Understanding the costs of data capture: paper, automatic and with the Internet

CIMTECH 2000

OVERVIEW

Organisations have sometimes been surprised and disappointed when they re-engineer a forms-based data capture process but fail to achieve their anticipated savings.

This paper explains:

- how capture costs are built up from data entry **plus** dealing with problems
- an example of costs for an automated process, and for dealing with the paper forms that are left after you bring in an Internet process
- four techniques for investigating the costs of your current process.

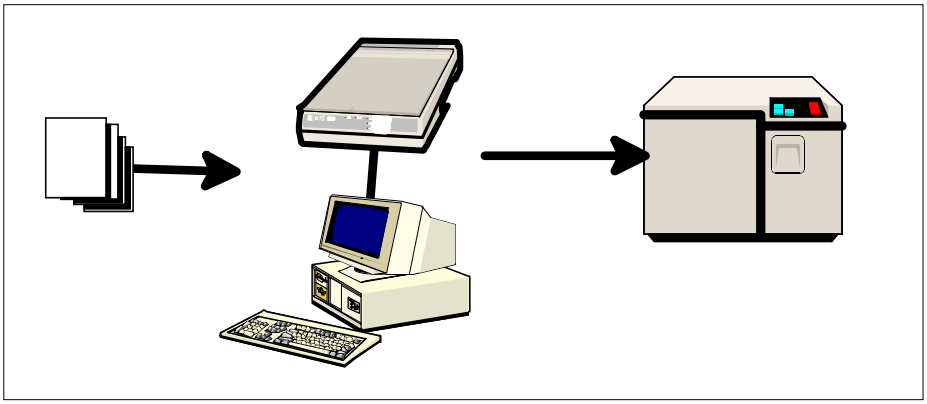
Overall, you will be able to understand your own process and find out whether automation will achieve what you expect from it.

1. THE COSTS OF DATA CAPTURE

1.1 Definition

The cost of data capture is the overall cost of turning incoming documents into data used in a business process.

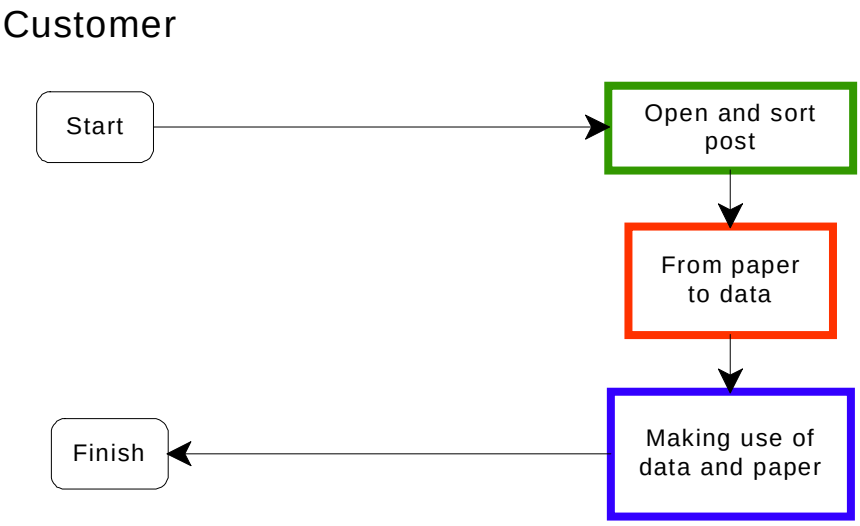
If you are considering automated data capture in your organisation, you may have seen schematics like this:



Documents arrive at a scanner, are captured, and end up on your organisation's database.

The problem with these schematics is that they leave out one major element of your costs: the staff who operate the technology. I will not be considering the technology in this paper. Instead, I will be considering the human element: what your staff have to do at each stage to achieve the overall business purpose.

On the next page, there is another schematic of capture:



1.2 Outline

This paper will describe:

The costs	what people have to do at each stage when turning paper into data
How the numbers work	an example of costs for an automated and for an Internet process
Investigating your costs	four techniques for investigating the costs of your current

2. OPENING AND SORTING THE POST

2.1 Opening the post

Figure 2.1 Incoming post, courtesy of the Open University (author's photo)

Your documents do not arrive at the capture process by themselves. The post has to be opened and sorted.

When you start scanning your documents, rather than reading them, you have to be more careful when you open the post.

For example, one of my customers complained that the scanner would not work. I found that they were opening the envelopes with an automatic letter opener which sliced each form into two pieces. Then the data capture department was sticking the pieces back together with sticky tape which was too thick for the scanner.

2.2 Sorting the post

Your incoming post probably includes some of these:

- incoming cheques and other forms of money
- items marked 'personal and confidential'
- post for a different office

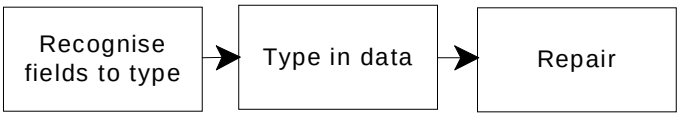
and various other items which may or may not have an impact on the process you are automating. Somewhere you will need to have a post sorting station like this:

Figure 2.2 Form sorting desk, courtesy of Open University (author's photo)

Your post team will need to know what to do with all your forms. Are you going to get them to remove the forms that are too torn or crumpled to go into the scanner? What about forms which arrive on the old version of the stationery? What are you going to do if an incoming form is attached to a letter? Are you going to deal with it as a letter or a form first? What if the letter modifies the contents of the form?

3. FROM PAPER TO DATA

Once you have got your forms to your data capture staff, the next step is turning paper into data.



