AIIM Market Intelligence

Delivering the priorities and opinions of AIIM's 65,000 community

aiim

Industry Watch

Content Analytics

- research tools for unstructured content and rich media

Underwritten in part by:

IBM®

Send to a friend ▸

aiim®
Find, Control, and Optimize Your Information

aiim.org | 301.587.8202

About the Research

As the non-profit association dedicated to nurturing, growing and supporting the ECM (Enterprise Content Management) community, AIIM is proud to provide this research at no charge. In this way the education, thought leadership and direction provided by our work can be leveraged by the entire community.

We would like this research to be as widely distributed as possible. Feel free to use this research in presentations and publications with the attribution – "© AIIM 2010, www.aiim.org"

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 partially made possible by our underwriting companies, without whom we would have to return to a paid subscription model. For that, we hope you will join us in thanking our underwriters, who are:

IBM®

IBM
3565 Harbor Blvd,
Costa Mesa, CA 92626
Phone: 800-345-3638
Email: arichar@us.ibm.com
www.ibm.com/software/ecm/compliance

Process Used, Survey Demographics and Terminology

While we appreciate the support of these sponsors, we also greatly value our objectivity and independence as a non-profit industry association. The results of the survey and the market commentary made in this report are independent of any bias from the vendor community.

The survey was taken by 527 individual members of the AIIM community between February 9th and February 26th, 2010, using a Web-based tool. Invitations to take the survey were sent via e-mail to a selection of the 65,000 AIIM community members.

Survey population demographics can be found in Appendix A. Graphs throughout most of the report exclude responses from suppliers of ECM products or services.

About AIIM

AIIM (www.aiim.org) is the community that provides education, research, and best practices to help organizations find, control and optimize their information. For more than 60 years, AIIM has been the leading non-profit organization focused on helping users understand the challenges associated with managing documents, content, records and business processes. Today, AIIM is international in scope, independent and implementation-focused, acting as the intermediary between ECM (Enterprise Content Management) users, vendors and the channel.

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 been involved in their evolution from technical solution through business process optimization to the current enterprise-wide adoption. Doug has also worked closely with other enterprise-level IT systems such as ERP, BI and CRM. Doug has an MSc in Communications Engineering and is an MIET.

aiim®
Find, Control, and Optimize
Your Information

© 2010
AIIM - Find, Control, and Optimize Your Information
1100 Wayne Avenue, Suite 1100, Silver Spring, MD 20910
Phone: 301.587.8202
www.aiim.org

Table of Contents

About the Research:

Introduction:

Search and Research:

Levels of Adoption:

Business Drivers for Content Analytics:

Rich Media and Digital Asset Management:

Business Drivers for Rich Media Analytics:

Projected Spend:

Conclusion:

Appendix 1: Survey Demographics:

Appendix 2: Glossary:

Underwritten in part by:

Introduction

The term “Content Analytics” has been coined to cover a range of search and reporting technologies which can provide similar levels of business intelligence and strategic value across unstructured data to that conventionally associated with structured data reporting. Sophisticated content search across text and rich media file-types, combined with trend analysis, content assessment and behavioral reporting, has created the opportunity to track and manage unstructured content and digital assets with the same levels of capability as BI reporting of structured content - with associated business benefits of content optimization, asset management, pattern detection and compliance monitoring.

As is usual with new technology, levels of awareness of both the technology and the naming terminology vary considerably. We have measured this in the report and provided a glossary of the main terms in Appendix 2. We have also explored the user-perceived limitations of conventional search, and the potential savings that could be achieved by application of content analytics to a number of business scenarios such as fraud detection, asset protection, healthcare research and market monitoring.

