The Data Game Changer - Part Two

Choose the right tools to transition from Migration to Management

An Experian Data Quality white paper

Experian™
Data Quality

July 2014

Table of contents

About Author

Johny Morris has over 25 years experience in Information Technology spending the last 15 years consulting exclusively on Data Migration assignments for various industries such as Financial services, utilities and telecommunications.

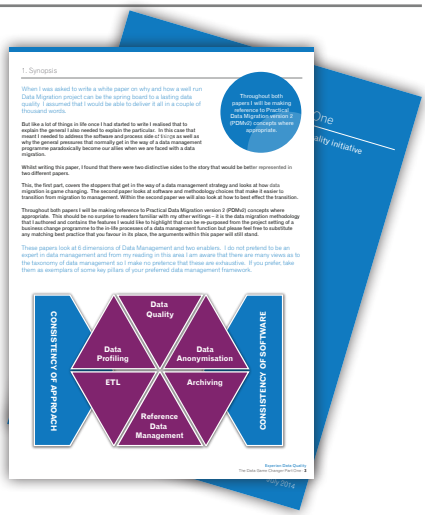
His book “Practical Data Migration” defines the leading Data Migration methodology – PDM. Johny co-founded the specialist data migration consultancy, Iergo Ltd and is on the expert panel of Data Migration Pro.

Author
Johny Morris

Data Quality Specialist

1. Synopsis

In Part 1 of this paper I looked at how the mere fact that you are performing a Data Migration provides you with a unique opportunity to get your Data Quality process up and running.



Data Quality is an issue that you just can't duck if you are going to perform a data migration smoothly. Here we look in more detail at the kinds of choices you need to make to cross the Rubicon from a desire to improve to actually getting a commitment to change.

There are two sides to this – the first is software and making a software choice with an eye to the long game. The second is process and choosing ways of working that can seamlessly transition from data migration project to in life data management

Software

Within this section I will look at each of the components of software that fulfils the purposes outlined above and set it against the Experian tool set for illustrative purposes.

