

Successful migration of large data sets

Graham Every discusses the processes involved in high-volume data conversion and offers guidance to ensure a successful outcome.

As with all projects, when it comes to migrating large data sets into new systems or processes, 80% of the work is in the planning. It's a truism, but the more thought that goes into the planning stages, the less chance there is of large-scale problems coming to light late in the process and the greater chance there is of a successful conclusion and of client satisfaction.

The importance of planning can be most clearly seen by splitting the conversion process into seven separate stages, of which only one is the conversion itself:

- Identifying the client's requirements
- Analysing the data to be converted
- Creating the specification
- Creating the conversion program
- Converting sample data
- Converting the full date
- Overcoming real-world challenges.

I'll discuss each of these stages in turn, commenting on the processes involved and the pitfalls to avoid in each case. I'll conclude by considering the issues of concern to the budget-holders — the timescales and the costs.

Choosing between industry-standard and bespoke DTDs

It is worth comparing the benefits of industry standards such as DITA or DocBook with those of a bespoke application. In most cases, the industry standard will need to be customised to meet the client's requirements. In practice, this often means that the client has to compromise an internal process to fit the requirements of the application, or that the benefits of the DTD are lost. A bespoke application need involve no such compromise, and an experienced data conversion consultant will have a library of tried-and-tested applications that they can use to make creation cost-effective.

In either case, the client will need to have access to a technical consultant to monitor and maintain the DTD or schema once the conversion process is in place. They may choose to have you migrate the maintenance skills to an employee or to outsource to you when changes are required. Since neither industry standard nor bespoke options are maintenance-free, the flexibility of a bespoke DTD or schema is often the best solution. However, in the real world, it is more than likely that the client will want the project to use the industry standard, and this is often the deciding factor.

Identifying the client's requirements

This is the stage where the project is scoped in broad terms. What material is to be converted? Is there an in-house DTD or schema that will need to be followed? Is the data migrating to the web, to CD, to print? These are the obvious questions, of course.

To increase the likelihood of client satisfaction and a successful project, though, it is important to find out about the client's internal processes and take these into account too. These might range from the practical to the political. If

something 'has always been done that way' and it can be incorporated into the conversion with no great difficulty, then do it. Successful data conversion is something that fits into an existing process, not something that requires an existing process to be scrapped. If your proposal requires the restructuring of an internal process that has been in place since the year dot, you will never get full buy-in — or cooperation — from those involved in or affected by the conversion. And that is a recipe for an exhausting, painful and compromised project!

Analysing the data to be converted

This stage scopes the project in more detail. I insist that the people who work with the data on a daily basis are involved here because, quite simply, they have more knowledge of the data and the processes than the management team. More than once, they have identified previously unknown issues that have a massive effect on the conversion process for projects on which I have worked.

So, identify all sources of data and gather samples of each of them. (Do double check this list with the client. You would be surprised how often something is forgotten and, clearly, the later addition of a piece of legacy data adds time and money to the conversion process.) Obtain a copy of any existing DTD or schema too, together with any associated notes.

Once you have all the data, analyse the *electronic versions*. At paragraph level, check for consistencies of styles and/or typography. At in-line level, identify how special styles are marked-up, whether that be through character style or emphasis. (Ensure that your identification is unambiguous: for example, not all italicised text will necessarily refer to a book title, so you will need to find additional evidence.) While this stage is important, do not try to get more than you can from the documents: it may be that queries will need to be put in place to be handled in post-conversion clean-up. Once you have drawn your conclusions, it goes without saying that they need to be sense-checked with the client.

Once you and the client have analysed the data, this is the point at which, if it is needed, the DTD or schema is created or the suitability of an existing DTD or schema is discussed.

Creating the specification

When writing the conversion specification, include:

- Mapping between styles or typography and the Elements in the DTD or schema

- A detailed breakdown of what will be converted, and, just as importantly, what will not be converted. Once it is complete, as with all specifications, agree it with the client and remember it is the last chance you have to ensure you have set the client's expectations correctly.

