Data Management and Migration

Achiever software

Contents

Data Management and Migration **1**

Introduction ... 3

Identifying and consolidating data sources ... 3

Choose your sources carefully ... 3

Data cleansing ... 3

Mandatory fields ... 4

Duplicate records ... 4

Cleansing procedures ... 5

Using outside expertise ... 5

Cleansing address data ... 5

De-duplication rules and procedures ... 6

Database structure ... 6

Formatting data ... 7

Field drop-down options ... 7

On-going maintenance of quality ... 7

User buy-in and data ownership ... 7

Periodic data cleansing exercises ... 8

Use of third party data tools ... 8

Impact of integrating other applications ... 8

Data ownership between applications ... 9

Legislative considerations ... 9

Data Protection Act ... 9

Freedom of Information Act ... 9

Mailing Preference Service ... 10

Summary ... 10

Appendices ... 11

Further information ... 12

About the authors ... 12

Introduction

In the 21st Century, effective business is all about customer service and to provide good customer service requires efficient database systems and good quality data.

"Any computer system is only as good as the data it holds"

This assertion comes from hard experience and sums up the absolute importance of good data. A sophisticated new database, however much it costs, whether it is a financial application, CRM, ERP or medical records system, is of little use if the data it is providing is not accurate.

Maintaining the quality of an organisation's data is a challenging and time consuming task for system administrators, and requires the constant of users and the backing of managers. However, it is well worth striving for good data quality and, if successful, it can lead to many business benefits that affect the performance of the company and the all-important bottom line.

In this white paper we will look at some of the pitfalls for data quality and consider ways to avoid or at least reduce their impact.

Identifying and consolidating data sources

Although the maintenance of data quality is a continuous task, there are certain times at which there is an increased risk that quality could be seriously affected. The initial loading of data into a new system is one such point and the data chosen to populate a new corporate information system and the mechanisms used to load it can have a large bearing on the eventual quality of the data.

Choose your sources carefully

Very often a new corporate database system will be replacing an existing incumbent system that has be out-grown or is based on dated technology that is no longer fit for purpose. It is only natural that the default assumption is that all of the data in the old system will be transferred into its replacement. This is the point at which it is very important to consider whether this is the right approach.

It is probable that it is appropriate to load a large proportion of the existing data into the new system but not always. It might be that the data is structured in a format that does not map easily to the structures of the new system or it might be that restrictions in the old system did not allow effective data quality management and so the quality of imported data might be low. Whatever the circumstances, consideration should always be given to whether the data should be loaded, if so how much of it and what cleansing and checking will be required before it is imported into the new system.

Data cleansing

There are several types of data problem to which systems are prone. The first can be generically termed "entity integrity". This relates to the contents of a particular record, how complete its data is and the quality of the data that it contains.

© Interactive Software Limited

A contact record is not a lot of use if there is only a surname and email address but no job title, telephone number or organisation details.

Mandatory fields

There is a usually a minimum amount of data required for a meaningful record, whether it is an organisation, an address, a contact, a sales order or a medical diagnosis. Very often these records are not complete and their usefulness is reduced as a result. Setting fields as being mandatory in an application is very useful in the longer term but if data has been imported from a previous system then often there are already records that do not have values in these fields.

This situation is usually addressed by a proactive information gathering exercise such as a telephone data-capture campaign where each contact or company with such missing is data is called to check the accuracy of the data already held and to capture data that is missing. This can be time consuming but the data gathered can be valuable. Once a field is populated on all records it can be set to be mandatory, assuming the application or database allows such controls to be set by the system administrator.

It should, however, be noted the above exercise is usually only successful if it is accompanied by longer-term considerations such as training users to enter appropriate values and ensuring that all users who create such records actually have this information to hand at the time.

Free-text fields that allow the user to simply enter a value, but which are set as mandatory will soon be filled with full stops, spaces and “n/a” values if users are not educated about the business need for this information and the implications of such omissions.