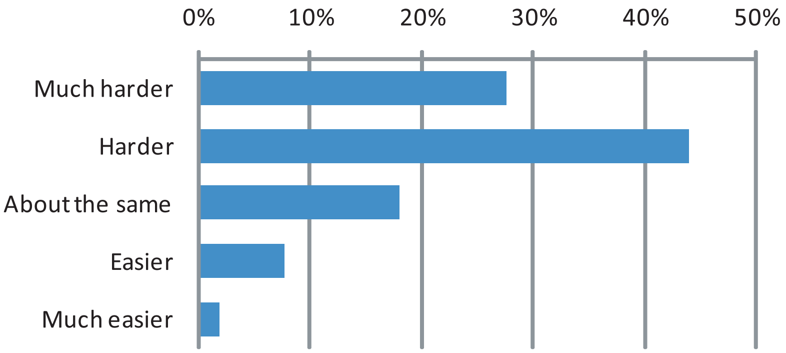
Key Findings

- For 72% of respondents, it’s harder to find information owned by their organization than information not owned by them – i.e, on the Web.
- Of the 47% who find they frequently need to use Advanced Search options, more than half would like something more effective.
- 70% would find advanced analytic functions “extremely useful” or “very useful.”
- For most content types, our respondent’s ability to “research” is 3-6 times less than their ability to “search”, particularly for rich media files, but also office documents and emails.
- E-discovery, Digital Asset Management (DAM), Web Analytics and De-duplication are the better known technologies compared to Sentiment Analysis, Copyright Detection and Digital Forensics.
- There are strong plans to adopt DAM, Faceted Search, E-discovery and Content Assessment in the next 18 months.
- The biggest obstacle faced regarding content decommissioning is, “not clear which content is valuable and which is not”. There is also considerable “Fear of the compliance and regulatory impact of deleting information.”
- Only 15% have an automatic way of finding and deleting duplicates in their content stores, with just 8% able to analyze them automatically for relevancy and to delete irrelevant content.
- 50% would find it of “high” or “very high” commercial value to be able to link a customer/citizen/staff-member search across structured (database) data & unstructured documents & case notes. 44% would find it of “high” or “very high” commercial value to be able to automatically redact (blank out) sensitive information across forms, etc.
- 81% of those who have digital assets to manage are not using a dedicated Digital Asset Management system but 14% of our respondents are planning to implement one in the next 18 months. 48% store digital assets and rich media on ad hoc file shares.
- 59% would find it of “high” or “very high” commercial value if they could use a faceted search across multiple metadata tags to cross-reference categories of rich media. 50% would find it of “high” or “very high” commercial value if they could detect unauthorized use of their assets across the web.
- Net spending on Enterprise Search, Digital Asset Management and Content Analytics is set for a considerable increase in the next 12 months.

Search and Research

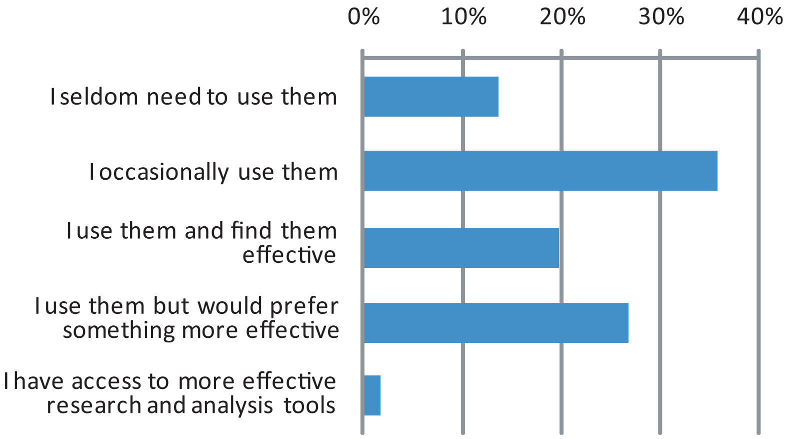
For most people, “search” is synonymous with a Google-style presentation, with results based on relevancy of matches across the web. There is no doubt that this can often be startlingly useful - although it can also be somewhat frustrating at times. In fact, many information workers would be only too glad to have the same search capabilities across their internal information repositories as they have on the web.

Figure 1: How easy is it to search information and documents held on your own internal systems compared to the Web?



However, as we move from the more common requirement of document “search” to the somewhat more demanding needs of “research”, straightforward search engine mechanisms are unable to provide the pattern matching, trend plotting and semantic analysis that may be required.

Figure 2: For your information research tasks, how effective do you find the “Advanced Search” options in standard search engines? (N=484, Non-Trade)



Of our survey sample, only 47% are regular users of the so called “advanced” search functions, and more than half of those would prefer something more effective.

Taking the idea of research versus search, we asked respondents how their ability compared across different types of content. We can see that rich media files such as audio, video and graphics lack tools for both types of activity. The differences are more clearly shown in Figure 4 where we show the ratio of search capability to research capability.

Figure 3: How would you rate your ability to research across the following content types? (N=450, Non-Trade, normalized for N/A)