The staff have to recognise what to type in - to understand the marks on the paper as figures or numbers to be put into the computer.

Then the figures or numbers have to be typed - keying the figure or number into the appropriate field.

Finally, your computer system is likely to validate the data. The staff have to decide what to do if the data keyed is not acceptable as an input. Repair is the process of dealing with the validation failures and any other problems on the form.

3.1 Recognition

Here is a typical box with a number written in it.

The process of recognition is understanding the squiggle in the box as a number. Humans are rather good at this, and most of us will interpret these squiggles as the number '3'. You might not even notice that the 3 has been written across part of its box.

Here is same box as a bitmap:

```
> 42 4d ce 01 00 00 00 00 00 00 3e 00 00 00 28 00
> 00 00 32 00 00 00 32 00 00 00 01 00 01 00 00 00
> 00 00 90 01 00 00 c4 0e 00 00 c4 0e 00 00 00 00
> 00 00 00 00 00 00 00 00 00 00 ff ff ff 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00 ff ff
> fe 1f ff ff c0 00 ff ff f0 0f ff ff c0 00 ff c0
> 00 00 0f ff c0 00 ff df c7 f3 ef ff c0 00 ff df
> d7 f1 ef ff c0 00 ff df ff f9 ef ff c0 00 ff df
> ff f9 ef ff c0 00 ff df ff f9 ef ff c0 00 ff df
> fe 39 ef ff c0 00 ff df fe 03 ef ff c0 00 ff df
```

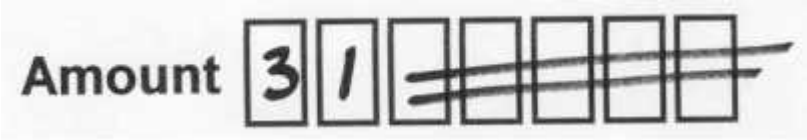
```
> fe 03 ef ff c0 00 ff df ff c7 ef ff c0 00 ff df
> ff f3 ef ff c0 00 ff df ff fb ef ff c0 00 ff df
> ff f9 ef ff c0 00 ff df fb f9 ef ff c0 00 ff df
> f9 f9 ef ff c0 00 ff df f8 f1 ef ff c0 00 ff df
> fc 01 ef ff c0 00 ff df ff 03 ef ff c0 00 ff df
> ff ff ef ff c0 00 ff df ff ff ef ff c0 00 ff df
> ff ff ef ff c0 00 ff df ff ff ef ff c0 00 ff df
> ff ff ef ff c0 00 ff c0 00 00 0f ff c0 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00 ff ff
> ff ff ff ff c0 00 ff ff ff ff ff ff c0 00
> 0000716
```

To recognise this as a number, the computer has to hunt for the dots that will create a number. When character recognition experts - OCR, ICR or whatever - talk about recognition, this is what they mean. Their equipment will look for the squiggles, separate them from the boxes, and turn them into numbers. They prefer to have the boxes printed in drop-out colours so that they do not need to deal with problems like my '3' touching the box it is in, but some recognition engines will sort this out for you.

When character recognition vendors talk about 90% recognition rates, they usually mean that 90% of *characters written in the boxes* will be recognised correctly.

3.2 Extra marks

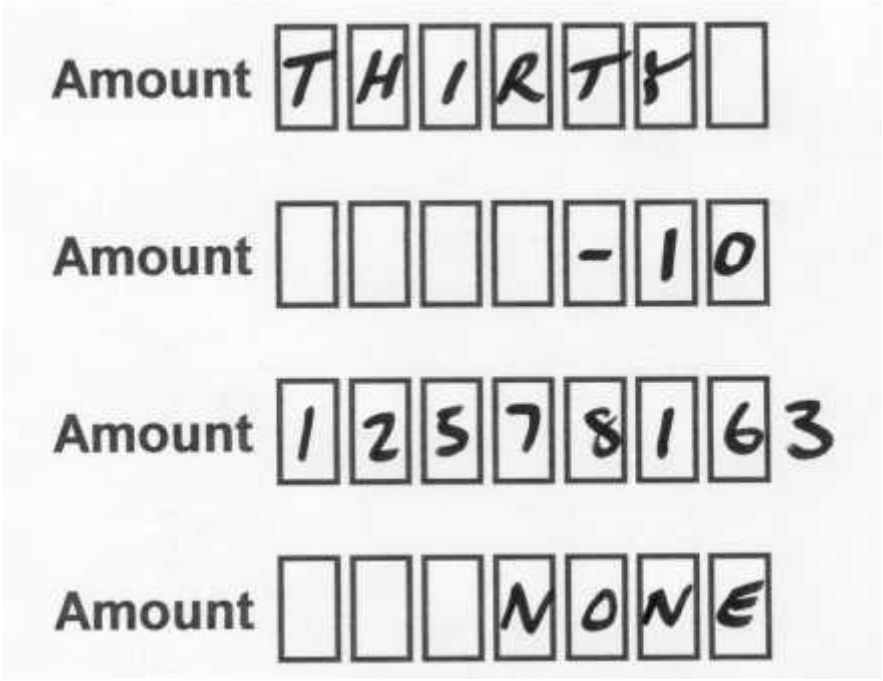
Unfortunately, customers do not always write clearly and neatly in the boxes.



This time, we have another '31' but the customer has felt obliged to indicate 'zero' with two horizontal slashes. Some OCR engines may fix this. Mostly, your engine will report this as a recognition failure.

3.3 Validation failures

Your computer system is likely to have validation on the input. For example, the amount field might only accept numbers in the range 0-9,999,999 and the validation will reject all of these:



Your staff will sort these out when they see them.