One final point on mandatory fields, do ensure that, where possible, pick lists or drop down lists (lists of pre-defined values from which a user can select) are populated with appropriate values and do not include a “Not specified” or “Not known” option. These are an open invitation for a user to quickly bypass the restriction and move on without providing the required data.

Duplicate records

The most serious data problem with most corporate databases is data duplication. Records that are entered into the database more than once cause havoc with data quality, statistical analysis and user confidence. This situation might have occurred due to user error, lack of validation when integrating with a second application/database, poor data cleansing prior to a data import process.

If a contact is duplicated in a database then users will not know which one should be updated or which historic or related data should be attached. The result is that data will be split between more than one record and there is no single place for a user to look to see the whole picture. Implications could be as serious as an order not being fulfilled, an appointment being missed with a prospective customer, a deadline being missed or even a patient’s condition being misdiagnosed. The real-life costs are often great but difficult to quantify.

Less conspicuous effects could be a reduction in perceived professionalism to customers and prospects. Nobody is impressed if they receive two or three commercial mailings from the same company at the same time, it is symptomatic of a lack of rigour and effectiveness and the tendency is to not even read the mailing. Even more relationship damage can be done if it is a Christmas card that is received more than once. Even though the message inside says “You are special”, the fact that there are two cards with different name spellings simply screams “you aren’t really important to us”.

© Interactive Software Limited

Whatever the cause, duplicated records are an anathema to effective database systems.

Cleansing procedures

When faced with a requirement to load existing data into a new corporate database application there are usually many constraints and conflicting priorities that might influence any decision to carry out a thorough data cleansing exercise. Project timescales, budget, available resource, trying to minimise system downtime, all of these weigh heavily on a project manger's mind when deciding what level of data cleansing can be carried out. It is often tempting to say "Let's get the data into the new system then we can carry out a big data cleaning project". This can be a very costly mistake.

> ***"Going live with a new system and then trying to carry out a data cleaning exercise afterwards is just asking for trouble"***

Firstly, performing a data cleansing exercise gives both the administrators and the users an opportunity to revisit the data being held, to consider whether any parts are no longer required or valid and to think through the data that will be required for the proposed working practices and system usage after go-live.

The transfer of data between systems is the perfect time to intervene and carry out a data cleansing exercise. Typically, the data transition between an outgoing system and its replacement requires a data extract from the old system, some restructuring work and a data load into the new system. The point of restructuring is an ideal time to cleanse the data in each record and perform a de-dupe between the records in the data set.

Using outside expertise

If sufficient experience or resources are not available in-house then it is well worth considering the services of a data cleansing specialist consultant or company.

Cleansing address data

One area of data that can be very effectively cleaned is address data. Addresses that have no postcode can have their postcode generated, those that have a postcode but not a complete address can have their full address generated and address components can be moved to the correct fields within the address, such as ensuring that a town name actually appears in the *town* field rather than address line 2.

There are many benefits of cleaning address data; the simplest is confidence that postal communications will actually reach their destination but in addition, profiling and categorisation are improved, geographic analyses will be more accurate and the consistent address data can be used when matching and deduping contacts and organisations. It is very useful corroborative information in such notoriously difficult matching tasks.

There are also software products that will allow the use of the Royal Mail's official address data (PAF data) to fill in addresses based on postcodes to improve data entry speed and accuracy. This can be seen as an important advantage of a new system when persuading users to "buy in" to a new system.

Once the records themselves have been cleaned up they can be searched much more effectively for duplicates.

© Interactive Software Limited

De-duplication rules and procedures

De-duplication or “deduping” is not an exact science. For small data sets it can be carried out manually although this is a thankless and time-consuming task that is prone to human error. Alternatively, it can be addressed as an automatic process with score-based comparisons. Identical or similar words are given weightings based on their relative importance and the importance of the fields in which they occur, so identical field values are given a high score while similar or “equivalent” words are given slightly lower scores and vaguely similar words are given low scores. Having compared all fields on two records the total score is compared to two user-set thresholds, the “definite match” and “no match” thresholds. Every record in a file is compared with every other record in the file and any record pairs with scores higher than the “definite match” threshold are flagged as matches and reported to the administrator and any record pairs that are between the two thresholds are reported as “possible matches”.

