

White Paper



Avoiding the Integration Tar Pit

Agile Integration for fast results

... a tool is useful to scale up a process you understand well on the small scale; and to deal with increasing complexity evolving from a system that you understand. Conversely, if you are unable to manage data quality on the small scale, your efforts to improve data quality through automation are likely to be in vain.

David Norfolk

Introduction

Almost all projects involve data integration these days, because nearly every IT project involves multiple applications and databases. If you don't plan and design the data integration parts of your projects, you are destined for a data integration tar pit, with a tangled mess of point-to-point integrations of exponentially increasing complexity.

It is also important to bear in mind that not all data integration platforms are equally effective; see, for example: the "Comparative Costs and Uses of Data Integration Platforms", a [Bloor white paper](#) by Philip Howard.

If you are going to avoid the tar pit you need good process and metadata management, institutionalised with the help of effective people/culture management and effective tools that promote:

1. Real-time data quality control;
2. Re-usability;
3. Rapid development;
4. Incremental adoption; and,
5. Rapid response to change.

These are the 5 characteristics of Agile Integration for delivering cost effective data integration which can respond immediately to changing business needs and which doesn't grow more complex (and expensive) to manage over time.

This paper is addressed to two classes of readers: the IT manager below CIO level, responsible for budgets, meeting deadlines and satisfying internal customers with many needs and demands; and the software applications vendor, who has to deliver into almost any environment and integrate with their customers' existing systems.

The Integration Tar Pit

What the Tar Pit is and why it matters

Data integration is an inevitable part of the foundation of most software development today. No one develops in a green field site any more—there is always some other system to integrate with. This means that almost every development project today has an, often unplanned, integration project attached to it. It may appear that the cheapest way forward is to treat each project and its integration as a little silo: you can concentrate on the immediate business need, perhaps buy a package that fits the business requirement, and put just enough point-to-point integration in place to satisfy the immediate stakeholders. This causes many of the same problems that isolated data silos cause. Even if it works at the time, it can lead to bigger problems later:

- New data requirements, unplanned for at the start, may be expensive to satisfy when you find that the data model implied by your initial project doesn't suit the needs of new business stakeholders. For example, you may assume that there is only ever one account per customer but that may only be true in a small part of the organisation;
- If the new data integration requirements don't closely match what your initial point-to-point integration provides, you will probably find that you can't re-use your existing integration assets because they were designed with only one set of requirements in mind. Ripping out and replacing work because it wasn't designed for reuse is expensive; reusing integration assets designed to be re-purposed or reused is cost effective;
- After a few new business requirements have been dealt with in a siloed, piecemeal way, you find that you have exponentially increasing complexity. You have the chaos of overlapping point-to-point solutions that are increasingly difficult and expensive to extend, frequently break down, and require expensive experts to repair;
- After this stage has been reached, you find that accommodating new business requirements becomes harder and harder, because you must first unravel the complexity of existing integrations and ensure that none of them break when you satisfy new ones. As a result, it becomes cheaper to produce entirely new systems for new business requirements and run them in parallel with your old systems,

rather than to add to your old systems. This is wasteful and expensive compared to doing it properly in the first place, but it will probably be your only feasible option.

Data shock

The basic description of the tar pit above is the optimistic scenario. In practice, your initial siloed integration often doesn't go smoothly—probably as a result of "data shock". Data shock is when you discover that the quality and integrity of the data isn't as good as everybody assured you it was. Since the integration wasn't planned but just tacked onto the project as an afterthought, you probably didn't look at the data in detail. When you do, you may find a host of problems:

- Duplication is a serious problem. For example, J. Smith, John Smith, Mr Smith. Mr and Mrs Smith, Smith, J, John Smith, J Smyth Esq etc. may be one customer or any number of customers;
- Fields you thought were present have been filled with meaningless placeholders, because doing so makes data entry quicker;
- Files are unreadable and not backed up.

All this can be fixed—at a cost, in people resources, chargeable time (especially if you resort to outside data cleansing contractors) and missed delivery dates.