It is likely that a recognition engine, with appropriate definitions and validation, will also refer them to the operator although there is a risk that the third of my examples would be captured as 1,257,816 rather than 12,578,163.

3.4 Mixed mode problems

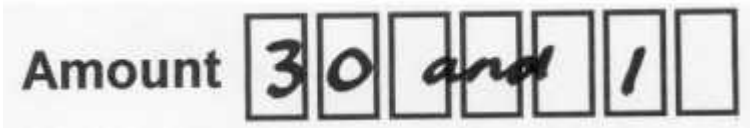
In this case, the customer has added a decimal point and two zeros. What happens here depends on the validations and the engine.

- You may have a recognition failure on:

- If you are in luck, you will get '31.00' back from the engine and you can then strip off the ".00" to put 31 into your database.
- If your engine and validations are not working in harmony, you may end up with '31100" in your database.

3.5 Multiple entries and messages

Quite often, your customer will do something like this:



Whatever did they mean?

A good OCR engine will give up on these - and is unlikely to count them as part of the 10% recognition failure. If the OCR tries too hard to understand the writing, you may get strange results: '300211'.

3.6 Repair

'Repair' is dealing with recognition and validation failures. A person has to view each one, and decide what to do at one of these levels:

Visual	You can repair the problem just by looking at the field. The 'extra marks' example is a visual repair.
Validation	You may need to know the validations as well, as in most of the 'validations' examples.
Internal fix	You can solve the problem with information on the document or within your organisation.
Go back	You have to obtain more information from your customer to solve these problems.

Let us look a bit more closely at the 'internal fix' and 'go back' levels.

3.7 Internal fix

I have concentrated on a numeric field, but frequently your customer will forget to supply some information you need: their account number, postcode, or previous order reference number. Quite often, you can look these up in your existing records.

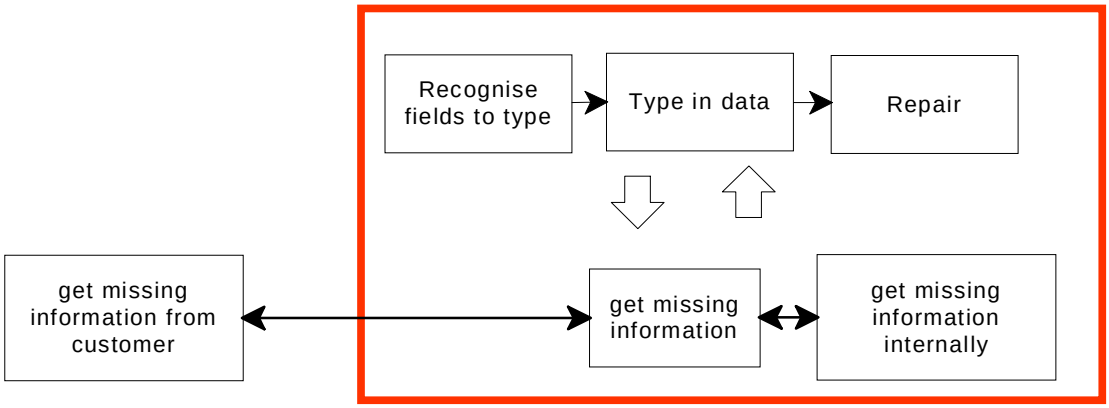
You may need to know some business rules to fix some validation errors. For example, your staff will probably know if you can deal with an amount of 12,578,163 - or whether the customer really meant something else, maybe £125,781.63.

3.8 Go back

Organisations are reluctant to send forms back to the customer, but you may have to - for example, if a signature is missing. You have to ask the customer what was meant, or for extra information to solve a problem.

If the customer has given both addresses on the form, then this could be an 'internal fix'. If not, we have a 'go back' example.

3.9 Combining processes



The diagram shows three separate processes of recognise, type in and repair. In fact, your staff will look ahead and choose a repair strategy before typing anything. They will read the marginal comments and deal with them first. And they will also know a lot of the validation rules. A human is unlikely to type '31100' in my 'wrong format' example. 'THIRTY' would be translated to '30'. A human would probably never realise that they have seen an 'extra marks' problem.

3.10 Repair and an automatic data entry process

When you use an automatic data entry process, there is much less flexibility in the process. Recognition always comes first: the engine has to try to reconcile the squiggles into letters and numbers. You can only proceed to the validation stage when you have actual data. The 'internal fix' and 'go back' problems, which should have been tackled before you tried to do anything with the form, may get left until the end or missed altogether.

3.11 Repair and the Internet

If you put your form out to your customer on the Internet, then you will solve some of these problems.

The visual problems, such as extra marks, go away. Your form will only accept data that is appropriate for the box.

