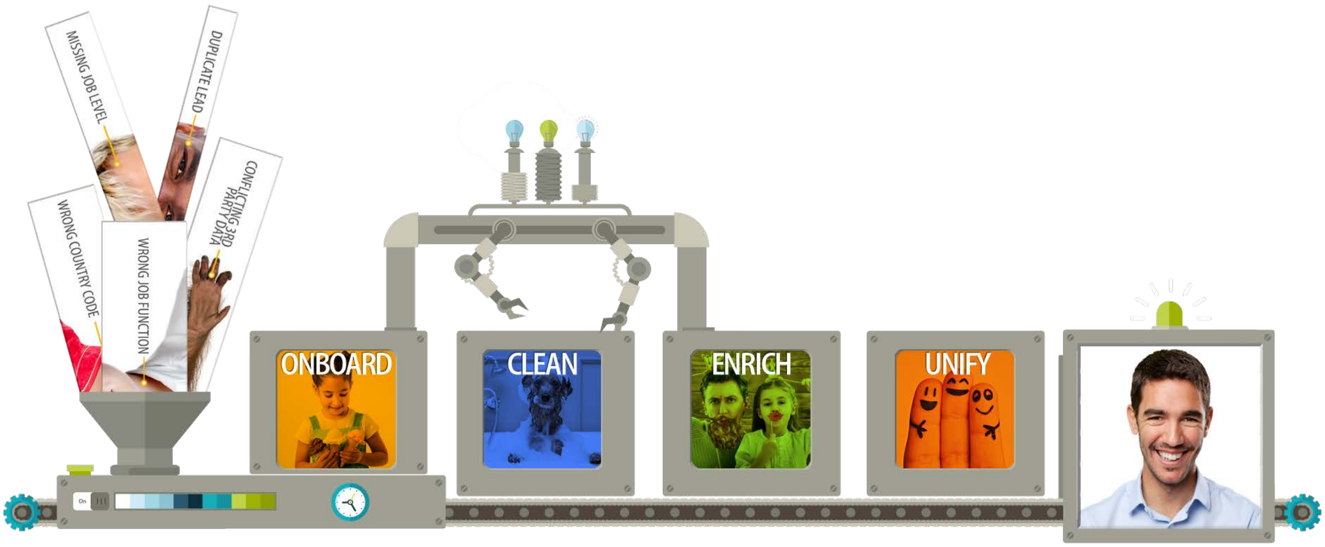
OPENPRISE™

The Complete Deduplication Survival Guide

For Marketing, Sales, & Support



The Complete Deduplication Survival Guide

for Marketing, Sales, & Support

Deduplication is one of the biggest data quality improvement processes that every marketing and sales operations professional has experienced firsthand. Data deduplication is simple in concept but can be quite complex in execution, especially when dealing with records distributed across multiple systems.

Data deduplication is not a one-time exercise. An effective dedupe program is implemented as a continuously running program. This is because duplicate records can trickle in from multiple sources, such as list imports, broken sync between systems, and manual record creation.

In this guide, we'll share detailed data deduplication how-to's and best practices, covering what you need to consider before, during, and after a deduplication project. We'll address people, process, and system issues.

Table of Contents

1. Why the Need to De-Dupe

Data geeks like us do it for fun, but most people spend money and effort on deduping because of the huge ROI the comes from the following:

Preventing Multiple Reps from Calling on the Same Leads

This is probably the number one driver cited by our customers for data deduplication. When you have duplicate accounts, contacts, and leads, you can easily end up in a situation with multiple account reps and sales development reps calling on the same lead or account. This problem is more acute with a large sales team and with round-robin system for distributing new leads. You can end up with multiple reps working on the same account for an extended period of time that can result in sub-optimal account engagement and commission dispute. Worse yet, having an SDR calling on an existing customer can also make your company appear clueless and sloppy.

Linking Trial Users to Other Program Leads

Many software and consumer service products offer a free trial, so anybody can sign up via self-service and kick the tires. This is a proven lead generation tactic that can be extremely productive for the right product and buyer persona. Whether the trial signup process is handled by your product or your marketing automation platform, you can easily generate a massive number of duplicate leads from the trial program. If your trial has a time constraint, you likely have leads that have signed up multiple times using different email addresses. Trial users are valuable “mid-funnel” leads where you need to maximize your conversion rate. In order to accomplish that, you need to correlate trial user’s activities across all records and programs, which means you must dedupe.

Automating Sales, Marketing, and Fulfillment Processes

There are many good reasons why you need to automate your sales, marketing, and fulfillment processes. There are plenty of great software solutions that can help you automate the workflow and transactions across different systems and departments. However, automating business processes when your database has a large number of duplicates can cause more trouble than it’s worth. Duplicate records can cause duplicate transactions and processes, creating confusing and repetitive touchpoints with the customer and propagating the duplicate data into your finance, order management, and help desk systems.

Saving Money and Improving Performance of Marketing Automation Platforms

Most marketing automation platforms are priced according to the size of the database. A large number of duplicates directly costs you money in terms of the license fees you pay. If your duplicate count is especially large, say over 20% of your database, it can also negatively affect the performance of your marketing automation platform. Processes that used to be “real time” may lag significantly, creating issues with the Service Level Agreement you have between your marketing and sales organizations.