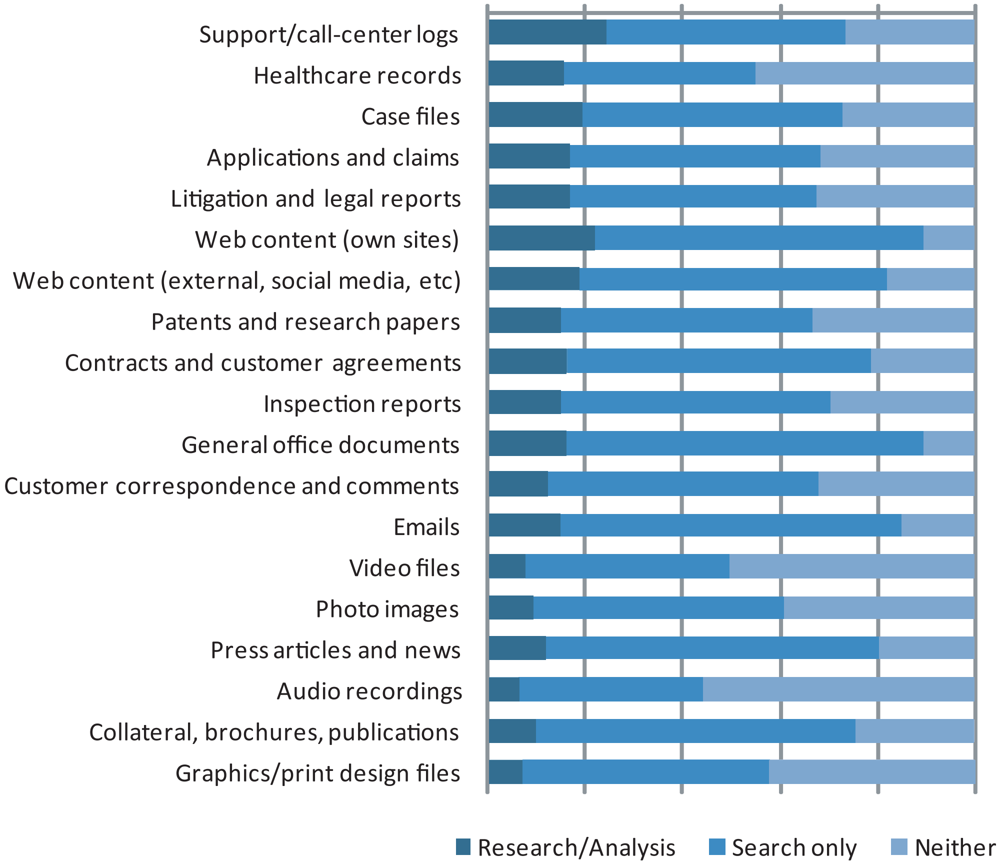
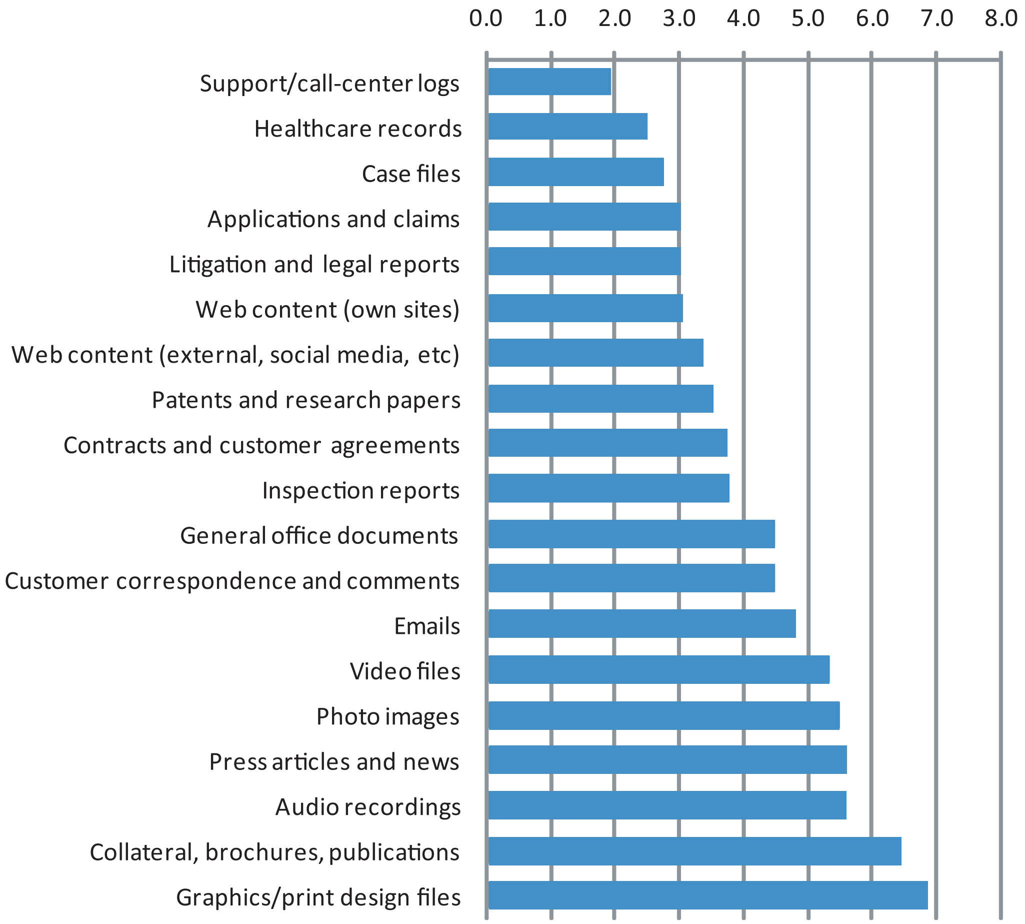