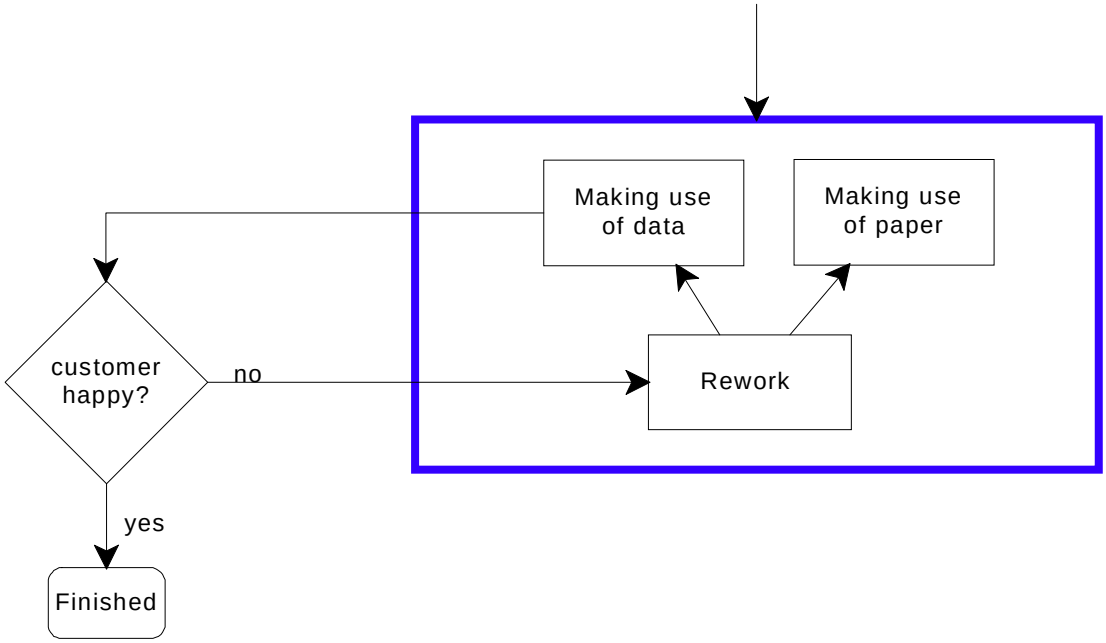
Validation errors will be fewer. Many validations can be done immediately, so that the customer corrects the problem for you.

However, “internal fix” problems that you could sort out from data you hold are a worse problem on the Internet. Your customer is unlikely to have access to the same information.

Finally, “go back” problems where you have to approach your customer for more information are also likely to be worse. It may be that your form asks for a mandatory field that your customer cannot understand, or (for some reason) refuses to supply at first. If your Internet form rejects incomplete data, you may never find out that your customer attempted to fill in your form. If you are unlucky, that might be a customer who goes away.

4. MAKING USE OF DATA AND PAPER

Once you have taken the data from the paper, your final step is to make use of the information.



4.1 Making use of the paper

Even though you have lifted the data from the paper, the paper itself often has its own roles:

- As an instruction to later parts of the process. For example, you might use the order form as a pick list to create and pack the order.
- As a record of who did the work. For example, your staff may sign or mark the form to take responsibility for processing it.
- To control workloads. For example, your supervisors may judge the amount of overtime they need based on the height of the pile of unprocessed forms.

If you decide to remove paper from your process, you will need to find out all the things you use the paper for so that you can develop an alternative method.

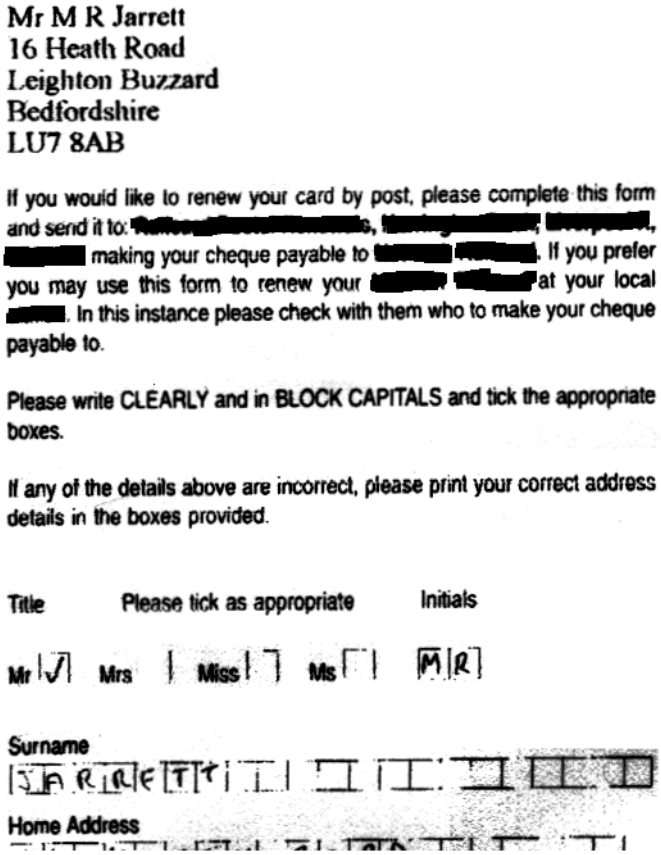
4.2 Is the data useful?

The final question to ask is: what are you actually capturing at the moment? When you bring in your automated process, you may find that costs go up because you start to capture data that you did not capture in the manual process.

This may seem strange. If you have a box on the form, and the customer has completed the box, why would you ignore it?

Customer tells you what you already know

Consider this example.



Mr M R Jarrett
16 Heath Road
Leighton Buzzard
Bedfordshire
LU7 8AB

If you would like to renew your card by post, please complete this form and send it to: [redacted], making your cheque payable to [redacted]. If you prefer you may use this form to renew your [redacted] at your local [redacted]. In this instance please check with them who to make your cheque payable to.

Please write CLEARLY and in BLOCK CAPITALS and tick the appropriate boxes.

If any of the details above are incorrect, please print your correct address details in the boxes provided.

Title Please tick as appropriate Initials

Mr [√] Mrs [ ] Miss [ ] Ms [ ] [M][R]

Surname

[J][A][R][R][E][T][T]

Home Address

The customer here has filled in his name (and goes on to fill in his address) even though the name and address are the same as the ones already printed at the top of the form. There is no need to capture anything. 'Change' of name and address boxes are notorious for this - in one study, I found that half of the entries in the change of address box were the same as the current address. Often, the form asks the customer to fill in the whole name and address even if only one part of it

has changed: you may be processing 50 to 100 characters of information when the only change is a single digit in the post code.