This is what the integration tar pit is. At first, it looks innocuous and the path towards it is paved with good intentions; "satisfy just the business requirements", "don't bother wasting money on anything the business doesn't need now", "buy something cheap and effective and just customise it". Then when you get into the tar pit, you find that things aren't as simple and easy as you expected—as you add new integrations to satisfy new business needs, everything gets more complicated and more expensive. Then, when you try to rationalise everything and get out of the tar pit, you find you can't—no one is quite sure which integrations are important, which siloed systems depend on other systems and no one has an overview of what it all does. In the integration tar pit, you don't have time to worry about real-time data quality or re-usability; and rapid development is a dream. Incremental adoption and rapid response to change are impossible. There must be a better way.

The Integration Tar Pit

Agile integration, a better way

“Agile” in IT systems development, has come to mean many things, some quasi-religious. In the context of data integration, its meaning is simple:

- The ability to embrace change without increasing complexity and cost;
- The ability to meet changing business needs without being held back by technology;
- The ability to accommodate change with less consequent downtime;
- The ability to adapt to new requirements quickly.

In practice, agile integration comes from understanding and modelling data so that you have a framework for building integration assets that address the business needs of today without compromising any business needs that may emerge tomorrow. It comes from building for reuse with an integration layer that can accommodate present and future needs without constantly increasing complexity. Also, most importantly, it comes from using integration tools that support metadata management, have a capable integration layer architecture, and feature asset reuse and transparency. An appropriate integration tool makes good practices easier to implement, share and institutionalise. It helps break down silos and give everyone access to integration assets which they can understand and reuse.

Best practices for integration agility

The issues that an integration project has to address include:

- How fast can you deliver integration and how do you know that you've "finished"?
- How can you manage changes in the integrated data (or subsets of it) over time without costs increasing exponentially?
- How do you manage new or changing technology platforms (such as "cloud computing", or "Software as a Service") without making them into new information silos?
- How can you manage the metadata (semantics) associated with the data without this becoming tool-centric?

The general answer, for a reasonably mature business- and metrics-focused company, is to avoid practices that increase complexity and limit future integration; and to buy powerful tools with open interfaces that enable its business integration processes. This is easy to say, but unless you have overcome your

fundamental data quality issues, tools (however good) are unlikely to help. If you are unable to manage data quality manually (at least on a subset of your environment), then the risks associated with automated integration increase materially. In other words, a tool is useful to scale up a process you understand well on the small scale; and to deal with increasing complexity evolving from a system that you understand. Conversely, if you are unable to manage data quality on the small scale, your efforts to improve data quality through automation are likely to be in vain. Nevertheless, as illustrated in the mini case study "Climbing out of the tar pit", it is still possible to address impending data integration issues if you recognise them in time. The ideal is to avoid the tar pit by following good practices from the beginning. However, this may not always be possible—the tar pit is an anti-pattern, something that looks attractive and is easy to fall into after following a path paved with good intentions. So, good practice has to accommodate tar pit avoidance, the avoidance of data shock and recovery from the tar pit anti-pattern.

Case studies

Case study 1 - good practices for integration

The subject of this story services a loan portfolio worth \$1.48 trillion with over 8.9 million loans, managing and distributing taxes from escrow accounts. It also provides property data services from a property database with almost 120 million parcels within 1700 counties and 12,000 taxing authorities. The subject is an advocate of Continuous Process Improvement and Six Sigma and manages projects across multiple departments. It sees the constantly increasing scope of integration as an important characteristic of business automation today, not just across local silos but also across international corporations and across company boundaries. In order for integration tools to facilitate this sort of organically growing integration, they must support:

- Incremental adoption;
- Changing deployment scenarios; and,
- Reuse of local integration globally.

The issue this company sees is that people aren't always thinking of where their integration will grow to, so they aren't building in flexibility and aren't efficiently reusing their small integrations at larger scale.

Message: Start with just what is necessary for current business needs but build with future development and uses in mind. Don't just build a box, build reusable box components.

