

THIS ARTICLE
ORIGINALLY APPEARED IN

database
marketing

Database Marketing is the only UK magazine that covers the tools and techniques used for both business-to-consumer and business-to-business customer management today. Every month, it addresses critical topics like customer retention, profiling and segmentation, data selection, site location and campaign management through a combination of regular software reviews, articles and opinion. If you want to know more about tools like data cleansing packages, OLAP analysis software and GIS, this is the magazine to read.

Not afraid to mix data warehouses with targeting or statistics with geodemographics, *Database Marketing* bridges the gap between sales, service, marketing and IT to inform both those that work directly with these tools, techniques and data, as well as board level executives that have to decide which systems and services to choose for their company.

Why not register for a free trial copy?

For a sample issue:

Contact 0115 989 5445 or email
info@dmarket.co.uk.

Visit www.dmarket.co.uk for more
information and to register online.

Tim Drye compares the different solutions available for customer data cleansing – online, bureau or desktop software – and discusses which one might work best for you.

Cleansing Solution

Everyone in the business of processing customer address information knows that perfect data is an impossibility. There is no customer database on earth that will ensure that every communication reaches the right person, creates the right impression, and generates the desired response – it is an impossible ideal. Unfortunately many people continue to sell products and solutions that perpetuate the myth.

The truth is that, even if perfectly accurate data were available for one campaign, within a short period customers' details will have changed in some way and their records will need refreshing again. Every organisation has to assess for themselves how close they can afford to be to the ideal. For every improvement in quality, there is an added cost to achieve it, and for every inaccuracy in the data, there is an increased risk of waste or damage to the brand, product or service. There is an optimum point where the cost of waste balances against the cost of quality and this will suggest the best approach for each organisation.

The outcomes

The test file was processed using a variety of solutions and the results indicate wide variation in the

internal processes used. The online cleansing sites are obviously particularly vulnerable to changes in the Internet connection speed, while the installed software would often benefit from hardware optimisation such as extra RAM. We can compare the different aspects of each solution to look at their respective advantages and disadvantages.

PAF match rates: the variations may come as a surprise, and demonstrates how flexible data processing can be. With this flexibility, there comes a need to assess how solutions handle difficult addresses like flats, and how this can be adjusted. Viewing record results can do this, but this fine detail is beyond the scope of this article. Little adjustment is possible in online solutions while you rely on the expertise of the operator with a bureau. It must be emphasised that each solution received exactly the same file. It is also apparent that different solutions have different criteria for PAF matches and improvement – post-coding changes have to be checked carefully as some packages can be rather heavy handed in their treatment of vanity addresses and also substitute incorrect postcodes. There is also the matter of PAF accuracy and recency to take into account.

Duplicates and suppression match rates: again

there is a significant variation in the counts achieved and some of the reasons are discussed below. Note that it is well worth considering the use of different suppliers for separate parts of the job. For example, you can remove merge files and remove duplicates in-house using a package like matchIT before submitting the file for online processing to pick off the suppressions.

Processing time: this varied from a few hours to over a day. With the right hardware, desktop software offers the quickest possible processing and does not require significant time to export and import data. Remember that online services have require little initial setup – both in terms of investment and resource – prior to processing.