2. People & Process Considerations

As with most any topic in Marketing, you usually start with people and process first, before you get to data and technology. Dedupe is no different. Below are some of the key people and process questions you should answer before embarking on the de-dupe problem.

Data Ownership

Deduplication can include Lead, Contact, and Account (using Salesforce.com terminology) data. These data often have different systems of record and owners. Lead data is often owned by Marketing and the system of record is usually the marketing automation platform. Contacts and Accounts are generally owned by Sales and the system of record is generally the salesforce automation platform. If user data is part of the project scope, then we add product and customer success teams as potential owners and your application and help desk platforms as additional systems of record. These data can be completely separate, partially synchronized, or fully synchronized.

The data owners must agree on the deduplication scheme and process. If the process requires significant time, effort, and budget commitment, then the data owners need to be fully committed for the project to be successful. For example, if sales insists on manually reviewing every account record merge, then the dedupe process must consider how to efficiently involve every account executive in the process.

Data ownership can be a multifaceted issue that include not just departments, but data hierarchies as well. For example, did you know it’s possible to have Contacts without Account affiliation in Salesforce? These “private contacts” are considered private data to most account reps. If your CRM allows for private contacts, should they be included in the dedupe effort?

One Time or Continuous Dedupe?

Is the deduplication process a one-time (periodic) batch process, a continuous process, or a combination of both? One-time processes involve a massive clean-up that happens periodically, from once a quarter to once every few years. For one-time processes, a manual or semi-automated solution is perfectly acceptable, as long as the solution proposed can accommodate the time and budget requirements. If the dedupe process is to continue as an ongoing process after the initial clean-up effort, then automation solutions must be in the discussion from the start. A manual or even semi-manual process is simply not scalable and manageable.

Given the people and process constraints you have, decide if a continuous process is realistic. If not, determine how close to an ideal state is acceptable, and consider if it can it be supplemented by smaller scale periodic batch cleanups. For example, if Sales insists on manually reviewing dedupe results and merging Account records, then in order to have a continuous dedupe process, you must get a Service Level Agreement from the Sales team on how quickly they can review the dupes.

Dedupe New Records, Some Records, or All of Them?

It’s a lot easier to prevent new duplicates from being created than to remove existing ones. It often makes the most sense to separate the two objectives into separate processes that work together to create a comprehensive dedupe solution. The new dupe prevention process can run continuously, supplemented by a full dedupe process that runs less frequently, say once a quarter. For example, if list loading is a major source of lead input for you, then by simply preventing dupes from being created while loading a list can help tremendously in slowing the growth of dupes in your database. While the full dedupe of your database may involve more stakeholders and take longer to figure out an acceptable process, you may be able to implement a dupe prevention process for quick list loading, as Marketing has full control over that data and the process of list loading.

Here are some ways you should consider to reduce the problem:

1. If you understand the major sources of your dupes, try to manage those sources first to “stop the bleeding”.
2. If multiple databases are involved that are separate or partially synchronized, consider first deduping each database or parts of the database separately, before combining them one at a time and de-duping in phases.
3. If a certain subset of the data is more complicated to de-dupe due to data quality, people, or process issues, consider leaving them out in the first phase of the deduping effort. Deduping success doesn’t require achieving perfection. Having a 95% clean database this month is better than doing nothing and never getting started.

Which System to Dedupe Against

If the data you’re looking to dedupe exists in multiple systems, you need to make a decision about where the dedupe should happen. In general, when data is synced between different systems, one of the systems is the master. This is the system we recommend you dedupe against in most cases.

Note that the master system may not be where the data originated. One common example is Salesforce.com Contacts vs. Marketo Leads. The lead record may have originated in Marketo, and then synced with Salesforce. As long as the person remains a Lead, Marketo is considered the system of record and dedupe should be done against Marketo. Once the lead is converted to a Contact, Salesforce now becomes the system of record. In this case, it’s best to dedupe a Contact directly against Salesforce.

There may be constraints that will limit your choices. For example, you may not have purchased API access (e.g., available only in the Enterprise subscription of Salesforce.com) for the system you wish to dedupe. You may not own or do not have authorization to change the data in certain systems. In these situations, you may have to dedupe against the secondary system and let the data synchronization propagate the changes to the system of record.

3. Pre-Deduplication Checklist: Salesforce.com & Marketing Automation Data

You are excited about getting rid of those pesky duplicate records in your Salesforce.com and marketing automation solution (well, *we* get excited about this kind of stuff). You want to jump right in because those duplicate records have annoyed you for too long and you want them GONE! GONE! GONE!!!

Before you get started, a little bit of planning can save you a lot of frustration as you progress. Here is a handy dedupe checklist to help with your project planning.

1. Document your deduplication logic
2. Verify your data sync status and arrangement
3. Check for data verification rules and other automation that may interfere
4. Check for bad data and business processes that may interfere

Document Your Deduplication Logic

Deduping is one of those seemingly simple tasks that's actually fairly complex in execution. There are many nuances you're probably unaware of unless you have done it many times before.