A "build once use many times" strategy can save a lot of time, effort and money in the longer term as business demands grow. Integration solutions must support **rapid development of new integrations** and the **incremental adoption** of one integration project at a time provides rapid ROI and fast business advantage. Having easily reusable and modifiable components also delivers the advantage of **rapid response to changing requirements**.

Case study 2 - the critical importance of the integration layer

The subject of this story is a large privately-owned international real estate development and management company running over 1000 projects in over 100 cities; it has offices in 16 countries. This case study highlights that many integration vendors "oversimplify and misrepresent" the integration issue while overstating their own capabilities. For example, many integration tools can't support integrating data from a range of SaaS and on-premises solutions (in this case, for transaction processing and investor data; marketing contacts/CRM; marketing fulfilment; expense reporting; vendor payments). Over 80% of the data this organisation needs in order to function is external data, so it must have robust connectivity and high reliability. An agile integration solution must support:

- Complex and layered integrations, both within and between companies (not just atomic "a to b" data movements);
- Integration across firewalls, to both hosted and on-premises applications;
- Bi-directional integration;
- A partnership between IT and business data users;
- The ability to change creatively as the business changes.

The biggest issue is that integration is not a pure IT project, the business is a major stakeholder too. The people building the solution using such an agile product must understand both the business processes and the usage of data, both the corporate needs and the IT requirements. They must always design for flexibility and think beyond merely moving data—because integration processes and data structures are constantly evolving, as are business needs. The people behind the integration effort are important—a tool can't do everything. One clever thing that this company's integration architects did, which was made possible by the flexibility of their toolset, was to reuse their file structure metadata across multiple filetypes. This allowed them to reuse maps and minimise the amount of effort needed when making changes.

Case studies

Message: A flexible architecture, and an active partnership between IT and the business users during design, delivers a robust, resilient integration solution that can weather change well.

What the **integration layer** delivers in this case is broad connectivity, flexibility, reusability, and rapid response to change.

Case study 3 - climbing out of the tar pit

This story is somewhat longer than the others, reflecting the fact that escaping the tar pit is harder work (and more expensive) than not falling into it in the first place. It comes from an American finance company which provides a service for U.S.-based mortgage companies, which are responsible for confirming the amount of property tax due on a real-estate parcel and when it is due for payment. The subject charges less for its services than the competition because it automates the process and can also remit funds together with the appropriate documentation. However, this means managing a major integration environment, with assessor and tax data and funds transfer documentation in thousands of different jurisdiction-dependent formats.

The subject began its journey towards the tar pit through the purchase of a solution intended for use by “business-savvy” business users: “the product was sufficient for our needs then, but as our volume and staff grew and our organisation began to specialise and differentiate the staff, its limitations were quickly exposed”. Support for this product was costly and it was almost impossible to upgrade. The historical product supported conventional data analysis well, but its overheads were significant, especially at larger scale: processing for one customer needed two data dictionary files, two transaction files, three map files, a special file for custom functions, a BAT and an INI file for production running, and possibly more files for normalised-record formatting. The real killer was that all files created were binary files, which was a huge problem for cloning or mass-changing them for reuse. Fortunately this company really only experienced the edges of the tar pit, because of its prompt reaction to the issues. The subject climbed out of its tar pit by implementing—at high speed and in time for its most critical quarter—a new, more-agile integration tool that facilitated reuse of its existing integration assets.

The key characteristic of the new product in this situation, apart from its ability to cope with the anticipated volumes and many different record formats, was speed of implementation with “no surprises”.

The process of getting out of the tar pit mustn’t become an end in itself and mustn’t impact production processing and production schedules adversely. The subject had several good solid business reasons for choosing this new tool to pull them out of the tar pit:

- The new tool was available to test and evaluate ahead of planned introduction—with no barriers from vendor licensing controls or support schedules;
- There was no need for vendor support or database set-up as part of installation;
- All of the old integration maps could be reused (converted to work with the new tool).

The fourth quarter of the year is the busiest time of the year for the subject: “we live or die during this time, and this is how we are measured by our clients. Nearly half of our processing for the entire year is done within a 2 ½ month period”. The subject was able to ramp up for this critical fourth quarter, using an entirely new tool, while training a new person on the new software, in a 2-month period.