Using such software requires a lot of experience and some understanding of the data. Set the top threshold too high and your software only finds a few of your duplicates. Set it too low and it identifies too many “false positives”. Set the bottom threshold too low and you have a mountainous task of manually checking spurious duplicates.

“False positives” (record pairs identified by such software as duplicates when they actually are not) are a major headache but sometimes unavoidable. A contingency plan will need to be devised for dealing with them if any arise. This is where two contacts, organisations, addresses or other records are identified as duplicates and then merged. If these are address records then the result could be two organisations incorrectly sharing a building according to your database – annoying but not necessarily critical.

However, if it is two medical patient records that are merged then this could be really serious with the medical history and any diagnoses for both people being attributed to one of them. The context of the database is therefore very important when setting the dedupe rules/thresholds and a data specialist consultant can help with these decisions.

On the whole, though, a data cleansing and de-duplication process carried out before data is imported into a new system can really pay dividends in terms of improved buy-in by the users, a smoother UAT (user acceptance testing) phase, easier and more productive training and perceived professionalism and quality improvements for your customers, partners and prospects.

Database structure

Modern database systems usually provide more flexibility than older system when it comes to data structures. It is important when selecting a new corporate information management system that its data structures closely match your business data requirements. This does not necessarily mean that it should match your existing data but that it should match your proposed new data and processes. As already stated, there will often be a necessary data restructuring process when implementing a new system and this is usually to the benefit of all users of the new system.
If your business deals exclusively with contacts as representatives of organisations, generally known as “B2B” (Business-to-business), then your database should reflect this.

If your business deals largely with individual members of the public, freelancers, self-employed consultants or individual patients then your data structures should allow contacts with no organisation and possibly no address. These are generally known as “B2C” contacts (Business-to-consumer). It might be that you need a contact to have many addresses or relationships with many organisations at many addresses. Whatever your business model, your database should support this without difficulty.

© Interactive Software Limited

Although the underlying database can easily support such a variety of data structures, many applications available today are aimed at either B2B or B2C type data and their features do not lend themselves to a mix of both. Be careful when looking for a new system that its preferred model matches your data requirements.

Formatting data

Data formatting can be carried out during a data take-on exercise to standardise data before they are loaded but these rules should also be followed over time in the general usage of the system. The system should make it easy for an administrator to implement such business rules into the system on an on-going basis. These rules might also change over time and this should be possible to implement.

The two main changes that can make a big difference to a system's data in the long-term are the appropriate use of mandatory fields and the careful selection of options for field drop-downs.

Field drop-down options

Available options in field drop-downs are essential to ensuring that appropriate classification can be carried out of the data. Appropriate options must always be available or items will either not be selected (and records not classified) or, if such a field is also mandatory, it will actually be incorrectly classified. This can destroy the quality of such data.

The decision about what are appropriate drop-down options is sometimes made more difficult in an environment involving integrated systems. When a field is shared and used in both systems then the drop-down values made available need to be consistent and appropriate for users of both applications.

On-going maintenance of quality

Poor quality data reduces the confidence of users in that data and, in extreme cases, can lead to the gradual adoption of local department- or team-centric unofficial databases with smaller, focussed data sets that have been manually cleaned by the users and hence have their confidence.

This is the nightmare scenario for a corporate Information Officer or system administrator. Small unofficial decentralised databases cannot be managed, cannot be included in a corporate backup or security policies. Also, being "under the radar" cannot be monitored for compliance with the Data Protection Act or the Freedom of Information Act which leaves the organisation exposed to legal implications.