The first thing you need to do is to think through and document your deduplication logic in both Salesforce.com and your marketing automation solution. There's no such thing as "generic dedupe logic." If a vendor tells you it has a proprietary algorithm that can magically dedupe your database, you should run away because your database will probably end up being ruined.

Your dedupe logic must accommodate the following:

1. The current state of your data
2. The sources of dupes
3. The systems and syncing technologies involved
4. The controls and automations that are in place
5. The people and the process it touches (see our last blog on this topic: Dedupe Project Considerations: People & Process)

There are three key parts to the deduplication logic you must figure out:

1. How to identify the duplicates

2. How to select the surviving/winning record
3. How to merge the non-surviving/losing records, accounting for system restrictions

Verify Your Data Sync Status Between Salesforce.com and Your Marketing Automation Solution

Whether you're deduping Leads, Contacts, or Accounts, chances are the data set you're trying to dedupe exists in multiple systems. For B2B marketers who are looking to dedupe leads and contacts, the data usually lives in both your Salesforce.com and marketing automation platform, and possibly others systems such as help desk, finance, and customer success platforms. Chances are they're synchronized to some degree. We often see these scenarios:

1. Salesforce.com and marketing automation systems are fully synced, so every record exists in both systems.
2. Salesforce.com solution has more leads than the marketing automation platform because only marketable leads that are CAN-SPAM compliant are in the marketing automation platform.
3. The marketing automation platform has more leads than the salesforce system because only marketing qualified leads (MQLs) are pushed into Salesforce to focus the sales team on hot leads.
4. A combination of scenarios 2 and 3.

Depending on how many systems are involved and how your syncing is currently implemented, you may be able to get away with just deduping one database and letting the changes ripple through via sync, or you may need to independently dedupe the individual databases.

If you have the option to dedupe only one database, which one is better? In general, deduping should be done on the system of record for each dataset. For example, deduping of leads should be done in your marketing automation platform while deduping of accounts should be done in the sales system.

Just because your databases are supposed to be fully synced doesn't mean they actually are. Many things can cause syncing algorithms to miss records, and they can accumulate to substantial amounts quickly. Here are typical reasons why your data syncs may not be perfect:

1. Many syncing and data automation technologies only fire one time when the record is created. Subsequent changes to that record may not get synced.
2. Many syncing technologies don't handle deleted and merged records properly, so when one record is deleted in one system, the other systems are unaware that the record no longer exists.
3. When syncing is interrupted, backed up, or runs into errors for whatever reason, not all syncing technologies can gracefully recover and resume, which can result in records being out-of-sync.

4. If you have bi-directional syncing, the syncing logic may not be built or configured to be exactly the same in both directions. Sometimes syncing works better or is supposed to work only in one direction.

It’s worthwhile to audit your various databases to verify if they’re actually in sync.

Check for Data Verification Rules and Other Automation That May Interfere

Your systems may have data verification rules and other types of features turned on that can interfere with duplicate removal. If so, you’ll see these symptoms when you try to merge records:

1. A record you deleted in System A is not deleted in System B and therefore becomes orphaned in System B.
2. A record you deleted in System A is not deleted in System B and is later recreated in System A by System B.
3. A record you updated in System A is partially updated in System B because some data field updates were blocked by System B. The records remain out-of-sync forever, or get back in-sync the next time System B initiates the sync.

Here are some examples of interfering automations:

1. Your Salesforce.com Contact record may have required fields that the Lead record doesn’t have. In order to merge a Lead with a Contact record, the Lead record must be converted to a Contact first. The conversion will fail if required data is missing.
2. If you have duplicate blocking turned on in SFDC, the above scenario will also fail while attempting to convert a Lead to a Contact.
3. Salesforce.com is the system of record for Account data. Therefore, any change to the company information for a Lead in Marketo that is a Contact in SFDC, won’t propagate to SFDC. In addition, the Marketo company data will get reverted the next time SFDC initiates a sync.

If you are doing a one-time deduplication exercise, you can work around these conflicts by suspending part of the process for the time being. However, if you are setting up a continuous deduping process, then you need to rationalize which technology does what so they don’t step on each other.

Check for Bad Data & Business Processes that May Interfere

Your system may have bad data that can also prevent successful deduping. Some of these bad data situations may be quite a head-scratcher on how they came to be, but they are definitely out there. In some cases, they've been explicitly allowed by your business processes.

For example:

1. A record owner may no longer be a valid user in the target system. Subsequently, any attempt to update or merge such a record will be rejected.
2. If you allow a Contact record in Salesforce.com to have no Account affiliation (a.k.a. Private Contacts), then any attempt to merge a Contact without an Account to Contacts with Accounts will require additional logic on Contact-to-Account matching.
3. A record may contain an old data value that has since been changed to a picklist. Any attempt to update such a record will require additional logic to reset the outdated record.

Any skilled crafts man will tell you, "Measure twice and cut once." Deduplication is no different. Proper planning upfront can save you a lot of pain, frustration, and damage control later on.

4. Identifying Duplicates in Salesforce.com and Marketo

Once you've taken into consideration people and process and have your checklist ready, the next order of business is to write down your dedupe logic, and the first part of the dedupe logic is how you identify duplicate records in your MarTech database, whether that's Salesforce.com, Marketo, Pardot, Eloqua, or something else.

Which Data Fields to Use?