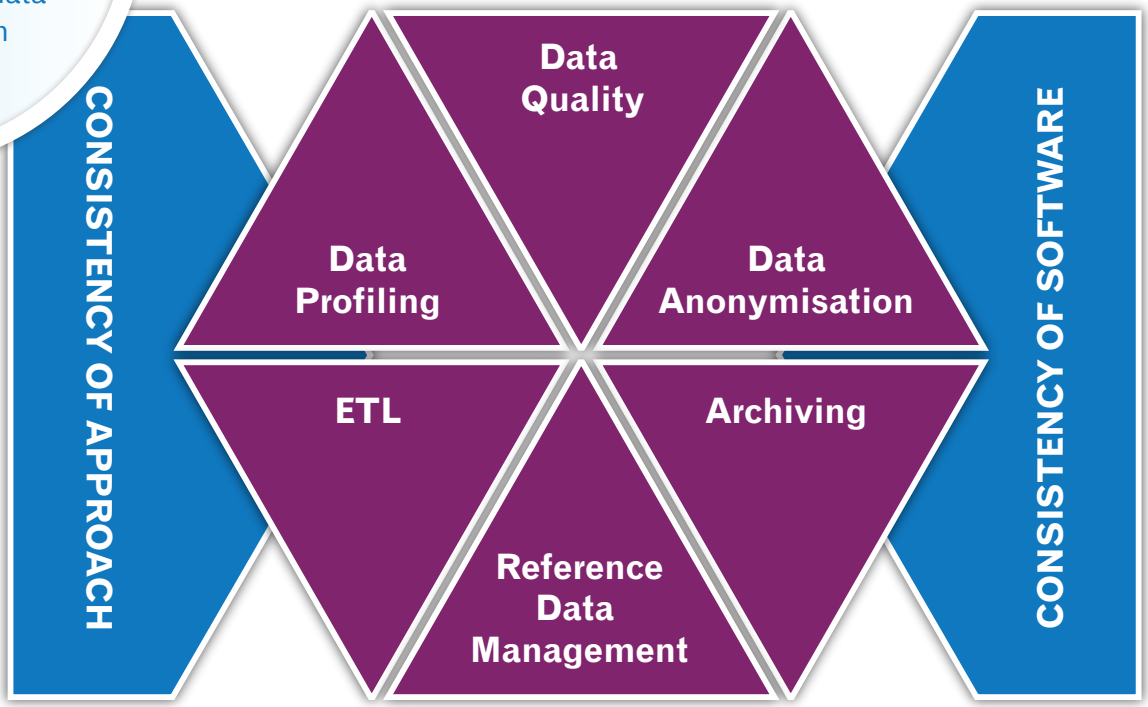
&

Method

Having the right software is only good if you realise them in productive use and how you actually deploy them. Having the right method in place will ensure you are in good steed for a successful migration and at the same time launching your data quality initiative.

Recap

Here is a recap of the 6 dimensions and 2 enablers of a data migration from part 1



Data Quality

The application of existing business rules to current data to uncover the breaking of those rules in the form of redundancy, synonyms, homonyms, and downright bad data. The existence of these rules can be a consequence of Data Profiling.

Data Anonymisation

The masking of data items to render to protect data subjects or business critical data from deliberate or accidental release into the public domain.

Archiving

The storing of Useful (usually historic) data that is not necessary for the normal functioning of a particular system but for which there is a business need (albeit infrequent).

Reference Data Management

The identification and consolidation of multiple definitions and redundant instance representing the same external object type into a single useful list.

ETL

Extract Transform and Load, the three steps that move data from one data store to another.

Data Profiling

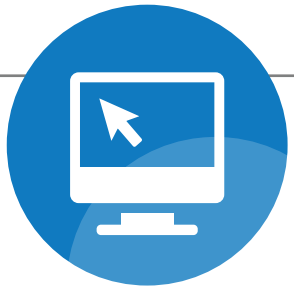
The discovery of the underlying relationships and rules in existing data including the breaking of those rules in the form of redundancy, synonyms, homonyms, and down right bad data.

Consistency of Approach

a single set of steps that are capable of being extended from the Data Migration to the new Business As Usual (BAU) world.

Consistency of Software

a software set that conserves re-usable components and is as suited to the BAU as the Data Migration world.



Within this section I will look at each of the components of software that fulfils the purposes outlined above and set it against the Experian tool set for illustrative purposes. For each headlining I will look at potential re-use of the tool in a post migration BAU setting.

- **Data Profiling**
- **Data Quality**
- **Reference Data Management**
- **Anonymising**
- **Archiving**
- **ETL**

2.1 Data Profiling

Data profiling is the discovery of rules in existing data. Every data store has its own internal rules that make it what it is – a CRM, an accounts package etc. And, in all but the very simplest spreadsheet, we expect every data store to have discrepancies between those rules and the data we find in the data store. By this we mean CRM packages with duplicate customers and customers without essential contact details or miss-posted accounts.

We also need to perform cross system profiling (also known as heterogeneous data profiling). Are all the customers in the CRM represented in the sales ledger? Do they have matching core data items (like names)? If they aren't or don't, which of the two systems holds the correct values?

However there is also another function which we commonly look to our profiling tool to support – Data Discovery. When we are performing a data migration, we need to reach into our legacy data stores and drag out essential operating data that in pre-migration business as usual we never needed to see. Data discovery is the ability to find the pieces of the data jigsaw puzzle which we know must be in the legacy somewhere but which we have never had to bother about until we come to move the data. Until you come to try doing this for real it would seem simple but take it from me – when you are faced with undocumented legacy data systems that have been allowed to grow in a haphazard manner over a number of years some of these "bolt-on" fixes can have been put in the most unlikely places.

2.1.1 Heterogeneous Data Profiling

There are a number of ways of managing this. The simplest is to create a profiler that can make multiple connections to multiple data stores at the same time. This has the advantage that it is relatively simple to set up and can work on near real-time data refreshes. A second way is to export the data from a number of sources into an industry standard repository (like a data warehouse).

This can be a bit of a cart before the horse solution though where, issues of data quality that prevent data being loaded in the first place get in the way of getting your data into a repository to discover all the data quality issues.

Then of course there is the most complicated, least efficient but still probably the most common way of proceeding – use tools appropriate for querying data store A to export some interesting data (like all the customer ID's) and use them to drive queries in a tool appropriate for data store B. This is messy, error prone and slow. It also depends on the operator knowing the questions they want answered. This is less data profiling than data surveying.