I have chosen name and address because so many forms ask for them. But the problem of asking for information you already hold is much more pervasive than just name and address. On the form I used for the example above, every entry was the same as the previous year's data.

Information arrives more than once

The second hidden cost I want to remind you about is the cost of getting the same information from two sources. For example, some of your customers may place orders twice: once by telephone or Internet, and then a second time on an official order form. Although you may try to discourage them from doing this, you do not really want to put them off placing an order with you. So how are you going to reconcile two versions of the same data?

All boxes on the form are not equal

I have been discussing forms as a whole. But when you consider the individual boxes, you may find that:

- Some are crucial to your business: perhaps the signature box, or the credit card number. We will assume that you capture these.

- Others are for more general purposes: market research, future mailing lists, a description which helps to confirm the product reference number. Your manual process might only capture these for some forms, or for certain special events, or if other information is missing.

- Some may just be left over from an earlier version of the business process and are no longer used at all. If you simply start to capture all the boxes on the form when you introduce your automated process, you may be using a lot of staff time on recognition and repair for boxes you are not going to use.

It is important to distinguish between these boxes when you design your automated process. A validation which accepts almost anything on a vague box about 'please indicate your leisure interests' might be inappropriate for a box, similar layout, asking 'please indicate your choice of product colour'. If a box is only used under certain circumstances, it might be more efficient to ignore recognition errors on it until the point it is needed.

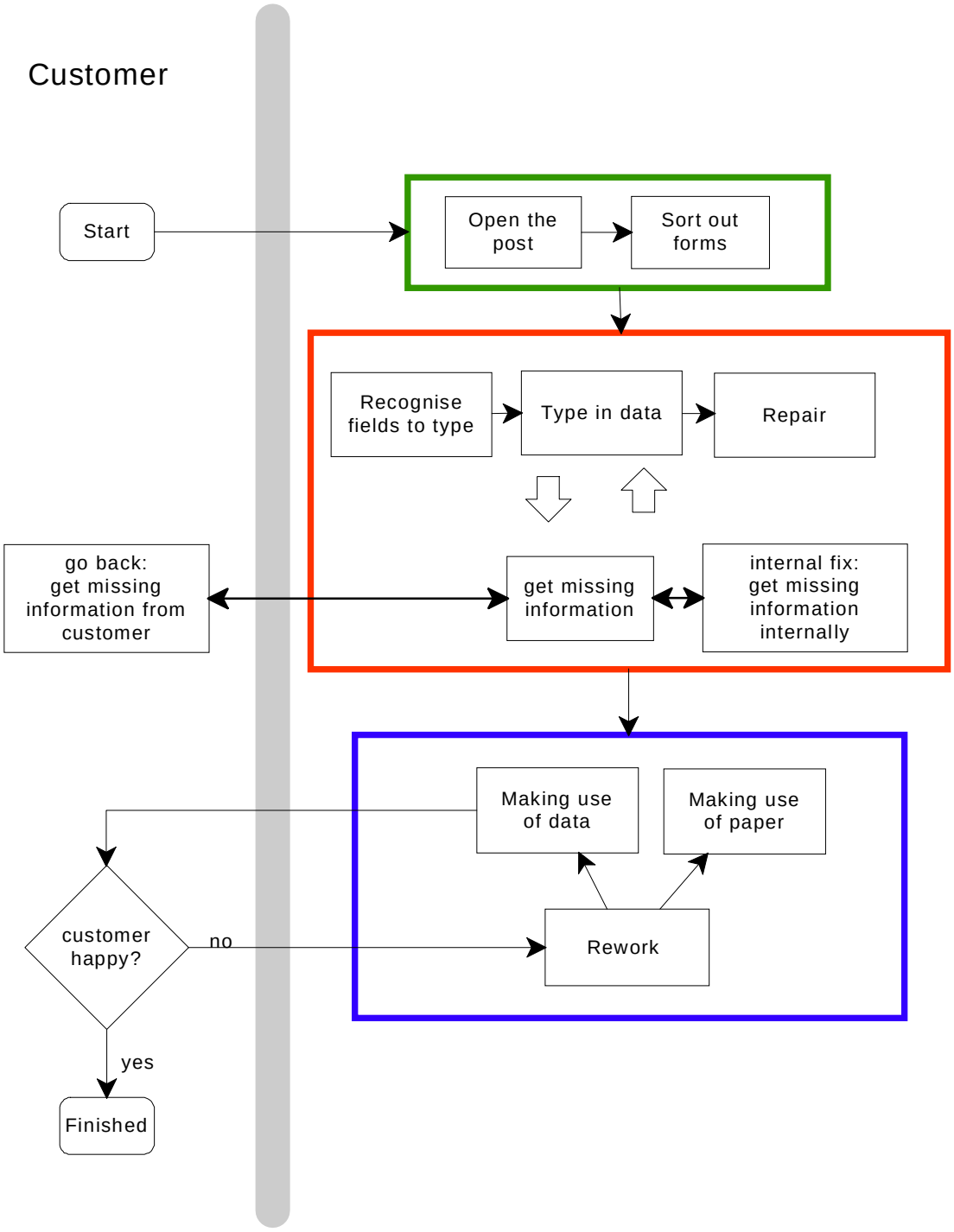
4.3 Rework

So that leads me on to my final category of problem: rework. If you miss problems at the initial capture stage - whether because your staff got it wrong, or your automated process got it wrong - then you have to solve the problem later.