The most common data field used by B2B marketers to identify duplicate records is email address, which makes a lot of sense. However, that's just the starting point. Here are a few more options to consider, which can improve your ability to catch those more elusive duplicate records.

Mobile Phone Number

Mobile phone number has evolved into a unique identifier, largely due to these four reasons:

1. We're now able to keep our mobile phone number when switching phone companies.
2. Interstate long distance charges have pretty much disappeared.
3. Large metropolitan areas now have overlay area codes so we have to dial the full 1 + (area code) + phone number even if we're calling within our own area code.
4. Company-issued phones are now rare since most people don't want to carry multiple phones.

So now when people move to a new job or even move across the country, our mobile phone number stays with us and it's becoming part of our identity. Deduping based on mobile phone number can help you identify contacts across different company affiliations and contacts with different email addresses.

Before you can dedupe records based on the phone number, however, you should normalize the phone number format first. We recommend normalizing all your phone numbers to the international format.

Domain

If Company is one of the data fields to dedupe on, whether for Account record or Contact record deduping, consider using domain matching first because company names can be tricky to match on, which we'll cover next. "Domain" is a more exact way to match companies. For readers not familiar with what a "domain" is, here's an example:

Website: www.usa.acme.com

Email : jdoe@usa.acme.com

The full domain for the above website and email address is "usa.acme.com"

The root domain for the above website and email address is "acme.com"

Some of the additional considerations when using domain as a dedupe field include:

1. In most cases, root domain is the best matching option. The full domain is suitable if you wish to keep divisions of large corporations separate as different accounts.
2. You can extract domains from both email and website using a data automation tool.
3. Before you extract a domain, it's best to clean up the email and website data so that you don't end up extracting the domain "acme.con" from a bad email address "jdoe@usa.acme.con" with an invalid suffix.
4. Filter out email addresses from ISPs (Internet Service Provider), free email providers, and email anonymizer services. Openprise provides a list of these email domains in our Open Data Library.
5. A company can own multiple domains and the domains used for website and email maybe different. Data providers like Dun & Bradstreet and Orb Intelligence can append your company master record with additional domains.

Company Name

For Account record dedupe, Company Name is usually the secondary match field after domain. Company Name is often involved in deduping Contact records as well, because your Contacts can be affiliated with multiple companies.

Before you use Company Name as a dedupe field, it's best that you do the following first:

1. Clean up the company name. Instead of trying to match on "Toyota Motors USA Corporation", "Toyota Motors (USA)", "Toyota Motor U.S.A.", "Toyota Motors USA Corp.", your match rate can drastically improve if you clean up all these names to and standardize to just "Toyota Motors USA".
2. Normalize the company name across its alias. For example, "Toyota Motors USA" may also appear in your database as "Toyota Motors Sales", "Toyota Motor Sales USA", and "Toyota Cars". Normalizing them all to "Toyota Motors USA" is best. For companies with multiple divisions like "Toyota Forklifts" and "Toyota Financial Services", you need to further decide if these business units should be treated as an alias of the parent account, or separate accounts.
3. Consider using a data service to normalize the name or use a unique identification code like a DUNS number as the golden standard.

Address

If you would like to dedupe on address, you must first clean up and normalize the address. There are just too many different variations of address formats in your database. Pick one of the mapping services as your golden standard and be consistent, whether it's Google, Bing, Here, TomTom, or USPS, etc. Run your address data through these services to fill in gaps, correct mistakes, and standardize on format.

How Many Dedupe Fields to Use?

There is no universal correct answer to this. It depends on your business and your database. That said, here are a few things to keep in mind:

- Use as many matching criteria as you can, even if some of it may only yield incremental results. The only incremental cost is processing resource and time. But with the right data automation technology, the incremental cost is trivial. For example, you may try all of these matching criteria on a contact record:
    - Email
    - Mobile phone
    - First name + last name + company name
    - User ID
- Matching on a combination of data fields may be required if matching on a single field doesn't provide sufficient uniqueness. A couple examples:

- A contact maybe affiliated with different companies in different roles, such as a broker or a service partner.
- After a contact has moved to another company, you may want to preserve the old contact record to properly archive the historical data associated with the opportunity, which allows for more accurate analysis of win/loss and ideal customer profile.

To Fuzzy or Not to Fuzzy?

Fuzzy matching can be a very powerful tool when it comes to identifying duplicate records. However, no machine algorithm is perfect, so anytime you use fuzzy matching you're going to be trading off between false positives and false negatives. False positives are incorrect matches found. False negatives are matches that were missed. The general rules of thumb are:

- The more you can clean up and normalize your data, the less you will require fuzzy matching. So clean up your data as much as you can before deduplication.
- Experiment with the fuzzy factor to decide which configuration yields the best tradeoff of false positives vs. false negatives that fits your business process.
- If you are short on time, start with exact matching only, then introduce fuzzy matching and gradually increase the fuzzy factor. This is the most conservative approach.
- There are different fuzzy matching algorithms. Some algorithms are better suited for certain types of data. If your data automation tool provides different algorithm options, experiment with them.
- You may apply different fuzzy algorithm and factor on different data fields within the same matching criteria, for example:

FirstName with fuzzy = 0.6

AND LastName with fuzzy = 0.8

AND Domain with exact match

5. Determining the Surviving Records

Once you've identified the duplicate records in your MarTech databases, the next step in the dedupe logic is to identify the surviving/winning records. These are the records you will keep. The other records are the non-surviving/losing records. The non-surviving records will be merged into the surviving records or simply discarded. Determining the surviving records is the most complex part of your dedupe logic.

