OPENTEXT

MARCH 2013

Discovery

Discovering the Power of Information

MARK BARRENECHEA
OpenText President and CEO

OPENTEXT CEO WHITE PAPER SERIES

OPENTEXT

Contents

Introduction

The OpenText Enterprise Information Management (EIM) White Paper Series is a set of publications from OpenText on the topic of Enterprise Information Management.

EIM is the discipline of discovering, managing, extracting value from, and building applications on top of unstructured enterprise information. At OpenText we know these Enterprise Information Management practices as the next generation of enterprise software.

To help present the topic of EIM, it will be described and detailed in the following white paper series:

- Enterprise Information Management (EIM)
- The Social Enterprise
- It's all Connected
- Focused on the Value
- The Journey
- Enterprise Content Management (ECM)
- Business Process Management (BPM)
- Customer Experience Management (CEM)
- Information Exchange
- ***Discovery***
- Mobile and Cloud
- Security
- Governance, Compliance, and Risk Management
- Information Flows
- Customer Case Studies

These white papers will be delivered in a series starting in the fall of 2012 and completed by the spring of 2013.

Discovery in EIM

In the first white paper in this series, *"Enterprise Information Management (EIM): the Next Generation of Enterprise Software,"* we defined Enterprise Information Management as "the discipline of handling all unstructured data within and between an enterprise and other organizations." The white paper also noted that unstructured information, even in this era of Big Data, still represents 90% of the information in most organizations.

But 90% of how much information? In a recent **Information Week Survey**, almost 50% of organizations claim to have more than 500 terabytes (TB) of data within their organization and 25% of organizations claim to have more than five petabytes (PB).

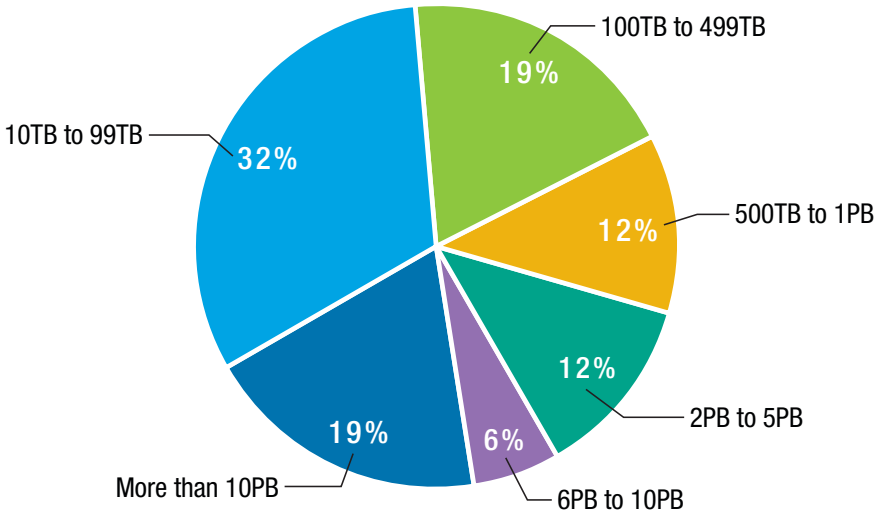


FIGURE 1:

Total Amount of Data in an Organization

Source: Information Week 2012 Big Data Survey of 231 business technology professionals, December 2011

The word "all" is significant in the definition of EIM above, especially when considered in light of the volumes of information held in most organizations. *How* can all of this information be effectively managed? What *kind* of information is all of the unstructured information? And *where* does it reside?

Email, Microsoft® Office® documents, searchable and non-searchable images, web and intranet pages, and increasingly social and collaborative content make up the majority of the unstructured information in the enterprise. This information lives in email servers, archives, file servers, Enterprise Content Management (ECM) systems, Business Process Management (BPM) systems, websites, bespoke applications, and of course, on our desktops and mobile devices. As well, large volumes of unstructured information live in legacy systems, often in legacy formats.

What is Discovery?

Discovery technologies and applications help solve some of the biggest challenges organizations face due to the deluge of electronic information created, consumed, and stored today. It is a category of EIM offerings that helps organizations capture, combine, and transform data across information silos into formats that can be analyzed for deeper business insight.

Business insight is gained by organizing, combining, and visualizing information in new and repeatable ways to identify relationships, risks, and new opportunities for growth. Leading-edge capabilities like progressive search, semantic analytics and navigation, and categorization help to mine, extract, and present the true (and often hidden) value of enterprise information.

Why EIM Needs Discovery

Discovery technology is the workhorse that powers information governance. Effective information governance programs demand that organizations balance the competing priorities in the creation, management, retention, and deletion of business information. This requires a keen understanding of the information that is under management, as well as the ability to act on this content. Whether these actions are tagging content, changing permissions, applying litigation holds, or moving content, Discovery technologies enable organizations to perform actions on millions and millions of documents.

As well as helping to govern information, Discovery provides a window into Big Data through the enrichment of content and content analytics. Uncategorized content can be automatically analyzed to help discern relevant topics, summaries, sentiments, and relationships to deliver more enriched information.

Discovery aggregates and integrates content across an organization within a larger EIM context – providing access to content that is trapped in silos – and synchronizes this content to reduce the burden on end users. Discovery technologies deliver non-intrusive solutions that eliminate the need for business users to sort and classify growing amounts of content, or even in some cases, recognize what they are searching for. Managers and employees rely on information to do their jobs, and having content indexed and accessible significantly improves efficiency, productivity, and employee satisfaction.

Discovery technology is also the honey that makes your web properties sticky and enables users to find what they're looking for, helping to monetize content and generate revenue through longer visits and more completed transactions.

