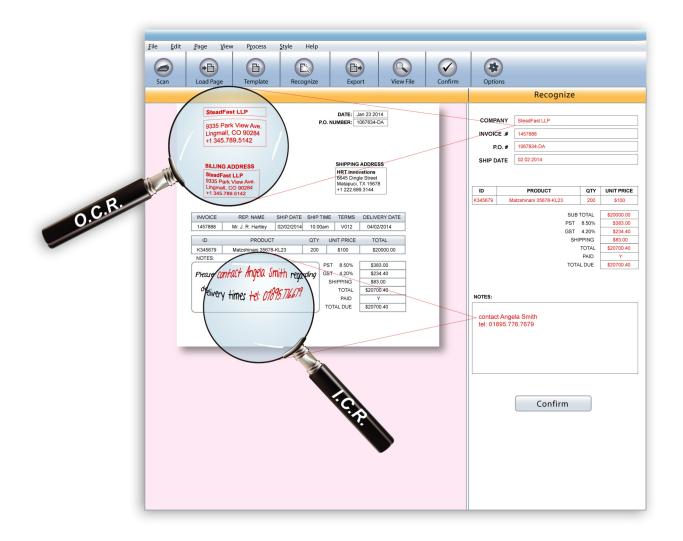
Forms Processing - user experiences of text and handwriting recognition (OCR/ICR)



Sponsored by:





About the White Paper

As the non-profit association dedicated to nurturing, growing and supporting the user and supplier communities of ECM (Enterprise Content Management) and Social Business Systems (or Enterprise 2.0), AIIM is proud to provide this research at no charge. In this way the education, the entire community can leverage the thought- leadership and direction provided by our work. Our objective is to present the "wisdom of the crowds" based on our 70,000strong community.

We are happy to extend free use of the materials in this report to end-user companies and to independent consultants, but not to suppliers of ECM systems, products and services, other than Parascript and its subsidiaries and partners. Any use of this material must carry the attribution - "© AIIM 2012 www.aiim.org / © Parascript 2012 www.parascript.com"

Rather than redistribute a copy of this report to your colleagues, we would prefer that you direct them to www.aiim.org/research for a free download of their own.

Our ability to deliver such high-quality research is made possible by the financial support of our underwriting sponsor, without whom we would have to return to a paid subscription model. For that, we hope you will join us in thanking our underwriter for this support:



Parascript LLC 6275 Monarch Park Place Longmont, CO 80503 USA Phone: (+1) 303-381-3100 Website: www.parascript.com

Process used and survey demographics

The survey results quoted in this report are taken from a survey carried out between 09 March 2012 and 29 March 2012, with 324 responses from individual members of the AIIM community surveyed using a Web-based tool. Invitations to take the survey were sent via email to a selection of AIIM's 70.000 registered individuals. Respondents are predominantly from North America and cover a representative spread of industry and government sectors. Results from organizations of less than 10 employees and suppliers of ECM products and services have not been included, bringing the total respondents to 255.

About AIIM

AllM has been an advocate and supporter of information professionals for nearly 70 years. The association mission is to ensure that information professionals understand the current and future challenges of managing information assets in an era of social, mobile, cloud and big data. AllM builds on a strong heritage of research and member service. Today, AIIM is a global, non-profit organization that provides independent research, education and certification programs to information professionals. AllM represents the entire information management community: practitioners, technology suppliers, integrators and consultants. AIIM runs a series of training programs, including the Capture training course www.aiim.org/Training/Capture-Course.

About the author

Doug Miles is head of the AIIM Market Intelligence Division. He has over 25 years' experience of working with users and vendors across a broad spectrum of IT applications. He was an early pioneer of document management systems for business and engineering applications, and has produced many AIIM survey reports on issues and drivers for Capture, ECM, Email Management, Records Management, SharePoint and Social Business. Doug has also worked closely with other enterprise-level IT systems such as ERP, BI and CRM. He has an MSc in Communications Engineering and is a member of the IET in the UK.

© 2012 AIIM 1100 Wayne Avenue, Suite 1100 Silver Spring, MD 20910 +1 301-587-8202 www.aiim.org

© 2012 Parascript LLC 6273 Monarch Park Place Longmont, CO 80503 +1 303-381-3100 www.parascript.com

user experiences of text and handwriting recognition (OCR/ICR)

Forms Processing

Raiim White Paper

Table of Contents

About the White Paper:

About the White Paper	2
Process used and survey demographics	2
About AIIM	2
About the author	2

Introduction:

Introduction	. 4
Key findings	4

Drivers and Adoption for Capture:

Drivers and Adoption for Capture: 4
Non-Adopters - Forms Scanning6
Non-Adopters - OCR7

