PARASCRIPT®

Research Study

Capturing Dark Data & Handwritten Information as Part of Your Information Governance Process

What’s inside?

6273 Monarch Park Place, Longmont, CO 80503 USA
T: 303.381.3100 | toll free: 888.772.7478
F: 303.381.3101 | info@parascript.com

www.parascript.com

About ARMA

ARMA International is a not-for-profit professional association and the authority on governing information as a strategic asset. Established in 1955, the association has approximately 27,000+ members, including records and information managers, information governance professionals, archivists, corporate librarians, imaging specialists, legal professionals, IT managers, consultants, and educators, all of whom work in a wide variety of industries, including government, legal, healthcare, financial services, and petroleum in the United States, Canada, and more than 30 other countries around the globe.

About Parascript

Parascript LLC, a leading document capture company, develops solutions that read information from forms and documents. The company's advanced recognition technology processes virtually any document format and text type (handprint, machine print, cursive, marks and more), providing fast, reliable access to information and transactions. Parascript's software is used by Fortune 500 companies, postal operators (including the U.S. Postal Service), major government and financial institutions. It is online at www.parascript.com

About The Study: Process Used and Survey Demographics

The purpose of the study from ARMA and Parascript on Capturing Dark Data & Handwritten Information As Part of Your Information Governance Process was to better understand the role of records management professionals, to gauge their understanding of information governance (IG) and dark data, determine their organizations' adoption of IG and dark data, and solicit questions that they were looking to answer.

The survey was conducted over three weeks from May to June 2014. ARMA distributed the survey via three email communications to its audience of records management professionals, and Parascript managed the survey. Over 200 ARMA members responded – 50% indicated they are managers/directors, over 20% operations/end user, and 16% 'other' and the remainder were divided between CEO level and vendor or consultant.

Introduction

In today's vast digital world, organizations are scrambling to better control information to meet legal, regulatory, and business requirements. Adding to the challenge is an increasing number of information sources, including online, email, and print. From each of these comes growing volumes of records, receipts, forms, email, and desktop documents and more, further amplifying the amount of information that is out there for companies to discover, classify, archive, and manage.

Among all of these documents and records is the continuous need for businesses to better leverage data and improve information governance. Companies need to accurately, efficiently and securely capture this intelligence to enable it to be effectively leveraged. The 'intelligence' within records and documents can include everything from printed information to handwritten information and signatures—holding a wealth of knowledge for the organization.

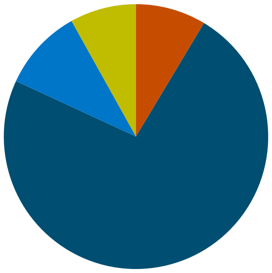
Organizations today must leverage this information, while simultaneously balancing the need to protect it through better information governance policies. Capturing valuable, yet often times hidden, data on documents can help organizations gain a competitive advantage by increasing business performance, reducing risk and better managing costs. By accurately and automatically capturing information, organizations can route data more quickly, store and retrieve details more effectively, and identify risk points such as unauthorized transactions.

But how does one identify and access this data? What is this information on forms and documents that many organizations are missing out on? How does locating this data fit into the greater context of information governance?

In this research study with ARMA, we set out to answer these and other questions.

FIGURE 1

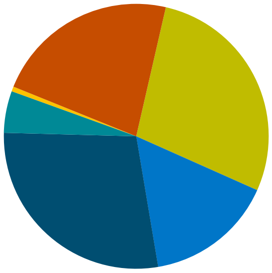
How do you define information governance, as it relates to your organization?



🟧	8.6%	Management of a document through its lifecycle for the purpose of compliance
🟦	73.2%	Management of a document through its lifecycle for the purpose of classification, archival, retrieval, security, discovery, and compliance
🟦	10%	A new name vendors gave to records management
🟨	8.1%	Other

FIGURE 2

What line of business is responsible for Information Governance?



🟨	28%	Does not have a centralized governance strategy
🟦	16%	Operations
🟦	28%	IT
🟩	5%	Finance
🟨	1%	Sales & Marketing
🟧	22%	Legal
⬛	0%	R&D

First, What is Information Governance?

Information Governance is the management of a document through its lifecycle for the purpose of classification, archival, retrieval, security, discovery, and compliance.

