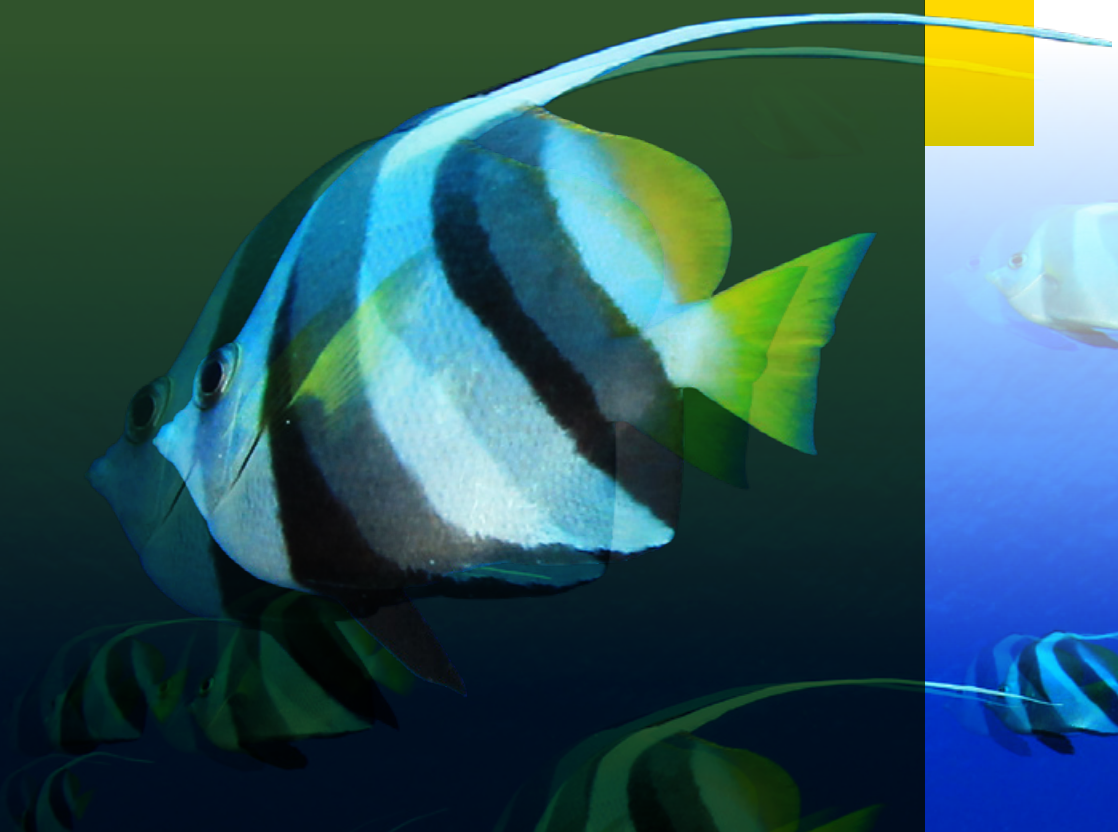
HOPEWISER

The Dark Art of Deduplication.



Content

INTRODUCTION

Quite often we don't understand how many duplicates are within our customer data until we try to move the data between systems. If this isn't sorted straight away then a number of issues will occur. As the old adage says Rubbish In, Rubbish Out.

Duplicates can be disguised in a number of ways, the most visible form being a direct carbon copy of another record. These are the easiest to spot.

However the most common (and harmful) type of duplicate data is a partial duplicate. These are records that may contain the same name but the address is just slightly different e.g., Warrick Avenue instead of Warwick Avenue or the same address but different people, Jo Bloggs and Joanne Bloggs.

Sometimes partial duplicates are created by human error, for example; when a customer or member of your team enters information by hand. Someone may inadvertently create a new record for a prospect, customer, or company that is already contained in your CRM or other database.

Other times it is because you have disparate data silos within your organisation with different data collection points and methods.

Either way duplicate data can have a massive effect on the success of your Sales & Marketing campaigns. Here are just a few of the pitfalls of duplicate data;

1. Waste of your marketing budget. For a database of 50,000 records, if 5% were duplicates then based on delivering 2.5k unnecessary brochures at a cost of £4 per brochure, the waste for each marketing campaign is £10k. And that's not mentioning the damage to your brand reputation if multiple brochures land on your customer's doorstep.

