

**Pitfalls and Perils in Planning Data Quality Projects**

helpIT SYSTEMS

- deduplication
- addressing
- suppression
- duplicate prevention
- inquiry

- **Pitfalls to avoid when planning a project**
- **Practical methods to ensure objective success**
- **Know your enemy!**

**Steve Tootill**  
CEO, helpIT systems

### **Steve Tootill**

Steve Tootill is founder and CEO of data cleansing specialist helpIT systems. Steve has over 30 years experience in the IT Industry, starting as a programmer. After gaining a BSc in Mathematics from Imperial College, London, Steve spent many years with Computations Pty Ltd, an Australian software house, latterly as Software Director of Computations UK. Following a period as a consultant, Steve became head of IT at Family Assurance, where he realised that there was a dearth of packaged matching software, especially of software that would make use of all the data that a company held about its customers. helpIT systems was established in 1991 to remedy this, with Prudential Life of Ireland as the first user of its first deduplication package.

Now dividing his time between New York and London, Steve oversees helpIT's operations in the US and UK, with particular responsibility for product strategy.

**Mr. Tootill, aka Mr.**

- **Toothill**
- **Toutil**
- **Tootle**
- **Tootal**
- **Tutil**
- **Twotill**
- **Foothill**
- **Toohill**
- **Toosti**
- **Stoolchill**

deduplication  
addressing  
suppression  
duplicate prevention  
inquiry

addressIT  
suppressIT  
filterIT  
findIT

Implementation  
Data Cleansing  
Software

over 1000  
Customers  
Worldwide

With my surname, I've got used to a lot of misspellings over the years... more often than not, people put an H in my name, even when I spell it out correctly!

## About helpIT systems

- **Established over 15 years in UK**
- **Global client base**
  - over 1,000 companies worldwide
- **Offices in UK and US**
- **Experts in data cleansing**
- **Proprietary fuzzy matching technology**
- **Used for contact data management**

**Steve Tootill**  
CEO, helpIT systems  
SteveT@helpIT.com

Services: deduplication, addressing, suppression, duplicate prevention, inquiry

Client logos: QAS, Barnardo's, Britannia Building Society, Christian Aid, hp, m, Learning Tree International, Microsoft, AVIVA, PUBLISHERS CLEARING HOUSE, ET

AUTO

Perhaps my name was one of the motivations for me founding helpIT systems, providing software that initially focused on fuzzy matching of names and addresses.

The graphic illustrates a few of our better known customers in the UK and the US. Around  $\frac{2}{3}$  of our customers are in the UK, most of the rest in the USA, although we do have customers in around 15 countries in all.

Our software includes dedupe, customer data integration, address/postal code verification, suppression against industry standard files for MPS, gone aways and deceased names, as well as components for preventing duplicates at point of entry and intelligent inquiry.

So, my company and I have seen a lot of data quality projects involving our software, our competitors' software, in house developments and a combination, with a lot of things done right and a lot done wrong.

## Common pitfalls

- **IT and Business don't work together**
- **Ineffective planning**
- **Passing the buck**
- **Square data, round boxes**
- **Data decay**

deduplication  
addressing  
suppression  
duplicate prevention  
inquiry

addressIT  
suppressIT  
filterIT  
Data Cleansing Software  
Customers Worldwide

AUTO

I'm going to talk about the pitfalls, common fallacies and resulting project failures, then conclude with pointers for success. I welcome comments about your own experiences and questions along the way.

So, the key pitfalls that I'm going to talk about are those listed here.



The slide features a light blue background with a pattern of overlapping circles. In the top right corner, there is a dark blue circular logo for 'helpIT SYSTEMS' with a grid of dots. Below the logo is a legend with five items: 'deduplication' (red dot), 'addressing' (purple dot), 'suppression' (yellow dot), 'duplicate prevention' (green dot), and 'inquiry' (orange dot). The main title 'Gartner workshop survey results' is in a large, bold, dark blue font. Below the title, there are several smaller text elements: 'matchIT', 'addressIT', 'suppressIT', 'filterIT', 'findIT', 'Data Cleansing Software', and 'Over 1000 Customers Worldwide'. The central text provides two Gartner document links. The first link is for a presentation with a bar chart, and the second link is for more relevant research.