Outsourcing:

Outsourcing7
Drivers for Outsourcing8

Capture and Recognition Strategies:

Capture and Recognition Strategies	9
Central vs. Distributed	9
Data Capture and OCR	9
Levels of OCR/ICR	10
OCR Performance	11
Currency of OCR/ICR Software	11

Hand-Writing Recognition:

Hand-Writing Recognition12	2
Drivers12	2
ICR Adoption13	3

Conclusion and Recommendations:

Conclusion and Recommendations	14
Recommendations	15
References	15

Appendix 1: Survey Demographics:

Appendix 1: Survey Demographics	16
Survey Background	16
Organizational Size	16
Geography	16
Industry Sector	17
Job Role	17

Underwritten in part by:

Parascript LLC	18
AIIM	19

Introduction

Forms processing and character recognition is not a new technology. In fact the first application was in mailing address recognition for sorting machines over 40 years ago. From that day to this the challenge has been to scan, clean, analyse and match the characters fast enough to provide a sufficiently accurate recognition at a suitable document throughput rate – and these two factors are always a trade-off. This kind of image processing is very compute intensive, and also lends itself to multi-processing. In view of the advances over the last 5 years in multi-core processors and the sheer compute power in even the most basic of servers, we can see that recognition technology is likely to have made dramatic strides in both performance and accuracy. Sophisticated character analysis algorithms have been steadily refined whilst throughput keeps up with the fastest modern scanners. On-the-fly layout recognition can quickly adapt to mixed form types whilst superfast look-up arrays have extended character-validation to full-word context checking. In addition, multi-pass voting techniques improve ambiguous matches and the long-term goal of handprint recognition and even cursive script recognition is now well within the capabilities of some systems.

However, because the core technology has been around for a long time, existing users can easily be complacent about the performance of their existing capture suites, and more importantly, may not recognize new potential applications that can be capture-enabled if the latest recognition and server technology is used. If the existing capture operation is separated from the downstream line-of-business process, or if it is outsourced, the process owner may not be aware that such new possibilities exist.

In this report we will review the different levels of forms-processing in use and the issues and potential benefits of character recognition. We will explore the awareness and take-up of the latest technologies, particularly ICR (Intelligent Character Recognition), as generally applied to hand-writing, and compare this to the more traditional OCR (Optical Character Recognition), which is usually restricted to machine printed text.

Key Findings

- 88% of survey respondents scan forms but only 32% do text recognition. 55% workflow scanned images and manually re-key the data.
- Localized decision-making is given as the main reason for non-adoption of forms scanning, followed by a lack of designated owner.
- 42% have half or more of their forms with handwritten data fields. For 38% half or more of their forms have hand-written name and address fields, and 32% have hand-written free text or open-ended data fields.
- 12% use ICR to recognize hand-printed constrained field entries. 6% use ICR to recognize hand-written script and free-form entries.
- An average productivity improvement of 34.8% was considered possible if recognition of hand-written text could be automated. 36% of respondents would expect a 50% or more improvement.
- In 26% of organizations, hand-written script fields play a key role in the efficiency of their business processes. A further 3% consider them to be quite important.
- Non-users of recognition cite hand-filled form fields and technology reservations. But 43% haven't evaluated ICR lately.
- 60% of outsource users have never been offered hand-writing recognition, or have never asked. 13% feel that their outsource does not have up-to-date technology.
- 42% of users last updated their recognition software 3 or more years ago. 13% last updated 5 or more years ago.
- Across OCR and ICR, 44% are achieving a 95% or better non-intervention rate per scanned form. 61% are getting 90% or better.

Drivers and Adoption for Capture

Every business has forms. The bigger the business, the more forms. Each form will relate to a process, and each process will involve employees entering and processing the data on each form. As we know from previous AIIM reports¹ scanning forms and moving image files rather than paper will improve productivity and speed up response. It provides an electronic workflow that can be readily monitored and managed, thereby eliminating bottlenecks in the process and improving transit times.

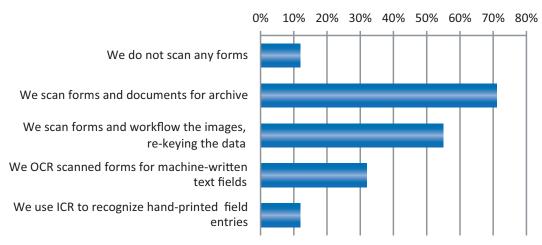
The next step in productivity improvement is to remove the need for keying the data from each form into the

Forms Processing

user experiences of text and handwriting recognition (OCR/ICR)