Message: Ease of implementation is essential if software is to help get you out of the tar pit.

Obviously, **real-time data quality profiling** is important here: with a hard deadline around mission-critical processing, bad data couldn’t be a system-crashing surprise. Also a mistake while converting the tax data for its clients, the largest companies in the US mortgage lending sector, could prove very costly—tax offices could hand out penalties if the tax data returned was inaccurate. However, the critical aspects of the tools used to escape from this tar pit are **re-usability** and **rapid development**.

Conclusions

Avoiding the tar pit

More mature, business-focused companies will want to integrate at the business process level and, possibly, use decoupled service-based architectures to make integration easier and more agile; the ultimate goal is to integrate at an abstracted, metadata level maintaining an integration layer which exposes both legacy systems and new developments as well-defined, well-documented, business-level services. This should pre-emptively avoid the integration tar pit. However, all integration projects come down to data integration in the end, even if the higher-level business abstractions and frameworks, which provide a context for data integration, are still important. The second mini case study, “The critical importance of the integration layer”, explored practical tar pit avoidance further. Nevertheless, although tar pit avoidance is not an impracticable dream, it may well be difficult for companies overtaken by circumstances. Merger and acquisition activity is the classic opportunity for unexpected integrations, with externally imposed but often infeasible deadlines. Many mergers/acquisitions that entirely made sense in business terms have been highly unsuccessful—because the cost of integrating their disparate business automation solutions was higher than expected or integration simply couldn’t be delivered on a timescale that made business sense.

Agile integration

Agile data integration, and the incremental adoption of new developments that is part of the agile approach, seeks to prevent problems. It achieves this by supporting re-use of existing integration assets wherever possible, thus providing more rapid development of integration solutions and a rapid response to change, while also controlling costs. Incremental adoption or delivery is an important part of agile development. It lets you deliver near-continuous increments of benefit, thus maintaining interest and “buy-in” from stakeholders. It keeps any surprises small and non-critical, thus reducing project risk and making the accommodation of change easier, because when change is needed, you don’t have to change an entire, monolithic, edifice. You deliver incremental adoption by “eating the elephant in small pieces”; that is, by dividing an integration project up into small, coherent integrations, each satisfying a discrete requirement or need. Nevertheless, avoiding the

tar pit requires special characteristics in both company culture and tools:

- Planning must have a longer term, as well as a short term, perspective;
- Abstraction and modelling skills must be available;
- IT and the business mustn’t operate in separate silos. Integration projects mustn’t become an end in themselves but a means to deliver business benefit. For this to work in practice, both IT and business stakeholders must be involved in modelling and in deciding which abstractions should be implemented;
- Data integration tools need to be able to take advantage of metadata abstractions, provide strong visualisation of integration structures and enable the reuse and re-purposing of integration. However, the metadata management must not be tool-centric.

Data shock prevention

Pre-emptively avoiding data shock is comparatively easy, providing that you can institutionalise good practice using suitable tools. It does assume that you have analysed the semantics, allowed values and format of your data and that the metadata model holding this knowledge is both accessible and maintained. This involves a facts-based trade-off of your investment in metadata discipline (developing and maintaining a metadata model takes discipline and costs money against the business benefit from managing and responding to change more cost effectively). In other words, thinking about the semantics and structure of your data is useful but don’t physically implement any models that can’t be shown to deliver actual business benefit. Appropriate investment in tools can help to institutionalise the discipline needed and reduce the short-term overhead of metadata management while ensuring that the longer-term business benefits are realised. Once this is in place, the fundamental principle for avoiding data shock is to quality-assure the data stream on its way in, in real time, although data streams can be complex and you’ll still need to profile your existing data. You should expect data quality issues with an integration project and should automate the profiling of your existing data at the same time as you apply automated data profiling and multiple business rule checks to new data as it enters the system. The principle

Conclusions

“ You should expect data quality issues with an integration project and should automate the profiling of your existing data at the same time as you apply automated data profiling and multiple business rule checks to new data as it enters the system. The principle is that you don't tolerate poor quality data in the system (where it can be processed and probably damage the business) and you stop new bad quality data getting in at the earliest stage possible. ”

is that you don't tolerate poor quality data in the system (where it can be processed and probably damage the business) and you stop new bad quality data getting in at the earliest stage possible.