## Gartner workshop survey results

The bar chart in the presentation is available from:  
[www.gartner.com/DisplayDocument?ref=g\\_search&id=587907&subref=simplesearch](http://www.gartner.com/DisplayDocument?ref=g_search&id=587907&subref=simplesearch)

More relevant research is also available from:  
[www.gartner.com/DisplayDocument?ref=g\\_search&id=497089&subref=simplesearch](http://www.gartner.com/DisplayDocument?ref=g_search&id=497089&subref=simplesearch)

This chart is from a survey of participants in a recent Gartner workshop in which “groups of attendees worked to identify the major risks and potential reasons for failure of their business intelligence projects.”

The equal second risk factor identified was “Addressing Data Quality Issues”, with the top risk factor “Lack of Sponsorship/Engagement Outside of IT”.

“Many workshop participants reported that their BI initiatives were sponsored, funded and led by the IT organization. As a result, they struggle to obtain the levels of engagement from both business management and users that are needed to define requirements and validate results. This often leads to misguided development efforts whereby the IT organization delivers tools and applications that do not meet the business need.”

Although the workshop addressed a wider brief than data quality, these key findings are 100% relevant to our subject today.

**IT and Business  
don't work together**

- **Lack of business involvement**
- **Lack of executive commitment**
- **Insufficient funding**
- **Not establishing effective success criteria**
- **Lack of agreement on scope**

**helpIT**  
SYSTEMS

- deduplication
- addressing
- suppression
- duplicate prevention
- inquiry

Background text: Data Cleansing Software, filterIT, addressIT, suppressIT, filterIT, findIT, Customer Worldwide

## MANUAL

Quoting from another Gartner research paper: “Business users must be involved alongside the IT department to identify and resolve data-quality issues and ensure ongoing governance and stewardship of the data being consumed by BI applications.”

In order to ensure the right level of commitment from the business areas, the executive management of the organisation must be committed from the top to resource the project properly – both in terms of allocating experienced users to the project and funding it adequately.

When it comes to establishing the right level of funding, the context is not simply how much has been invested to date in systems, but more importantly how much has been invested to acquire the customer and prospect database and the potential it has for contributing to increases in income.

It is crucial that experienced users help frame the goals of the project, particularly the criteria by which the success of the project will be judged.

Lack of agreement between business representatives and IT on the detailed scope will doom the project to failure.

## Ineffective planning

- **Over scoping – “what do we want?”**
- **Oversimplification – lack of analysis**
- **Underestimation – extrapolating from simple cases**
- **Expecting projects to “evolve”**
- **Skimping examination of potential solutions**
- **Not scheduling effective trials**

On the other hand, simply agreeing the scope doesn't guarantee success! Feet must be kept firmly on the ground here, as a series of shorter projects is far more likely to succeed than a single large project - some of the speakers yesterday touched on this point. Prioritisation of requirements according to the business gains they allow can enable tasks to be allocated to smaller projects as appropriate.

Every requirement has a cost and a benefit. Detailed analysis of every requirement is necessary to ensure that the corresponding development costs as well as the benefits can be estimated.

With data quality, the devil really is in the detail. It's OK to illustrate requirements with simple examples, but not enough to estimate development effort using just these cases.

Don't be tempted to allow your project to evolve, with requirements being fleshed out and added after the project is under way. To avoid the necessary detailed requirements specification taking too long, define projects that are as short and self-contained as possible.

Having identified your requirements, you should evaluate more than one potential solution – in a data quality project, this is likely to comprise partially or solely of software from external vendors. It is important to evaluate any candidate internal solutions with the same stringency as you use for external solutions. When you are developing a system in house, if possible, use proven, reusable components to develop around.

You must schedule effective trials for the chosen solution. This applies to data quality projects as much as baggage handling systems and the dangers of not doing so are apparent.