User buy-in and data ownership

Users must have confidence in the quality of the data held in a system if they are to be expected to use it effectively. Using it effectively also means adding to the data and maintaining the existing data over time.

© Interactive Software Limited

The ultimate owners of the data are the users of the system and they need to recognise that they have a responsibility to take ownership of it and to maintain it. It also means that they need to be empowered to do so. They should be formally given a share of the responsibility for data quality and also the tools and security clearance to enable them to effectively carry out that responsibility.

One common technique that helps users to recognise their responsibilities is the allocation of ownership of each key entity in the database. These entities are usually single records that might be customer contact records or supplier company records or medical patient records. Each has its own rules about completeness, relevance, confidentiality (both within the organisation and externally with partners and suppliers, for example) and users need to be aware of these rules in their business dealings and when working with the records. This individual record ownership helps to focus the minds of users to their responsibilities. It is not a magic wand but it can certainly be a useful tool.

Periodic data cleansing exercises

Some organisations think it is worthwhile to periodically schedule a formal data cleansing operation. These might be annually or bi-annually but are aimed at reinforcing the on-going data quality efforts. They also periodically reinforce the message that data quality is important to the organisation and involving end-users in the work can help to deter them from ignoring the data implications of laziness during subsequent data entry.

Use of third party data tools

The general defences against poor or incomplete data can be bolstered by the use of third party tools. Such tools can integrate with an application or work alongside. Such tools include PAF systems that automatically format and correct address data at the point of entry. It should be noted that PAF tools only need to be implemented. Other automated record duplication search tools can alert system administrators to possible duplicates before situations cause issues.

The use of such tools is not sufficient without the concept of data quality being adopted by users of the system. This battle for quality is more a battle for hearts and minds than for technical solutions.

Impact of integrating other applications

The integration of applications introduces a level of complication to issues of data quality. The most obvious difference is that records shared between applications, for example a CRM and an accounts application or a patient management system and a medical test scheduling system, are used for different reasons in the two systems.

Because of the different uses to which the data is being put in the different applications the fields available and their data quality settings may well be different. A field that is mandatory in a CRM's quotation record may not be required at all when it is sent through to the accounts application for invoicing.

© Interactive Software Limited

Similarly, the drop-down options appropriate for a field on an account applications purchase order line item may not be appropriate when it is passed through as a stock item record into an ERP system's stock control module. Careful consideration needs to be given to such issues by system designers and administrators.

Data ownership between applications

Even when a record has a similar usage in two applications, it is rare that both applications are given equal rights to update the record. It is typical for one application to be the dominant application for a certain record type. This means that such records are only inserted using one particular application or that these records, when they reach a certain status or stage should only be updated using another application. This concept is usually termed "ownership".

An example of ownership might be organisation records, when first identified by a company are inserted into a CRM application, usually as sales prospects. It is normally not appropriate to enter such organisations into a financial accounting application until a sale has taken place and the organisation has therefore formally become a customer. The CRM owns the organisation record and has full update responsibility until the organisation becomes a customer. Once it becomes a customer ownership might be passed to the financial accounts application and updates to the record might be controlled primarily by accounts.

Legislative considerations

All UK organisations are constrained by two pieces of legislation regarding data, the Data Protection Act and the Freedom of Information Act. The former is aimed at ensuring that only appropriate information is held about individuals, that it is only held for as long as it is of direct use and that communications can be opted out by the people to whom they relate. The latter is aimed at making available to a person all information held on them by an organisation. These pieces of legislation relate to all contact data held by all organisations. The main implications of each of these are as follows.

Data Protection Act

All contact records need to have opt-out options allowing the recording of personal communication preferences. Further, all mailings and communications generated from your system need to honour these preferences. Failure to comply can result in financial penalties to the organisation. It also makes perfectly good business sense to honour a contact's personal preferences. Sending a marketing mailing to a contact that has opted out will not, generally, result in a sale.