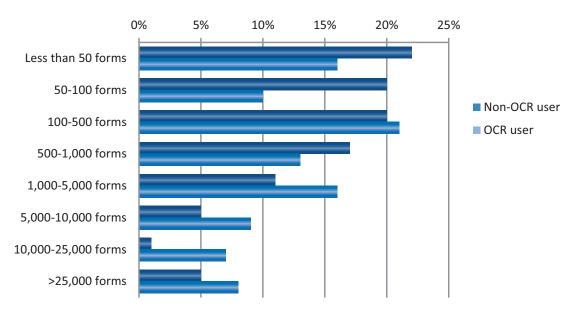
process application. Although this would seem to be an obvious step, it is not as widely adopted as might at first be thought. Whereas 88% of respondents to our survey scan forms, only 32% use OCR to recognize machine-written text, and only 12% recognize any level of hand-written text. As we will see later, most processes derive considerable value from handwritten forms data.

Figure 1: How do you pre-process forms coming into your business unit, or generated within it? (Check all that apply, including outsourced services) (N=255: part 1)



Many potential users consider that there is a distinct point below which there are insufficient forms coming into the business to justify OCR automation. In fact, the correlation is not as strong as one might think. There is a degree of inversion at 1,000 forms per day, but 26% of OCR users are processing 100 forms per day or less. There are also organizations processing many thousands of forms per day who are not using OCR.



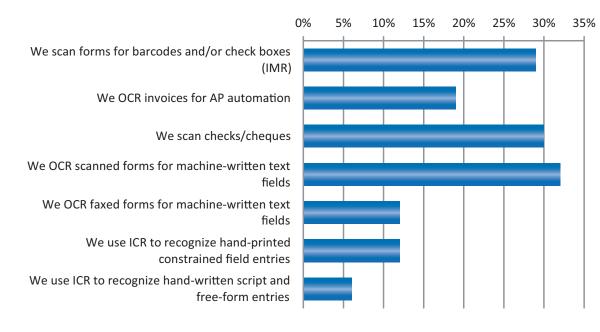


Overall, 29% or our respondents are processing more than 1,000 forms per day (250,000 per year), rising to 42% of the largest organizations.

Looking in more detail at different levels of IMR (Intelligent Mark Recognition), OCR and ICR activity, we can see that around half of those deploying OCR for machine-written text use it for invoice automation (AP, Accounts Payable) although the most popular forms scanning application is actually check (cheque) scanning.

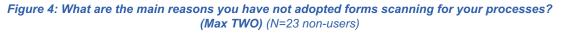
Forms Processing user experiences of text and handwriting recognition (OCR/ICR

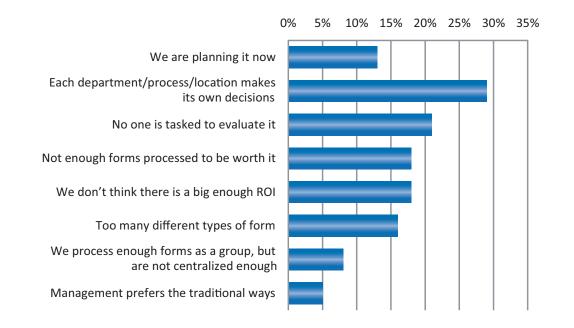
Figure 3: How do you pre-process forms coming into your business unit, or generated within it? (check all that apply, including outsourced services) (N=255: part 2)



Non-Adopters - forms scanning

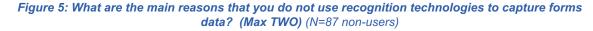
The main reason given for not adopting forms-scanning is localized decision-making within individual departments and processes, and this is particularly the case in mid-sized businesses. Even where a common, centralized requirement for scanning can be established, leadership will often fall between IT, records and facilities management. There is also confirmation here that a certain level of forms throughput is considered necessary before the technology becomes cost effective – a situation that can be improved by centralizing mail deliveries into a single address.

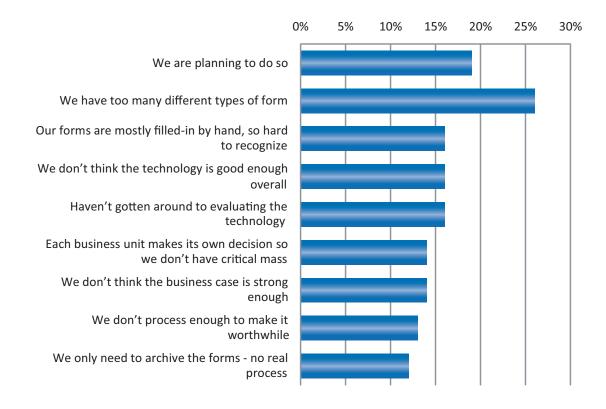




Non-Adopters – OCR

The main reason given for not adopting recognition software is the feeling that it is difficult to accommodate multiple forms layouts. Although a certain amount of pre-setup is required for each form, a modern capture system will make this very easy. Multiple form types can then be fed from the scanner in a mixed feed, and the capture software will automatically separate and detect each form-type and its layout, and use the correct template to find the fields. The next most likely reason is the difficulty of dealing with hand-written fields, and we will deal with that in detail later.