The tar pit anti-pattern

However, sometimes it is too late to avoid the tar pit. Luckily, however, an anti-pattern not only describes a state and its consequences, it also documents a path back to good practice. Unfortunately, getting out of the tar pit may not be easy and it certainly won't be as cheap as avoiding it in the first place. The mini case study “Climbing out of the tar pit” explored this process. Fundamentally, good tools that are easy to install and use are key to getting out of the tar pit—and it helps if they are freely available for evaluation, pilot studies and training.

Further Information

Further information about this subject is available from <http://www.BloorResearch.com/update/1069>

Bloor Research overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the right story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the right story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter "noise" and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channel.

Founded in 1989, we have spent over two decades distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

About the author

David Norfolk
Practice Leader
Focus Area: Development/Governance



David Norfolk first became interested in computers and programming quality in the 1970s, working in the Research School of Chemistry at the Australian National University. Here he discovered that computers could deliver misleading answers, even when programmed by very clever people, and was taught to program in FORTRAN. His ongoing interest in all things related to development has culminated in his joining Bloor in 2007 and taking on the development brief.

Development here refers especially to automated systems development. This covers technology including acronym-driven tools such as: Application Lifecycle Management (ALM), Integrated Development Environments (IDE), Model Driven Architecture (MDA), automated data analysis tools and metadata repositories, requirements modelling tools and so on. It also covers the processes behind them and the people issues associated with implementing them. Of particular interest is organisational maturity as a prerequisite for implementing effective (measured) process and ITIL (v3) as a framework for automated service delivery.

David is a past co-editor (and co-owner) of Application Development Advisor and associate editor for the launch of Register Developer, and is currently executive editor for GEE's "IT Policies and Procedures" product. He has an honours degree in Chemistry and is a Chartered IT Professional, has a somewhat rusty NetWare 5 CNE certification and is a full Member of the British Computer Society (where he is on the committee of the Configuration Management Specialist Group).

His early career involved working in database administration (DBA) and operations research for the Australian Public Service in Canberra. David then returned to his UK birthplace (1982) where he worked for Bank of America and Swiss Bank Corporation, at various times holding positions in DBA, systems development method and standards, internal control, network management, technology risk and even PC support. He was instrumental in introducing a formal systems development process for the Bank of America Global Banking product in Croydon.

In 1992 he started a new career as a professional writer and analyst. Since then he has written for many major computer magazines and various specialist titles around the world. He helped plan, document and photograph the CMMI Made Practical conference at the IoD, London in 2005 and has written many industry white papers and research reports including: IT Governance (for Thorogood), Online Banking (for FT Business Reports), Developing a Network Computing Strategy and Corporate Desktop Services (for Business Intelligence), the Business Implications of Adopting Object Technology (for Elan Publishing).

He has his own company, David Rhys Enterprises Ltd, which he runs from his home in Chippenham, where his spare moments (if any) are spent on photography, sailing and listening to music.

Copyright & disclaimer

This document is copyright © 2009 Bloor Research. No part of this publication may be reproduced by any method whatsoever without the prior consent of Bloor Research.

Due to the nature of this material, numerous hardware and software products have been mentioned by name. In the majority, if not all, of the cases, these product names are claimed as trademarks by the companies that manufacture the products. It is not Bloor Research's intent to claim these names or trademarks as our own. Likewise, company logos, graphics or screen shots have been reproduced with the consent of the owner and are subject to that owner's copyright.

Whilst every care has been taken in the preparation of this document to ensure that the information is correct, the publishers cannot accept responsibility for any errors or omissions.



2nd Floor,
145-157 St John Street
LONDON,
EC1V 4PY, United Kingdom

Tel: +44 (0)20 7043 9750
Fax: +44 (0)20 7043 9748
Web: www.BloorResearch.com
email: info@BloorResearch.com