Built for purpose tools – like the Experian Data Quality Platform - may use a more sophisticated fourth option. They take data from various

data sources and import it into it's a special repository. Here the data is scanned and multiply indexed. The benefit is that the resultant repository runs at lightening quick speeds.

The result of this process is that the tool offers inbuilt heterogeneous profiling, seamlessly ignoring system boundaries in the legacy and presenting the user with a single view across all the data sets – but a view that retains the original, familiar data structures.

So now we have your data in a profiling tool what do we do next?

These tools have many standard profiling features like

- **Fill rates** – which columns are full, which are empty, and which are part complete
- **Data typing** – often a text field will be filled with only numeric values perhaps the order reference number
- **Uniqueness** – columns with mostly unique values are often lookups for other tables
- **Potential foreign key relationships** - in many legacy systems not all foreign keys are reflected in the schema. The software will seek out matches to columns profiled as potential lookups and suggest relationships
- **Data patterns** – do some columns contain data patterns that suggest a field's use that may be at odds to its name? Like post code or national insurance numbers? (This of course work both ways with field names that suggest post codes having data patterns that suggest something else)
- **Custom data patterns** – it is all very well exposing data patterns but what about seeking out patterns that are only standard to your business like part numbers?

These are some of the basic features – and as this is not technology review I'll move on to where modern, built for purpose tools, really start to earn their crust – data discovery.

Because the repository is optimised for retrieval and seeks out potential foreign keys as an out of the box trick, finding navigations across heterogeneous data stores is a point and click activity. Better than that, the speed and the use of real live data sets allows the results of searches of even millions of rows to be performed in seconds with the results available for inspection by our business colleagues. Having used tools like this it is difficult to imagine what it was like back in the day when the turn around time for a query to be written and run was hours and the discovery of links across data stores was measured in days.

2.1.2 Re-use of Profiling Post Migration

On the surface it would appear that data profiling is a stepping stone technology. Once we have used the tool to get us the data we need then this aspect of it can be forgotten. However this is not necessarily the case. In complex data migrations where we are phasing our migration – maybe by country or by customer type – we will obviously start the process over. In these circumstances it is not uncommon to find that accompanying the natural split of the organisation over these phases there are legacy data stores unique to each phase.

Perhaps Spain has its own loan book for a category of loans that don't exist elsewhere and this is held in a locally built database that is being subsumed within the new system. When we get to Spain this loan book database will have to be brought into the repository and profiled alongside existing profiled data.

There is also a role for the re-use of profiling in an archive solution – but that I will leave until I get to archiving.

2.2 Data Quality

The application of known data quality rules to legacy data is an integral part of data migration and also a key focus for data management.

Fixing things in the current setting based on current understanding is an important lesson to take from this white paper. It reinforces the practical experience of working collaboratively, with support of the appropriate tools, to prioritise and fix business data issues in a work day environment not just in the rarefied atmosphere of a project office. If we can do it during the build up to the migration then we can do it after the migration as well.

As well as loading the data quality tool with data quality rules as they are discovered and pointing it at the legacy data, there is the issue of constantly re-checking existing data quality rules. We need to do this for two reasons – firstly, corrections to legacy data can take time, especially when they are conducted by front line staff alongside the day job. We want to be able to track that we are on course.

2.2.1 Re-use of Data Quality Post Migration

Data quality, both technically and as part of a formal process, is the easiest aspect of the data migration suite to justify. Here we have the chance, possibly for the first time, to embed best practice and a sustainable framework of constant improvement.

This is doubly assisted by the compelling event of the migration. On the one hand we have had to do something about data quality or else the new system would quite simply not have gone live. On the other, the time pressures will mean that we will not have fixed all the data quality issues we will have unearthed. So if handled correctly we will have the ideal situation. We will have software that has been proved in the most exacting of circumstances. We will have a collaborative approach to prioritising and solving issues. We will have a list of all the issues with the data that we did not have time to fix. All we lack is the will to move forward with this.