Organizations across the globe are unleashing the power of their information with Discovery solutions to improve decision-making and drive business outcomes to create competitive business advantage.

Why Discovery Needs EIM

Within the context of an EIM solution, a centralized **Enterprise Content Management** (ECM) repository plays a fundamental role in helping organizations realize the value of Discovery solutions. While Discovery technologies work to identify content that has value, an ECM system enables this content to be accessed, managed, and archived. Moving large quantities of unmanaged content into an ECM repository not only reduces storage costs; it enables content to be classified and retention policies applied, allowing for eventual disposition and further cost and risk reduction.

For Discovery applications, ECM with its associated Records Management (RM) capabilities provides a single source of the truth, which is critical when Discovery is used to provide insight and decision-making support. ECM with RM capabilities also delivers a single source of policy that Discovery applications can extend into the information silos found in most organizations.

If we consider the time, money, and effort invested in **Customer Experience Management** (CEM) – another category of EIM offerings – Discovery technologies can be used to effectively measure and apply analytics to understand the impact and return on investment in CEM technologies. CEM programs deliver volumes of information in many formats. Clickstreams, social media interaction, and eCommerce transactions are well complemented by Discovery applications to help organizations process, enrich, analyze, and help to visualize critical customer-based information. CEM activity and data is becoming both a showcase and proving ground for Discovery technologies.

Who Should Care About Discovery?

Discovery technologies can be applied to problems that impact both the top and bottom lines in an enterprise. Embedded in the basic definition and use cases for Discovery are the concerns of the key stakeholders, including the Chief Legal Officer (CLO), Chief Compliance Officer (CCO), Chief Marketing Officer (CMO), and Chief Information Officer (CIO).

Completing proactive and reactive audits is a requirement to fulfill the litigation and compliance obligations of the CLO and CCO. Organizations require the investigative capabilities of Discovery to search, analyze, and collect information for eDiscovery. Additionally, Discovery, with its ability to help users find and act upon content, helps to enforce information governance policies on large volumes of content.

For the CMO, Discovery provides a deeper understanding of the customer, measures the effectiveness of marketing, and monitors the mobile, social and web trends that impact their brands. Discovery can be used to explore variances and relationships between data, including third-party information. By analyzing this information, organizations can learn more about customers to understand their needs and deliver a more tailored, satisfying customer experience. Insight into customer needs, personas, and behavior gives organizations the ability to make faster, more accurate and cost-effective business decisions. Discovery helps CMOs monitor their campaigns and brands, profile their target markets, gather and analyze competitive intelligence, and launch more effective marketing programs to drive revenue and increase customer retention.

The CIO is increasingly wearing many hats in the enterprise and has a lot at stake when it comes to exploiting the capabilities of Discovery technologies. Thanks to the combination of consumer technology and the "Consumerization of IT", the CIO is significantly more concerned about user satisfaction and productivity, and has become the voice of the end-user. Discovery improves operational effectiveness by automating and integrating processes and technologies. Reducing complexity and cost by rationalizing applications and decommissioning legacy applications is top of mind with many CIOs. To this end, many CIOs are using Discovery as a tool to identify and assist in the defensible deletion of content to control storage costs.

The CIO is also the owner and architect of the EIM vision within the organization. For this reason, it's the CIO specifically who focuses on Discovery as part of a larger vision to manage *all* of the information in the enterprise. Discovery capabilities are integrated and actionable, and to be effective, should not be treated as an afterthought in a technology stack.

Discovery Use Cases

Discovery applications have been designed to solve defined business problems. The applications support myriad use cases and best practices to maximize business impact. The technologies typically fall into two camps: those that impact the bottom line by supporting good information governance practices, and those that impact the top line by improving productivity, providing actionable insight, and monetizing content.

Discovery and Good Information Governance Practices

Many different stakeholders are often concerned about distinct aspects of information governance, which is why it's so hard to find common ground when it comes to the applications and policies that govern information. This is why projects become skewed to focus on one particular set of issues, especially when legal, IT, and records and information management departments have not been engaged. The following table illustrates key stakeholders and their differing information management concerns.

STAKEHOLDERS	INFORMATION CONCERNS
Legal	Risk
IT	Cost
Line of Business	Value and Productivity
Records Management	Risk and Value

In order to achieve balance in your information governance programs, policies and systems should address the concerns of each stakeholder group listed above. This is especially true when we consider that the business can be looked at in two ways: 1) using content to make money; and 2) the impact of policies and systems on end-user adoption and productivity.

The goal of information governance is to strike a balance between the competing priorities in the creation, ongoing use, retention, and disposition of content.

Discovery can help organizations strike this balance in a series of applications that address the risk, cost, and value of information all while ensuring the end-user productivity is maintained or improved.

eDiscovery: Litigation Hold

The requirement to place litigation holds on content is based on the Legal Discovery obligation to suspend regular deletion policy on documents and materials that are related to the subject matter of anticipated litigation. This means that for content under Records Management control, where content is regularly deleted as per policy, it's necessary to be able to immediately suspend the disposal of this content.

Records managers and litigation support personnel need to be able to respond quickly and easily to requests from legal based on the information at hand. When legal first anticipates litigation, the information that needs to be placed on hold can range from the vague to the very precise; the simple to the complex. The capacity to defensibly adjust the criteria of the hold as more information about the matter becomes apparent is also critical.