Hardware: all the systems were operated from the

| Solutions and suppliers | | | | | | | | | |
|---|--------------------------------|---|---|---|---|--|--|---|--|
| Supplier Solution | Meta-morphix | Experian | UKChanges | Absolute Data | lequalsP Cygnus | helpIT systems matchIT suite | DQGlobal DQMatch | Capscan Matchcode | AFD Software Refiner |
| Type of solution | Bureau | Online | Online | Online | Integrated desktop verification, deduplication and suppression | Integrated desktop verification, deduplication and suppression | Desktop Deduplication | Desktop PAF verification | Desktop PAF verification |
| Sample verification results | | | | | | | | | |
| Matched PAF records | - | 408314 | - | 214440 | 489516 | 477245 | - | 487701 | 478924 |
| Improved records | - | 73781 | - | 269493 | 20131 | 27580 | - | 16467 | 22040 |
| Total verified records | 519099 | 482095 | 461094 | 483933 | 509647 | 505689 | - | 504168 | 500964 |
| Sample deduplication results | | | | | | | | | |
| Duplicates found | 16806 | 6112 | - | 23428 | 12855 | 22564 | 24648 | - | - |
| Sample suppression results | | | | | | | | | |
| TBR match | 2438 | - | - | 993 | 1408 | - | - | - | - |
| Deceased match (all available data) | 4895 | 2290 | 5621 | 2862 | - | - | - | - | - |
| MPS match | 15852 | 15256 | 18135 | 15866 | - | - | - | - | - |
| Technical requirements to process sample | | | | | | | | | |
| Additional technical requirements to operate the solution | High capacity email connection | High capacity internet connection | High capacity Internet Connection | High capacity Internet Connection | None | Prior import into Foxpro data format | A unique reference field attached to each record | Ability to process and format data correctly prior to operation | None |
| Time to process | 24 hours | 23 hours | 7 hours | 11 hours | 4 hours | 23 hours | 25 mins | 2 hours | 27 hours |
| Memory requirement | - | - | - | - | 130Mb | 256Mb | 74Mb | 56Mb | 87Mb |
| Comparison of solution functions | | | | | | | | | |
| Ease of setup | - | High | High | Medium | Low but supported by training | Medium but supported by training | High | Medium | High |
| Level of direct control and customisation of process | - | Low | Medium | Medium | High | Medium | High | Medium | Low |
| Handling of bad record formats | Accommodated | Accommodated | Accommodated and Reported | Reported/ Amended by bureau | Accommodated/ Reported | Accommodated | Accommodated | Reported | Accommodated |
| Expletive screening | Standard | Only available through the bureau service | Not available online | Available but not standard | Standard | Can be user supplied | - | - | - |
| Additional features of each solution | | | | | | | | | |
| | | Easy access to a wide range of customer information, very straightforward and simple desktop interface. Customer electoral roll validation provided as standard | Deduplication is unavailable within the online service, a variety of file splits are made available for exporting the results | A high level of customer service from the technical support team, customer Electoral Roll validation provided as standard | Offers very good control of the file output and multiple file processing, comprehensive reports and data review. Processing speed suffered from lower than recommended RAM in test PC | Comprehensive reporting and interactive management of record deletion and output. Processing speed suffered from lower than recommended RAM in test PC | The ability to integrate the DQMatch capability into existing customer databases | Very fast at processing high volumes of data | Easy to use, point and click setup and processing. Refiner V1 designed for up to 100K records. V2 (release April '03) expected to handle 100K+ records. Deduplication available, not tested. |
| Single user first year cost | | | | | £20,950 | £11,272 | £2995 | From £3000 | £1100 |
| Single user second year cost | | | | | £8950 | £4282 | £599 | From £1500 | £900 |
| Data Processing cost | £6253 | £6117 | £6686 | £3363 | | | | | |

same hardware platform. If internal resource and skills are to be kept to a minimum, then emailing a zipped file to an external bureau will take care of all requirements. Note that bureaux will usually only deal with larger files to cover the additional human and service costs. Online systems and easy-to-use desktop systems provide a useful low cost entry level.

It must be stressed that the more sophisticated desktop software requires platforms with higher specifications to achieve their best results. This particularly affected the match/IT suite; with higher specifications, much improved results can be obtained.

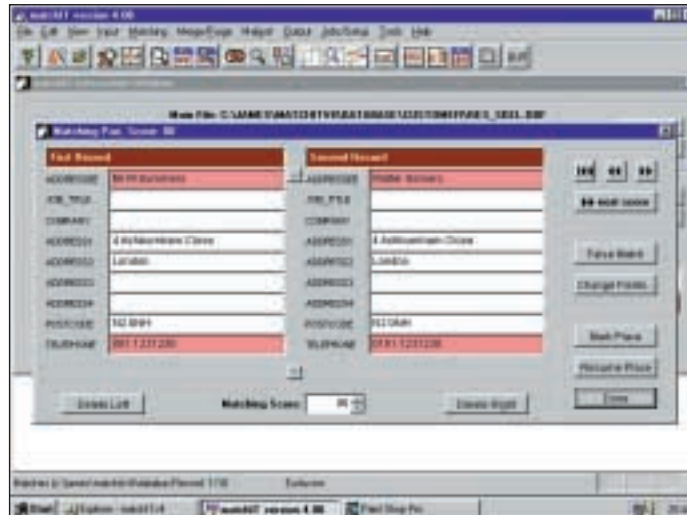
Ease of setup: The criteria here are more subjective but serve to compare the ability of a new user to process a job without making additional changes to settings. This requirement is particularly useful when data processing is infrequent and an additional responsibility for users. Reliable default settings allow very quick set up, though it means sacrificing some control and is appropriate to smaller scale jobs. An easy set up allows less skilled personnel to manage the processing as long as the possibility for lower data quality is acceptable.

Level of control: Larger volume, higher quality processing requires the ability to modify matching criteria, import and export settings and much more to suit the sort of processing and the end application for the data. Here, the integrated packages are particularly strong as long as personnel with appropriate skills are involved. This also applies to bureau outsourcing where both the software and skills should be in place.