Outsourcing

As we might expect, the largest organizations are nearly 3 times more likely to outsource their forms-processing than the smallest, and they process around five times as many forms overall. Mid-sized companies reflect a balance between these two, although they are likely to outsource a higher percentage of their forms.

Figure 6: What percentage of your forms throughput do you outsource? (N=240, excl. 10 Don't Know)

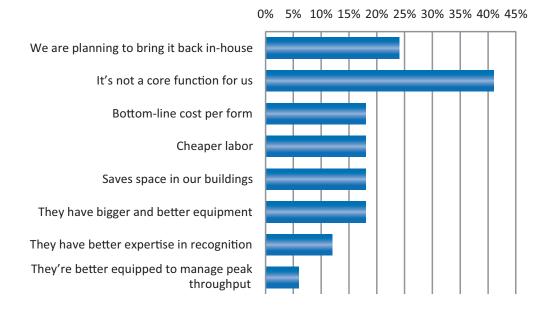
Size of organization	Average no of forms/day	Use outsource	% of those outsourcing 25% or more of forms	% of those outsourcing 50% or more of forms
10-500 emps	1,216	14%	58%	50%
500-5,000 emps	2,349	27%	77%	59%
5,000+ emps	6,787	41%	74%	39%

Forms Processing user experiences of text and handwriting recognition (OCR/ICR)

Drivers for Outsourcing

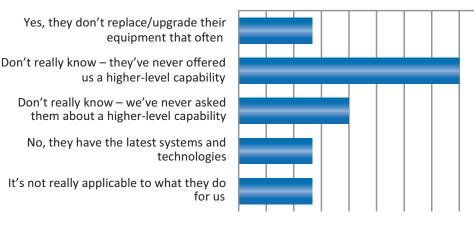
The main driver for full outsourcing is that it is not a core function. The possibility that an outsource might have better equipment or expertise is not such a key issue, and nearly 25% have plans to bring forms processing back in-house.

Figure 7: What percentage of your forms throughput do you outsource? (N=17, fully outsource users)



There is also a suspicion that once an outsource contract is in place the level of potential OCR/ICR capability is not upgraded, with 40% not being offered a higher level of recognition – and 20% not asking about it. In addition, 13% feel that their outsource may be using old or outdated equipment. This indicates that outsourcers may be missing out on the potential of an increased level of capture and a greater value-add, particularly with regard to hand-writing recognition.

Figure 8: Do you think that the level of ICR/OCR technology at your outsource is hampering the degree of handwritten recognition they can do for you?? (N=17, fully outsource users)



0% 5% 10% 15% 20% 25% 30% 35% 40% 45%

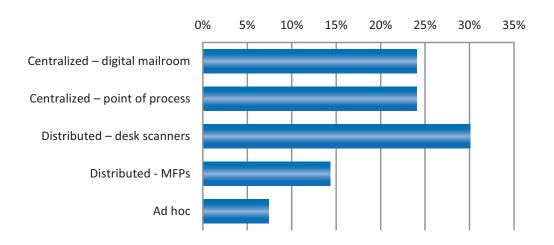
– user experiences of text and handwriting recognition (OCR/ICR)

Capture and Recognition Strategies

Central vs. Distributed

Although there has been some movement over the last few years away from centralized scanning towards distributed scanning using desk-top scanners and MFPs, more recently we have seen many organizations, particularly larger ones, move back to a centralized model, primarily to create a digital mailroom concept where incoming forms and mail are scanned on entry to the business and distributed electronically. As well as keeping paper out of the business, this can also justify a greater level of investment in associated capture and recognition technology. In this survey, 24% are using a digital mailroom for their forms scanning, and overall, 4% more are centralized compared to distributed. Those using a digital mailroom are more likely to funnel the majority of their forms through it.

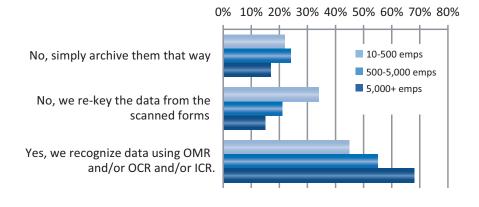
Figure 9: What is the primary forms-scanning mechanism in your business unit for process input? (N=216, excl.17 using outsource and 23 not scanning forms)



Data Capture and OCR

Taking a closer look at how scanned forms are subsequently processed, we see that around 20% are simply scanning direct to archive, either before or after a paper-based process. Even amongst the largest organizations, 17% are scanning forms but have no capture functionality.