Discovery applications deliver simple and cost effective methods for reducing the risk of spoliation, which is the overt or accidental destruction of evidence. Combined with the Records Management capabilities of an ECM system to capture and manage information, litigation hold procedures can preserve content in place, ensuring that it will not be deleted or modified. The ease at which this can be done ensures that organizations can afford to be broad and inclusive in the application of litigation holds and that they reduce the risk of spoliation and associated sanctions.

eDiscovery: Early Case Assessment, Collection, and Culling

Legal teams can utilize Discovery capabilities to connect to disparate data sources and quickly index content and metadata. A quick, first assessment can provide valuable information on the basic facts, key custodians, and the volume of content that will need to be collected, processed, and reviewed. This information can then be used to determine a strategy for the matter, as well as providing the preliminary cost and risk information that can shape the decision to settle or proceed with the matter.

After a case strategy has been determined, the same Discovery technologies can be used to perform the required actions upon the content under index. Based on search results, content can be collected from various data sources and preserved. These data sources include desktops, file servers, Microsoft® SharePoint®, live email servers, and more. This is a particularly critical step when working with unmanaged content where no in-place litigation hold is possible.

LITIGATION HOLD – A TACTICAL ADVANTAGE

In the recent *Apple Inc. vs. Samsung Electronics Co. Ltd.* patent dispute, an inability to preserve email and other data almost resulted in the jury receiving adverse inference instructions against both parties. An adverse inference instruction means a judge will tell the jury to assume missing documents were favorable to the other side. In this dispute, the magistrate judge initially ordered an adverse inference instruction against Samsung for its failure to stop the automatic deletion of email. However, the judge later found Apple had also failed in its data preservation duties and ordered identical adverse inferences instructions against both parties. Subsequently, the judge agreed to omit both instructions to the jury. What would have happened if one side had their litigation hold procedures and capabilities in order?

TAKE-AWAY: Having your litigation hold capabilities in order can lead to tactical advantage in litigation.

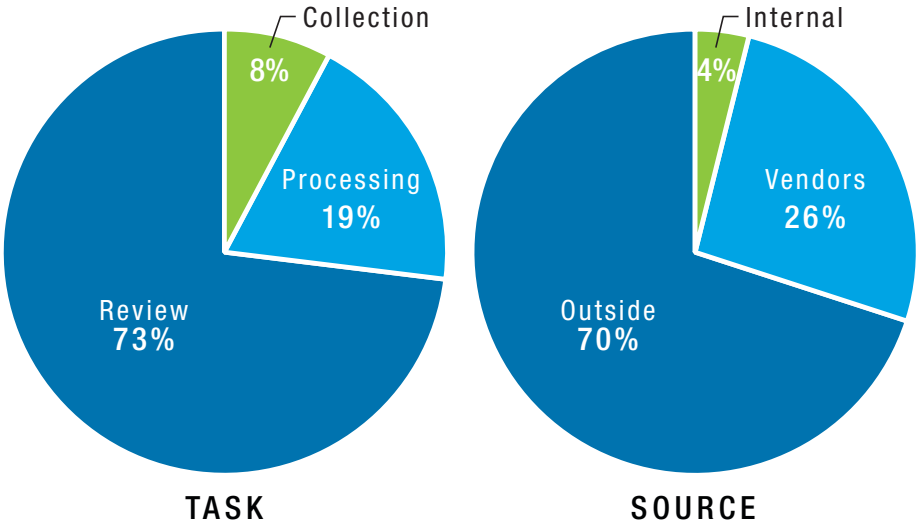


FIGURE 2:

Where eDiscovery Costs Come From

Source: "Where the Money Goes", Nicholas M. Pace and Laura Zakaras, Rand Institute for Civil Justice, 2012

NOTE: Values reflect median percentages for cases with complete data, adjusted to 100 percent.

Once content has been collected for preservation purposes, it can be merged with the content that has been preserved in place in ECM systems. This content can be further searched, analyzed, culled, and deduplicated (eliminating duplicate copies of repeating data). The benefit of further culling is realized through the reduction of the volume of content being sent for review. Since the review process accounts for almost 75% of the cost of eDiscovery, any action that can reduce the volume of content for review has an immediate and significant impact on cost.

eDiscovery: Metadata Extraction and Processing

The Rand study **"Where the Money Goes"** notes that processing is the second most expensive step in eDiscovery, constituting almost 20% of the total cost. Processing is the preparation of content so that it can be loaded into an eDiscovery review application. It involves the extraction of metadata, the separating of email and attachments, the extraction of content from archives, and the identification of duplicates and near duplicates. This must occur while the organization maintains and demonstrates that content and metadata have not been modified throughout this process.

The capabilities of Discovery technologies are particularly suited for processing. This is because indexing organizational content aligns closely with processing content. Indexing depends on the ability to understand location and extract metadata and full text from documents, email, compound documents, and archives. While they are indexing content, it follows suit that organizations are also processing content, which means that they are replacing some or all of the cost associated to the second most expensive step in eDiscovery. Processing is volume-based and typically charged on a per gigabyte (GB) basis. Charges can range from $100 - $1000 per GB, so the potential cost saving in using Discovery capabilities can be significant when cases involve hundreds and thousands of gigabytes of content.

PUTTING THE "DISCOVERY" IN EDISCOVERY

There is a reason that Discovery technology has seen so much innovation and been widely adopted for eDiscovery. The return on investment can be tremendous. A quick summary of the ROI Discovery can provide includes:

- Reduced Spoliation Risk – through comprehensive and immediate application of litigation holds
- Lower Collection Costs – from reduced IT involvement and simplification of the collection and preservation of Electronically Stored Information (ESI)
- Reduced Review Costs – through culling and a reduction in the amount of content that must be sent for review
- Significantly Reduced Processing Costs – from the full extraction of native files and metadata to create review-ready load files
- Managed Risk – by giving the Legal/ eDiscovery Team immediate access to ESI to assess risk and the viability of cases

Discovery and Auto-Classification

For the last five years, Records and Information Management (RIM) professionals have been stuck between the proverbial rock and a hard place. The rock is the mounting volume of unmanaged information, for which the only viable strategy is to classify that information. Content needs to be classified or understood in order to determine why it must be retained, how long it must be retained for, and when it can be disposed of. Managing the retention and disposition of information reduces litigation risk – along with discovery and storage costs – and it ensures that organizations maintain regulatory compliance.