Data about a person should only be stored for as long as it is directly useful. It is not acceptable to leave information stored against a contact "just in case" it is useful in the future. Flexible query tools are a necessity, therefore, to stay on the right side of the law.

Freedom of Information Act

Any person has the right in law to request within "a reasonable period" (generally accepted to be about 48 hours) to receive a copy of all information held about them by your organisation. You are allowed to charge a "reasonable" administration fee (generally accepted to be about £10-£15) but the work and costs involved in collating such information can be very extensive if your organisation employs multiple databases.

© Interactive Software Limited

Having a single primary database of all contacts in a single application makes this task very much easier and if users avoid keeping local copies of data (even temporarily) then gathering such information is at least possible.

Mailing Preference Service

Any organisation that carries out direct mailing should cross-reference its target contacts list against a central opt-out contact list administered by the MPS called the MPS Consumer File. Avoiding sending unwanted mailings to people on this file makes good business sense as it contains contact names representing the least effective target list.

Honouring such preferences shows professionalism and a respect for the customer. It also reduces spend on natural resources and postage and can give a boost to a company's "green credentials".

Summary

Ensuring good data quality and implementing processes to maintain it can lead to significant benefits to an organisation. Bad data reduces efficiency, can cause friction between departments, reflects badly on your organisation and reduces profitability.

Raising and maintaining data quality can take considerable effort but the benefits are felt throughout the organisation from board-level to office workers and partner organisations. Honouring personal contact preferences and being able to respond to requests for information in accordance with the Freedom of Information Act are made less difficult and keep you compliant with the appropriate legislation.

When upgrading your software applications and database systems it is a good idea to cleanse data before transferring it into the new system. This can reduce workloads and improve adoption after go-live.

Periodic reviews of data quality are a useful way to maintain quality and to promote the ethos of data quality within an organisation. Ongoing cost savings are difficult to quantify but can be very significant.

The most important thing is to maintain an ethos of good data quality. If users of your database understand the importance of data accuracy and completeness and they have appropriate tools and training then your data will continue to be one of your organisation's biggest assets.

© Interactive Software Limited

Appendices

© Interactive Software Limited

Further information

1. **Maintaining data quality and integrity** white paper available from Achiever Software
2. Data Protection Act - www.opsi.gov.uk/ACTS/acts1998/19980029.htm
3. Freedom of Information Act - www.opsi.gov.uk/ACTS/acts2000/20000036.htm
4. Mailing Preference Service - www.mpsonline.org.uk/mpsr/

About the authors

Founded in 1992, Achiever Software is a UK-based Customer Relationship Management (CRM) solutions developer that has extensive experience of implementing Corporate Information Management solutions in the Private and Public sectors.

Achiever Software's innovative partnership approach to the delivery and implementation of the Achiever Anyware application facilitates the creation of a solution that suits the unique business process and data requirements of each customer. Achiever's flexible relational data structure allows solutions to be designed to suit diverse business requirements across multiple departments and industry sectors including Public Sector, Education, Contact Centres, Medical Information Systems and Financial Services.

Achiever is aimed at organisations requiring a Corporate Information Management solution capable of managing sophisticated and complex business requirements whilst allowing them to maintain system ownership.

Achiever's strong integration capabilities allow it to maximise the benefits of a customer's existing business systems to create a complete solution providing benefits such as improved visibility of information, increased productivity, and reduced administration work.

Data security is paramount. Achiever is designed to ensure that only authorised users have access to sensitive and confidential information whether on screen or in reports.

Achiever has been successfully implemented in organisations including Surrey County Council, Papworth Trust, Warwickshire College of Education, Momenta (AEAT), VisitBritain, Skillset and NECC.

Achiever Software
Ashted Lock
Aston Science Park
Dartmouth Middleway
Birmingham
B7 4AZ

T: 0121 380 1010
F: 0121 380 1011

© Interactive Software Limited