Figure 4: How would you rate your ability to research across the following content types – ratio of "search" to "research" (higher number indicates poor ability to research compared to search)



Healthcare records are an interesting example here in that Figure 3 indicates a lacking of basic search, but Figure 4 shows that some of our respondents have some useful research tools, and similarly with general case files which can be quite complex to analyze, but yield useful results. Graphics and print design files, on the other hand are fairly easy to find, but few have the ability to analyze their content - for example, seeking out obsolete logos.

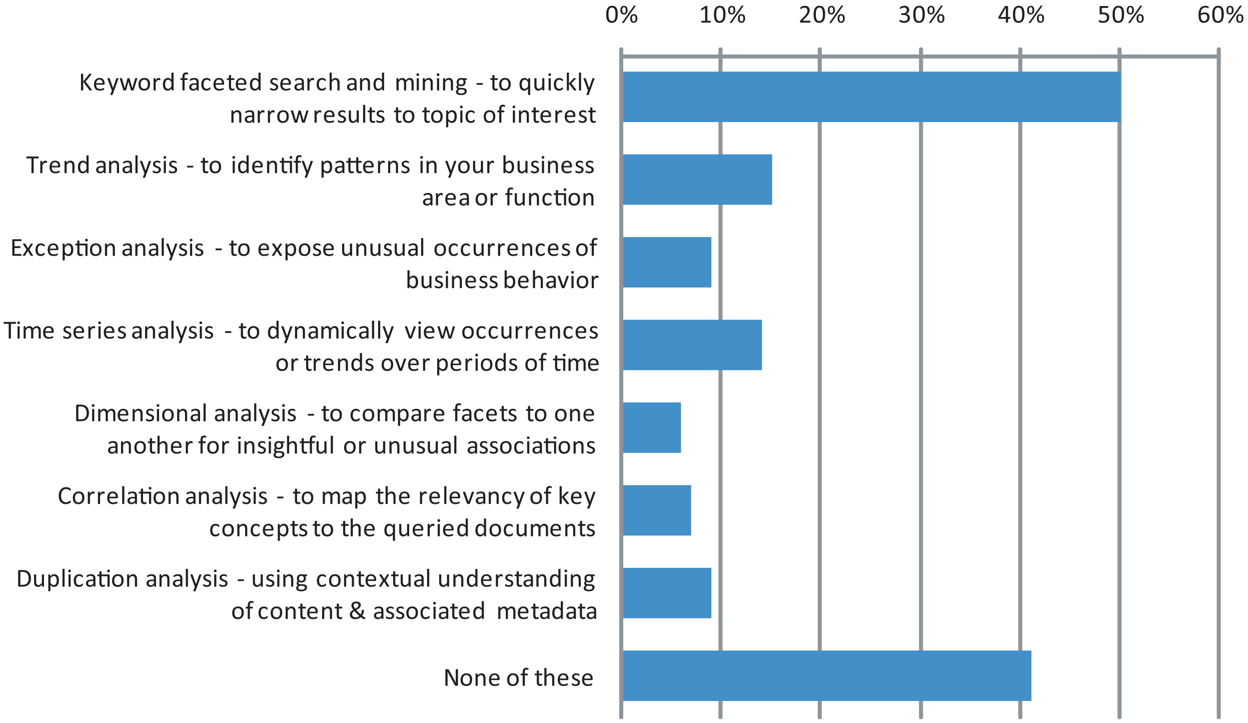
Looking in Figure 5 at the types of analysis capability that researchers might wish to access, we see that users would definitely like to better exploit the keyword metadata, and use faceted drill-down to quickly refine results (although there may be some confusion here with conventional keyword search).