## Passing the buck

- **Handing over ownership of the problem**
- **Expecting the incumbent supplier to fix it**

deduplication  
addressing  
suppression  
duplicate prevention  
inquiry

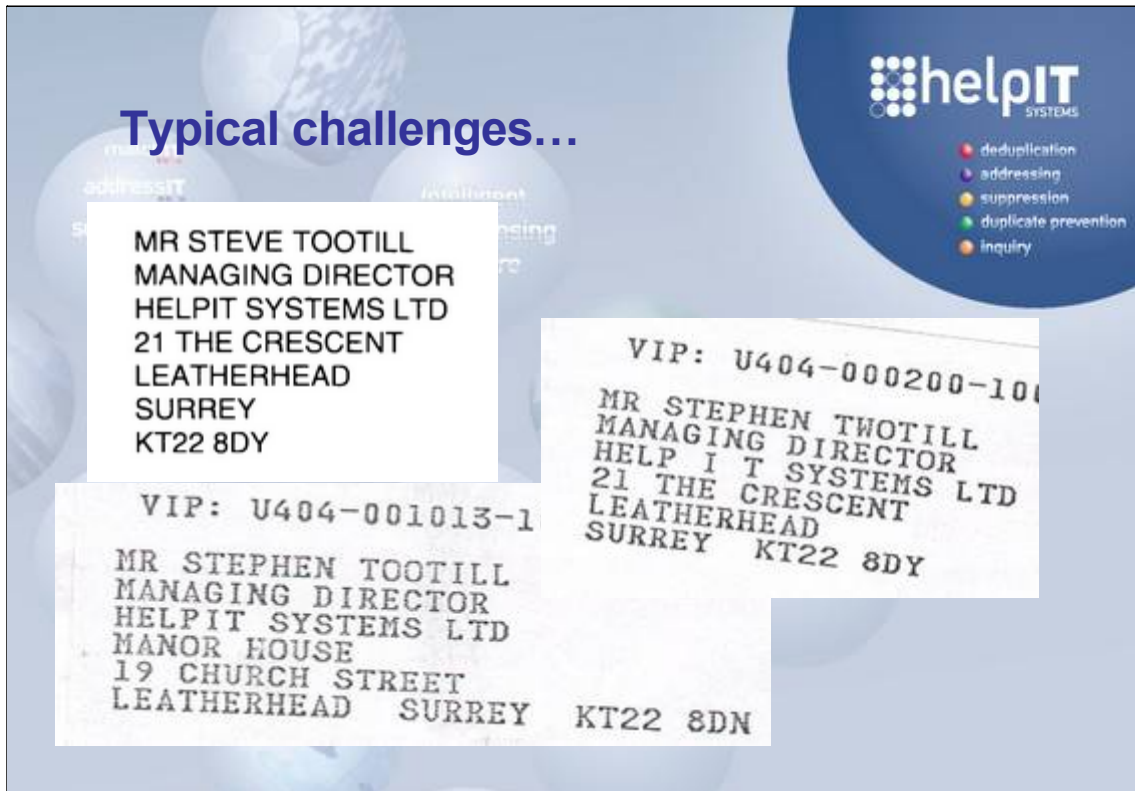
Data Cleansing Software

over 1000 Customers Worldwide

addressIT  
suppressIT  
filterIT  
mergeIT  
auditIT

However, using a solution from an external vendor does not mean that you should abdicate your role. According to Bill Hostmann of Gartner, "companies that say 'we can outsource the whole darn thing,' should think again. Service providers can complement a skills shortage and provide 're-usable methodologies,' but they are no replacement for knowledgeable employees." At helpIT systems, we saw an example of trying to hand over ownership of the problem to an external vendor with the NHS National Program for IT, where some NHS managers seemed solely concerned with how to spend the money rather than helping to establish realistic requirements.

Your current in house systems or external software suppliers are likely to specialise in your business applications, rather than data quality. Expecting them to resolve your data quality issues without specialist assistance is rather like expecting your office management team to make your desks and chairs.



These examples of duplicate mail that I received illustrate the scale of the challenge facing a development team needing to match records within and across systems.

Here you can see a short form of a name in Steve/Stephen, a phonetic match in Tootill/Twotill and a non-phonetic fuzzy match in HELPIT and HELP I T. Of course, this is a very limited example of the kinds of match that effective data cleansing will have to cope with, particularly with regard to company names.