2. Hinders personalisation. Your campaigns may pull important campaign data for personalisation from the wrong customer record, killing your conversion rates and burning through your ad spend.

3. Lost Productivity. Spending time sorting through your data and dealing with duplicates will lead to lost productivity. Editing data by hand is a time-consuming task, often taking days or weeks in large databases.

4. Inaccurate Reports. Duplicates held in disparate databases eliminates your Single Customer View which results in an incomplete picture of what an average client really spends? Which then reduces an understanding of upsell/ cross-sell potential.

Given all these pitfalls then doesn't it make sense to cleanse your data regularly, in particular when migrating from one system to another? This paper demistifies the dark art of deduplication, which will help you make an informed decision on how best to proceed with your deduplication task.

DEDUPLICATION - AN ART OR A SCIENCE?

We all think we know what duplicates are. But do we really? Yes we know duplicates are records that reference the same information, held as two (or more) different records. But as many of us have found to our cost, when dealing with data, most duplicates escape the net because they aren't identical.

More likely it is data entered in slightly different ways, which then makes it more difficult to identify. Mistyping, mishearing and misreading all contribute. Some people use the shortened version of their name sometimes, but not always. Some use their middle name, but will sometimes use their first name. Plus, vanity elements of an address, such as housename and locality are sometimes used, but not always.

Adding to this, some systems enter data in a flexible non structured way and some allow data to be added twice, due to offers for new customers, or by mistake, when they have forgotten they have already entered information.

In addition to these problems there are also many ways to hunt for duplicates. So given all the above, which way is right for you and how do you choose?

First you need to access which items you normally have available within your data. For a B2C company that is - Name & Address, an email & telephone. For a B2B company if might be Name, Address, Email, maybe a phone number, but this could be a direct line, reception and/or a mobile.

Then you can choose from the different methodology available:-

1. Basic sorting on fields to position information together

This is the very start point of any deduplication. Items have to be 'moved' together to help locate the matching ones. This will help find items that are exactly the same, depending on what item(s) were used to sort on.

Going further to help find more duplicates, then white spaces could be removed before sortation to help identify duplicates where an extra space is added, such as Pear Tree Cottage and Peartree Cottage.

However, if the information has been added to different fields, then this will not bring those records together. For instance, Flat 1, The House compared to Flat 1 The House. Also, if two records have different information, then they will not be sorted together, such as Hopewiser House, 1 High Street compared to 1 High Street.

Clearly items that are misspelt or are mistyped will not be found and matched together.

2. Phonetic/Soundex Matching

This can be performed across the whole information.

Phonetics/soundex generally creates a four character key for each word to match against, based on certain criteria, such as ignoring certain letters (e.g. vowels).

This is useful, especially when using limited amounts of data, but taking a full address, forename and surname, does create a lot of information. If there are extra words, then the algorithm to find the matches become more and more complex and requires a lot more scanning of all the data.

For instance, if one record had a housename, whilst the duplicate record did not, then immediately the sets of keys are offset. Or in the case of the Pear Tree Cottage and Peartree Cottage you get different amounts of keys, because a key is created for each word.

In general, phonetic matching lets you search a list for names that are phonetically equivalent to the desired name, e.g. Chris and Kris.

However, at the same time Robert and Rupert are the same when using a soundex routine. Therefore this system cannot be fully relied upon.