The hard place records managers find themselves in is the realization that end-users have no desire and very little aptitude for classifying the information that they deal with on a daily basis. Users see the process of sorting records from transient content as intrusive, complex, and counterproductive. This can lead to project failure, endless training loops, rogue IT projects, and general end-user and management dissatisfaction.

The obvious solution to this problem lies in automating the classification of content. End users are notoriously inconsistent in applying retention policies to content like email, social media, and documents that tend to be numerous, ambiguous, and often not critical to their daily responsibilities. Typically, the biggest obstacle in end-user driven classification is that users opt out completely, or only classify a small subset of documents. An Auto-Classification solution provides a consistent, programmatic approach to applying retention polices on content.

To date, automated classification has not been widely accepted by records managers, compliance processionals, and legal departments – mainly due to their lack of confidence in the defensibility of allowing technology to determine the retention and disposition of content. Due to the following confluence of events, however, many organizations are reconsidering Auto-Classification.

- **eDiscovery becomes standard in court:** Up until a few years ago, governments, regulators, and the courts had a mistaken impression that the gold standard for categorization and classification of information was human review. The past few years of eDiscovery case law and best practices have demonstrated that perfection, when dealing with the truly massive volume of electronic content, is simply not possible. What the courts and regulators are looking for is due diligence and reasonable, good-faith efforts in dealing with electronic content.
- **Organizations target legacy content:** Most organizations have determined that they have vast stores of unmanaged legacy content that represents cost and risk when it comes to litigation and general IT agility. They have also realized that end users will never be coerced into trying to help solve the problem, making automating classification or retention policies a viable solution.
- **Technology assisted review gains visibility:** In response to the high costs of having lawyers review every document in eDiscovery, some litigants are turning to a combination of technology and process to reduce the number of documents lawyers are required to review. Technology assisted review is particularly relevant right now because of a number of high-profile cases where it has been challenged and accepted in the courts. This process can reduce the volume of documents that must be reviewed by up to 80%, while maintaining accuracy rates comparable to human reviewers. This acceptance of both technology and process in the courts has set the stage for Auto-Classification within the enterprise.

If records managers, and compliance and legal departments are now more willing to accept Auto-Classification to deal with the issues associated with managing very large volumes of information, they are only willing to do so if some basic criteria are met, based on the answers to the following questions:

How do we do it? A transparent process is required, with the basis for automated classification decisions readily understood, fine-tuned, and explained. This entails tools and processes for easily identifying the exemplar documents for training, the documents to test the training process, and feedback that clearly identifies methods for improving accuracy.

How do we prove it? This process involves the ability to facilitate adequate sampling to demonstrate both classification accuracy (precision) and completeness (recall) that aligns with the risk profile of the information. It's important to identify and create a statistically sound sample set for the content that is to be classified. As well, it's critical that the application support the very rapid review of documents to mark them as correctly or incorrectly classified. This QA exercise demonstrates the ongoing accuracy, as well as diligence and a consistent programmatic approach to classifying content that's critical should the courts or regulators question the efficacy of your retention management.

	ACCURACY	BUY-IN	RELATIVE ACCURACY
HUMAN (REFERENCE)	90%	50%	45%
MACHINE (CHALLENGER)	60 - 90%	100%	60 - 90%

FIGURE 3:

Auto-Classification – A Consistent Defensible Approach

Auto-Classification will be critical for organizations as they move toward managing all their information, especially the high-volume, low-touch content that is found in legacy content, email, and social media. It will require work, care, and feeding by Records Management professionals, but will be more than worth it when the benefits of retention and disposition are felt. End users and management will also appreciate that productivity is not impacted by additional requirements to classify content.

The Case for Content Remediation

Every organization has a problem with legacy content.

Legacy content is typically made up of files in the form of Microsoft Office documents, archives, image documents from capture applications, and a huge amount of email in loose and archive formats. This content can be found in numerous nooks and crannies of an organization, including in fileservers, desktops, departmental deployments of ECM, back-up tape, and unmanaged archives. Legacy content is made up of a mixture of valuable content that should be managed as records, and typically, a large percentage of content that could be deleted according to the organization's policies for the deletion of transitory content. The problem is separating the wheat from the chaff.

Content remediation is a combination of the technology and processes needed to organize, analyze, defensibly delete, and bring under management the legacy and unstructured information large organizations currently house. The content remediation process combines a number of Discovery capabilities. Initially, legacy content is analyzed to determine if any content is subject to litigation holds. This is a critical step in ensuring that evidence is not destroyed. This content is typically segregated for preservation purposes or migrated into an ECM system as a safe repository for longer-term retention.

Then, using Auto-Classification capabilities, retention policies are associated to content. This includes understanding the difference between what constitutes a record and what constitutes transitory content. Finally, Discovery's capacity to take action on content is applied to migrate the records that will require ongoing management for retention and disposition into an ECM system. What is left over in the unmanaged data sources is the transitory content that could be deleted according to the organization's policies for the deletion of transitory content.