aiim

Industry Watch

Content Analytics - research tools for unstructured content & rich media

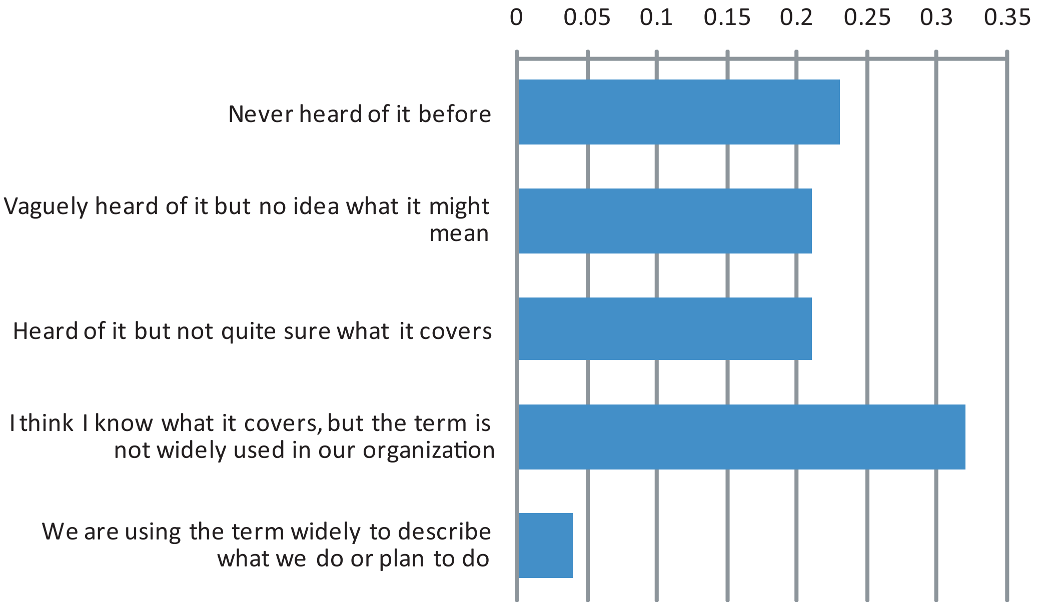
Figure 5: Do you have access to any of the following analysis capabilities for research of unstructured/document/media content? (Tick all that apply)?



Terminology

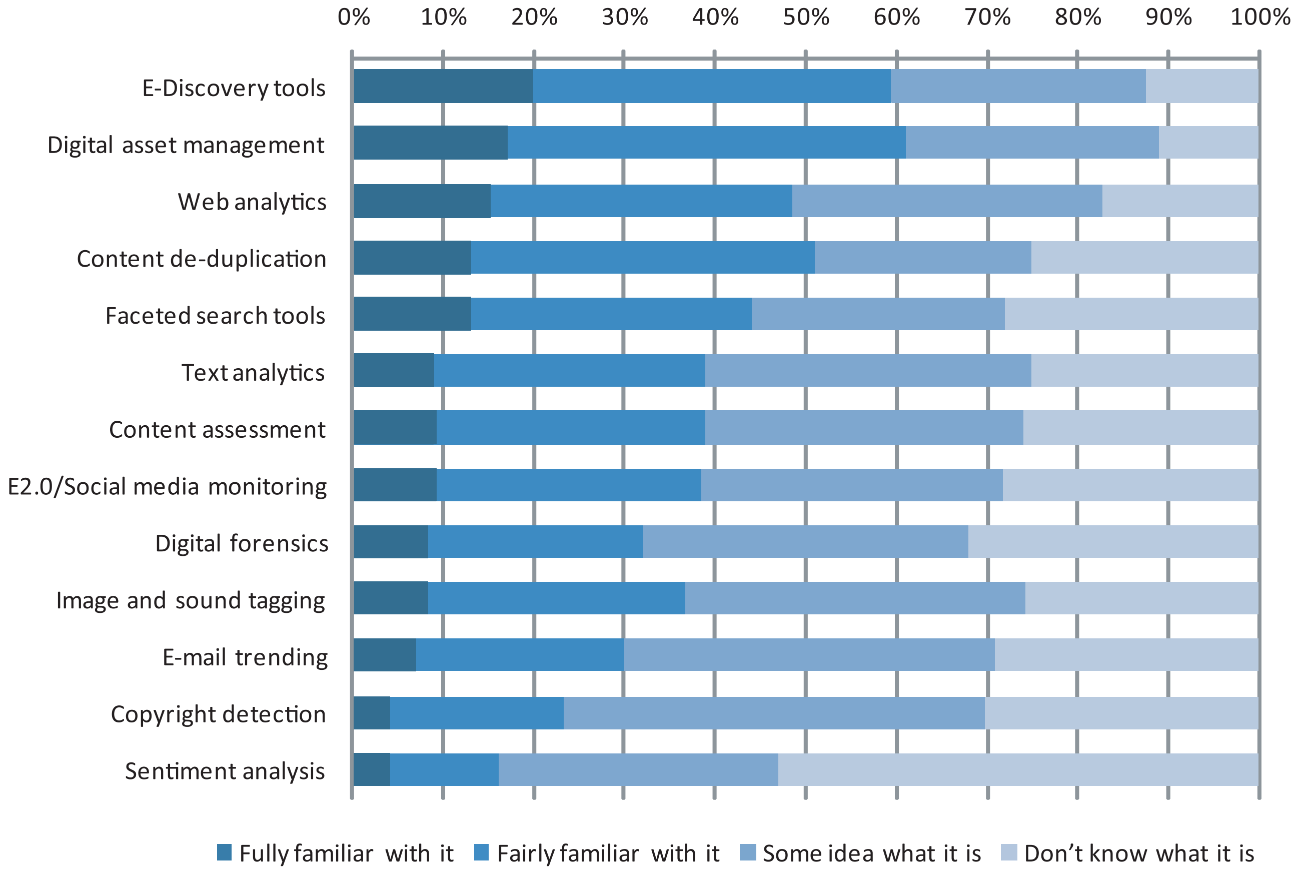
As mentioned in the introduction, the term "Content Analytics" and many of its constituent parts are relatively new. This makes it quite difficult to measure the existing installed base.

