Tackling the ticking time bomb

Data Migration and the hidden risks

An Experian white paper

Experian™
Data Quality

Table of contents

1. Introduction

Data migration is intrinsic to most big IT projects and yet as an industry we have a poor record of tackling it.

If you are faced with responsibility for an IT project where there will be some element of data migration then this white paper written by industry expert Johny Morris will help guide you past the pitfalls that await the unwary and even how to add value as a consequence of performing this necessary task.

About Author

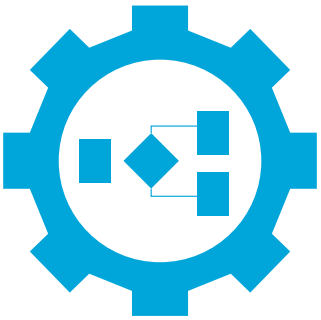
Johny Morris has 25 years plus experience in Information Technology spending the last 15 years consulting exclusively on Data Migration assignments for the likes of the BBC, British Telecommunications plc, and Barclays Bank. Johny is the author of the Practical Data Migration (PDM) method and book of the same name. Johny co-founded the specialist data migration consultancy, Iergo Ltd (www.iergo.com) and is on the expert panel of Data Migration Pro.

Author
Johny Morris

Co-founder
Iergo Ltd

2. What kinds of projects?

Data migration is the moving of data from our old computer systems to new computer systems but it is not an activity we perform in its own right. **However if you are about to embark on:**



- A technology refresh
- A merger or acquisition
- A de-merger or buyout
- Outsourcing or in-sourcing
- System replacement due to end of support
- Regulatory requirement to produce a consistent view of customer

Then you are probably going to be involved with a data migration.

3. Why so many bad news stories?

How can data migration be so hard? We are all expert enough to download huge video files from the internet. We all routinely copy files to our data sticks and pass them around.

The more adventurous of us may even synch our calendars on both our phone and desktop. So why is it that with all the technical competence at our disposal we routinely fail to move data that has been sitting happily in our old systems?

And the figures are appalling. If you go it alone, without appropriate software support, training or external methodologies, the chances are greater that you will fail than that you will succeed (source Data Migration, Philip Howard, August 2011, Bloor Research).

Data migrations go wrong for four principal reasons

- Underestimating the scale of data issues in the legacy system(s)
- Overestimating the ability of the technologists to fix the problems unaided
- Poor prioritisation and management of data issues
- Misunderstanding what you signed up for

You will see that "Mapping problems" is not on the list. Although we sometimes hear of mapping being a problem on closer inspection it is usually because of problems in other areas – like assuming that a legacy field is populated with consistent values then finding out that half the values are missing.

4.1 Underestimating the scale of data issues

We get to trust our existing systems to the extent that they become almost invisible to us. And yet under the cover of that old familiar skin unknown numbers of errors lurk. Looking into old systems is like being on an archaeological dig. Each layer reveals a new set of issues. There are long forgotten system crashes, misapplied patches, changes to validation, company re-organisations, botched updates, long forgotten initiatives, poor end user training and (my favourite) end user inventiveness that populates under used fields with user determined and undocumented values.

These problems are compounded if our data migration is driven by a merger or acquisition where even the limited knowledge we have of our own systems is missing.

The scale of the issues facing us is normally a complete unknown when we start our project. Will we be pleasantly surprised or horrified at what we find when we peal back the covers? Because we routinely use these systems our tendency is to underestimate the time and effort that will be involved.

4.2 Overestimating the ability of the technologists to fix the problem unaided

Most of the data problems you find are quite within the competence of your technologists to fix. When it comes to dates in the wrong format, names and address that need to be re-arranged according to the post office address formula or the proper representation for phone numbers, our programmers and analysts are in their element. But, and it is a big but, there are always some residual problems that they cannot fix or to put it another way, they cannot fix without help from the business.

Let us look at a real life example. A supplier in a heavily regulated industry was half way through their migration to a new, all embracing computer system. The commercial customer migration had been successful but when they started on the domestic customers something like 20% of their domestics had commercial tariffs. From a technical perspective this is simple. Get a list of the erroneous tariffs, cross reference it to the list of correct domestic tariffs and in the data migration replace one with the other.

From a technical point of view it may be simple but from a business perspective things were a lot more difficult. There were over payments and under payments. There were differences in sales tax on commercial and domestic tariffs. There were potential regulatory

problems. There were PR problems. There was the potential for the call centre being swamped if 20% of the customers phoned in. All of these issues and more took months of interdepartmental meetings to resolve and were certainly above the pay-grade of the data analyst in the ITC department who discovered the underlying problem.

