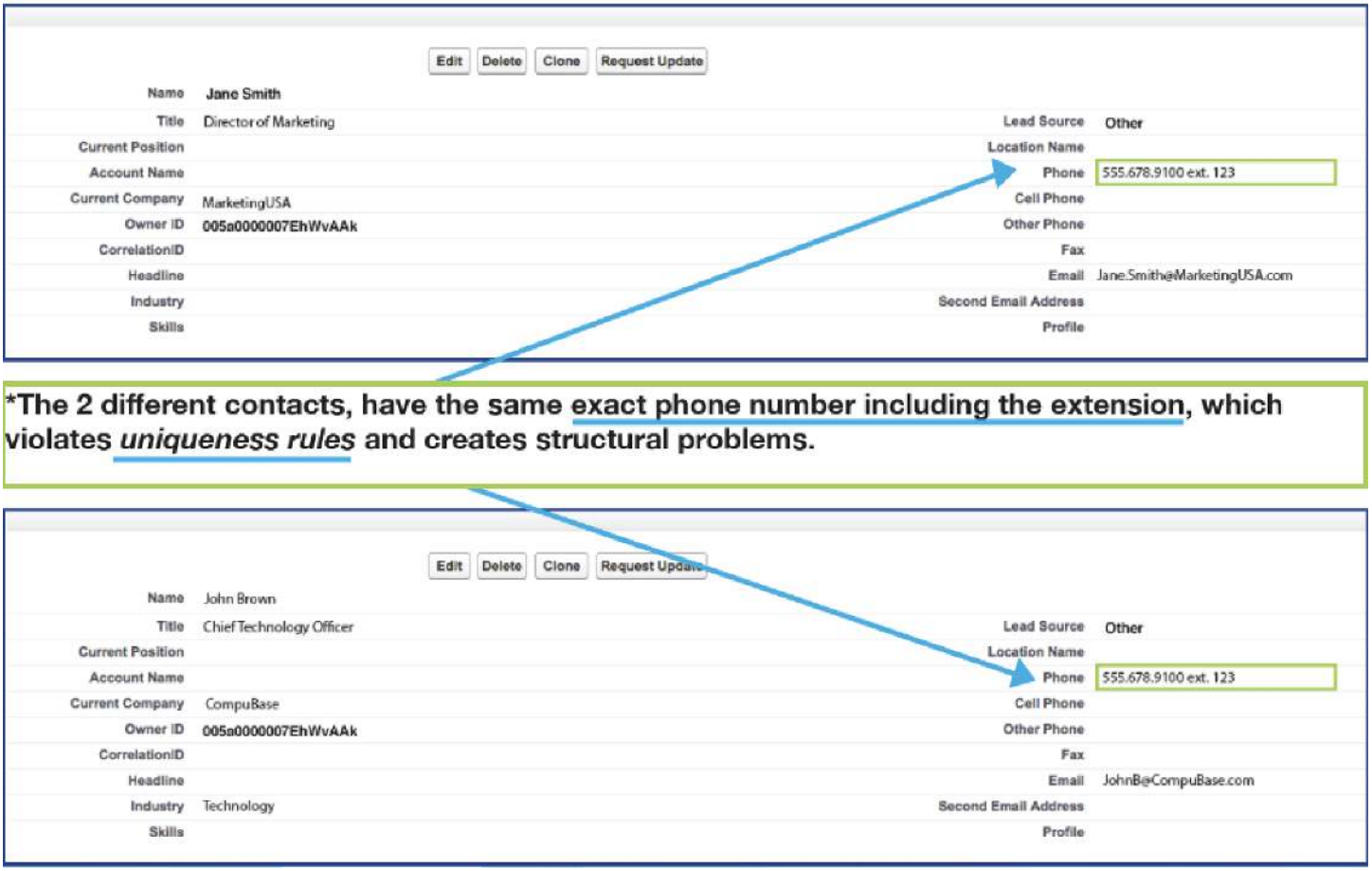
Embedded values - multiple values entered into one attribute field

Misfielded values - value is in the incorrect field

Violated attribute dependencies - one field depending on another field incorrectly, like city and zipcode

Duplicated records - the same contact is shown twice due to data entry or merging errors