Handling bad records: Data can become corrupted in any number of ways. This either prevents a single record from being processed and the remainder of the job continues as normal, or the rest of the file gets disrupted and the job aborts, in which case the file needs to be reformatted to remove the bad records. With low kill levels, it is better to localise the errors and let the job continue. If high levels of quality are paramount, then remedying the faults and re-running the job is better as long as the right staff and software are available to manipulate the data.

Expletive screening: One of the benefits available through the bureau route and with the integrated packages is the suppression of false records: most bureaux have an in-house "swears" file they use for this. This includes profane words and expressions, and also obviously fictitious names like "Mickey Mouse". This is particularly important if the customer data is sourced from the Internet where joke registrations are common. Bureaux are strong here because their wide experience of different formats and sources makes their reference files particularly complete.

Flexibility: Integrated desktop software gives a high level of control over matching and other processing, but also requires investment in internal skills, hardware and reference data licences if you want to do it



The match/IT Verify Matches window provides a clear and easy way to decide which dupes to delete, and to cut and paste between records.

all in-house. Unless you work for a very large company indeed, online solutions and bureaux will have access to a wider variety of up-to-date reference files than you will normally be able to justify. Again, mixing and matching suppliers for different parts of the process is common, perhaps to take advantage of low suppression costs.

Costs: A direct comparison is difficult: online and

The test process

To provide an objective assessment of the different types of solution, a sample file of 527494 consumer records was created that contained a number of errors commonly found in address data. This included duplicates, poor formatting and sample records containing names such as "deceased" or "the occupier", expletives and obscenities, company records and temporary addresses such as hotels.

The file was then processed using a standard 500MHz PC with 128Mb RAM. As available, each solution had to import the data, verify the address against the PAF file and code the address quality, find duplicate records that were present within the sample file and identify those records that matched external suppression files. Suppression files used in the tests were restricted to the MPS (Mailing Suppression File) and deceased registers. These included the Read Group's Bereavement Register and Mortascreen from Smee & Ford. In addition, the Intact site made additional deceased files available: "Deceased Register", "Experian Mortality" and "Registry Trust Deceased".

bureau processing costs for the test file and annual software costs for the desktop solutions are given. Software solutions require an initial investment followed by ongoing licence fees and support irrespective of the volumes of data processed. They also need additional hardware, personnel and other fixed overheads associated with an in-house resource. On the other hand, outsourcing cost is directly related to volume, and so for very high volumes can become very expensive. The key criteria in deciding when to switch from an outsourced to an in-house solution, are: How many records will be processed annually? What value

do the customers represent to the business? What level of data quality is required to support these customers and future prospects? As well as the costs presented this requires an understanding of the internal IT resources currently available.

Assessing match rates

Looking across the table of results shows a wide variation in the different figures obtained, particularly for deduplication and suppression matches. To understand why this occurs, you have to first appreciate the different types of matching that can take place and how choices in the software settings and the way in which the software operates can affect the number of matches. For example, if numbers of dupes are relatively low, this suggests that the process is only matching identical records. If numbers are high then the decisions are based on much broader matching criteria such as fuzzy or phonetic matching.

Different real world goals require different approaches to deduplication. For example, some informative or cold mailings, perhaps distributed for the government or to generate product trials, only need to ensure that identical records are not sent out in order to save money. In this instance any communi-

cation sent to the wrong address is still likely to serve a purpose and will cause little damage to the brand.

But for a sensitive promotional campaign where the organisation wants to inspire confidence in their procedures, for example selling insurance or other financial services, low quality data can damage the overall impact of the communication and the brand. As a result, looser criteria need to be used when matching to similar records.

When looking at cold prospect mailings, I usually recommend a conservative approach when identifying prospects where there is a large amount of available data – if there is any doubt then I would screen the record out. In this context identifying high match rates and duplications are better. However if data is in short supply for a particular niche it becomes more important to understand where the lines are being drawn, and ensure that only those duplicates where a match is particularly good are excluded.

What do you need?

Volume

Where volumes are very high, two main issues need to be addressed. Firstly a system has to be automated to ensure that operations can be dealt with in the available time. It is then a question of how automatic decisions are made; settings for the cleansing process need to be adjustable to the specific requirements of the dataset and its application, for example, whether to dedupe at household or individual level. In addition, the need to export and import large volumes of data can become an obstacle in applications, and particularly for online cleansing. Very high volumes may well require an outsourced solution outside the scope of this review.