People make mistakes all the time. You may find out that the capture process has gone wrong when you deal with the information. Or possibly your customer will point out the mistake and allow you to put it right.

To simplify the calculations, I have grouped all of these into one overall category: rework.

4.4 Overall capture schematic



5. HOW THE NUMBERS WORK

5.1 Typical mixture

Let us now look at how repairs influence costs for dealing with forms. There will be a range of difficulty, but I will simplify them into two types:

Easy — The simplest forms - less data and no complications. Time to process = about half the time of an 'ordinary' form = .5 unit

Ordinary — The run-of-the-mill work. 'Typical' forms with an average amount of repair. Time to process = 1 unit.

We also identified three types of problem forms:

Internal fix — Repair from information you already hold. Same basic time as an ordinary form, plus about the same again to deal with the problems. On average, these take about twice as long as an 'ordinary' one = 2 units.

Go back — Need to go back to the customer during initial processing. Same basic time as 'ordinary' plus time to contact customer, retrieve information, and finish. A working assumption: 4 times as long as an ordinary form = 4 units.

Rework — Someone points out a problem later. Has to be re-worked plus there will be additional care and attention needed. Some of these will take a long time to sort out, but to keep the calculations simple we will assume 2 units more than a 'go back'. Total time = 6 units.

The mix of work will vary according to the application, but the table on the next page shows a typical split and the effect on process time. For each type, I have worked out:

- typical percentages in the overall mixture
- the time for processing the form
- split overall time into post handling, keying, and problem handling.

Table 5.1 Time to process typical work mix

	% forms	Units of work compared to 'ordinary'	Post handling	Keying	Problem handling
Easy	20	.5	.03	.47	0
Ordinary	45	1	.03	.97	0
Internal fix	20	2	.03	.97	1
Go back	12	3	.03	.97	2
Rework	3	6	.03	.97	5

5.2 An automated process

Let us now consider the effects on these times of an automated capture process. You will replace your keying time with automatic recognition.

As you are now be scanning your forms, you have:

post handling time — probably doubles - now 6% of unit time

scanning time (new) — typically 10% of the previous unit time

The original keying time is replaced by:

recognition failures — 10% of the data will need repair because of not being recognised - 10% of previous keying time for the form

However, from the earlier discussion of errors you will recall that there are other reasons for repair. The automated data capture product will not deal with 'extra marks' or 'wrong format' errors. So we need to add:

validation failures — maybe 10% of previous keying time

extra marks — maybe 10% forms will have this problem throughout, but to keep it simple we will use 10% of previous keying time per form

On top of that, we will actually increase the time for dealing with our problem categories of documents. With the manual process, we could recognise the problems and start to deal with them immediately. With an automated process, these problems get left until the recognition and validation problems have been identified. This increases the time to deal with each problem: the 'problem penalty'.

There is also a probability that problems get missed during the data capture stage, as some forms will no longer be getting the same scrutiny as they used to. Therefore some work has to be put right at the later stage. The effect is spread among all the forms, but to keep the calculations simple I have just assumed that 'ordinary' forms are affected.

The overall effects will depend on your particular process, but as a first estimate let us assume:

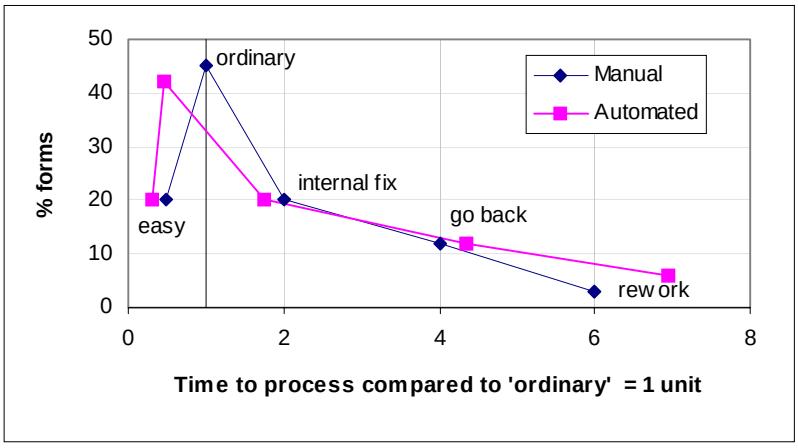
- time to deal with problems increases by 30% (problem penalty).
- 3% of forms move from the 'ordinary' into the 'rework' category

Table 5.2 New automated process

Type	% forms	Previous keying time	New post opening time	Scanning time	Data failures	Previous problem time	Problem penalty	Total time
Easy	20	0.47	0.06	0.10	0.14	0.00	0.00	0.30
Ordinary	42	0.97	0.06	0.10	0.29	0.00	0.00	0.45
Internal fix	20	0.97	0.06	0.10	0.29	1.00	0.30	1.75
Go back	12	0.97	0.06	0.10	0.29	3.00	0.90	4.35
Rework	6	0.97	0.06	0.10	0.29	5.00	1.50	6.95

5.3 Comparison of costs: manual and automated process

The chart compares the manual process with the automated process.



This is rather a successful implementation of an automated data capture process. The new process is a bit more problematic for the difficult forms, and there are a few more of them. But 85% of our forms are being dealt with more quickly.

The table below puts this into numbers.

Table 5.3 Manual process and automated process

	Manual process			Automated process			
	% in category	time to process	total time (% in category x time to process)	% in category	time to process	total time (% in category x time to process)	time saved
Easy	20	0.5	10	20	0.3	6	0.4
Ordinary	45	1.0	45	42	0.5	19	0.6
Internal fix	20	2.0	40	20	1.8	35	0.1
Go back	12	4.0	48	12	4.4	52	-0.1
Rework	3	6.0	18	6	7.0	42	-1.3
Total time			161			154	

The bottom line is that the time saved on dealing with the easy and ordinary forms is offset by the increased time on the problem forms. Our new technology has achieved (161-154) / 161 = 4% saving.