The third example shows a different address for helpIT systems, although it is obviously the same company and the same individual. Specialist software using Royal Mail's Postal Address File and possibly external business datasets is required to establish which is the correct address.

**Passing the buck**

- **Handing over ownership of the problem**
- **Expecting the incumbent supplier to fix it**
- **Adopting path of least resistance**
- **Undervaluing in house knowledge**

Over 1000 Customers Worldwide

helpIT SYSTEMS

- deduplication
- addressing
- suppression
- duplicate prevention
- inquiry

Addressing Software

Further ways of passing the buck continue with the temptation to take the easy option. This can often be to bow to unreasonable requirements, or an insistence that a user department is allowed to continue doing things the way it always has done.

It is important to remember that you know more about your business than any external supplier, so it is vital to use this knowledge in your project – even if the real business experts are always the people with the least time to spare, they have to be allowed to make time for proactive work on your project rather than simply reactive day to day work. Ask yourselves how you would cope if an expert user left or was seriously ill – you would manage, so you must be able to manage if they devote maybe 80% or 90% of their time to your project, leaving 10% or 20% to consult to the people taking up their burden.

## Round data, square boxes

- **Use of old systems for new products/methods**
- **Merging unlike data (differing structures and standards)**
  - Title: Mr; First name: John; Surname: Smith
  - Name: Mr J T Smith
  - Name: JOHN T SMITH BSC
  - Name: John & Mary Smith
  - Name: Mr & Mrs J Smith
  - Name: Mr John Smith & Mrs Mary Smith
- **Uncontrolled data proliferation**

Often, data quality projects have to cope with problems due to legacy systems being used to hold data for which they were not designed. To illustrate this problem, I remember in the days of the forerunner of our matchIT dedupe package, we had a customer building a marketing database for a cross channel ferry company, when they were preparing to meet the challenge of the Channel Tunnel. The system was comparing every record for a postcode with every other record for that postcode, and making very rapid progress on that key, when it suddenly seemed to grind to a halt. On inspection, we found that out of the half million records in that file, some 25,000 had the address and postcode of the company's main administrative office. The problem was that when a ferry ticket was bought over the phone for collection at the port, the operator only had the passenger's name and vehicle registration number, but their system insisted on an address and postcode being entered – so they entered their office address. After we'd updated our software to allow for unnaturally large clusters like this, we found out how useful the feature was when another early user, a large life insurance company, were matching their prospect database with their master file – they apparently had 15,000 people all born on 3<sup>rd</sup> March 1933. This was due to a quotation system that insisted that a date of birth be entered, even when no date of birth was required for that particular type of policy. So, the users were told to enter 3/3/33...

More commonly, there are the challenges posed by having to reformat data into different fields to be able to amalgamate data from different systems e.g. having to split a name into separate fields. This is easy enough for simple single names with no qualifications, but as you can see from these limited examples, can get very complex. Often, you have differing numbers of address lines in different data sources, town fixed in a specific line or not etc.

Sometimes, these challenges are ducked and disparate data definitions are perpetuated across systems, leading to duplication of record maintenance instead of a single update.

**Data decay**

- **Expecting clean data to stay clean**
- **Underestimating change**
- **Ignoring compliance**

helpIT SYSTEMS

- deduplication
- addressing
- suppression
- duplicate prevention
- inquiry

addressIT  
suppressIT  
filterIT

Over 1000 Customers Worldwide

Data Cleansing Software

Data decays for many reasons:

- People die and move
- Postcodes and standard postal addresses change to allow for population shifts and new methods
- Data that was adequate for the purpose that it was originally collected is not adequate for new requirements
- Lapsed customers and old prospects have limited or no value after a while

With an average of 9,000 people moving house and 2,000 people dying every day consumer data decays by at least 3-4 % per annum and for business databases, the rate of change is much higher. Allied with the other reasons for change, the scale of change is extremely high.

One other example of change stems from legislation, which continually requires new and more onerous standards for maintenance and retention of personal data.

**Major Fallacies**

- **“If we build it, they will come”**
- **Measuring effectiveness by numbers**
- **“Answers should be black and white”**
- **“All data should be standardised”**
- **“Non-standardised data is poor data”**

**helpIT SYSTEMS**

- deduplication
- addressing
- suppression
- duplicate prevention
- inquiry

## AUTO

There are a lot of misconceptions that can bedevil your data quality project. I've listed a few here and we'll look at each of these in turn.