Peel the Onion Logic

When coming up with your surviving logic, you'll feel as if you were peeling an onion (and we don't mean the "tearing up" part). Every resolution logic step you write down leads to another question. Here's a very typical example of figuring out the surviving logic in a Marketo Leads and Salesforce.com Contacts + Leads dedupe project.

If you have both and Leads and Contacts within a group of dupes, then the Contacts should survive.

1. If there are more than one Contact within the duplicate group, then the Contact that is associated with an Opportunity should survive.
2. If there are more than one Contact dupe associated with Opportunities within the duplicate group, then the Contact record associated with the Opportunity with the most advanced stage should survive.
3. If there's no Contact within a group of dupes, then the Leads that have signed up for a free trial should survive.
4. If there are more than one Lead within a group who have signed up for the free trial, then the Lead who has completed certain tasks within the free trial should survive.
5. If no Lead has signed up for a free trial, then the Lead from the most trusted lead source should survive, based on a ranked list of lead sources.

As you can see, this can get pretty involved quickly.

It's All About Your Logic

Now you can see that there is no such thing as a proprietary or secret sauce algorithm any technology vendor can provide that can magically figure out which one of your records should survive. Every company's logic for this is different, depending on how it conducts its business and what its data sources are. There is no way around this. You must document your own surviving record logic, then you need a flexible technology to execute your logic.

We often hear people say there is no consistent logic in some cases because it involves human judgement. We always challenge that claim. Human don't make random decisions. When a human is making one-off dedupe decision between a set of records, she is applying some consistent logic in her head, whether she realizes it or not. Document that logic.

Test and Iterate in a Safe Environment

Your initial logic is likely to have gaps because you're not done peeling the onion yet, which is OK. Come up with the most complete logic you can think of, then test it. Make sure the deduplication technology you use can support testing outside of your system of record, like Marketo and

Salesforce.com. In order to come up with the complete deduplication logic, you'll need to go through at least a few iterations of:

- Running your algorithm
- Reviewing the dedupe results
- Making adjustments to your dedupe logic
- Rinse and repeat

This type of iterative development and testing is best done within your data management tool or a sandbox. Update your system of record only after your dedupe algorithm has been fully tested.

6. Clean and Normalize Before Deduping

You can't just jump into a deduplication project with a dirty database. A dirty database can greatly hinder how well your dedupe logic performs. Cleaning and normalizing the data fields involved in your dedupe logic is highly recommended. For example:

- Clean up bad email addresses like "jdoe@acme.con" so it will match with "jdoe@acme.com"
- Clean up company names like "Acme Corp." and "Acme Corporation" so they will match
- Extract domains from URLs and email like "acme.com" to use as matching criteria
- Normalize phone numbers so that "415.555.1212" will match with "+1 (415) 555-1212"
- Normalize lead source names so "Dreamforce 2016" and "DF16" will match
- Clean up and remap old status values like Lead and Opportunity status

You May Need to Integrate More Data Sources

In the example above, you see that in order to execute that logic sequence, you'll need more than just your Lead and Contact data. Specifically:

- Opportunity data from Salesforce.com
- Opportunity Contact Role data from Salesforce.com
- A ranked list of Opportunity stages from Salesforce.com
- A ranked list of Lead Sources from Marketo or Salesforce.com

In addition to other data sets from your Salesforce.com and marketing automation platform, you may even need data from other systems like help desk, product database, and finance systems.

If your deduplication logic requires data from other data sets, you'll need a data integration tool to pull the data together. A data automation tool like Openprise combines integration, cleansing, normalization, and deduplication capabilities all in one, which can greatly simplify your dedupe project and save money on multiple tools.

7. Merging Duplicates

Once you have identified the duplicate records and figured out which surviving records to keep, the last part of your deduplication logic is to merge the non-surviving records into the surviving record. In some cases, you may want to simply discard the non-surviving records. That simple scenario requires no further discussion.

First Establish a Default Logic, Then Add Exceptions

Chances are you have more than a handful of fields in the data you are looking to merge, perhaps hundreds, and we have seen even thousands. In order to scale, you should first establish a default merge logic that will be applied to all data fields. Once you have a default logic, you can then define exceptions for specific data fields. The most common default logic is “fill if empty”. We will discuss the various merge logics next.

Types of Merge Logic

Fill If Empty

This is the most common merge logic, thus the most popular default logic. This logic says that if any data field in the surviving record is empty, then attempt to fill it with a non-empty value from one of the non-surviving records. You also need to provide additional logic on what sequence to sort through the non-surviving records. Here is an example of three records in a duplicate set, with the non-surviving records sorted with the more recently updated records on top. The merge logic is fill if empty using latest modified record.