5.4 An Internet process

So how about considering an Internet process instead?

Let's assume that you offer both an Internet form and a paper form to your customers. For the moment, we will ignore the cost of providing the Internet option, and look only at the impact on the cost of staff who deal with the paper forms.

It is likely that the customers who opt for Internet are those who are coping well with your paper form. We will be optimistic and assume that 80% of your 'easy' forms now arrive by Internet, and 60% of your 'ordinary' forms.

"easy" forms	20% of previous levels
"ordinary" forms	40% of previous levels

Against this, we have:

double filing	Some customers will send in paper and Internet – perhaps 10% of your Internet customers
problem penalty	With a more complex process, there are more things to go wrong. Some problems will take longer to sort out. As with the automated capture process, I have assumed that problems take 30% longer to solve.
forms change category	Some forms will move from the 'ordinary' to the 'rework' category. As with the automated capture process, I have assumed that 3% of forms change category.

We will also ignore any additional costs for helping your customer to solve problems with the Internet forms. We will assume that they don't give up on your organisation, but instead persist is sending in the 'internal fix' and other problem categories of forms in the same numbers as before.

The new mixture of forms is shown on the next page.

Table 5.4 Lower levels of paper forms after Internet option available

Type	% forms	Internet rate	% on Internet	% on paper	Double filing	Change categories	Total on paper
Easy	20	80%	16	4	1.6		5.6
Ordinary	45	60%	27	18	2.7	-3	17.7
Internal fix	20			20		3	23
Go back	12			12			12
Rework	3			3			3
	100						61.3

Even allowing for the people who are sending their forms in twice, we are now handling only 61.3% of the forms on paper that we processed before we introduced the Internet option.

In the old paper process, it took us 161 units of time to deal with 100% of the forms:

Table 5.5 Old paper process: total time to deal with forms

Type	% forms	Previous process time	Previous problem time	Total time per form	Time per form x % forms
Easy	20	.5		.5	10
Ordinary	45	1		1	45
Internal fix	20	1	1.00	2	40
Go back	12	1	3.00	4	48
Rework	3	1	5.00	6	18
					161

Table 5.6 Dealing with paper forms after introducing Internet option

Type	% forms	Process time	Problem time (30% increase)	New time per form	Time per form x % forms
Easy	5.6	.5		0.5	2.8
Ordinary	17.7	1		1	17.7
Internal fix	23	1	1.3	2.3	52.9
Go back	12	1	3.9	4.9	58.8
Rework	3	1	6.5	7.5	22.5
	61.3				154.7

Putting the two tables side by side shows that the new process has not achieved the savings that we expected.

Table 5.7 Old paper process compared to fewer, more complex paper forms after introducing Internet option.

	Old paper process			Dealing with fewer, more complex forms		
Type	% forms	Total time per form	Time per form x % forms	% forms	New time per form	Time per form x % forms
Easy	20	.5	10	5.6	0.5	2.8
Ordinary	45	1	45	17.7	1	17.7
Internal fix	20	2	40	23	2.3	52.9
Go back	12	4	48	12	4.9	58.8
Rework	3	6	18	3	7.5	22.5
	100		161	61.3		154.7

The total cost is now 154.7, which means:

(161-154.7)/161 = 4% saving.

5.5 The richer mixture

In many organisations, the less experienced staff deal with the easier forms and the more experienced tackle the problem forms.

Unless you take steps to tackle your problem forms, introducing technology can speed up the easier forms while worsening the problem of your more difficult forms.

Look again at the example of the changes to your paper process after introducing an Internet option.

Type	Old mixture of forms (%)	Easier and problem forms as % of mixture	Paper forms dealt with - Internet option available	Form types as % of paper forms	Easier and problem forms as % of mixture
Easy	20	65%	5.6	9%	38%
Ordinary	45		17.7	29%	
Internal fix	20	35%	23	37%	62%
Go back	12		12	20%	
Rework	3		3	5%	
	100		61.3	100%	

So although the Internet option in my example meant a small overall saving of 4% in staff time, it has also altered the balance of staff. Previously, I needed 35% experienced staff in my data capture department. Now I need 62%. Let's hope that I didn't move my experienced staff to the Internet forms helpline.

6. 4 TECHNIQUES FOR INVESTIGATING YOUR COSTS

I have spent a lot of time explaining what might go wrong and how those problems might end up giving you trouble on your own project.

Against that, there are many success stories and the opportunity to remove tedious work. So how can you find out what the risk factors are on your own project?

6.1 The data you need

My calculations used:

- categorisation of forms into 5 types: 'easy', 'ordinary', 'internal fix', 'go back' and 'rework'
- percentage of each type in the overall workload
- post handling time
- keying time
- recognition rate for properly completed boxes
- percentage of validation failures
- scanning time
- percentage of recognition failures because of improperly completed boxes ('extra marks' and other errors)
- penalty percentage on dealing with problems in a different order because of automated process
- increase in rework category.

and I also warned you about capturing data you already know, data that arrives twice, or data which you may not use.

6.2 Four techniques to help you

The four techniques are:

1. Cohort tracking
2. Exceptions analysis
3. Usage analysis
4. System calibration

6.3 Cohort tracking

Cohort tracking will define:

- your own types for categorisation of forms
- your own percentage of each type in the overall workload
- the post handling time

The steps are:

1. Make an appointment with your post room.

   Find out when they open the post, and what the pattern of post is from day to day and week to week. Choose a typical day for your first cohort.