Value and quality

The value of your customers and their purchases sets the benchmark for the level of investment in the skills needed for your data processing. At one extreme, for high value niche marketing and a lot of b2b marketing, you cannot trust software alone to make the final decisions about matching data. For example we advised an international watchmaker that sold products worth over £10,000 each that it was worth every record being evaluated manually.

Software was used to provide access to reference data and show a range of possible duplicates, but human operators then made the final decisions. Many b2b communications mean making cleansing decisions involving similar customer values and database volumes. In these situations, the data and customers are valuable enough to warrant investing in the internal skills required to integrate and operate desktop solutions linked directly to bespoke database systems. Other applications cannot take this approach because of the volumes of data involved and also because the

The different options

Normal bureau

Meta-morphix received the data via a zipped-up email attachment and returned the results in the same way.

Online solutions

Experian Intact: a client software package is downloaded from the Intact site which then controls all the interfacing with the Experian servers, for both download and upload of data, export of results and the review of ongoing processing.

Absolute Data & UKChanges: all processing is controlled through a secure website where jobs can be set up, audited and run.

Integrated desktop solutions

Cygnus from lequalsP: a high end desktop solution that gives a high level of functionality particularly in the manipulation of data both on import and output, and offers a very clear graphic interface for workflow management.

Match/IT from help/IT systems: provides a suite of different data processing solutions controlled via standard menus. Based on the Visual FoxPro database.

Standalone desktop solutions

DQMatch from DQGlobal: provides a high level of control for the processing and deduplication of data, and has the capability to be integrated with other data sources to provide verification and suppression.

Matchcode from Capscan: offers a batch processing tool for matching high volumes of address data to the PAF (Post Office Address File).

Refiner from AFD: a point-and-click solution that gives access to PAF verification for small to medium-size companies ■

revenue generated means that the costs of manual methods cannot be justified.

Internal resources

The final consideration in choosing a cleansing solution is the level of resources you want to support internally, and to balance the people skills with the technical capability. For high value data, it can be tempting to invest in technical solutions to apparently replace the need for highly skilled personnel. You can install in-house integrated systems, like Cygnus and matchIT, but remember that these will need data processing experts to run them otherwise decisions can be made that inadvertently scramble the data rather than improve it. As customer data takes a higher profile, finding properly trained and experienced staff is becoming harder and more expensive.

Integrated systems, particularly Cygnus, can handle very high volumes of data, but they do require manual import and export of the data rather than feeding automatically from a source database. You also have the alternative of integrating desktop products like Matchcode and into your internal systems so that you can use them directly from your own applications. The different functional modules of integrated solutions also tend to be available, for example, suppressIT can be bought as a standalone product instead of the matchIT suite if suppression is what you need to do.

In all these decisions it is vital to match personnel

skills to the technical capability. So if you only want to support low levels of skill internally, particularly if you have low value and low volumes of data, then either use external expertise via online solutions or the simpler point and click solutions like AFD Refiner. Low volume, low value data needs a solution that is easy to use, and has a low set-up cost, for which online cleansing sites are perfect. As volumes get bigger and quality issues become more important, then employing traditional bureaux like Meta-morphix gives you access to high levels of dedicated data processing skills.

Keeping your eye on the goals

It can be seen from the results and the discussion above that there is no one solution that meets everyone's requirements. First you need to identify how valuable data quality is to your business and ensuring that the solutions match the value generated. Remember, there is no business where ignoring data quality is the best solution. This is particularly the case with data acquired from the web, wherever you are on the spectrum of requirements, I would urge you to assess your current processes and make sure they are still delivering the best value that is available. A small amount of work could pay dividends in terms of money saved and increased data quality.

Tim Drye is managing director of DataTalk
(timd@DataTalkOnline.co.uk).

THIS ARTICLE
ORIGINALLY APPEARED IN

database
marketing

Database Marketing is the only UK magazine that covers the tools and techniques used for both business-to-consumer and business-to-business customer management today. Every month, it addresses critical topics like customer retention, profiling and segmentation, data selection, site location and campaign management through a combination of regular software reviews, articles and opinion. If you want to know more about tools like data cleansing packages, OLAP analysis software and GIS, this is the magazine to read.

Not afraid to mix data warehouses with targeting or statistics with geodemographics, *Database Marketing* bridges the gap between sales, service, marketing and IT to inform both those that work directly with these tools, techniques and data, as well as board level executives that have to decide which systems and services to choose for their company.

Why not register for a free trial copy?

For a sample issue:

Contact 0115 989 5445 or email
info@dmarket.co.uk.

Visit www.dmarket.co.uk for more information and to register online.