This was an exceptional case but even where you do not find such mammoth problems you should still plan to find some problems that just have to go back to the business to resolve and unfortunately the business side is just not equipped to turn these around quickly on top of their normal workload.

4.3 Poor prioritisation and management of data issues

If there is a necessary time lag on the business side getting questions answered, this will be exacerbated by failing to have in place arrangements for managing data quality issues or even worse failure to pre-warn our business colleagues that these data issues are coming their way.

The worse that happens is the dreaded responsibility gap opens up. This is where the technical side, having exhausted what they can do, push a large pile of data items at the business side who are overwhelmed, under prepared and under supported. They push back. Between the two sides a gap develops into which many a data migration project has plunged.

We need to arm both sides with the appropriate tools and structures which will ready them for the data preparation task and unite them into one big virtual team.

Those of us that have been involved in data migrations before will be aware that we will uncover more data issues than we can fix in the time available before go live. We will need to prioritise but we need to do it in a way that takes account of the business as well as the technical drivers. Left to their own devices it is all too easy for the technologists to make erroneous assumptions about what is most important.

4.4 Misunderstanding what you signed up for

Just as we have to build a common team it is also essential that all the parties know what is expected and, just as importantly, what is not expected of them. In most medium to large scale data migrations there will be a systems implementer (SI) or a vendor as well as the client's own IT competency carrying out different aspects of the migration. There may even be multiple teams from multiple vendors. It is vital that we know what everyone is doing.

This may seem obvious but we could share any number of case studies where the purchaser has misunderstood the scope of the implementer's activity and vice versa. There are a number of reasons for this but the two most common are the rigidity of modern procurement contracts and the impact of underestimating the scale of data issues.

Underestimating the scale of data issues we have already touched on and when this is compounded by poor prioritisation and overestimating the ability of technologists to solve the problems unaided it's easy to see that we have a recipe for going into the project with a misjudgement of what each party can realistically do. The technologist cannot fix all the problems unaided and the ones they cannot fix tend to be the most intractable and time consuming. This is the first misunderstanding.

"Those of us that have been involved in data migrations before will be aware that we will uncover more data issues than we can fix in the time available before go live."

The second is the unforeseen consequence of tighter and tighter procurement contracts drawn up by our buyers. The history of IT projects is long and marked with some spectacular failures. The learning from this has been to draw up well specified output based contracts where the supplier is tied to delivering a working system at a given point in time for a fixed price. In response the suppliers have had to withdraw responsibility for activities where they cannot predict the outcome with sufficient exactness to be able to price the risk of failure into the contract. Experienced suppliers know that the issues of data quality in legacy data are often understated but for any one client they cannot know by how much. They cannot therefore quantify the effort needed to fix the issues. The result of these two forces – the purchaser looking for certainty in the contract and the supplier looking to remove unquantifiable risk - has been contracts that focus more on the "Final mile". In other words the responsibility for discovering and fixing legacy data issues is a client side one. Classically the supplier will present a template to the client for them to source or a staging database for them to populate.

To be fair to the suppliers if you are migrating to a complex product (like SAP for instance) then that final mile can be a tortuous one that requires considerable expertise so there is still plenty for them to do.

The end result is that the supplier will do the data migration – but only of data that matches the staging database they provide. Anything other will be rejected and it will be up to the client to get it fixed and re-submitted within a timeframe that will not compromise the contracted end date.

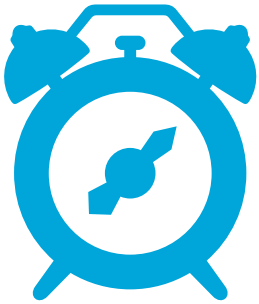
So the responsibility for selecting data sources, profiling and performing some initial transformations lies with the client – but often, dazzled by the promise that the new system brings, this is overlooked or underestimated. There are many variants to this scenario. Sometimes the supplier will perform profiling of suggested sources and report potential errors back to the client. Sometimes the supplier will develop the target but expect the client to understand how to generate the staging database. Sometimes the supplier will perform extended transformations of data so that it will load but just as often we have seen cases where the contract and the supplier will refuse responsibility to perform any data enhancement. We have even seen this extended to include defaulting in missing values but that is not the norm. More often than not the supplier will use their experience to willingly go beyond the contract to help the client select and prepare data for loading but in almost all cases the ultimate responsibility for the data selection and preparation tasks invariably falls back on the client.