2. Count the unopened envelopes.

   You may need to start selecting at this point. If you have thousands of items, it will be hard to track them all, so you may need to compromise by picking every tenth, or every hundredth envelope.

3. Watch your envelopes being opened.

   Note how many items are removed for separate handling
   Note how many of the items arrive with or without supporting information. Establish how much work gets done on your items.

4. Mark each item you are interested in.

   You could use a small rubber stamp, or a dot of highlight pen, on the reverse of the item.

5. Take note of some identifying number for each item

   This could be the customer number or some other identification.

6. You now have a cohort of items.

The idea is to check back to see how many have been dealt with and what happened to them. Who sorts them, and why? How many are finished with by the end of day 1, of week 1, of month 1? Did any of them turn into complaints?

By tracking your cohort through the process, you should be able to create your own split of your documents into the 'easy', 'ordinary', 'internal fix', 'go back' and 'rework' - or a more complex selection of categories.

Depending on the overall time for your process, you may need to check on your cohort over days, weeks or even months. If you do not do it, you may miss the really time-consuming examples that eliminate any savings from an automated process.

If the typical cohort shows encouraging figures - low percentages in the 'go back' and 'rework' categories - then you may need to track more cohorts, choosing them from the extreme or unusual: end of the month, the day after bank holiday, extra orders after a big promotion.

6.4 Exceptions analysis

Cohort tracking identifies the overall fate of a batch of items. Exceptions analysis is its opposite: it looks at the work of a single day, or even of a single hour, and identifies the full range of problems encountered during that particular session of processing.

Exceptions analysis will define:

- keying time
- percentage of validation failures
- percentage of recognition failures because of improperly completed boxes ('extra marks' and other errors)

This time, you start with the forms as they arrive at the point of capture. You may need to enlist the help of the staff who deal with the form. For each form and every field on the form, you note:

- if you find the writing on the form difficult to read
- whether there are extra marks
- if any field is completed in a way which might fail validation
- if anything mandatory is missing
- if there are internal contradictions, for example two items both completed, when only one is allowed
- if there are any marginal notes, additional information, or accompanying letters to explain entries on the form.

By the end of this, you should be able to identify:

- the percentage of forms with extra marks
- the rate of validation errors
- the rate of internal fix errors
- the rate of 'go back' errors

Of course, you could do the exceptions analysis on the cohort. But often the cohort gets split up during its various sorting phases - and you find that you can only look at exceptions with particular members of staff who do certain types of work. The cohort tracking gives you the overall mixture, whereas the exceptions analysis gives you the detail on elements of the mixture.

6.5 Usage analysis

The third step is usage analysis. This tells you how you use the information you ask for on the form.

Do not try this step until you have your cohort and exceptions information. If you use a blank form, you will find that there are unrealistic expectations about the standard of accuracy in the information written on the form.

Armed with the data about what actually gets put onto the form by your customers, you convene a meeting - or it may have to be a series of meetings - of your computer team, your processing team, and the overall owner of the business process. For every box on the form, you find out:

- what gets put into the computer system by the processing staff
- what the computer system does with that information
- the consequences for the business of the exceptions you have identified

The meeting should discover the discrepancies between what you ask for and what you need.

6.6 System calibration

By the end of the cohort tracking, exceptions analysis and usage analysis, you should be able to judge whether your data process is sufficiently straightforward to allow you to consider an automated process. My broad criteria would be:

- 20% or fewer forms have exceptions
- you are using all the information you ask for, and
- more than 90% of forms fall into the 'easy' or 'ordinary' categories.

You can now prepare a realistic pack of test forms to discuss with your chosen vendors. The pack should include the same mixture of types of documents as your overall mix, discovered during the cohort tracking. Remember to put them into the appropriate envelopes: the folding is important.

During system calibration, you will discover:

- effect on post handling time
- scanning time
- recognition rate for properly completed boxes
- percentage of recognition failures because of improperly completed boxes ('extra marks' and other errors)
- penalty percentage on dealing with problems in a different order because of automated process
- increase in rework category.

Test the proposed scanner by getting the forms out of their envelopes, sorting them as required by your vendor, and scanning them.

Ask your vendor to provide you with the recognition rates and recognition failures on your test pack of forms. Do not accept rates produced on any other test pack.

Finally, check the output data you are likely to obtain. Will this be sufficiently accurate to prevent any increase in rework?

7. THE BOTTOM LINE

Every data capture application has to accept that people will make mistakes. Your new automated system will just expose the problems that your current manual system avoids by using the perceptual and judgmental skills of your staff. Your new Internet system forces your customers to solve problems that your staff deal with at the moment.

Often, these skills come as a surprise. Your post-room staff and data entry clerks are often some of the lowest-paid employees in the business. Despite this, they are making hundreds of decisions every day. They may be relatively tiny decisions, but the cumulative effect can be fundamental to your business.

If you want to save costs by implementing new technology, you must:

- find out how your forms are filled in
- find out what your staff actually do
- understand what your business does with the forms and data.

What you learn should help you to win, not lose, from implementing your new technology. Or you might decide to improve some of the old technology. How about some hand-cream for the staff who open the post?

8. THE AUTHOR

Caroline Jarrett is an independent consultant. She specialises in improving forms, and in designing efficient processes for dealing with forms.

Caroline can be contacted at:

Effortmark Limited
16 Heath Road
Leighton Buzzard
Bedfordshire
LU7 8AB

Tel: 01525-370379
Fax: 01525-372363
Mobile: 07641-124431

Caroline.Jarrett@Effortmark.co.uk