As a result, content associated with litigation holds is clearly identified and safely preserved, critical company records are placed under secure management and retention control, and finally, the defensible deletion of legacy content saves the organization in storage costs, as well as downstream eDiscovery costs and risk.

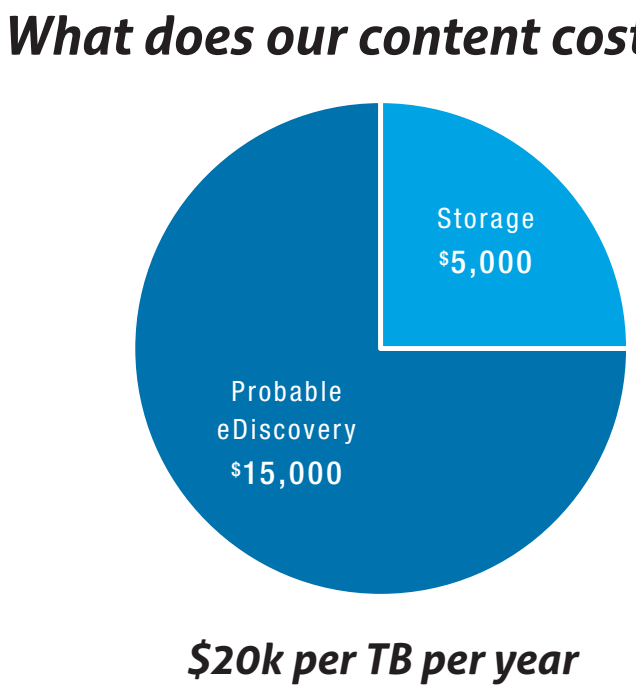


FIGURE 4:

What Does our Content Cost Us?

A recent article, **Hoarders: The Corporate Data Edition** proposed that content is costing us at least $20K per TB per annum. These costs are based on a hard cost of $5K for storage and a probable cost of $15K for eDiscovery based on a probability of review of 0.1%.

The Case for Retention Management

Until recently, most applications that deal with unstructured content in the form of captured documents haven't been built with retention in mind. These applications typically include ECM functionality without integration with Records Management or Business Process Management systems. Where volumes are very high or content is particularly risky, organizations are realizing that defensible deletion remains a critical strategy to reduce application and storage costs, and minimize compliance and litigation risk.

While the long term plan might be to move the application to an ECM platform that supports Records and Retention Management, the shorter-term option is to manage the retention of content by extending policy from the system of record.

Through search, it is possible to know and understand the location, content, and metadata of documents. Based on criteria in the metadata and the capability of Discovery applications, organizations can control the retention of content in these high-volume sources of information. In managing the retention of content, permissions can be changed to ensure documents are not deleted. Discovery allows for the deletion of documents once they have reached the end of their retention period.

Discovery and Legacy Decommissioning

Keeping legacy systems in place incurs extra costs for IT to maintain and support both hardware and software. As the technology ages, the expertise required for using the systems diminishes, and often, only a few people within an organization retain specific knowledge about these systems. These resources are costly and could be better applied to strategic projects.

As organizations implement updated, compliant, and effective solutions for their enterprise needs, legacy systems need to be retired. Many of these legacy systems contain data that is critical in meeting regulatory, litigation hold, or business-critical reference requirements. This problem is particularly acute in highly regulated industries such health sciences, financial services, energy, and government.

Decisions must be made as to what data has to be migrated to a new system for ongoing operation. Often, there is significant data remaining and organizations are caught out by the need to maintain the legacy system (along with associated costs) and the requirement to maintain accessibility to this data.

The solution for decommissioning legacy systems is based on Discovery capabilities that can connect to data sources, extract structured and unstructured information, and move that content into an ECM system where Records Management retention rules can be applied and the content can be "wound down" according to company policy.

Rip and Replace?

Organizations have made large investments to embed business logic into their applications. Unfortunately, Records and Retention Management was not typically part of that investment. Whether it's a loan origination application or a clinical trial application, it's difficult to justify replacing an application simply to provide Records Management.

Today's Strategy

Manage the retention of content using Discovery capabilities. Unified access to the content can be provided for casual users of the content housed in these systems.

Tomorrow's Strategy

Consider standardizing on an EIM platform that combines BPM and ECM to manage the business logic, as well as the retention and disposition of the associated content. By doing so, you can significantly reduce your vendor footprint, as well as maintenance and support costs.

DIGITAL STATES AT RISK! MODERNIZING LEGACY SYSTEMS

Almost 70% of States surveyed observed that more than 40% of their systems are considered Legacy Systems

"A Legacy System is not solely defined by the age of IT systems (e.g. 20 years) as there are many systems that were designed for continued upgrades, but the term also focuses on elements such as "supportability," "risk" and "agility," including the availability of software and hardware support, and the ability to acquire either internal or outsourced staffing, equipment or technical support for the system in question. The term may also describe the system's inability to adequately support "line-of-business" requirements or meet expectations for use of modern technologies, such as workflow, instant messaging (IM), and user interface."

Digital States at Risk!: Modernizing Legacy Systems A NASCIO.org Survey Dec. 2008

A Case for Monitoring

Discovery is increasingly playing an important role in monitoring the organization for compliance issues. With the slate of new regulations being enforced in the wake of the financial crisis (Dodd-Frank), in response to globalization (FCPA and UK Bribery Act), massive data breaches (PCI), and the move to electronic health records (HIPAA), organizations are attempting to improve their ability to respond to issues and proactively prevent compliance issues.