**Fallacy: If we build it, they will come**

- **“We can figure out what the users want”**
- **“They’ll make time to learn how to use it”**
- **“They’ll be delighted to drop the old methods”**

helpIT SYSTEMS

- deduplication
- addressing
- suppression
- duplicate prevention
- inquiry

Over 1000 Customers Worldwide

Data Cleansing Software

Who’s seen a movie called Field of Dreams?

To quote Gartner’s Research Director Bill Hostmann again, "Too many IT departments build a data warehouse based on the assumption that once it is built, users will automatically see the benefit," he said. "BI applications require a clear and intimate understanding of the business itself and it is only by working on business and IT issues in tandem that the real value of BI is realised."

Hostmann said that some companies believed that if they built a BI system – without considering business, user or training requirements – the value would be so obvious that users would clamor to learn and use the system.

What can seem like an antiquated, cumbersome set of procedures to an IT analyst can be comfortingly familiar to a user, who can be surprisingly ingenious in making something work to the best of its (usually limited) potential. Under these circumstances, dropping the old system for one designed without sufficient input from users is often an impossible sell.

## Fallacy: Measuring effectiveness by numbers

addressIT  
suppressIT  
filterIT  
Data Cleansing Software

helpIT SYSTEMS

- deduplication
- addressing
- suppression
- duplicate prevention
- inquiry

- **You can't simply measure data quality by:**
  - Percentage match rate against Postal Address File
  - Duplicate record counts
  - Cost of suppressions using industry suppression files
  - Processing speed

Over 1000 Customers Worldwide

Accuracy is the only worthwhile measure. For address quality, where an address can match against more than one address on the Royal Mail PAF file, records can be matched wrongly and postcodes and addresses updated wrongly as a result.

The false reporting of a potential duplicate is just as damaging as missing a duplicate match. A simple count of duplicates is meaningless – it is the numbers of true and false matches reported by the software that is significant.

Cost of suppression against chargeable suppression files such as disConnect from Equifax, GAS and TBR from The Read Group is easy to report – but the worthwhile measure is the overall saving on wasted mail and the increase in mail reaching the correct address through using files such as National Change of Address (NCOA). Of course, it is very difficult to measure the value of preserving the image of your brand by ensuring that customers do not receive wrongly addresses, duplicate mail or mail for previous occupants or deceased members of their family.

The only effective way of measuring accuracy of address verification and dedupe software packages is to compare the differences in the resultant files produced by the leading candidate solutions. For dedupe or CDI, this can be done easily for comparing two products by matching a good test file in each product, then matching each product's resultant single customer view file in the other product. This easily enables you to see if either product is missing good matches, or finding false matches. Here, you should allow for any grading that the products might do, by using a realistic threshold for the grade of match when determining dedupe and reporting thresholds.

Finally, effectively improving data quality of large volumes of data can take software hours or even days. The important measure here is whether it can fit in the windows of time available, not how fast it runs.

**Fallacy: Answers should be black and white**

- **“If the data can’t be proved right, then it’s wrong”**
  - What about business addresses with floor numbers?
  - Or vanity addresses?
- **“It’s either a match or it isn’t”**

helpIT SYSTEMS

- deduplication
- addressing
- suppression
- duplicate prevention
- inquiry

addressIT  
suppressIT  
filterIT  
findIT

fulfillment  
Data Cleansing  
Software

Over 1000  
Customers  
Worldwide

The most obvious example that gives the lie to the first statement is a business address containing a department name or floor number that is not on the Royal Mail PAF file. It can’t be proved right, but the data will be poorer without it.