Surviving Original	John Doe	jdoe@acme.com		
Non-surviving 1	J. Doe	jdoe@acme.com		VP Marketing
Non-surviving 2	John M. Doe	jdoe@acme.com	Acme Inc.	CMO
Surviving Merged	John Doe	jdoe@acme.com	Acme Inc.	VP Marketing

Always Replace

This is exactly the same logic as the one above, except it doesn’t require the surviving record data field to be empty. It applies the merge logic to all the records in the duplicate group, including the surviving record, picks the value that meets the requirement, then replaces the value in the surviving record, empty or not. Common examples include:

1. Always take contact information from the last modified record
2. Always take lead source from the oldest record

Here is an example of three records in a duplicate set sorted by latest modified date on top. The merge logic for email is to use the latest modified date. The merge logic for lead source is to use the earliest modified date. The default merge logic is “fill if empty”.

Non-Surviving 1	J. Doe	jdoe@acme.com	Acme Inc.	Webinar
Surviving Original	John Doe	jdoe@looney.com		Dreamforce 16
Non-Surviving 2	John M. Doe	jdoe@tunes.com	Tunes Corp.	Free Trial
--				
Surviving Merged	John Doe	jdoe@acme.com	Acme Inc.	Free Trial

Append

With most merge logic you are throwing away some data you believe is not as good as the ones you are keeping. In some cases, you want to keep them all. This is common with unstructured data like notes or multi-value categories and segmentation data. For these data fields, use the append logic. Here’s the same example above, but instead of keeping only the earliest modified lead source, we want to append lead source.

Non-surviving 1	J. Doe	jdoe@acme.com	Webinar
Surviving Original	John Doe	jdoe@looney.com	Dreamforce 16
Non-surviving 2	John M. Doe	jdoe@tunes.com	Free Trial
--			
Surviving Merged	John Doe	jdoe@acme.com	Webinar, Dreamforce 16, Free Trial

Based on a Formula

For numerical or binary data fields, it often makes sense to apply a mathematical formula, such as:

- Pick the maximum or minimum value
- Calculate a sum or an average value

- Set to “true” if only all records are true

Here’s the same example above with two numerical fields: behavioral score and demographic score. The merge logic is to pick the highest demographic score, but sum the behavioral score.

			Behavior	Demographic
Non-surviving 1	J. Doe	jdoe@acme.com	15	100
Surviving Original	John Doe	jdoe@looney.com	10	10
Non-surviving 2	John M. Doe	jdoe@tunes.com	50	50
Surviving Merged	John Doe	jdoe@acme.com	75	100

Do Not Merge

This one is simple, for some data fields you just do not want to merge.

8. Manual Review and Overwriting Results

So far we’ve covered all the required planning, setup, and automated de-dupe steps, including the first three:

1. Identifying the duplicates
2. Picking the surviving records
3. Merging the non-surviving records into the surviving records

Frequently, the next step is to manually review and overwrite the automated dedupe results. We have a very clear and strong position on this. Beyond the initial setup validation and some exceptions discussed below, we believe manually reviewing dedupe results is completely unnecessary and a waste of resources, and hopefully we can convince you of the same.

When Manual Dedupe Review is Not Required

If There Is Logic to It, You Can Automate It

The most frequently given justification for why people insist on manually reviewing dedupe results is that they want to introduce a human decision to the process. Our question to that response is always, “What consistent logic is that human using to make those decisions, or is it just random judgement?” Almost always, when a human reviews and overwrites the automated dedupe results, he or she is applying a consistent set of logic. If the automated de-dupe results require a large amount of manual correction, then this is due to one of two causes:

- The automated dedupe logic is incomplete, and/or
- The data is so poor that a human has to do additional research and append the data to in order to make a decision

If a human reviewer is applying additional logic that the automated dedupe algorithm isn't using, then the simple answer is to document and incorporate that logic into the algorithm.

If the data isn't good enough to support the necessary logic, then we recommend you append the missing data first by either using third-party data providers, or data from your other systems. It might sound counterintuitive why you should spend money to append data that you may throw away after de-dupe, but if your data quality is poor and unable to support your de-dupe logic, then you may end up keeping the wrong data and throwing away the good data. That can be more costly than the money and effort spent on appending your database before deduping.

If There Is No Logic to It, It Won't Make Any Difference

If a human reviewer is indeed making ad-hoc, intuitive, random, eye-test decisions, then you're better off *not* doing it because the net result will be worse, and you would have wasted precious human resources while gaining nothing, if not causing additional damage to your data.

While it's true that on any one specific record, a human action could potentially make it better, there is also equal chance that the human action could make it worse. If the human decision is indeed "random" because the claim is that there is no consistent logic, then statistics say that over a large enough data set, which most marketing databases will fit that bill, the positive and the negative impacts of the human action are a wash. With 10,000 records or more, the net effect of the human action is most likely ZERO. It's just like if you were to flip a coin enough times, you get 50/50 heads/tails.

There are more valuable tasks you can use your human resources for than doing a painful task that yields zero results, and that your human resources are guaranteed to hate doing.

It Will Not Scale, Even If It Gets Done at All

Manual dedupe review is such a painful and slow process that most people will procrastinate, and procrastinate, and procrastinate. They will do anything else first if they have a choice. Any dedupe project that requires manual review, especially the ones that involve a large number of people, like the sales team, to participate in will simply never get done. We don't like it, but it's simply human nature. You have two options:

1. Spend the time and effort to “push on the rope” for a long time and never get the project done, or
2. Move the project forward without the manual review and spend the effort to deal with any complaints and re-mediations after the fact.