Amongst smaller organizations 45% are using data recognition rather than data re-keying, but more than a third are scanning their forms and re-keying the data from the scanned image. Only 15% of larger organizations are working this way, with two-thirds doing some form of data capture.

Even where data capture *is* being used, the most common application is for archive indexing (64%). Similar indexing procedures for workflow routing, and for search-term extraction are also common. 40% use the data captured from the form to partially or fully populate the process, with data capture to financial processes likely to be the most popular application.

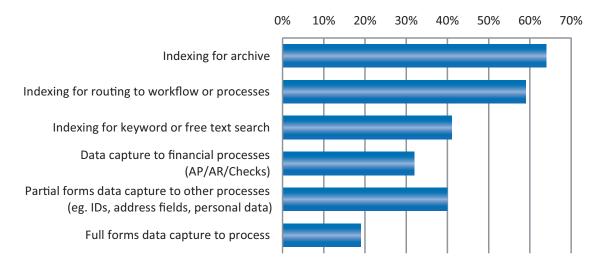
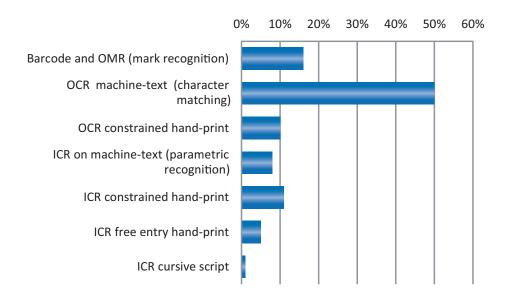


Figure 11: What uses do you make of captured forms data? (N=102 capture users)

Levels of OCR/ICR

Traditional OCR of machine text using character-matching techniques is by far the most popular method in use. Along with the much simpler optical mark or barcode recognition, this is the highest level of sophistication for two thirds of our recognition users. We can add to this 8% who are using the parametric analysis techniques known as ICR to capture machine text. 27% are recognizing hand-writing in some form, mostly as constrained hand-print where the person filling out the form needs to keep within a box or marker for each character. 6% are utilizing unconstrained hand-writing recognition and/or cursive script, which tend to be much more challenging for the recognition software.

Figure 12: What is the highest level of recognition that you use? (N=102 capture users)

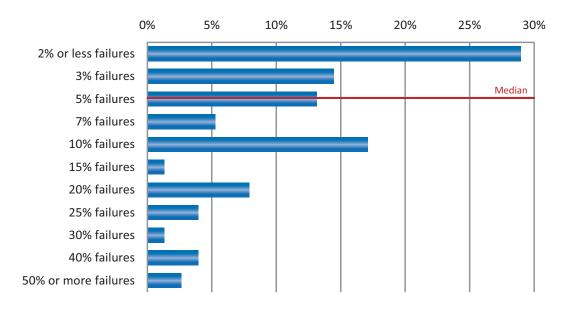


Forms Processing – user experiences of text and handwriting recognition (OCR/ICR)

OCR Performance

An important aspect of any data capture operation is the level of "straight-through" recognitions that don't require QA intervention or manual re-keying. This can be measured on a character-by-character, field by field or form-by-form basis. We compared users' assessment for field-by-field and form-by-form and found little difference in reported results.

Figure 13: For your OCR/ICR technology, what recognition failure rate/QA intervention rate would you say you are getting on a form-by-form basis? (N=75 OCR/ICR users, excl. 21 Don't Know)

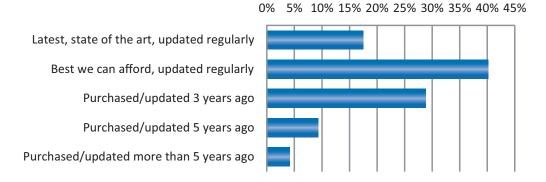


More than a quarter have a very low intervention rate of just 2% or less. 56% are achieving a failure rate of 5% or less on a form-by-form basis – ie, a 95% of forms processed without intervention. 79% achieve a fail rate of 10% or less. There is evidence of some complacency in monitoring recognition performance as only 25% regularly calibrate the performance of their scanners and OCR using standard test pieces.

Currency of OCR/ICR Software

Obviously, in this survey we do not know how difficult the forms are to recognize or how many fields they are on each form. We can, however, take a view on how many are using the latest recognition software. We can see from Figure 14 that 58% are up-to-date – at least to the capability they can afford - 29% last updated 3 years ago, and 13% updated 5 or more years ago. This could explain the long tail in recognition failure rates, and also has a bearing on users' experience of hand-writing recognition.