Altogether underestimating the scale of data issues, overestimating the ability of technology to fix the problem, poor prioritisation and management of data issues and mutual misunderstandings leaves a time bomb ticking away in the heart of our biggest projects. This set of inter-related problems explode when we try to load data in earnest blowing apart our project plans, compromising our delivery date and severely curtailing our return on investment.

5. Defusing the time bomb

Just are there are four major sources of trouble when it comes to data migration, there are also four key steps that should be taken to alleviate them:

- Landscape analysis and profiling
- Early business stakeholder engagement
- Data quality management processes
- A proper project charter



5.1 Landscape analysis and profiling

Data migration is the building of a bridge from old systems to new over which data flows. When engineers are building a bridge they pay just as much attention to each end. Classically within IT projects we just have not done that. We have been so mesmerised by the promise and novelty of the new that we have neglected the old. Our focus is on buying in expertise of the target not on rediscovering the legacy. But our legacy may represent twenty or more years of effort. Some well directed some not. We cannot know, without looking, what all those keystrokes have contributed to the quality of the data in there. We can anticipate from experience that there will be all sorts of dubious data to deal with. If you are dealing with anything but the very smallest of legacy data sets then we should also know that manual efforts just will not be good enough. To review the millions of entries in large data stores across many years we need tools capable of doing the job. This is even more the case when we have to combine data from multiple sources. How well will it fit together both at the model level and at the detailed, customer by customer and account by account level? How do we de-duplicate a million customers?

These tasks gain extra degrees of complexity when we are dealing with a merger or de-merger driven data migration. Here we may not even have access to subject matter experts (SME) who know the data.

For these jobs specialist data profiling tools have been created. These reduce weeks or months of potentially hand coded scripting to the work of hours or days. Specifically designed to look for rules hidden in the data they direct our effort to the anomalies – the fields that are always to supposed to have a unique value but which are empty or contain duplicates, curious data patterns in fields that are not supposed to be in use and so on.

The research above suggests that without access to a built for purpose profiling tool you have a 45% chance of missing your go live date.

With a tool, and the good practices that go with it, you will have the knowledge that will allow you to properly scale and plan the remainder of the migration.

"These tasks gain extra degrees of complexity when we are dealing with a merger or de-merger driven data migration."

5.2 Early business stakeholder engagement

The profiling and landscape analysis phases will generate an awfully large number of questions and potential issues to deal with. As we saw earlier, the majority of these can be dealt with by the technologists but there will be some, usually the most intractable, that will need to be referred back to the business to be solved. We know we are going to have to do this, so it is important that we seek out and bring into our processes the subject matter experts (SME) that we are going to need as soon as possible. It is only human nature but if you only seek help when you have a potentially project wrecking problem on your hands and time is running short don't be surprised if you encounter reluctance to engage. It therefore makes sense to start profiling early when there is less time pressure and to have a data quality management process in place before the big issues appear.

5.3 Establish data quality management processes

Within our favoured methodology (Practical Data Migration or PDM) one of our earliest steps is to establish a board composed jointly of technologists and SME to look at known problems that will impact the migration together.

To perform this function properly you need to access modern data quality software. This allows side by side working where data can be examined swiftly and rules created and tested on live data by a process of joint SME – technologist engagement.

Often the solution to data quality issues lies with the business correcting errors in the source data. Where this is the case it is important that the data quality rules can be re-run efficiently to track that you are on target to get your data right by the go live date. And of course, it is also true that often whatever is causing the bad data in the first place may not have been fixed (after all why fix something when the replacement is only months away?) Built for purpose data quality software allows you to easily monitor that the source data once cleaned up is not degrading again and to take remedial action where it is.

5.4 A proper project charter

As with many journeys the first steps are the most important. There needs to be a clear understanding of the responsibility boundaries of each of the participants. It needs to be shown that there are no gaps. Here are some specimen questions:

- Who is responsible for the detailed design of the target including reference data design (e.g. the code values for products, the detailed chart of accounts)? Often the supplier is contracted only to deliver a working system but it is the responsibility of the client to create this level of detail.

- Who is responsible for retiring legacy systems? If we do not retire systems then we are not performing a migration which by definition is the permanent movement of data from one platform to another. Retirement can be logical not necessarily physical. When an enterprise is demerged from its parent it may have to give up the use of all of its existing systems but of course these carry on working to serve the parent.