From our experience, Option 2 is much better. It gets results and the effort is generally less.

When Manual Dedupe Review is Required

Verifying Algorithms

When you set up the dedupe algorithm, you absolutely should review the results to ensure the algorithm is correct and complete. It’s an iterative process because hopefully we have shown you thus far that a robust de-duping algorithm is not a trivial development and almost never as simple as you think at the start. However, the purpose of the review is not to manually overwrite the results produced by the algorithm, but to provide the feedback to improve the algorithm.

When the Data Set is Small and Remediation is Costly

The classic example here is CRM Account records, especially strategic/named accounts. This is a small dataset, typically no more than a thousand records, with ownership distributed to a rather large set of account reps, so each owns about 20 or so strategic account records. Any type of automated deduplication here can create costly (normally not financial, but political) consequences. Whereas the better alternative is to just identify the dupe for the sales team and let the account reps take the manual merge actions for their own small data sets.

8. Legitimate Dupes

Not all duplicate records are created equal and not all duplicate records should be removed. There are many business situations where duplicate records should be kept. We like to call these “legitimate dupes.” Here are some of the more popular cases of legitimate dupes we see.

Brokers and Channel Partners

Many businesses sell through partners. In the Consumer Packaged Goods (CPG) business there are brokers and distributors. In the semiconductor business, there are design partners. These channel partners have multiple relationships with retailers. In the Salesforce.com world, these partners are Contact records that must be associated with multiple Accounts. Until recently, SFDC only allowed each Contact to be associated with a single Account, so one common way businesses use to sidestep this limitation was to create multiple Contact records for the same partner. In this type of situation where you want to preserve the partner-to-customer relationships within the Contact-to-Account construct, make sure you only dedupe Contact records within each Account and not across

Accounts. This is a very broad category covering many types relationships across different industries.

If you decide to keep these partners as legitimate dupes, you should consider “syncing” (vs. merging) these legitimate Contact dupes, so all the duplicate records for the same partner have the same data all the time.

Preserving Historical Context

When a Contact leaves Account A for Account B, should you:

1. Create a new Contact for Account B, or
2. Change the Account assignment for Contact from Account A to Account B?

There is no universal right answer to this, but we do see a clear split in preferences between marketing and sales teams.

Marketing often prefers to reassign the Contact from Account A to Account B. This preserves the engagement history with the person so marketing so that they can properly attribute their campaigns to the revenue that is ultimately produced. Also, this reduces the number of dupes, and marketers hate dupes!

Sales, in general, prefers to create a new Contact for Account B. This is because Sales sees each Opportunity and Account in its own separate context. Any previous conversation with Contact at Account A doesn’t have much relevance to a new conversation with a Contact at Account B. Keeping the old Contact record with Account A also preserves the proper context for Account and Opportunity history.

As more marketing and sales technologies become available to help Marketing and Sales analyze ideal customer profiles, preserving the historical context about how a deal went down, and who were involved at the time can be a powerful justification for keeping historical data that represent dupes of the same person.

If you take this historical preservation need a step further, some might even argue that every time a Contact changes job, gets promoted, or relocates, you should also create a new duplicate Contact. This way you can see that at the time an opportunity was closed years ago the Contact was a manager in marketing, although he is currently an executive in business development.

In Salesforce.com, things get even more complicated if the Contact that left Account A is now just a Lead at a company that isn’t even an Account yet. Of course, you can’t un-convert a Contact in Salesforce. In this case, you have a dupe across Contacts and Leads.

Business Units and Divisions

For Account records, every sales organization has a different take on what a duplicate Account is. There is absolutely no universal right answer here. It all depends on how your sales organization is organized. Here are some common examples of tricky situations:

- **Business units by location:** Toyota Motors USA, Toyota Motors Canada, Toyota Motors Mexico
- **Conglomerate business units sharing the same name:** GE Healthcare, GE Transportation, GE Appliances
- **Conglomerate business units not sharing the same name:** Alphabet, Google, Waymo, Verily
- **Keiretsu:** Mitsubishi Steel, Bank of Tokyo, Meiji Mutual Life, NYK Line

One common example that is related to this is ship-to vs. sold-to addresses being modeled as different Accounts.

9. Merge Blockers

When you merge duplicate records, depending on the system you are working with, there can be many situations where merging will not be allowed. Here are the most common ones we see in Salesforce.com and how to remediate them.

Merging a Lead with a Contact

You can't merge records of different object types. Although Leads and Contacts are both fields about people, they are separate objects in SFDC.

The remediation is to convert all the Leads to be merged to Contacts first. During the conversion, make sure to assign the Leads to the right Accounts. In addition, confirm that all the required fields for the Contact objects are filled in. Otherwise, the Lead-to-Contact conversion request would be blocked as well.

Merging Contacts with Other Contacts

Often times, duplicate contacts may exist in different Accounts, especially if they're created by different owners. However, Contacts in different Accounts cannot be merged.