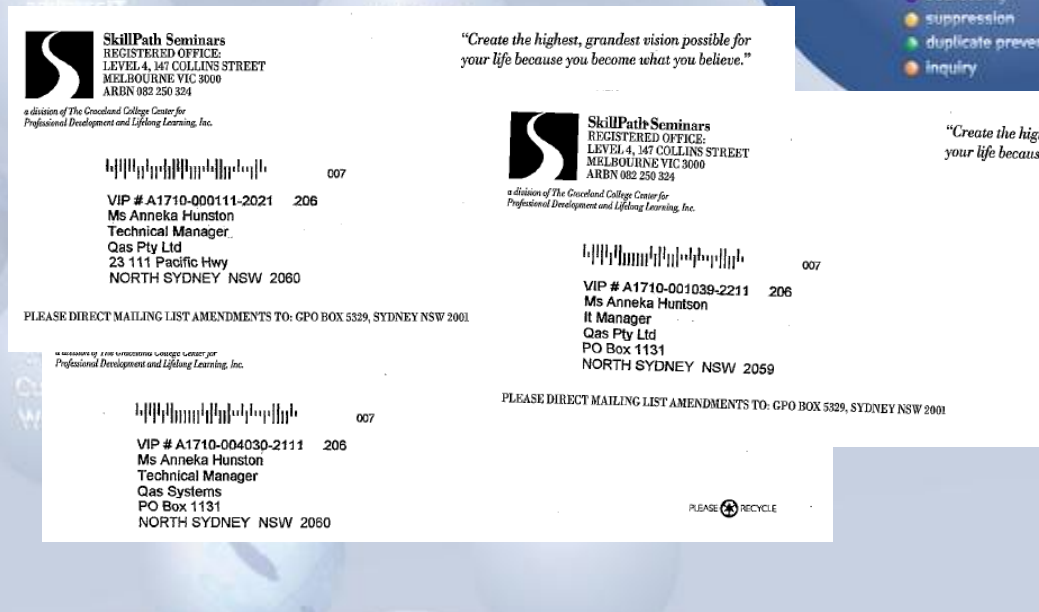
Similarly “vanity addresses” may not conform to Royal Mail standard, but that is the address that the customer used when they filled in the order or inquiry form, so it’s the address that they prefer you to use.

Given the almost infinite numbers of ways that people can enter names and addresses into databases, there will always be a “grey area” of records that may or may not be duplicates. The most effective matching software grades matches to allow the user to choose a level of comfort according to the nature of the requirement – for a cold mailing they might err on the side of calling it a match (“overkill”), for a CDI project, they might err on the side of not calling it a match (“underkill”), or want to use underkill initially and then review all the matches in the grey area manually. It is essential that your project allows for the grey area explicitly, in one of these ways.

## Fallacy: All data should be standardised

helpIT  
SYSTEMS

- deduplication
- addressing
- suppression
- duplicate prevention
- inquiry



This builds on the point that answers aren't always black or white. Different data standards can be required for different purposes.

This is a real example sent to me by our Australian distributor, QAS. Here, mail should go to the PO Box address but deliveries of goods to the street address.

Variation in the spelling of the surname, how the company has been keyed, job title and postal code also illustrate what can happen with real data.

## Fallacy: All data should be standardised

- “We’ll adopt a single set of standards across the board”
- “All addresses can conform”
- “Non-standardised data is poor data”
- “Poor addresses can be improved automatically”

deduplication  
addressing  
suppression  
duplicate prevention  
inquiry

Some address data is so poor that it can not be reliably corrected using Royal Mail PAF. It is better to measure the completeness of such addresses using parsing algorithms that are independent of the PAF file and to make a decision whether to keep each such address or drop it, or use phone or email or other research to correct it, than to expect its automatic correction.

## Resulting Failures

- **OVERRUNS, abandonment**
- **Non-delivery of benefits**
- **Recurrence of problems**
- **Uncontrolled duplication of data**
- **The enemy has moved on**

These failures all stem from the flaws and fallacies that we've discussed. Let's look at each of these headings in some detail.

## Overruns, abandonment

- **Failure to plan properly**
- **Insufficient contingency**
- **Not establishing expected results**
- **Not agreeing when it's finished**
- **Creeping scope undermines urgency**

### FIRST TWO BULLETS:

These stem directly from flaws in the planning process.

A key to successful software development projects is the use of test-driven development. With this approach, you don't just write the documentation up front, you also detail inputs and expected outputs for different parameter settings up front. By making sure that complex tests are included as well as simple cases, this forces agreement up front on how sophisticated the functionality needs to be and eliminates misunderstandings. These expected results should cover boundary cases as well as typical values.