- Who is responsible for the archive solution? In any but the simplest of migrations a large amount of historical data that may be required for occasional processing (e.g. tax records retained just in case there is an inspection) will not be loaded into the target. Where is this data going to reside? Who is responsible for the design and build of this archive?

- Where is the transformation and enrichment going to be performed and by whom? By the client between extracting data and placing it in the staging database? By the supplier after the raw data is placed in the staging database? Some by the client and some by the supplier and if so who decides where?

- How are you going to manage the efficient passing of design changes from supplier to data migration team, especially towards the end of the project when there can be a swirl of late fixes falling out of user acceptance testing?

There is a helpful check list of all the items to look for in the book Practical Data Migration – 2nd Edition.

By now it should be clear that when it comes to data migrations you will be faced with having to perform significant amounts of activity, to a high degree of accuracy, within a tight timeframe. It is therefore imperative that you give the technologists the right tools to do the job.

Generically there are four types of software. Each type corresponds to a specific phase in the migration process. These software types are:

- Profiling software
- Data quality software
- Mapping software
- ETL software

6.1 Profiling Software

We saw that we are often ignorant of the exact state of the data in our legacy systems. The first step therefore is to shine a light on that uncertainty and discover what is really going on in our legacy data stores. This is what profiling tools do. Reading through the source databases they rapidly produce a set of reports on what they encounter. The best of breed produce reports on scores of dimensions of data issues: They show us where columns have mostly unique values. They show us columns that have identifiable patterns (e.g. bank phone numbers or credit card numbers). They show us which columns seem to have matching data with other columns (i.e. where columns appear to be related and so possibly a lookup table join). They show us where a column has data types inconsistent with its purpose (numbers in a first name field). Altogether they give us a rich picture of what we have in our legacy data.

Our recommendation is to start profiling as early as possible. Do not wait until the target is fully defined. The target is likely to arrive later than planned and subject to change right up to go-live. In theory there is a risk that you will try to fix problems that you ultimately do not need to fix. In practice this is not the case. There will be enough data issues that have to be fixed, whatever the target will ultimately look like, to be getting on with. On the other hand, as we have seen, when it comes to getting responses back from the business these are necessarily slow. Move as much activity up the time line as possible. Tools that allow maximum collaboration between the technologist and the expert business user are best suited to this task. You need to be able to review real data sitting side by side.

6.2 Data quality software

If profiling software discovers the rules in legacy data (and their violation), data quality software takes business rules and applies them to legacy data. This is a task that commences once the target starts becoming clearer. In the best case the software allows rules that are discovered during profiling to be carried forward into the data quality phase. For this kind of reuse to be effective you need a single, integrated, software set. Again prototyping software that allows agile rapid development of rules is what is needed here.

6.3 Mapping software

Mapping software is, naively, what we first think about when we consider data migration. Linking fields from source to target, enhancing and transforming existing data en route is the end point of our journey. As we have seen however the bulk of our work should have occurred prior to this. Once again software that allows for a collaborative approach, where real data can be seen transformed, is more efficient and less risky than the traditional forest of mapping spreadsheets.

6.4 ETL software

This is the software for the final mile. Often the target will have specific load software optimised to load data into their environment. With SAP it is the BAPI, IDOCS, LSMW etc. These are often within the purview of the supplier and mandated by the target. Mapping, from a client perspective, is therefore often only to the staging database to be picked up by the ETL software.

6.5 Why buy – can't I just use existing coding resource to do all this?

It is true that everything profiling, data quality, mapping and ETL software can do could be done by writing programmes in the traditional way, however the speed and sophistication of the specialised tools could not be replicated in anything like the time you will have available to you. A modern profiling tool will perform hundreds of checks in a single pass that it would take a programmer months to replicate and that is before we get on to more esoteric features like fuzzy matching. The cost benefit case is clear. If in doubt about what this is all about ask for a demonstration, the results will be startling.

'The cost benefit case is clear. If in doubt about what this is all about ask for a demonstration, the results will be startling.'

7. Appropriate methodology

As we all know technology is a wasted expense if we do not know how to exploit it properly. With fast profiling and data quality tools come great leaps in productivity but only if they are used within an appropriate framework.

Obviously there needs to be training in the application itself but also in utilising it in a data migration setting. A profiling tool on its own can create a library of metrics. Every column in the source data base will be analysed in over a score of different ways. How do we sort the wood from the trees? When we have a bunch of problems refined, how do we get answers and action out of our hard pressed business colleagues? How do we know that we have covered all the bases in our data migration strategy?