Creating the conversion program

Once you have written or customised the DTD or schema, it needs to be tested with a small set of data. Once you are happy with the converted test data, ensure it is what the client requires too. The point at which the client sees the converted test data is often the point at which you hear their unspoken requirements for the first time. If the conversion is to meet with the client's complete satisfaction, these unspoken requirements need to be accommodated, even if they weren't part of the specification. In my opinion, if the data conversion doesn't meet the client's precise requirements, it hasn't been successful and will only serve to increase the suspicion in which data conversion is held in some circles!

Converting sample data

Before beginning to convert all the data, I strongly recommend converting a large sample set. Doing this gives you — and the client — the strongest possible reassurance that you have picked up all the potential anomalies and that the DTD caters for all eventualities before the entire data set undergoes conversion. Naturally, you need to adjust the conversion program in the light of any findings.

Converting the full data

Once you are happy you have completed all the preparation, you can convert the data, check it has gone through successfully and complete any post-conversion clean-up that is necessary. It is rare that some manual intervention is not required, but the work you put in before the conversion should keep it to an absolute minimum.

Be aware, however, that it is often at the clean-up stage that companies end up investing absurd sums of money. This is largely due to two scenarios. In many cases, it is because the client has developed its in-house processes but has not fed these developments through so they can be included in the conversion scripts. If not that, then it is the effect of an unspoken or avoided requirement that was not picked up during the creation of the conversion script.

Overcoming real-world challenges

The preceding may have given you the impression that data conversion is a relatively straightforward process and that success is almost guaranteed as long as you follow all the procedures. Welcome to the real world! While it

is true that any data can be converted, GIGO rules supreme: if you put garbage in, you will get garbage out. The quality of the data to be converted will have a direct correlation with the quality of the conversion result, so it's a matter of degrees of acceptability. If the client knows and, crucially, accepts that data conversion is not a magical cure-all for the data's shortcomings, then all is well. Of course, this is unlikely to be the case.

There are essentially two ways to improve the quality of the end result, and you and the client need to decide which method will be the most cost- and time-effective. The first method is to allot a period of time to post-conversion manual clean-up. If it is a relatively small amount of legacy data that is failing the quality control tests, this is likely to be the solution. If the problems present themselves on a larger scale or if the conversion project is a rolling one, it is likely that adjusting the source data is the correct route. This might involve re-formatting the text or, in extreme circumstances, it might require an interim conversion.

When it comes to costs of this, and indeed data conversion as a whole, it is, rather unhelpfully, a case of 'how long is a piece of string?' The only guidance I can give is that cost is inversely proportional to the quality of the incoming data. The more work involved in preparing the data for conversion, the greater the cost will be.

Conclusion

Converting large data sets can, and should, be a fun and rewarding experience. Unveiling a successfully completed conversion can draw gasps of amazement at what has been achieved. However, to get that gasp of amazement, you need to have listened to the client at every stage and adapted the conversion processes to fit their processes wherever possible. You also need to have set expectations so you have, in their eyes, over-delivered. Automated data conversion is often the solution for organisations looking to maximise the value of their data but being a realist, not a salesperson, is the best way to help the client realise these benefits. **C**

Graham Every has worked in the field of automated data conversion for over 17 years. His consultancy, Graham Every Ltd, provides large and small volume automated XML data conversion services to businesses and organisations worldwide. He is also a well-respected trainer and has spoken at numerous conferences.
E: graham@grahamevery.co.uk
W: www.grahamevery.co.uk

McFelder.com
Technical Translation Agency

The full service provider

- Translation
- Desk Top Publishing
- CAT Tools used for quality and quotation purposes
- Highly competitive pricing
- VAT exempt for UK clients
- Impressive Client List

Call us today on:
0870-068-1079 or
+34972575921
www.mcfelder.com
global@mcfelder.com