

DATA RECOGNITION REVOLUTION

Enter the Electronic Eye

Any history of the 20th Century, will recognise the invention of automatic data processing and computing as key transformation agents in creating the modern world.

The computer is really a second-half of the century story but the development of automatic data processing spans the entire century. It is only recently that one could say that automatic data processing was now becoming automated data processing.

In 1890 Dr Hollerith invented the punched card in order to automate the production of US Census information. The punched card - literally holes in dollar bill sized pieces of cardboard - became the input and storage media for the pre-computer era tabulating machines and subsequently it was the main input media for most of the first three generations of computer systems. It was as late as the 1970s that punched cards ceased to be ubiquitous in computer installations. Up until then, everything that went into a computer had to be converted into something the computer could read quickly.

By the 1980s on-line terminals and PCs enabled direct keyboard input without the need to convert information into holes in cardboard and paper-tape. By 1995, around 80% of information input into computers of every kind was keyboarded.

Much of that information originates on forms and documents - some

handprinted, some machine printed. The question arises - if computers are so smart why can't they read and recognise hand and machine print thus saving the chore of transcribing information from paper through keyboard to computer?

Over the last 100 years, billions of pounds have been expended researching and developing machine readability and data recognition. Nothing sounds simpler. Put a sheet of paper under a scanner, read what is written, recognise what it means and then process the information. In science fiction it has been trivial for 100 years - the electronic eye sees and comprehends everything. In real science, nothing has been more elusive.

The good news is that given infinite financial resources it is now possible to read and recognise everything electronically or electro-optically. The bad news is that cost-effective reading and recognising are not yet widely available. We are in the transition stage between having most of the basic science and rolling-out an ever-increasing number of tools that will take us to mass market reading and recognising.

The history of 100 years of reading and recognising is illuminating. Before the relatively low cost electronics and opto-electronics of the 1970s, reading and recognition had been highly specialised. The electrical engineering era had produced punched cards and papertape, shoehorning information into a format that computers could read; the banks had found magnetic ink character recognition (electro-magnetics) with the stylised numbers and characters seen on the bottom of cheques as the most effective way of processing and clearing cheques quickly; the utilities with their enormous volumes of remittances (the cheque with gas bill stub) had

found optical character recognition again with stylised but more human readable fonts that could be read and recognised at very high speeds; the supermarkets found bar-codes that could be read and recognised any whichway but which could only cope with a limited amount of information; and latterly the National Lottery found mark sensing which could detect marks on documents. Unfortunately no-one has yet found the single all-purpose, all-singing, all-dancing read and recognise tool that like spreadsheet or WP would sweep the world markets.

Cheques and remittances can be read and recognised quickly and relatively inexpensively because the size and quality of the documents, the printing, registration and colours are rigorously controlled. The first law of read and recognise today is control the media (standardise, rationalise, simplify and monitor) just as it was with Dr Hollerith. The second law of read and recognise is if you can't control the media, you must be able to handle and repair rejects and doubtfuls, and have controls that pick-up false positives (where for example a bad '8' is recognised as a '6').

The third law of read and recognise is understand the technologies. Don't confuse text scanning of printed material produced in standard fonts (relatively simple) with intelligent recognition of the same content meticulously handwritten in copperplate (different technology, orders of magnitude more difficult). Read and recognise tools are all relatively narrow in their performance capabilities. Read and recognise systems are usually bespoke for very specific applications. There are around 300 so-called recognition 'engines' available and most systems use a number of different engines. Rule of thumb indicates machine-print is great, handprint (separated alpha and numeric characters) is always challenging, joined-up handwriting isn't commercial, mark sense is easy and barcodes

are wonderful.

The fourth law is about paper. High quality paper is always better than low quality paper. Crumpled, stapled and torn flimsies are the stuff of nightmares - the forms from hell.

There are three generic recognition technologies - omnifont which recognises extant print fonts; matrix-matching which compares what is read to a previously determined set and neural-net where neural computing techniques are used to 'teach' a computer to read by learning from previous examples. Many tools/engines use multiple technologies and systems are built from multiple engines.

A Data Recognition Revolution is now underway for a number of reasons: low-cost high performance micro computers, advanced low-cost opto-electronics, advanced scanner technologies (derived largely from fax developments) and improved software techniques have combined in the early 1990s to make data recognition a commercial proposition to most organisations. The most obvious application is capturing information from forms. Billions of forms are processed each year in the UK. Billions of pounds are spent processing forms. Everyone wants to process forms quicker, cheaper and more accurately than at present. Most people would also like to stop using forms and find a better way of capturing information. The UK Government has recently announced that it is abolishing 30 million forms per annum. The reality however is that forms usage is increasing in most organisations and technology is needed to control costs. If you can't abolish the forms the next best thing is likely to be to automate their processing.

The electronic 'eye' is here - not quite like science fiction but eminently

cost-effective for many information processing applications. Leading adopters are now automating. After 100 years automated data processing is set to take-off for the mass market. Increasingly married to forms and document processing systems, it will help to mark the change from 20th to 21st Century working practices. 100 years of investment has not been in vain.

...ends...