Figure 6: Prior to this survey, how familiar were you with the term "Content Analytics"? (N=454, non-trade/non-consultant)



In Figure 7, we asked about each individual technology, but without defining or describing what they are. Readers are referred to the Glossary in Appendix 2 for a description of each.

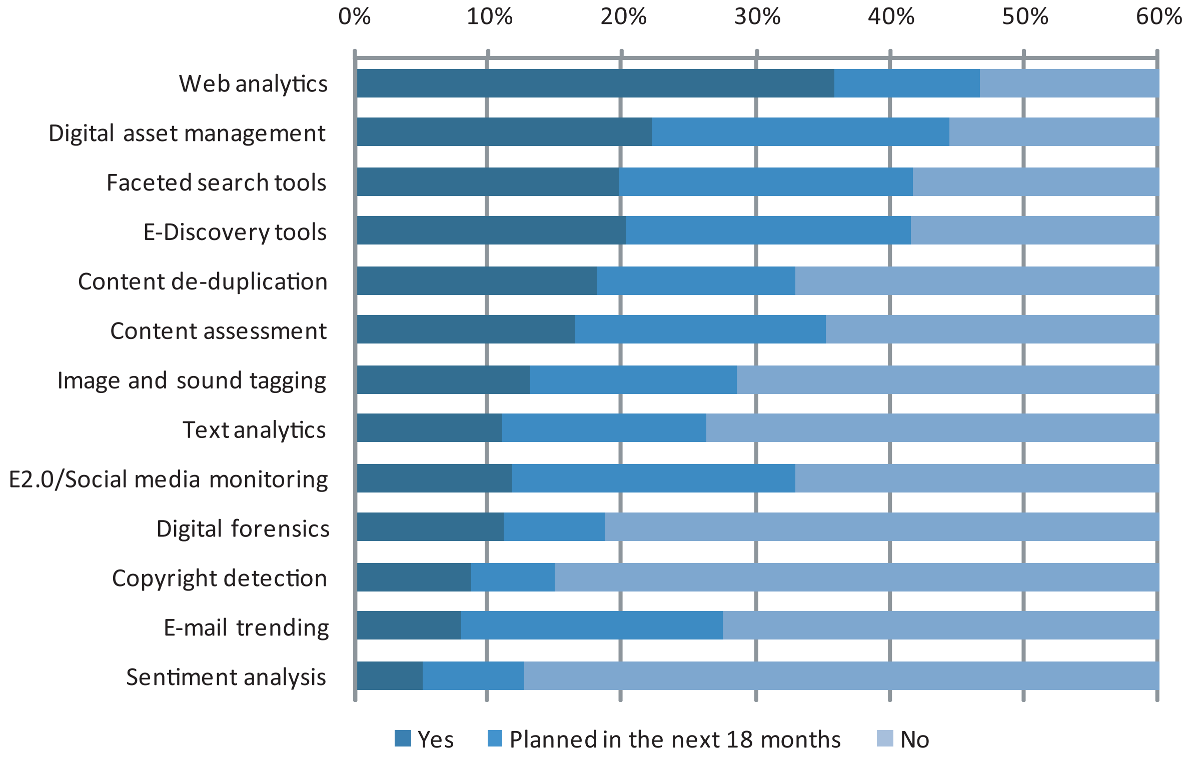
Figure 7: How aware are you of each of the following technologies?
(N=446, non-trade)



Levels of Adoption

E-discovery and Digital Asset Management are generally well known, followed by Web Analytics, which is the most widely adopted *(see Figure 8)*. Interestingly, Enterprise 2.0 and Social Media monitoring is an area of strong interest, but some would say this is a subset of more generic Sentiment Analysis. It is possible that “monitoring” is taken to mean a human overview rather than automatic monitoring and alerting.