Discovery can assist both in investigations and in proactive monitoring of an organization. It can identify risky data such as personally identifiable information (PII), credit card data (PCI) and personal health information (PHI). It helps to ferret out unmanaged records such as contracts and invoices, and it automates searches that highlight breaches in company policy. Discovery will increasingly play a bigger role in many organizations' compliance programs as the capacity to identify issues before regulators and whistleblowers, as well as the ability to demonstrate diligent monitoring and remediation activities, goes a long way in reducing sanctions and further investigation.

Discovery: Improving Productivity and Monetizing Content

Although there are a variety of Discovery use cases that support tangible benefits for the enterprise like saving money and reducing risk, potentially the most exciting and innovative area for Discovery will be applications that help organizations make money. The first organizations that understand the technology's ability to provide insight and productivity gains will find that they can differentiate themselves from the competition and significantly contribute to the top line.

The Case for Website Search

For most organizations, their web properties are their most important marketing tools, and very often, their most important sales tools. Organizations strive to create a compelling online experience and transform visitors into customers by connecting people to meaningful content. In order to enhance and nurture customer and brand loyalty, people need to find what they are looking for – not just what they are searching on.

A website search that lists a series of potentially relevant documents is simply not enough. Users expect relevant suggestions and to be able to narrow down a search result set to quickly find what they are looking for.

Based on these familiar activities, marketers also have growing requirements for website search. They want to be able control the placement of promoted links and advertisements based on the keywords used to search their sites. And they want to ensure that their websites are initially found, so they endeavor to improve Search Engine Optimization (SEO) through the addition of keywords and topics. Most of all, marketers strive to keep visitors on a website in order to market to and convert them into repeat customers. Marketers are now investing heavily in content marketing, and if their target audience can't find the information they're looking for, the investment in the creation of the content is wasted.

THE NEW REGULATORY REGIME IN THE U.S.

Within the past year, the US DOJ and SEC have announced:

- Several record-setting sentences imposed for FCPA related violations
- An expanded prosecutorial force that includes more DOJ and SEC prosecutors
- An FBI Strike Force utilizing techniques that include undercover FCPA sting operations and wiretaps
- An SEC Cooperation Initiative that encourages individual cooperation as well as corporate cooperation
- The targeting of particular industries (pharmaceutical, life sciences, and telecommunications) and regions (China and emerging markets in Southeast Asia)
- Intent to debar government contractors that are convicted of FCPA-related offenses

Website search has become significantly more sophisticated. By combining the Discovery capabilities of search and content analytics with a web server, it's now possible to enrich content. This includes the analysis and tagging (with metadata) of volumes of uncategorized content to identify relevant and insightful keywords, topics, entities, summaries and sentiments. Organizations can use Discovery technologies to guide searches, providing filters on the enriched content to include people, places, and things, allowing visitors to quickly narrow search results based on what they are truly looking for.

Discovery technologies are being used to improve SEO through the addition of keywords, topics, entities, summaries, and sentiments into metadata on web pages. This makes web pages more authoritative and optimized for search engine crawlers. Discovery technologies can also be used to manage promotions and ad placements through the use of search widgets that position content, links, and advertisements with relevant content at a time when the visitors' interests are piqued and they're most likely to buy or consume. This reduces bounce rates and helps to convert visitors into customers.

Discovery and Unified Information Access

Enterprise content is fragmented. It lives in multiple silos that reside in systems with duplicated functionality, un-integrated data, and limited end-user access. This is due in part to mergers and acquisitions, basic geography, and departmental solutions. At the same time, IT has become hyper-sensitive about the satisfaction of end users, knowing that they are now competing with the "Consumerization of IT" and consumer applications. Many organizations that previously would have had no issue with users accessing three or four different applications to complete a business process are now re-evaluating ways to increase both end-user satisfaction and productivity.

To help provide a cohesive end-user experience and consolidate information, Unified Information Access aggregates and enriches content from disparate sources. More robust than Enterprise Search, Unified Information Access provides end users with the ability to copy and move content from one source to another. As organizations rationalize their IT systems, reduce their vendor footprints, and unify access to information, they can continue to extract value from aging systems to improve end-user experience and productivity.

Why Big Data Needs Discovery

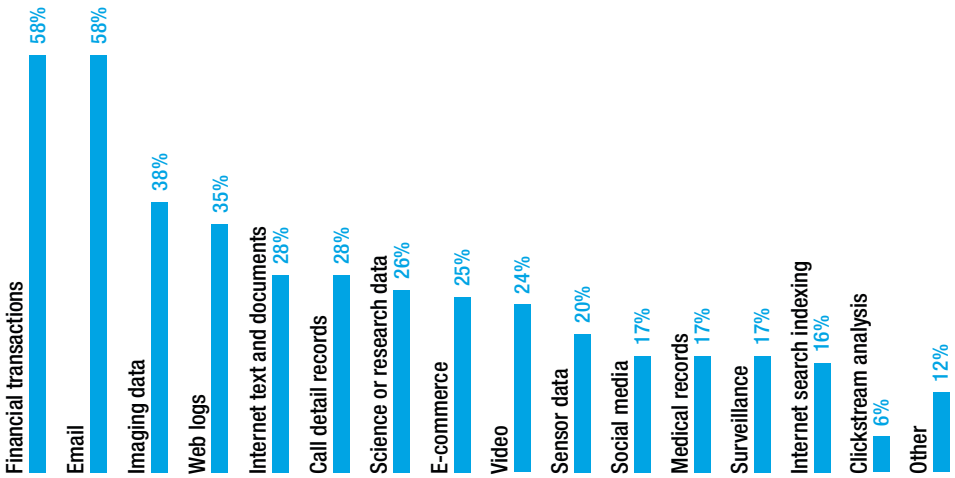
Big Data has typically included varied and large sets of text-based data like clickstreams, ecommerce transactions, social media feeds, RFID data, GPS data, and more. Big Data analysis has focused primarily on finding insight in these new and emerging types of data and answering questions that were previously considered beyond our reach. Due to the success of Big Data projects, whether we like it or not, companies and governments are expanding the definition of Big Data to include the vast stores of unstructured content that they have accumulated.