2.3 Reference Data Management

Managing reference data is a key part of many data migration. It may be standardising the terms of our product offerings. It may be the creation of a single customer list. These are common enough ambitions. With the use of appropriate technology we can search out and de-duplicate data in its multiple forms

2.4 Archiving

In any data migration but more especially in modern COTS package migrations it is almost always the case that you will only be able to migrate a minimum amount of data.

There will however be more data that needs to be held near online or offline. Some of this has defined periods when it will be needed (maybe prior year accounts at the next year end) but a lot will only be required for exceptional purposes (the classic example here is a tax inspection that could occur but is unlikely). Part of your data migration strategy should include making provision for defining and addressing the need for online access (data that should be loaded into the target at start up), near online (could be available by report from a reporting repository) or offline (can be accessible by ad hoc query).

The use of the Experian data management platform provides you with a readymade solution to near online and offline storage.

Because all the data from all the legacy systems can be held in the underlying repository, which itself can be backed up to a server, and all the drilldowns can be held in a project, it is possible to satisfy both foreseeable requirements by creating appropriate drill downs and ad hoc requirements by using the profiling capability described above to extract data from the legacy systems.

The data in the repository does not have to obey any referential integrity or other constraints. It can be just as bad as it was when it came from the source thus removing the cost of preparing data for loading into an archive when that data may never be called on to be used.

2.4 Anonymising

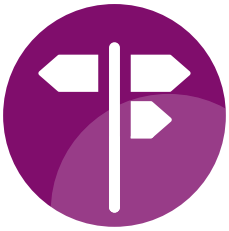
There is always an ongoing requirement for data for testing and training purposes.

There is also a conflicting requirement to protect the privacy of data subjects, especially where there is sensitive data being held. However, depending on the solution, simply replacing values in fields with randomly generated values may not work where those fields are necessary to maintain referential integrity across the application. Because the Experian solution masks values rather than randomises them the results of applying data masking to fields with identical values will always generate the same resultant value and so referential integrity will be maintained.

2.4.1 Re-use of Anonymising

There will probably not be a great deal of re-use of anonymised data created for testing the migration. The target will evolve in post migration business as usual and as each evolution needs to be tested and trained for it will need its own data anonymised. However the requirement for anonymising data whilst at the same time maintaining referential integrity will persist so the skills and techniques used will be transferable.

2.5 ETL



We touched above on the use of drilldowns to expose in real time extracts of data linked across multiple data stores so that the user can visualise what it is they are going to get.

Within tools like the Experian Data Quality Platform there is also the ability to perform sophisticated transformations of data, so joining, parsing and data enrichment are all available as well as complex selection and exclusion criteria. This is all done with point and click, drag and drop interfaces which reduce coding errors and massively reduce coding time when compared to traditional SQL. All this performed on real data that the user is familiar with. It is important to stress why this is so significant. This is, in effect a prototyping approach to data mapping. One of the common complaints (and cause of frustrating and time wasting rework) is the inability of a business representative to be able to visualise what the outcome of the traditional mapping spreadsheet will look like. Is this exactly what they want? Here in place of days (weeks?) of workshops, signoff's, coding, rejection, change control, rework and delay we have hours of joint effort.

The results can be either exported as data in csv or excel format for onward processing or the built in documentation generated can be the basis of their recasting in other tools. The choice of which route to take is dependent on how best to perform ETL in the environment in which you find yourself.

2.5.1 ETL In Practice

In most business transformation programmes where enterprise application software is being implemented, there is normally a supplier or systems implementer charged with actually delivering the new system. Extraction is generally the responsibility of the client and the final load is the responsibility of the supplier. Transformation is a responsibility that is shared between the two depending on the nature of the contract.

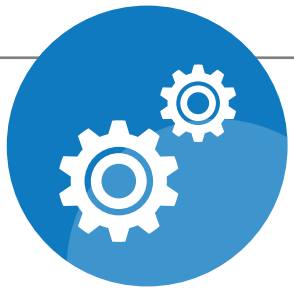
In the implementation of most COTS packages there is a prescribed load method and template (for instance the Data Migration Framework for Microsoft Dynamics). This will define a series of tables that must be filled for the entities being loaded (Customer, Supplier etc.). The validation for each of the fields in these tables and the referential integrity between them will also be tightly specified. There will also be a prescribed loading ETL (SSIS for Microsoft Dynamics). So the client has responsibility for extraction and the majority of transformation. The supplier has responsibility for a smaller amount of transformation and for the load. The architecture is therefore one of a landing stage between the extract/transform of the client and the load of the supplier formatted to match the load format prescribed by the COTS package. The exposure of files in an industry standard transfer file type (comma separated, pipe separated etc.) is often sufficient for the hand off between parties. It is therefore becoming less significant, from a client perspective, to be able to perform the full ETL.