In particular, information governance, or IG, is the set of multi-disciplinary structures, policies, procedures, processes, and controls implemented to manage information at an enterprise level. IG should support an organization's immediate and future regulatory, legal, risk, environmental, and operational requirements.

IG encompasses more than traditional records management. It incorporates privacy attributes, electronic discovery requirements, storage optimization, and metadata management.

A majority of executives, surveyed in the ARMA report, indicated they feel they are informed about information governance and agree with the definition of IG as management of a document through its lifecycle for the purpose of classification, archival, retrieval, security, discovery, and compliance (over 73%).

Yet many had questions, including: How do you get executives to understand Information governance? And, what is the clearest definition of a "record" that we can hang our hats on? **SEE FIGURE 1**

For the majority of ARMA members surveyed, IG also has established lines of business around records management. ARMA respondents indicated that IT, Legal, or operations, in that order, handles IG, and 75% say that same department also handles records management. The other 28% of respondents say they do not have a centralized governance strategy. **SEE FIGURE 2**

In another study by AIIM, Automating Information Governance - Assuring Compliance, over 35% or respondents cited loss of intellectual property or company information as one of the biggest risks to their company due to poor information governance. 41% view excess litigation costs or damages resulting from poor record keeping as being one of the biggest risks of IG failure.

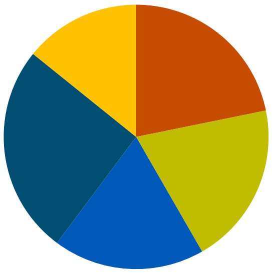
Meanwhile, many companies recognize the benefits of good information governance, the most significant of which is to reduce storage and infrastructure costs (nearly 60%), followed by exploitation and sharing of knowledge (50%).

As information grows in prominence in a company's architecture, a more expansive definition of this critical category is emerging, one that has broader implications and benefits, according to industry Analyst, Kevin Craine. While risk mitigation remains an important part of the mix, organizations are now also beginning to view information governance in terms of organizational cost and performance. Although information can carry risk, there is also an expense in using information inefficiently, or not using it in ways that improve the performance of the organization. And, conversely, there is opportunity to win significant gains in having better access to important information—from having easy access to customer feedback, the ability to classify and group relevant documents and signatures, or the ability to capture important feedback and comments in the margins of a contract, to name just a few.

The new model of information governance is focused on three areas: risk, cost, and performance. It makes sense to include activities to reduce the costs and risks associated with finding and using business information. But the real value of information governance may instead be found by building capabilities that increase business performance. Doing more with data to improve processes and products, boost brand satisfaction, and enable more effective, strategic decisions are just a few benefits of this expanded approach.

FIGURE 3

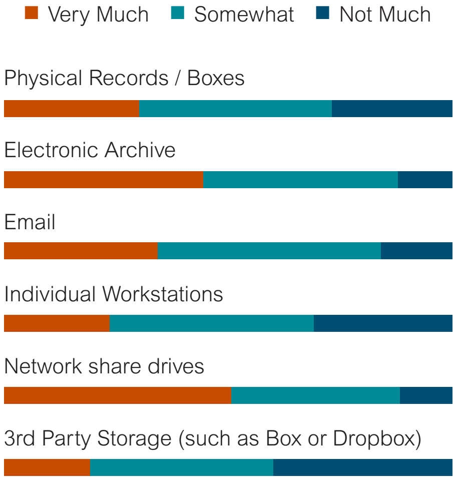
How do you define dark data, as it relates to your organization?



🟧	21.8%	Data my organization gathers and retains with no plan for utilization
🟨	19.9%	Data my organization gathers and retains "just in case" (i.e. for discovery)
🟦	18.5%	Data my organization gathers during a process, but doesn't utilize or retain
🟦	25.6%	It is just a buzzword for data that we don't use
🟨	14.2%	Other

FIGURE 4

How valuable is the dark data hidden in these locations?



What is "Hidden' or 'Dark Data'? And What Does It Have to Do With IG?

One way to do more with data is to better leverage information, which may otherwise be lost. This information may be referred to as 'dark data'. The use of the term 'dark data' is growing in popularity, prompting many to ponder its definition or ask if it is simply a buzzword? What is dark data?

Gartner defines dark data as a valid term to describe the 'information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes' (for example, analytics, business relationships, and direct monetizing).