Hand-Writing Recognition

As we outlined earlier, recognition of hand-writing is very compute-intensive and considerable advances have been made of late in terms of performance and throughput, although some vendors have made greater strides than others. More importantly, we need to look at the drivers and potential benefits of recognizing hand-written form fields.

Drivers

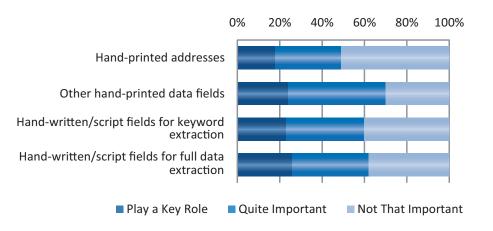
In most organizations, hand-written fields are prevalent on a significant number of their forms, with around a third having free-form, unconstrained fields on half or more of the forms they process.

Figure 15: How many of the forms processed by your business unit (or outsource) would you say have hand-written fields for: name and address, other textual data fields, unconstrained free-form data? (N=219 excl. 31 don't knows)

Hand-written fields	None	25% or more of forms	50% or more of forms
Name and address fields	20%	57%	38%
Other data fields	11%	62%	42%
Free text/open-ended data	12%	56%	32%

Not only are these hand-written fields prevalent, they are also important to the efficiency of the process. In 20% of organizations they play a key role, and in a further 40% they are quite important – including the free form script fields. As we enter the "big data" era, the contents of these "comment" fields is becoming even more important as organizations look to glean all kinds of information for product improvement, sentiment analysis, fraud detection, etc.

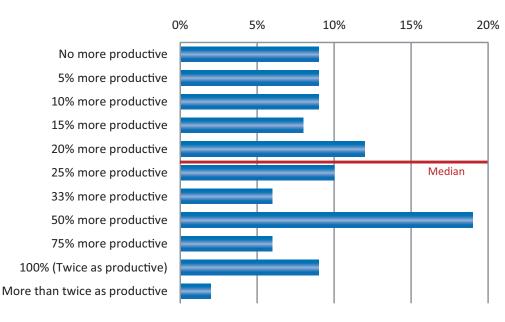




As we might expect, therefore, our survey participants estimate that they would achieve a considerable productivity saving if they were able to automate the recognition of hand-written text. The average estimate is 34.8% improvement, with a median at 23%. 36% would expect a 50% or more improvement.

Failm White Paper

Figure 17: How much more productive would you estimate that your admin staff would be (or are) if you could automate (or have automated) the recognition of hand-written text? (N=252)

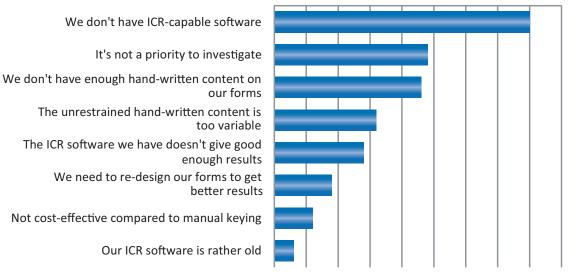


ICR Adoption

We asked the general survey sample for their assessment of current recognition technology for hand-written text. Overall, 20% are positive in their assessment, and a similar number feel it works well on constrained text. A third admit that they don't know, as they haven't evaluated it lately. Non-OCR users are likely to be much more sceptical of ICR technology – 10% positive compared to 31% of OCR users – and are much less likely to be basing their view on a recent evaluation – 43% have not evaluated it recently compared to 22% of OCR users.

Crucially, the main reason that most users are not doing hand-writing recognition is that they do not have ICRcapable software, followed by a lack of willingness to evaluate it. Doubt about the potential results is much lower down the list, and is centered on form design and content.

Figure 18: If you do not use hand-writing recognition, what are the main reasons? (max TWO)) (N=60 OCR but not ICR users)

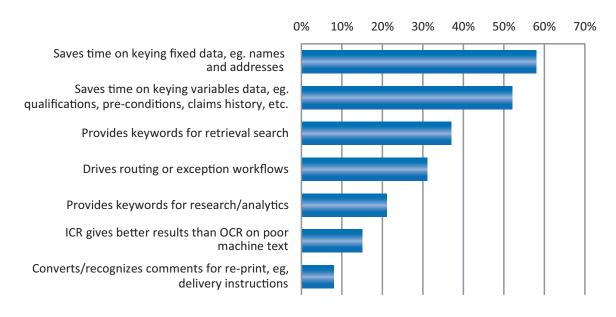


0% 5% 10% 15% 20% 25% 30% 35% 40% 45%

As we saw earlier in Figure 3, 12% of overall respondents are using ICR to recognize hand-printed constrained field entries, and 6% use ICR to recognize written script and free-form entries.