3. Elemental matching

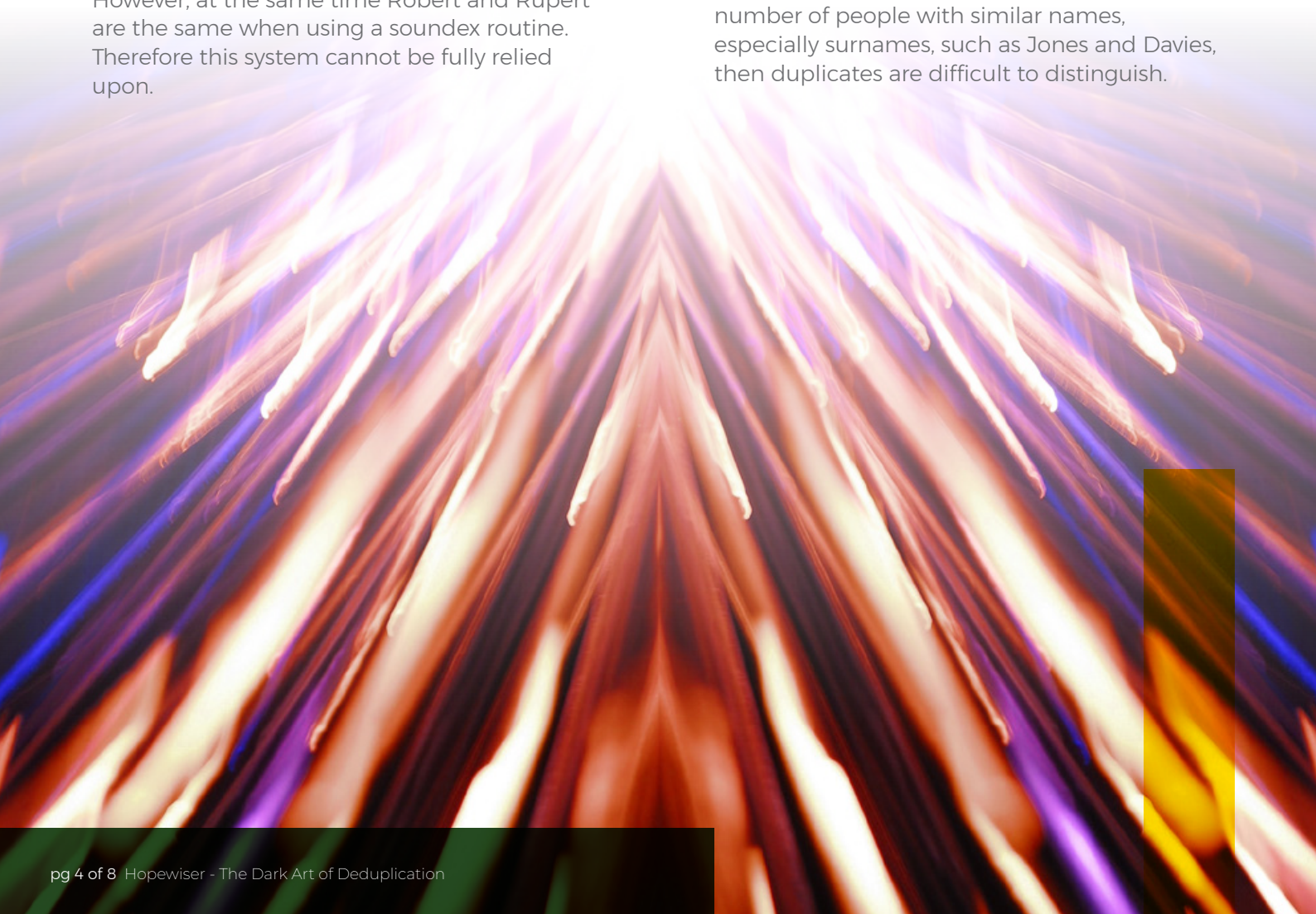
This looks at certain elements of the data, such as just taking postcode and premise. However, if there is a typo or fields put in the wrong place, then this will not work.

Just taking a postcode - a simple typo can make things very different: AB is the area code for Aberdeen, whilst BA is Bath, very different places and very far apart, but the same two characters with a huge number of postcodes the same. This is an extreme example, but worth considering.

Also, if the data is not up to date, then two records may be for the same premise, but the postcode may have been changed by Royal Mail. At one point Aberdeen was recoded by Royal Mail 3 times in 18 months.

Another example is when looking at Surname and Forename, with Postcode. As our culture becomes more diverse and people are not sticking to standard spellings of names, then misspellings on names or putting in surname and forename incorrectly is increasing.

Also where small communities have a large number of people with similar names, especially surnames, such as Jones and Davies, then duplicates are difficult to distinguish.