Similar to dark matter in physics, dark data often comprises most organizations' universe of information assets. Thus, organizations often retain dark data for compliance purposes only. Storing and securing data typically incurs more expense (and sometimes greater risk) than value.

'Dark data' is information that is not fully identified, captured, and leveraged. And while it is everywhere, it is especially prominent on paper documents such as forms, receipts, checks, and applications, and may include notes and annotations, signatures, and other handwritten and printed information. It is referred to as "dark" because often this information is not captured or leveraged, or viewed as usable.

Dark data is even held in boxes stored away in a company's closet or basement. There may not always be a pressing need to regularly access it, or managers may assume that automating the capture of this data may be too hard, so it may be pushed aside as something to do later. However, there are both risks in not knowing what information is being stored, and also potential value in capturing and utilizing that data. Dark data can be found in virtually every type of document, as well as in the handwritten, digital, and machine printed information therein.

But according to the ARMA study, the definition of dark data may still be unclear: 25% of respondents felt it was just a buzzword; and more than 21% agreed it's 'Data my organization gathers and retains with no plan for utilization.' Meanwhile, over 19% said it's 'Data my organization gathers and retains 'just in case' (i.e. for discovery)' and over 18% said it's 'Data my organization gathers during a process, but doesn't utilize or retain.' **SEE FIGURE 3**

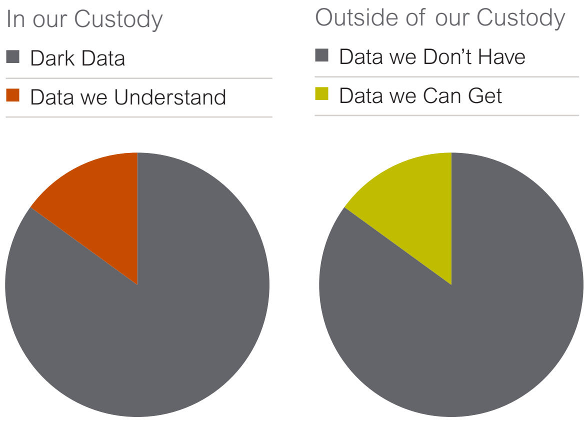
Network share drives, electronic archives, email, and physical records and boxes are most likely to have valuable dark data in them, according to ARMA respondents. **SEE FIGURE 4**

How Can Organizations Better Capture Dark Data and Other Information?

Regardless of what you call it, organizations are missing out on leveraging valuable information.

Consider the following graphic (from Anne Tulek of Access Sciences in an AIIM Webinar, Capturing the Value of Your Information and Data) that illustrates data that companies have custody or ownership of (left) and data companies do not (right). There is only a small portion of each that is actually currently being understood or utilized, or that companies could capture and use with some effort.

FIGURE 5
Dark Data and Big Data



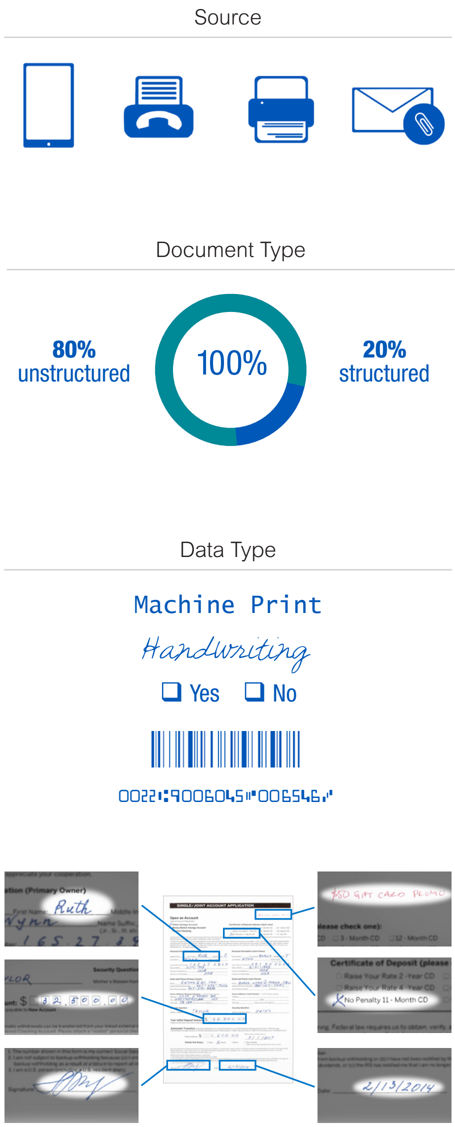
Another way to think about it is that Information Governance is the largest discipline, with records management as part of IG. Within records management are physical and digital records, and within that is some level of dark data, which can vary from organization to organization, that is not being captured.