7.1 Once in a business lifetime event

It is a simple fact that for most enterprises data migration is not a business as usual activity. Of course there are some organisations – data bureau, outsourcing companies etc. for whom data migration is a normal activity. These companies need to take a factory, production line approach to data migration but for the rest of us, our organisations may get involved in a major migration once every fifteen or so years. The chances are that the people who were involved last time will have moved on (or are still so scarred by the experience that they do not want to get involved again). The technology will certainly have moved on and yet data migrations are difficult, complex things requiring their own special skill sets so as well as appropriate technology you must have an appropriate methodology and access to trained experienced staff as well.

8. The benefits of a well conducted data migration project

Obviously the major benefit of a well conducted data migration project is the delivery of a fully functional software application that adds value to the enterprise from the moment it is turned on not a chaotic shambles for the first three months after go live as data is found to be missing or badly structured. However that is not an end to it.

8.1 Early wins

The great thing about starting your profiling and data quality activity early is that you can publicly slay a few dragons that have been hanging about for ages. If you are implementing a new Customer Relationship Management (CRM) package then you will need to de-duplicate your customer list. This is an activity that has probably been on a lot of folks' wish lists for years. So do it. Do it early. Publicise the success. Not only do you gain plaudits but you also build a lasting partnership with the business.

8.2 Data governance

Many, probably most, organisations these days have realised the value of information and the insight it brings. Data is no longer the lubricant that greases the wheels of profitable production, it is part of the profit generating product in its own right. Most forward looking organisations have instituted some form of data governance function to try to get a grip on this but how successfully? As a number of recent industry reports (backed up by our experience) show there are an awful lot of data governance initiatives out there that have started out boldly enough, only to collapse in a frustration born out of the natural clash between the structural goal of the benefits of good data governance and the short term drivers of everyday business. We may wish to end the scrap and rework cycle but it just never seems to be the right time. To rephrase St. Augustine of Hippo's famous aphorism "Make me make my data good lord, but not yet".

Increasingly what is being urged is that in place of the grand design to fix everything, look for the tactical opportunity to embed good data practices by using data governance principles to add value more immediately if more locally. And a good data migration experience is the ideal place. Procrastination is not an option when there is the compelling event of a data migration looming. You will have to deploy all (well most) of the tools of the data governance trade to succeed. You will be building your internal case study of the benefits of a systematic approach to data centric activity. You will have supporters in the business and, if you use a methodology like Practical Data Migration (PDMv2), you will have solved otherwise perplexing issues like who is the data owner.

You need to consciously think through the implications of linking data migration to your nascent data governance activity. It may alter your software tool selection for instance. You will need tools that are up to the longer term task. On the other hand it may give you reason to bring in tools in the first place, possibly on a lease basis, to establish their value.

8.3 Health warning

Better data governance and slaying a few data quality dragons are excellent things in their own right but they must not be allowed to interfere with your raison d'etre – shifting data of an appropriate quality to the right place at the right time. If you fail to hit the go-live date then your longer term goals may also suffer a fatal blow. Expect, given the time frame, to uncover more issues than you can solve. Make this a fact from the outset and use it to justify the ongoing data governance activity that should necessarily follow.

9. Learning points

- **Start your data migration project with a clear charter of roles and responsibilities based on an existing methodology that links business and technical activities**
- **Use appropriate built-for-purpose software to profile, manage data quality, map and perform ETL**
- **It is never too early to start profiling and working collaboratively on data quality issues in the legacy data. Do not underestimate the elapsed time it takes to correct data issues that require business side input**
- **See data migration as an opportunity to develop data governance competencies. Act tactically but think strategically**

About Experian Data Quality

Experian Data Quality has built up exceptional market coverage assisting customers with their unique data quality challenges. We provide a comprehensive toolkit for data quality projects combining our market leading software with a vast scope of reference data assets and services. Our mission is to put our customers in a position to make the right decisions from accurate and reliable data. The size and scope of data management projects varies considerably but the common factor in all ventures is unlocking operational efficiency and improving customer engagement. We see the potential of data. Whether it's in enabling ambulances to be sent to the exact location of an emergency or attributing charitable donations to the people who need it the most - data accuracy makes all the difference to service provision.

For more information please contact us:
T (61) 3 8699 0100 | E info@au.experian.com | W edq.com/au