For these ICR users, the biggest benefit is time saving on fixed data such as names and addresses, and these are relatively easy fields to validate against a names and address database. This is followed by variables data such as educational qualifications, medical pre-conditions, claims history, etc., where a degree of contextual validation can be applied. Extracting keywords from free text fields is also an important benefit for auto-classification and search.

Figure 19: If you use hand-writing recognition (ICR), what are the main benefits in your application? (check those that apply) (N=25 ICR users)



Conclusion and Recommendations

Until such time as we are all equipped with tablets and e-forms, or within easy reach of an always-on computer, paper forms will be the backbone of information gathering and process input. Scanning forms for archiving or image workflow is a widely accepted way of reducing storage space, improving access and speeding up processes. However, we have seen that many organizations are slow to take the next step, which is to replace manual data-keying with recognition software, and automatically transfer data into the routing or indexing engine, or better still, into the process itself. A frequently given reason is the difficulty of recognizing hand-written text, and we have found that hand-written address, data and free-format fields play an important role in most business processes – and an increasing one as organizations seek to exploit the "big data" they may contain. We also found that a significant proportion of business forms, no matter how well designed, still contain a significant number of hand-written field entries. Over and above that is the disconnected decision-making in many organizations that makes it very difficult to consolidate scanning and capture requirements across multiple departments or processes, particularly with regard to implementing a digital mailroom scenario.

For those organizations that use OCR to recognize machine-printed text, performance is on the whole very good, with the percentage of hands-free throughput in the upper nineties. It is generally acknowledged that the accuracy of OCR on machine text will usually be higher than human re-keying. However, we have found that many organizations have not upgraded or refreshed their OCR software for some years so may be falling behind the curve of what is now possible. For many, the output of data capture is used merely to automate the indexing for routing and search, rather than to feed the process, and there is still much scanning that takes place at the end of the process, allowing paper to hold sway within the process itself.

We have found that although users understand the potential benefits of hand-writing recognition in terms of a substantial improvement in process efficiency of around 30%, there is a level of both perception and complacency that is based on out-of-date evaluations of how well a modern ICR hand-writing recognition system can work – and indeed how much it might cost. In many cases, the agency for scanning and capture, whether it is an inhouse unit or an outsource bureau, is not exploring the possibility with the business process managers for automatically capturing these very useful free-format fields. As might be expected in such a demanding application, there is also considerable variation in the sophistication of ICR algorithms embedded within the main capture system products.

Recommendations

- Ensure that there is a clear responsibility in your organization for pursuing paper-free processes. Consult with process owners to consolidate requirements, particularly if a digital mailroom solution serving multiple business processes might be appropriate.
- If you are not currently scanning forms at all, re-evaluate the reasons and include all of the benefits of paperfree processes – visibility, accessibility, speed of response, mobilization.
- If you are scanning forms but not capturing data through OCR, evaluate savings in keying costs, speed improvements, and quality of data.
- Do not assume that you do not have sufficient forms to be cost-effective, nor that you have too many different types of form. Centralizing all mail processing can change the tipping point, and the cost/performance ratio of OCR technology has dramatically improved over the last few years.
- If the prevalence of hand-written fields on your forms has put you off automating your capture, or if you are currently using OCR for partial capture and ignoring valuable hand-written content, take a fresh look at hand-writing recognition and the latest ICR capabilities. ICR could also improve your machine-text recognition.
- Collect a number of examples of both typical and demanding forms, with mixtures of machine text and handwriting, and have different capture vendors show how well they can capture the data. Be prepared to provide supporting data for look-up and validation. Ask about mixed feeding of form types.
- If you are using a bureau or DPO (Document Process Outsource), ask them if they have an up-to-date ICR capability that could further improve the level of capture they offer. If you are a bureau or DPO, have you geared up your capabilities to offer the maximum value add as far into the process as possible?

References

1. AIIM Industry Watch, "The Paper Free Office – Dream or Reality" Feb 2012, http://www.aiim.org/Research/Industry-Watch/Paper-Free-Capture-2012

Forms Processing – user experiences of text and handwriting recognition (OCR/ICR)

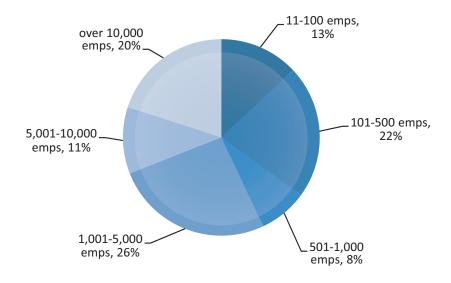
Appendix 1: Survey Demographics

Survey Background

The survey was taken by 324 individual members of the AIIM community between 09 March 2012 and 29 March 2012 using a web-based tool. Invitations to take the survey were sent via email to a selection of the 65,000 AIIM community members