In these circumstances the use of the combined extract and transform, will be more than adequate to satisfy the client end of the migration.

In very complex, high volume migrations where data loading is being staged and run in parallel with an application that is running live, a more sophisticated migration controller will be required in addition to the architecture described above with gating (i.e. controlling the volume of data being migrated), in-line migration reporting of volumes being loaded, workflow management of faults, rollback management, fallback management etc. This would need to be implemented separately but would accept data in the standard industry formats that can be generated by Experian's Data Quality Platform.

2.5.2 Re-use of ETL

Data Mappings in data migrations are almost always one offs. In phased migrations of course they can be re-used if the phasing is based on a geographical rather than functional decomposition. In other words if the migration phases are based on going round the business region by region rather than migrating say the CRM first then the HR function etc. However except in specialist system integrators who are established to perform a particular move from one COTS to another (say Peoplesoft to SAP) it may appear that once the migration is over the code is of no more use. However if the archive option suggested above is applied (using the data repository as the archive) then the drilldowns etc. remain available to subsequent queries and become part of the archive solution. This saves time and effort re-coding a solution. Given that the extract/transform completed in the Experian tool will produce load entities that are generally not a long way from the entities as they will become familiar to a post migration Business As Usual world, the exposure of data from the legacy in a format closely resembling now familiar structures, will make the process of performing adhoc queries on legacy data a whole lot easier to a generation of analysts who will now be more familiar with the post migration world than the pre-migration one.



The features above are only any good if you realise them in productive use and that productive use is limited by how you deploy them.

To deploy them successfully you need an approach that is likely to be successful. Of course you must be flexible. Your approach will evolve itself over time but I would recommend, as a starting point, adopting a tried and trusted existing methodology.

I will use my own approach (PDMv2) to demonstrate how this might be accomplished.

PDMv2 has multiple modules that cover everything from landscape analysis through partner selection, data mapping and finally legacy decommissioning. For this paper I will focus on the Data Quality Rules (DQR) module.

Data Quality Rules

3.1 Data Quality Rules

The DQR module comprises a set of forms, a series of processes structured in a particular way and a set of roles and responsibilities. Data management issues are bread and butter problems to any data migration. We may need to perform reference data management (perhaps for our product set and parts lists), we will almost certainly want to perform some customer or contact de-duplication. It is of course unnecessary to explain that Experian with its heritage of customer analysis is a leader in this field via its Single Customer View software.

But how to get them established in a sustainable manner?

3.1.1 Reach Out

Data quality works best when there is a willing participation of all parties. We need technical experts who understand the systems and business users who can provide the end user context. These parties have to be come together as a single team to evaluate, prioritise and agree solutions to the hundreds of issues that are the common place of data migrations.

We also need to leverage the informal network of contacts of our business colleagues to reach out to those specialists within the business who have the unique knowledge we will need to solve some of the obscure data issues we will surely face.

Within the DQR process we manage all of this by establishing a board (on large programmes this can be multiple boards) which meets on a regular basis (usually weekly). There is representation from the business, from the legacy technical side and from the target technical side. The whole is chaired by a data migration analyst. Every data related issue – whether data cleansing, data preparation, data discovery, data enrichment – is recorded and placed before the group for prioritisation and management.

Prioritisation is important. On every data migration with which I have been involved over the last 17 years there have always been more data issues than there is time to fix. We have the compelling event of the legacy decommissioning. This is great for concentrating the minds of our

business colleagues, and is usually a fixed date beyond which the business case for the new solution tends to be negatively impacted.

Expect hundreds of issues and expect to solve less than 50% in the life time of the project – leaving plenty in the kitty for working on post go live.