If the Accounts are actually duplicates, then the answer is to merge the Accounts first, which will merge the Activities, Opportunities, and Contacts as part of the Account merge operation. However, if the Accounts are not duplicates, then the only way to merge the Contacts is to first move the Contact to the same Account before merging.

It's important to validate the duplicate identification logic with Contacts. Simply using Email is insufficient. At the very least, Email and Account are both required.

We always recommend that Contact owners review these duplicates as well, as Contacts are highly sensitive records for the sales team.

Invalid Record Owners

Very often a record in Salesforce can be owned by a user that no longer exists. If you try to merge a record assigned to an invalid user, it won't be allowed.

The remediation is to ensure all records are assigned to valid users. This is best handled as a separate data quality maintenance process, independent of the deduplication process. However, if necessary, search for all invalid users and update the ownership to valid users.

Contacts without Accounts

In Salesforce.com it is possible to allow Private Contact records. These are Contacts not affiliated with an Account. Any attempt to merge a Contact without an Account to one with an Account will be rejected.

The solution is to ensure that all Contacts to be merged have an Account affiliation. This will require an additional step to assign all Contacts without Accounts to the same Account for the surviving record.

This is essentially a generic problem of invalid dependency that can occur when deduping other objects. For example, an Opportunity record must be associated with an Account.

Invalid or Missing Field Values

If the field value for the record is outdated, containing a value that is no longer part of the current picklist, any attempt to merge this record will also be blocked.

The solution is to ensure all field values are valid with respect to the current allowable values. This is best handled as a separate data quality maintenance process, independent of the deduplication process. However, if necessary, simply update the field with the valid value.

A related case of this is missing required field values. This usually happens when a required field is added after the record creation, and a default value isn't retroactively assigned.

The answer is similar to above, either keep up the data quality with a separate process, or just fill it in before merging.

Violation of Validation Rules

It is important to understand the validation rules that are active within Salesforce.com. Merge operations can fail due to validation rules that aren't related to required fields. Since validation rules are unique custom rules in each environment, it’s important for you to review the active rules and determine if merging the various objects would violate any of them.

10. Five Key Considerations for Effective Deduplication

If you have managed to read this far, you are now a de-dupe expert! To bring it all together, let’s highlight five key takeaways.

1. Stop the Bleeding First

Proper resolution of duplicates may require involvement of teams and change of processes that can take a long time to execute. Before you try to fix the existing bad data, look at what it will take to prevent the problem from getting worse. In other words, “stop the bleeding”. It can often be much easier and quicker to implement, without involving as many stakeholders. This will enable you to produce quick and valuable results to help rally the troops to support the bigger project of cleaning up existing data.

Procrastination and analysis paralysis only lets the duplicate problem build up and become increasingly more expensive and time-consuming to fix.

2. The Hard Parts are the Surviving Logic & Merge

Deduplication is not a simple and trivial task, despite what your solution vendors may say. Vendors like to talk about how comprehensive their duplicate identification algorithm is, but that’s the easiest of the three parts. The surviving logic and merge process are the way more difficult steps in comparison. Too many deduplication projects start with a bang but end with a whimper because of system and process issues when it comes to executing these two later steps.

Make sure you select a vendor and a technology that can actually handle the entire end-to-end process, and not just the identification part. You will need not only a tool, but deep system knowledge for the applications you use.

3. One-Time Project vs. Continuous Process

Deduplication often ends up being a one-time project. This is usually because of the complications and efforts required to handle the surviving logic and merge steps. Unfortunately, when you don’t have the right technology and vendor for the job, these two steps frequently end up being manual, which makes it feasible only as a one-time project. The benefits of making deduplication a continuous process are obvious. Making it a reality is quite doable as well. If you pick the right solution and vendor for the job, and put in the effort into the first bulk deduplication project and properly capture and implement your business logic and process, then you can simply keep the automated processes running. However, if you don’t put the necessary effort into making the process automatable, and decide to take the shortcut of relying on a manual resolution, then you’re simply putting in just a short-term fix and will watch your data deteriorate and must endure another large deduplication project down the road.

4. People > Process > Data > Technology...and in That Order

Duplicate records are caused by gaps and poor alignments between people, process, and technology. To solve the duplication problem, you first have to understand the root causes thoroughly, so you can design and implement the appropriate remediation. Otherwise, you'll fail to solve the problem, if not worsen it.

5. It Involves More Than One System

Most readers of this blog are likely working with CRM and Marketing Automation platforms. These systems are joined at the hip. Data is synchronized at some level and interactions between these systems can be complex. For such integrated systems, deduplication is fundamentally a cross-system problem. You must have stakeholders from both systems buy in and your solution must address both systems simultaneously, taking into account the constraints from both sides and understand the interactions and side-effects.

About Openprise

Openprise is a Data Orchestration Platform. We solve the garbage-in/garbage-out problem to make data-driven anything possible in Marketing, Sales, and Support. The Openprise automates critical data management processes including data onboarding, unification, cleansing, and enrichment. Openprise is designed from the ground up for CRM, so it has the business rules, best practices, and data built right in, and it seamlessly integrates with CRM solutions like Marketo, Eloqua, Pardot, Desk, and Salesforce, so you're up and running fast. For more information, please visit http://www.openprisetech.com/.