By definition, the project is finished when the software delivers the expected results, together with any other agreed deliverables e.g. documentation.

This approach also helps prevent creeping scope. Any additional functionality or requirements that is allowed after the project is planned and the expected results agreed, delays its likely implementation date. More worryingly, it also has the insidious effect of reducing the commitment and expectation of those working on the project, as they feel that once the initial target dates are rendered unachievable, there is no longer any meaningful date to which they should feel committed.

## Non-delivery of benefits

- **Due to oversimplification?**
- **Due to underestimation leading to simplification?**
- **Failure to agree what benefits are to be delivered?**
- **Lack of involvement and commitment from users?**

helpIT SYSTEMS

- deduplication
- addressing
- suppression
- duplicate prevention
- inquiry

Oversimplification of requirements at the outset of the project will lead to the new system being unable to cope with many of the challenges that it faces after it goes live.

Underestimation leads to either overruns, or functionality being cut back.

Obviously, any benefits which are expected by IT or business sponsors must be spelt out at the planning stage.

Lack of involvement and commitment from users during the project leads to lack of their availability for training and a reluctance to embrace the new system fully.

## Recurrence of problems

- **No data firewall in place**
- **No follow-up project scheduled**

\* Name:

Position / Job title:

Company Name:

\* Zip / Postcode:

\* Address:

Once the quality of the data already in the database has been improved, it is essential to keep it clean, by filtering all new entries.

Data entry by visitors to your web site poses a big challenge to the ongoing quality and integrity of your database, so it is essential to screen such data at point of entry.

The example from Viking Direct sent to an old home address shows an example of data that is maybe deliverable, but is poor quality. There is a misspelling of the surname and an error in the company name, but the funny one is that “Ken, Lime Tree Close” should have been “10 Lime Tree Close”.

## Uncontrolled duplication of data

- **New system doesn't allow things that the users could do before**
- **No agreement on alternative methods of achieving end result**
- **“I need a local copy to work with”**

### FIRST TWO BULLETS:

If the users are no longer able to do some things that they could do before, with no satisfactory replacement methods, then they will find ways of doing them that bypass and break the safeguards implemented by the new system.

Often, this involves them taking a copy of the end data into Excel, Access etc. and making changes there, which are then not reflected in the master database.

**The enemy has moved on**

- **Who is the enemy?**
  - Your competitors?
  - Legislators?
  - Your customers?
  - Prospective customers?
  - Or is it you?
- **All of the above, plus...**
- **Change!!**
- **Systems must be able to adapt**

**helpIT SYSTEMS**

- deduplication
- addressing
- suppression
- duplicate prevention
- inquiry

The slide features a light blue background with a pattern of overlapping circles and faint text including 'Data Cleansing Software', 'addressIT', 'suppressIT', 'filterIT', and 'findIT'. A dark blue circular logo in the top right corner contains the 'helpIT SYSTEMS' text and a legend with five colored dots corresponding to the listed services.

Competition forces you to launch new products, find new routes to market and improve efficiency, all of which challenge existing systems.

Legislators force higher standards for the records that you keep and your ability to keep them.

Your customers can't be relied on to provide their existing customer or account numbers when they deal with you.

Prospective customers are often too impatient to provide good data when making an inquiry, particularly via your web site.

Your own users will find the weakest point of new systems if it seems to make their own job easier, especially if they are on simplistic productivity bonuses.

The biggest enemy, due to competition, legislation and new ways of doing business, can be summed up as "change". If your systems cannot evolve to meet the challenge of change, they will decay and become a millstone around your neck.

## In conclusion

- **Get top-level commitment for business and IT resources**
- **Define business objectives with respect to time**
- **Use proven, reusable components to maximise chances of success**
- **Internal solutions should be evaluated in the same way as external**
- **Use test-driven development to assist project planning and control**
- **Establish terms of reference for a successor project**

# And finally...

filterIT  
addressIT  
suppressIT

Mailroom  
Data Cleansing  
Software



- deduplication
- addressing
- suppression
- duplicate prevention
- inquiry