Contradicting records - duplicate contact with different values

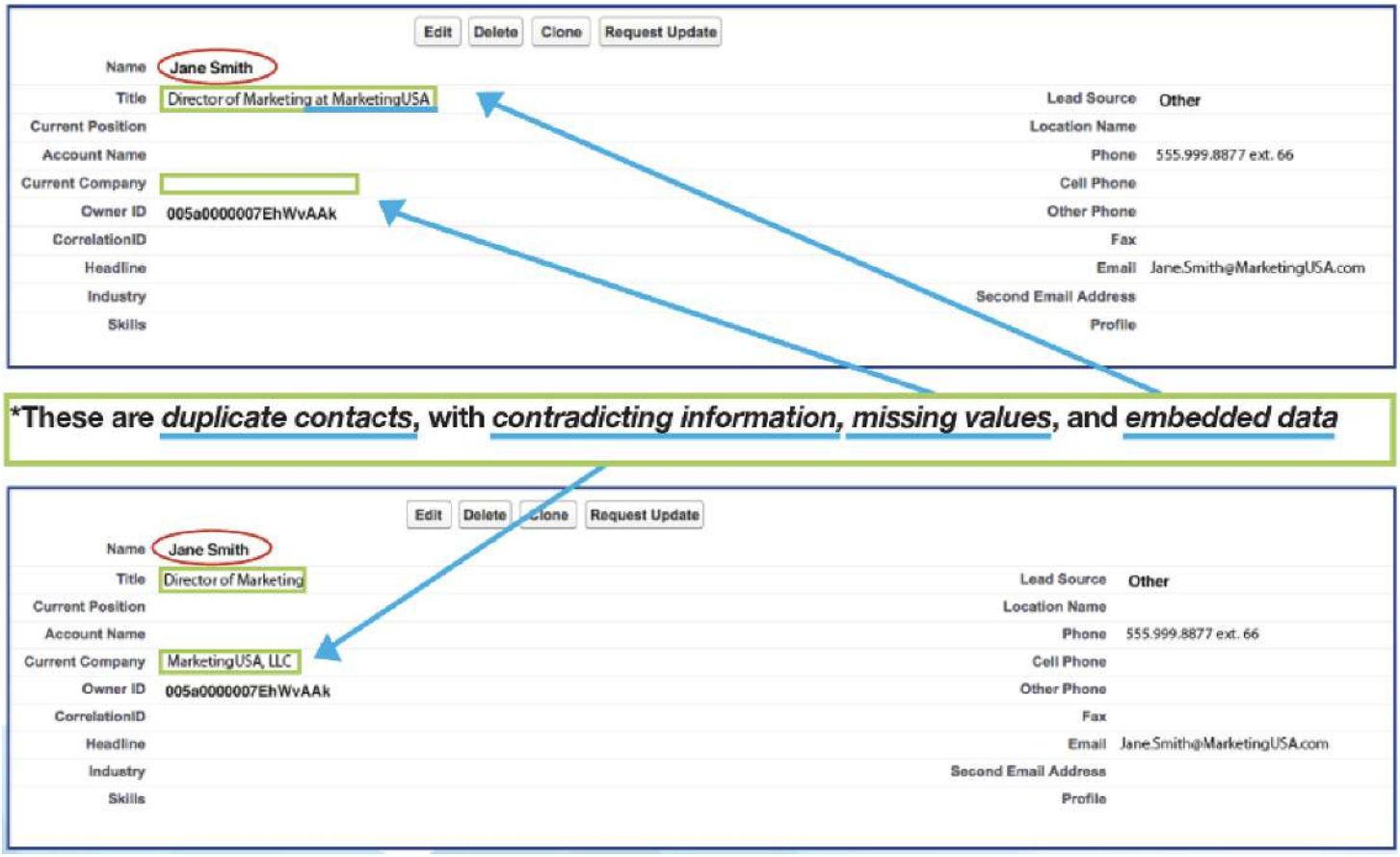


Multi-Source Problems

When multiple data sources need to be merged in a data warehouse, the need for data cleaning increases significantly. This is because the various data sources often contain the same data in different representations that overlap or contradict one another.

Schema Level & Instance Level

The following data model presents problems at both the schema level and instance level. On the schema level, the model was designed with different names for the same object (e.g. Source A uses the term 'Customer' while Source B uses 'Client'), creating structural conflicts. On the instance level, data conflicts are apparent regarding gender representations ("0"/"1" vs. "F"/"M").



Compiled Data

When looking to buy from a list broker, a key consideration must be where they source their data. Data that comes from the original source — i.e. self-reported information from the actual contact — is ideal. However, few data providers are able to offer data that comes straight from the source. Instead, what often happens is that the vendor will take large lists of contact data from a myriad of different sources and throw all of the lists together into their database. Accuracy and precision are ignored and the name of the game is quantity over quality.

5 Best Practices for Data Cleansing

As outlined earlier in this white paper, ignoring data quality issues can significantly impact your organization and impede growth. Identifying the origins of your dirty data is a step in the right direction. Resolve schema conflicts to ensure successful data cleaning, then implement these five best data cleansing practices for continued health:

1. Develop a Data Quality Plan
Knowing where most data quality errors occur and identifying incorrect data will help your team better assess the root problem and develop a project plan. A comprehensive data quality plan will impact many departments, so keep communication open and emphasize that better intelligence will save everyone time, money, and energy.

2. Standardize Contact Data at the Point of Entry
Check important data at the point of entry and standardize the method across all departments. This ensures that all information is uniform when it enters your database and will make it easier to catch duplicates.

3. Validate the Accuracy of Your Data
Validate the accuracy of your data either in real-time or by cleaning your existing database regularly to ensure it is complete and up to date. Do your research and invest in solutions that will clean and verify your data. Effective marketing occurs when high-quality data and cutting-edge technology are used in tandem to seamlessly merge various data sets.

4. Identify Duplicates and Normalize Data
Save your team time and implement a data hygiene tool that can effectively identify duplicates. The less manual work, the better.

5. Append Data
After your data has been standardized, validated, and scrubbed for duplicates, use a third-party vendor to append it. Reliable third-party sources can capture information directly from first-party sites, then clean and compile the data to provide more complete information for business intelligence and analytics. This will help you develop and strengthen your customer segmentation and send more targeted information to customers and prospects.

What Successful Data Cleansing Looks Like

Before implementing data cleaning, it's important to look at the big picture. What are your goals and expectations? How do you plan to execute it successfully? As you implement data cleaning, keep these tips in mind.

When done correctly, successful data cleaning implements these three key practices:

- Detects and removes major errors and inconsistencies in single data sources and when combining multiple sources
- Utilizes tools to reduce manual inspection and programming efforts
- Works in conjunction with schema-related data transformations and specific mapping functions, not solo

Just like any other habit, prioritizing data hygiene might take some adjustment, but once you make a habit of regularly cleaning and appending your data, all of your other marketing efforts become more efficient and streamlined. Do yourself (and your company) a favor. Spend the time to properly source and manage your data, and you're virtually guaranteed to see a drastic positive change in your campaign metrics as a result... all you have to lose is 500 bouncebacks.