THE HOPEWISER DIFFERENCE

Keys Based Matching:

A keys based matching process uses some sort of defined key or keys to help find the potential duplicates. In an ordinary database, a simple key is often added to each record for indexing purposes, which is just a number. Clearly, there should never be a duplicate of this number... Alternatively, a key could be a simple identifier, such as the Unique Delivery Point Reference Number (UDPRN), from the Royal Mail, which will pull data together, as long as the address has been matched (correctly).

However, there are a myriad of other ways to build a key or multiple keys to help identify duplicates, such as taking the postcode + premise. Any company or person building a keys based matching system needs to identify the elements they require in the keys, then how they want to process those elements (in full or reduced/standardised forms).

A key is a simplified version of the information, able to unlock the full record, which should be built in a consistent way.

So how do Hopewiser do it? Their Deduplication is based on all of the above, using a Keys Based Matching system. By using keys, it reduces the amount of data you are comparing, which is especially relevant in today's 'big data' world. The keys are not based on soundex, but have similarities. However, by creating keys with data fixed in certain positions, with some elements taking the best bits of soundex, then it allows the user to sort, potentially on elements, comparing relevant components of the data, rather than juggling huge sequences.

This Keys Based Matching allows for different types of deduplication, such as bringing together everyone at a single address for a brochure mailing, or at an individual level, if you want to create a single customer view using disparate databases.

Hopewiser concentrate on the address as the main key, since this is a core component of having something delivered, gaining credit or being insured.

Here are some examples of Deduplication scenarios using Keys:-

1. Household Marketing Campaign

A holiday brochure landing at an address. This is an expensive item to post out, but does not need to be individually addressed. However the client wants to make sure that only one brochure goes to each household.

Hopewiser recommends using as large an address "Key" as possible. This allows the process to look at all the elements of the address including sub premise level e.g. an apartment number, because you want each individual apartment to receive a brochure.

In this instance, you would use "Address Key" and "Extended Key". The address key contains the fundamentals of an address: town, street and premise. The extended key contains postcode and sub premise information.

An example of this approach and a demonstration of how a Keys Based Matching system might work is as follows;

Flat C,34 Forth Avenue,KIRKCALDY,Fife,KY25PS
Becomes: 00200745FRTHVN 0034, F C KY2 5PS

Apartment C,34 Fourth Ave,KIRKCALDY,KY2 5PS
Becomes: 00200745FRTHVN 0034, F C KY2 5PS

So while the addresses were spelt differently, used different abbreviations and spacing, they can be reduced to matching keys to identify the duplication.

Running this process would result in Unique Records, Parent Records and Child Records (which are the duplicates of the Parent Record). For this campaign the client would keep all the unique records and parent records. Eliminating the child records allows for a deduplicated dataset for use in the brochure mailing.

2. Private & Confidential Correspondence

In our second scenario, a client wants to send a personal letter to an individual. Hopewiser would use the Address Key and Name Key. The name key is generated based on an understanding of the surname, the forenames given, title and analysed to see if a gender can be identified.

This is done by analysing the "titles" within a database such as Mr, Mrs, Ms. However in the case of Dr. or other professional titles, additional analysis and checks will need to take place.

Now the client needs to decide how imperative it is to check everything and how important the communication is. For example you find Mr J Bloggs, 1 Church Lane, Manchester, M1 2WA and Joe Bloggs, Church Lane, Manchester, M1 2WA, both male, you can make an assumption it is the same person. However some clients need to check for certain, particularly in the Finance sector where the client is being sent private banking details.

Therefore it is difficult to create hard and fast rules. Hopewiser work with every client to really understand their data and can then create rules around them.

3. Creating a Single Customer View

When migrating data from disparate silos, the data has to be checked for duplicates. This is a fine art, as you need to be careful you don't merge records that are not duplicates. However, you want to find all the relevant duplicates across the databases. Hopewiser would run a "first pass" of the data using an Address Key, Extended Key and Name Key to find all the records that "look" the same across all elements.

Hopewiser runs further 'passes' by relaxing the rules of one or more of the keys. For example Flat C and Flat 3 may then be flagged as a potential duplicate because every other element of the address is the same (name and address details).