Is it Big Data or Big Content?

As found in the recent **Information Week 2012 Big Data Survey**, email was tied for first position in applications that drive Big Data needs. A closer look finds that three of the top five concerns around data inside organizations are concerned with email, imaging data, and Internet-based text and documents – largely what we consider to be unstructured content. These content sources have been identified as a primary concern because they represent a significant risk, from both a compliance and a litigation standpoint. Beyond these concerns, organizations are also realizing that these content types contain insight into the business and can provide transparency around activities like fraud, corruption, and data breaches.

Big Data Drivers

Which applications are driving big data needs at your organization?



NOTE: Multiple responses allowed.

FIGURE 5:

Big Data Confusion

Source: Information Week 2012 Big Data Survey of 231 business technology professionals, December 2011

These content types pose the following challenges, ones that Big Data projects have not had to contend with until now:

- **Processing:** Involves extracting full text and metadata from documents as well as recursively indexing compound documents and archives.
- **Normalizing:** Unifying and mapping metadata to common definitions.
- **Entity Extraction:** Refers to extracting and normalizing people, places, and things in any document for the purpose of refining searches.
- **Enrichment:** Adding additional semantic metadata like phrases, sentiment, and relationships as part of the metadata that describes the document.
- **Security:** Typically security is monolithic in Big Data, but not so in email and productivity-based documents.

As the definition for Big Data continues to change, the types of data, use cases, and the potential risks and benefits of Big Data will expand. Discovery capabilities, when combined with Big Data platforms like Hadoop® will become more common and an eventual baseline requirement for organizations that are looking to reduce risk and derive value from the huge amount of unstructured information being created inside the firewall every day.

Is Discovery Enterprise Search?

The answer to this question is an emphatic "No".

Enterprise Search is a term that is used so often that it's difficult to pin down a clear definition. It's in this absence of a definition that Enterprise Search becomes all things to all people. Unfortunately, Enterprise Search has garnered a lot of hype, a lot of expectations, and an awful lot of disappointing or failed deployments. This can be explained in part by the "Google®" phenomenon, where end users express their search requirements as "we want something like Google."

As Leslie Owens from Forrester so eloquently points out, simple, single-box search is not an effective way of getting information to users when they're searching for content within their organization. There is a reason why Amazon® knows what book we want before we do. They have used Discovery (and Big Data) technology to create an application that solves a specific business problem.

The idea that Discovery capabilities need to be tuned to solve specific business problems lies at the heart of a good EIM strategy. Rather than deploying generic search capabilities, organizations that deploy Discovery applications to solve targeted and specific problems will be more successful in their deployments and will be able to more easily correlate their return on investment.

"How can Amazon.com monitor my customer data so closely that it knows what book I want next, but after five years of daily use, my enterprise search engine doesn't get that I work in HR in the Chicago office?"

Semantic Technology in the Enterprise – April 2012 Forrester Research, Inc., Blog – Leslie Owens

Discovery Technology

The Platform

Discovery in Enterprise Information Management requires a platform to discover, analyze, and act on information sources throughout the organization. The platform provides a common set of features and functions that can be leveraged by multiple Discovery applications. A platform reduces the costs and complexity associated to having multiple indexes, multiple connectors, and separate hardware for the individual Discovery applications found in most organizations. A typical organization will have one or more indices for eDiscovery applications, one index for enterprise search, and multiple indices embedded in ECM applications.

The basic requirements for a platform include:

- **A unified index of enterprise information** to make it discoverable and to ensure analysis and decisions are based on precise and complete information.
- **Integration services** to provide authorized access to enterprise repositories, and the capacity to index and act on that content.
- **Indexing and content enrichment** must support multiple use cases and support the requirements for speed, completeness, and analysis. Indexing can vary from light-weight passes to extract basic file metadata for a quick understanding of the composition of large volumes of information to deep indexing, during which content is enriched with semantic analysis to identify facts and relationships, and detect sentiment, patterns, and trends.
- **Capabilities to facilitate the rapid development and deployment of applications and utilities** for Enterprise Information Management. Because many different enterprise information applications share common requirements, a rich set of APIs and a set of reusable widgets can be used to develop applications for specific use cases and user roles.

Discovery-based Applications

The role of the Discovery-based application is to address specific business problems with the following technology:

- The right set of features and capabilities derived from search, connectors, integration, content analytics, and visualization.
- A User Interface that has been created specifically for the personas that will use the application.
- Embedded processes to address business problems.
- Appropriate permissions, security, and roles to support the different personas that interact with the business process.

Discovery-based applications are already replacing generic search capabilities in most organizations. eDiscovery and Knowledge Management applications are some of the first applications that have been adopted. Unfortunately, most of these are stand-alone and require their own infrastructure, index, and support. The true value of both Discovery-based applications and a Discovery platform lies in the ability to solve specific business problems on a common infrastructure to control costs and support requirements.

The Future of Discovery

For all the current use cases and existing technologies, Discovery as a category is truly in its nascent stage. It has only been in the past few years that we've seen the combination of search, content analytics, and two-way connectors that make Discovery possible. As a core set of capabilities that can be extended into applications to solve real-world problems, Discovery is an area that will see significant investment, adoption, and innovation.