Figure 8: Are you using any of the following technologies in your organizational unit? (N=445, non-trade)

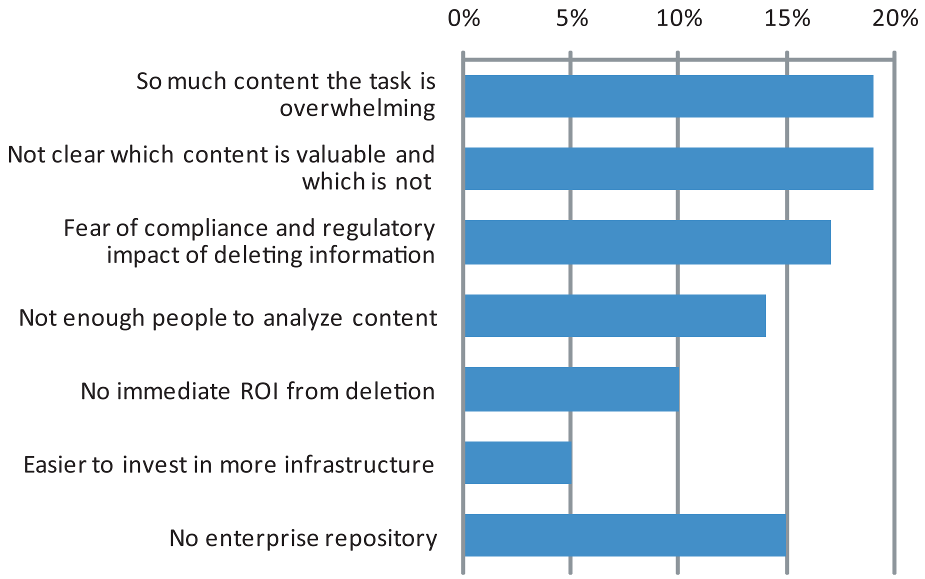


Within our survey sample, there are strong growth indicators for a number of areas including Digital Asset Management, Content Assessment and Content De-duplication.

Content Decommissioning

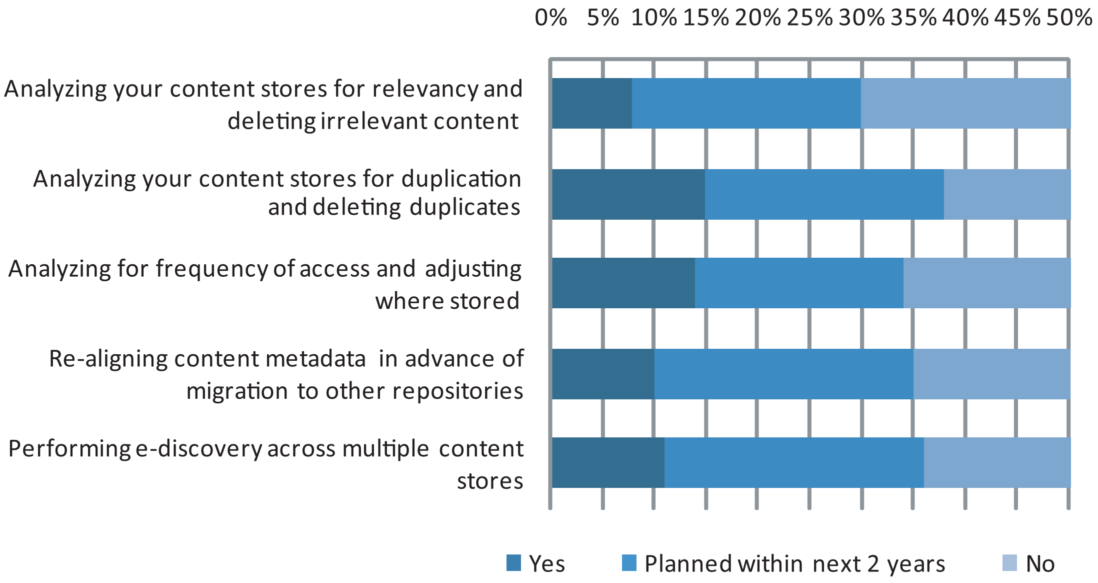
As more users approach their second or even third ECM system, whilst others need to aggregate content from one system to another following a merger or takeover, there is a strong interest in "clearing out" obsolete or duplicate content prior to migrating it to a new system.

Figure 9: Which is the biggest obstacle you face regarding content decommissioning? (N=421, non-trade)



Aside from the sheer enormity of the task, a lack of clarity in what content is valuable is the main obstacle, along with the fear of getting it wrong. Obviously, automated tools would simplify the task, but they do need to be sophisticated in the way they work. A number of respondents seem to feel that there is no immediate ROI from deletion, perhaps having failed to compute the total cost of ownership from maintaining little-used or duplicate content on their highest performing storage platforms, or more likely they feel that the cost of manual assessment and de-duplication would be huge.