This enables the client to work through these duplicates to ascertain whether they really are the same record. This is very important when trying to create a single customer view. Merging two client accounts that are not the same could be disastrous. However merging the "real" duplicates and having only one "account" for that person could mean they get more loyalty rewards, cohesive information when logging into their account and better customer service. It will also enhance data analysis and save significantly on costs for the organisation.

In addition email addresses change regularly and people generally have multiple ones, using them for different purposes, which is why this cannot be relied upon. However, a pass could use the email address to see whether it brings back clear duplicates. Other fields could also be used, depending on what you have in each database.

Given the current legislation for someone to have the right to be forgotten, then understanding where that person's information is held through-out an organisation is key.

4. Identifying Fraud

This is where Hopewiser would relax the rules of the keys from the start. By relaxing the rules, more duplicates are found. This means a client can spot patterns of usage at an address. For example many fraudulent applications have only slightly tweaked details to escape the net. For example the culprit may use the same name, same street, town, postcode but a different house number. Or the same home address, but multiple different names. By relaxing the rules and looking at "possible duplicates" it can highlight properties that are continually requesting credit.

5. Deduplicating a B2B database

Company names are regularly seen differently by people compared to the official name e.g. Sainsbury's, rather than J Sainsbury's PLC, Smiths, rather than WH Smiths, or companies have different types of store that are categorised by name, such as Tesco, Tesco Superstore, Tesco Express. Therefore, there are options on whether to use the whole of the company name to generate the key, just key words or other potentials.

CONCLUSION

With all the above scenarios, once you have defined the keys, then you can determine what elements/keys you want to use to group data together. These are defined as 'passes', so it is possible to do a number of passes of the data, using different elements to group the data together.

For instance, starting with a very defined pass, where the keys have to be the same, should group everything that is fundamentally the same. However, by continuing on and using different elements, such as reducing the number of characters which have to match or only using surname and address, then more and more data can be grouped together.

On the face of it, the basic idea behind locating duplicates is fairly simple, but as you can see, there are a large number of potential ways to do it, even within a keys based system. Depending on the data and the desired outcome, then how you do it to get the best result can be altered and would be complex, if doing it without relevant software tools.

Using Hopewiser's software, allows for easy to use options, but it is about understanding what outcome you want, how much you want to look at manually compared to automatically accepting, to really decide on the route.

Using keys based methodology and software, it is very easy to try a deduplication run, look at the results, tweak what is being looked at and see the results again, without each run taking a huge amount of time.

Deduplication is based on science, but it is definitely a (dark) art, but full of interesting results and data insights that you cannot easily see any other way. It can be a scary concept, especially when looking at large volumes of data, but the benefits are huge and it doesn't end there. Once you have located potential duplicates, then you can decide what to do next. Suppress, group, merge and purge for example.

Hopewiser are more than happy to help discuss and to work out the right path.

ASK

About Hopewiser

Suitable for organisations of all sizes, our address software is fast, robust and easy to use. Our rules-based solutions and services intelligently assess and match each address – that's why we're trusted by High Street Banks, Police Forces, and major Sports organisations.

The support we'll provide is backed by the specialists that have written the software in-house. This means you get access to their expertise. Our customers regularly praise the support team for its speed and problem solving capabilities.

We created the first address data software back in 1982 and were the first Royal Mail Value Added Reseller in the UK. The knowledge we have acquired and the sheer amount of data we have processed in that time, sets us apart from everyone else in the market. So if we do things a little differently from others, it might be because we have information that they don't.

And don't forget customer address data that isn't verified against an up-to-date, reputable source will always lead to some inaccuracies when it's captured by your organisation. So why not also view our Address Validation Services.

CASE STUDY

A council we were invited to present to, by an outside agency, who were creating a central data source of addresses, had a young lady going through all the data manually, by street. A long and arduous process, but one they felt would give perfect results.

They felt our time had been wasted in turning up. However, when we demonstrated with their data the first sets of duplicates, they realised the method being employed was flawed, because not all databases classed streets the same, such as some had blocks of flats as 'vertical' streets, whilst others had these as house names with the flats as sub-premises.

Alongside this, then small courts, industrial estates, etc. were also placed in either street or premise fields. Plus our processes highlighted records that were misspelt, placed in different localities and other issues. Alongside this, the speed with which the grouping had occurred changed their minds about our demonstration.

HOPEWISER