Enterprise Information Management practices and technology will continue to grow in popularity and importance for the next decade as organizations look to reduce cost and risk, and extract value from the untapped 90% of information that is unstructured inside organizations. Discovery technologies are a critical component EIM, with the combined ability to perform actions on massive amounts of enterprise information, aggregate and integrate information across enterprise technologies, and provide insight into volumes of content.

As EIM practices and technology mature, Discovery will continue to play a dual role. Common capabilities such as connectors, processing, indexing, content analytics, visualization, and the ability to act on content will be ubiquitously available for EIM technologies. At the same time, because of this ubiquitous availability, Discovery-based applications and features will be added to EIM Suites of software to address specific use cases and challenges. These applications will continue to impact both the top and bottom lines.

Discovery in the Cloud

In the rush to save money by moving applications and content to the Cloud, organizations have forgotten the lessons of the past (if they ever learned them). In the same way organizations created archives with no thought to how content could be deleted, or in the same way transactional imaging and BPM systems where developed with no thought to Records Management, organizations are doing it all over again in the Cloud. The first causalities are email and Microsoft Office documents that have been moved to the Cloud with no regard for retention, disposition, and eDiscovery. Other types of content will follow, including critical records like invoices and HR records. Discovery technology offers hope that technology might help save us from the sins of the past. The future will bring the ability to extend Discovery beyond the bounds of the firewall to discover, analyze, and act on content in Cloud applications, extending centralized policy, access, and eDiscovery to content sources stored in the Cloud.

Discovery technologies will continue to be a critical arrow in the quiver of EIM practitioners looking to unleash the power of information in order protect their organization from risk and cost, make end-users more productive, and improve decision-making by unlocking the value found in enterprise information.

For more information, visit:
http://www.opentext.com/2/global/products/eim-data-management.htm

Appendix 1: Where to Start

This white paper has outlined a large variety of potential use cases for Discovery technologies. This could represent a challenge when it comes to determining where and how to start if there are multiple use cases that are attractive to the CLO, CMO, and CIO in your organization.

For most organizations, the decisions can be broken down into a few, more simple steps:

1. **Document and assess the opportunities.** What will have the greatest impact on the organization: a top-line or bottom-line focused project? Does your organization's litigation profile include a large number of high profile cases? Is your marketing organization having difficulty attracting and maintaining visitors? Are poor decisions being made due to an inability to locate and analyze information?
2. **Perform a risk assessment.** Where does the largest cost and risk exist? This is typically the starting point, especially for information governance initiatives. Where is the largest potential for lost opportunity? This is the usual starting point for top-line projects, where time-to-market can make the difference in giving you competitive advantage.
3. **Create a business case.** What will the pay-back period be for the project? In today's economic climate, projects don't move forward without a scrutinized business case. Create a business case based on cautious and optimistic impacts on the business.
4. **Engage with and gain the support of the right executive.** Work with them to validate the business case and risk assessment.
5. **Develop a roadmap.** The nature of Discovery applications prescribes overlapping functionality applied to different use cases. Developing a roadmap of supported use cases will serve to manage expectations and also clarify whether a best-of-breed or a platform strategy should be used. Consider the impact a best-of-breed versus a platform strategy will have on long term maintenance, support, and infrastructure costs when developing the business case.

CREATE A BUSINESS CASE

OpenText has a team of Value Engineers that can work with you in developing a business case based on your existing processes, industry best practices and the positive impact best practices and technology can have on your business.

This service has been invaluable for many organizations as they rationalize changing technology and business process. Value Engineering and Business Case development is a complementary service for well-qualified customers.

For additional information, please contact your Account Executive.

OpenText Locations

AMERICAS

Canada:
- Waterloo, ON
- Richmond Hill, ON
- Ottawa, ON
- Montreal, QC
- Peterborough, ON
- Kingston, ON
- Calgary, AB

U.S.:
- Tinton Fallls, NJ
- Austin, TX
- Tucson, AZ
- Norcross, GA
- Irvine, CA
- Tallahassee, FL
- Chicago, IL
- New York, NY
- Rockville, MD
- Columbus, OH
- Burlington, MA
- Alameda, CA
- Bellevue, WA
- Tampa, FL
- Reston, VA
- Arlington, VA
- Rochester, NY
- San Antonio, TX

Brazil:
- Sao Paulo

EMEA

Germany:
- Munich (Grassbrunn)
- Konstanz
- Oldenburg
- Düsseldorf
- Kempten
- Hamburg
- Bad Homburg v.d.Höhe

Great Britain:
- Reading
- Wimbledon
- London
- St Albans

France:
- Paris

Sweden:
- Stockholm
- Gothenburg

Switzerland:
- Baden

The Netherlands:
- Hoofddorp

Ireland:
- Clonakilty

Spain:
- Madrid

Austria:
- Klagenfurt
- Wien

Czech Republic:
- Prague

Italy:
- Rome

Finland:
- Espoo

South Africa:
- Johannesburg

U.A.E.:
- Dubai

ASIA, PACIFIC

India:
- Hyderabad

Australia:
- Sydney
- Melbourne
- Canberra

Japan:
- Tokyo
- Osaka

Singapore:
- Singapore

Hong Kong:
- Hong Kong

Korea:
- Seoul

New Zealand:
- Auckland

Contact OpenText

Sales

North America: +1-800-499-6544
International: +800-4996-5440
E-mail: Global Sales

Partners

North America: +1-519-888-7111
International: +44 (0) 1189 848 000
E-mail: Global Partner Program

Media Relations

North America: +1-519-888-7111
International: +44 (0) 1189 848 000
E-mail: Public Relations