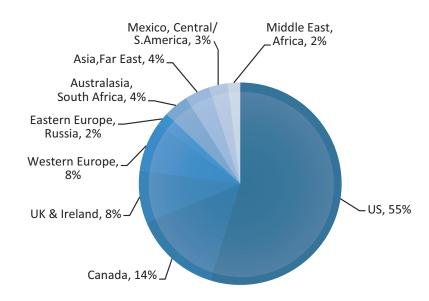
Organizational Size

Organizations of 10 employees or less are excluded from all of the results in this report. On this basis, larger organizations (over 5,000 employees) represent 31%, with mid-sized organizations (500 to 5,000 employees) at 34%. Small-to-mid sized organizations (10 to 500 employees) are 37%.



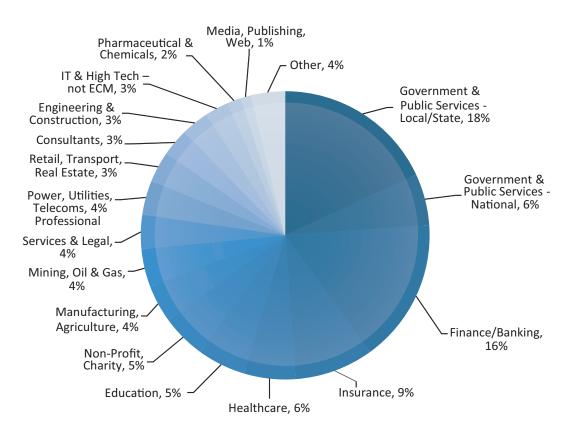
Geography

US and Canada make up 69% of respondents, with 18% from Europe.



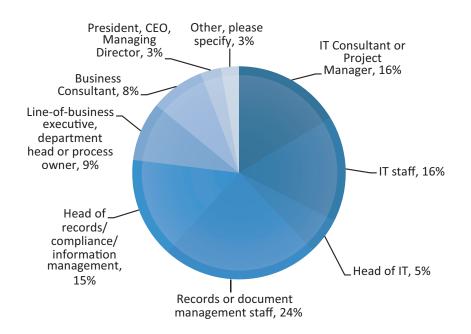
Industry Sector

Local government and public services represent 18%, and national government 6% - reflecting a long history of forms processing in the government sector. Finance, banking and insurance represent 25%. ECM suppliers and outsource bureaus have been excluded. The remaining sectors are evenly split.



Job Role

Records or Information Management disciplines make up 39% compared to 37% from IT. Line of business managers and business consultants make up 23%.



aiim White Paper

Forms Processing user experiences of text and handwriting recognition (OCR/ICR)

UNDERWRITTEN BY



Parascript LLC

Parascript is a global leader in developing cursive, handprint, and machine print recognition technology. Leveraging digital image analysis and advanced pattern recognition, its software enables critical business automation in areas like forms processing, postal and financial automation, fraud prevention and medical imaging. Parascript's award-winning technology draws on a proven 15+ year track record and processes billions of critical document images annually. Through its partner network, it enables companies to save money by eliminating costly manual data entry with automated recognition that accurately, securely and quickly turns characters into useful data. Parascript recognition technology reads all text styles, whole words or phrases, and deciphers poor quality machine print and handwriting unreadable by other recognition engines.

Parascript recognition products are designed to be configurable and supported by its worldwide network of integrators, original equipment manufacturers and value-added resellers. Fortune 500 companies, postal operators, major government, and financial institutions rely on Parascript products, including the U.S. Postal Service, IBM, Bell and Howell, Fiserv, Selex Elsag, Lockheed Martin, NCR, Siemens, and Burroughs.

Visit Parascript online at http://www.parascript.com.



AIIM (www.aiim.org) is the global community of information professionals. We provide the education, research and certification that information professionals need to manage and share information assets in an era of mobile, social, cloud and big data.

Founded in 1943, AIIM builds on a strong heritage of research and member service. Today, AIIM is a global, non-profit organization that provides independent research, education and certification programs to information professionals. AIIM represents the entire information management community, with programs and content for practitioners, technology suppliers, integrators and consultants.

© 2012 AIIM

1100 Wayne Avenue, Suite 1100 Silver Spring, MD 20910 301.587.8202 www.aiim.org

AIIM Europe

The IT Centre, Lowesmoor Wharf Worcester, WR1 2RR, UK +44 (0)1905 727600 www.aiim.eu