Capture systems and data recognition technologies enable organizations to do a better job at getting at dark data for information governance. They provide a way to digitize information that would otherwise be locked away on paper. It's hard to reduce risk, cut costs and improve performance when you are constrained by a paper-pushing process. Scanning helps reduce the burden of paper, while advanced recognition technologies "read" printed information on documents, extract that data, and enable a variety of processes and procedures that take information governance to a new level.

By utilizing advanced recognition technology to capture and automate access to information, companies will not only gain quick access to dark data, but can also greatly reduce risk and improve IG processes.

Consider the following graphic that shows how capture technology can open up the path to better access data. From scanning the form or accepting email attachments, to virtually pulling the data off the page, today's recognition technologies can process information regardless of the type of data—handwriting, cursive, signatures, print—or the type of form—application, purchase order, contract, etc.—or the order these documents are in. **SEE FIGURE 6**

FIGURE 6
Capture Technology Capabilities



OCR and ICR

Specific capture technologies that enable organizations to capture and better leverage data include OCR and ICR. OCR and ICR technologies improve information governance by enabling organizations to capture and digitize information that would otherwise be lost or overlooked in the scanning process.

Optical Character Recognition, or OCR, is the digital conversion of typewritten or machine-printed text into computer-readable text. Organizations use OCR to process forms, checks, and a variety of business correspondence. It is widely used as a form of capturing information (versus manually entering it) from passports, invoices, bank statement, receipts, business card, mail, or any number of printed records. It is a common method of digitizing printed text so it can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as machine translation, text-to-speech, key data extraction and text mining.

ICR (or handwriting recognition) is advanced recognition technology that allows fonts and styles of handwriting to be read by a computer. Using ICR, companies can readily capture important handwritten information.

ICR can be used to capture annotations, including subtle, hard-to-read handwriting to full-blown commentary, and keywords, often scribbled in the margins or in a comments field, including customer sediment, special service requests, and change of address.

More advanced versions of ICR can be capable of processing millions of documents, and can perform automated recognition on unconstrained handprint and cursive handwriting, as well as difficult machine print that may be encountered on forms. Advanced ICR can offer significant cost savings, greater recognition versatility and the ability to adapt to any style of form, along with high read rates, and accuracy levels. The technology can boast 97%+ accuracy rates in reading handwriting in structured forms.

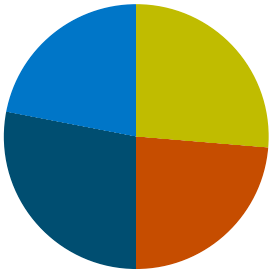
Why is capturing signatures and handwriting via ICR important? Vast amounts of 'knowledge' can be contained in handwritten data and signatures on forms. While ICR capabilities vary widely, some can support archival, retrieval and discovery of this important information—making it even easier for organizations to access handwriting and signatures. Consider, for example:

The benefits of having greater access to handwritten keywords and phrases:

- Companies can reduce compliance risks by controlling the flow of incoming documents and linking them with respective business transactions.
- Perform legal discovery against specific keywords or phrases.
- Identify keywords for archival, utilizing them to classify documents and generating metatags.
- Look for opportunities for customer service to engage a customer for a case study or, conversely, fix a brewing product problem based on handwriting in comments functions or other fields.

FIGURE 7

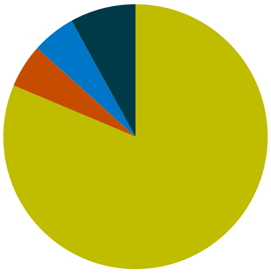
Are you applying OCR (optical character recognition) to imaged documents?



🟨	37.6%	No
🟧	33.8%	Yes – for indexing to support archive
🟦	40%	Yes – for classification / retrieval
⬛	31.4%	Yes – for business process

FIGURE 8

Are you applying ICR (intelligent character recognition) technology for handwriting recognition?