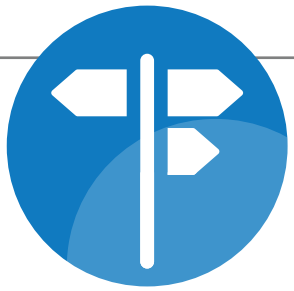
Start Early

Most projects commence in a spirit of optimism and camaraderie. There then comes a dip as the reality of project issues starts to impact us. Finally there is usually euphoria when we get across the line and deliver improved software to our business colleagues. Unfortunately for most projects it at the point of maximum squeeze, when data issues are mounting up in barely controlled piles, that data issue management commences. It is only at this point that the business side is reminded that that the data is theirs and therefore the data issues are theirs too.

A struggle develops with each side pushing responsibility for data issues back and forward across a ghastly divide that develops between them. Much better therefore to acknowledge that there will be a multitude of data issues and get the DQR group up and working from the very start of your project. As I pointed out earlier, there will be more than enough issues for you to work on that will need to be fixed whatever target you are trying to reach. Once the processes are embedded, data issues cease be the showstoppers they are on less well organised projects

It is a fact that whenever I am called in to a project that is failing and I ask to see the complete list of data issues they are never available. Some will be on the issues log. Some will be being fixed in mini projects. Some will be with the original discoverer who is beavering way on a local work around. If they aren't all in one place then you do not have any control of data readiness, which should be the key measures of a data migration project plan.

And of course when it comes to measurement then the right software is imperative which brings us back to being supported by the right software.



So you've been successful in your role as a data manager within the context of a significant business change programme. How do you push on from here?

Well plan your post migration activity carefully whilst you are in the course of the migration. By this I do not mean that you have a detailed post-migration gant chart at the ready or that you go into the migration with pre-conceived objectives that suit your post-migration agenda.

So the job in hand must come first. You must deliver and deliver well but look to what you can carry out of the project – high level management recognition across the silos and appreciation amongst the technologists. How are you going to use these precious items?

Will you have had the ear of all the C-level execs? Probably not. You may not have had air time with the Chief Finance Officer but you will have been engaged with their direct report – the finance director who manages the accounts function on a day to day basis. And you can bet they are reporting back to the CFO. You must show yourself to be business orientated, pragmatic but capable of delivering change. Not speaking in overly technical terms or uttering ivory tower platitudes about the benefits of better data governance. Do it and show it. Also don't forget your role as a listener. When it comes to where to start – well start with the pain points the significant players are feeling. It may not be what you expect, or even what you think is the most significant but it is a start on the journey. There is no reason why you can't lay the groundwork for later engagement at this crucial time.

If you use the DQR approach, or something similar, you will have built your team of Subject Matter Experts (or Data Stewards depending on your preferred nomenclature) and technical experts. More importantly you will have created a feel good factor around the data management principals you are trying to sell. You will have solved longstanding data issues that have been bugging the business for years. You will have embedded DQR's into the language of the diaspora of ex-project personnel who will return to their accustomed business roles. You will have educated the local technologists in the benefits of a close relationship with their business colleagues. You will have built those all important interpersonal ties between these parties which will make your ongoing initiative so much easier.

It is important to use this new credibility and visibility to move your agenda forward.

Remember what I said about data issues – there will be more of them than you will be able to deal with before go live. However you will have the DQR list of outstanding issues and, with the use of appropriate software, you will be able to cost and prioritise those issues.

You will of course have gained a real competency in your software tools. Profiling, data quality and ETL will be second nature. You will have educated large swathes of the business in the possibilities that come with using modern, business engaging software that allows prototyping and rapid turnaround times.

Also (and this may sound trite but can be significant with your peer technologists) you will have shown that the world of data governance is not just about proselytising, process re-writing and mini clean up operations. Data governance has sexy technology all of its own. Tech that is better, faster and more sophisticated than that which is in use in the rest of the ITC function. A little tech-envy is no bad thing when it comes to getting noticed.

Of all the things that you can carry over Data Quality is the big winner. Use the prescriptions above and you will have a vehicle (the DQR process) ready built to carry forward your initiatives, support from the top level down to the front line for your approach across all the silos, a list of prioritised outstanding items to be working on, tools that can find new data quality issues and measure your value add.

All it takes is needed is the enthusiasm to take it forward.

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