🟨	87.1%	No
🟧	5.7%	Yes – for indexing to support archive
🟦	5.7%	Yes – for classification / retrieval
⬛	8.6%	Yes – for business process

The benefits of having greater access to signatures:

- Companies can match, and validate, signatures on archived documents, including those for processing transaction orders or claims to prevent fraud, group relevant documents, and more.
- Documents that do not contain necessary signatures can be sent to exception handling for proactive follow-up by customer service.

But instead, many organizations are missing out on obtaining important information on forms. According to the ARMA study, 69% of respondents are using OCR to recognize printed information. **SEE FIGURE 7**

But 87% are not using ICR to capture handwritten information or signatures. **SEE FIGURE 8**

Additionally, while a 2014 study conducted by AIIM recognizes handwriting is prevalent on forms, and many companies could benefit from using ICR software to capture it, only 6% of respondents in the AIIM study say they are using the technology.

The same study found that most enterprise organizations have scan and capture systems already in place, but only half employ any kind of text recognition technology. And although 40% indicate that their inbound forms contain handwritten data fields, very few are attempting to capture this information, either deferring it to manual keying or ignoring it outright. Over a third (37%) said that they had never even considered handwriting in the mix.

The result is that in some cases information is either not captured at all, and is lost. Alternatives might be that information is:

- Keyed in from forms—tedious and not always accurate, or
- Less automated technology and/or processes—such as paying for recognition by the 'field' or number of documents scanned—may be employed and the process is not automated to the extent that the organization can benefit.

This trend may be changing, however. The study shows also that a majority of companies (67%) feel that handwriting and annotations will "play a key role" and are "quite important" in their future strategies.

This implies a great deal of opportunity to sharpen processes that can lead to lower operational costs and better discovery and retrieval of data. Today's highly advanced capture technologies offer greater simplicity and security.

Protecting Dark Data

In any organization there is lost information that is never noticed, never leveraged, and no one ever thinks about. The question becomes: How do you know what you don't know and how it is hurting you?

How can you know if your organization would benefit from the information on forms if it never leaves the pages of your organization's physical records, or if you can never readily access handwriting, signatures and other information? How do you know about the missed opportunity of dark data in the boxes of documents your organization is storing?

Today, advanced capture technologies greatly automate recognition of all types of information. And it does all of this with a great deal of accuracy and a high level of security safeguards that can be set to protect information for your organization. For instance, ICR can be used with advanced, automated redaction to 'hide' sensitive information by obscuring it, usually in black, to make documents secure for distribution or archival. Traditionally, documents are copied, redacted using permanent markers, and re-copied to make sure no information is still legible, which can be quite labor intensive. Automated redaction speeds up processes, provides greater accuracy, and reduces labor costs and manual errors.

In any organization there is lost information that is never noticed, never leveraged, and no one ever thinks about. The question becomes: How do you know what you don't know and how it is hurting you?

How Can You Better Capture Dark Data When It Comes to Information Governance and Archived Documents?

One way to do this is by auditing the information you currently gather as an organization and rating your need to access it as well as the organization's availability to do so. Looking at this information can help determine the need to have greater access to information, as well as identify opportunities to better leverage it. And once organizations implement capture technology, they soon discover many other opportunities they can quickly benefit from.

For example, you might consider:

- What information are you capturing that you are not leveraging?
- What information are you not capturing and leveraging from forms and records that you could be?
- How could you better leverage signatures and handwriting for document classification of contracts, records, forms, etc.?
- How could you better leverage signatures and handwriting for fraud prevention on payment, contracts, and other documents?

Organizations can only fully leverage valuable data by capturing it via advanced technologies. Organizations today have even greater opportunities to obtain more accuracy in recognition and automation than ever before with advanced OCR and ICR. Those that embrace sophisticated solutions will reap the benefits of not only enhanced information governance and record keeping but greater intelligence and efficiencies in leveraging business knowledge.

Being an IG Champion

As the need for stronger information governance policies and methods to better leverage information increases, companies can no longer afford to ignore valuable handwritten content. These 'missed opportunities' can mean a gap in intelligence and performance between companies that capitalize on them and those that do not.

The more data that can be brought together the more it can be analyzed to discern patterns and make better decisions. Capture systems, for example, can help uncover a wealth of overlooked information and dark data that might otherwise be lost. Ultimately, the power to harness data fosters better business decisions that enhance the quality and efficiency of processes, products and services, not to mention a company's bottom line.