Figure 10: Do you have any automatic ways of achieving the following? (N=422, non-trade)



Current adoption of this type of tool is very low, although around 25% indicate an intent to acquire them in the next 2 years.

Business Drivers for Content Analytics

In view of the low numbers of users as yet for content analytics, it is difficult to assess measured returns. However, we did present a number of possible use cases for the techniques, and measured the potential returns that would accrue.

We asked:

"How useful would it be if you could use questions like this across your content?

- What is most frequently occurring?
- Why is there a higher occurrence between these dates?
- What are the trends and why are they occurring?
- Is this a normal or an unusual result?
- What types of people, living where, are saying this and why?
- Can I find an image that matches this one?
- Has this been mentioned before in this context?
- How can I know what I don't know?"

32% of respondents would consider it to be "extremely useful" with a further 38% considering it "very useful" to be able to query unstructured content in this way.

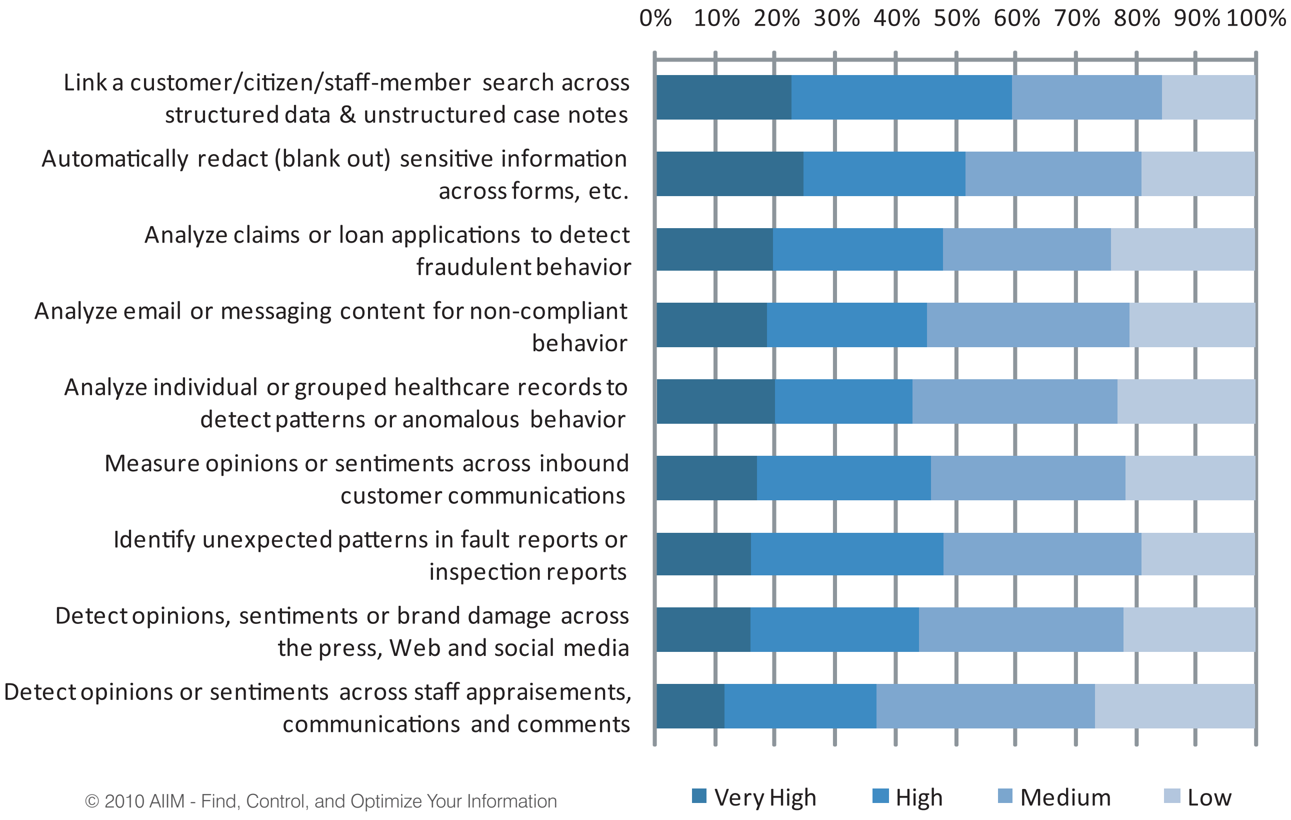
Extending that question to specifically cover federated search, we asked:

"How useful would it be if you could extend your information gathering to include private content hosted by another agency, partner or service as well as content hosted within your own firewall (federated search)?"

19% considered this to be "extremely useful" with a further 35% considering it "very useful".

To quantify things further we asked about specific commercial benefits.

Figure 11: What is or would be the commercial benefit to you if you could:
(N=417, non-trade, normalized)



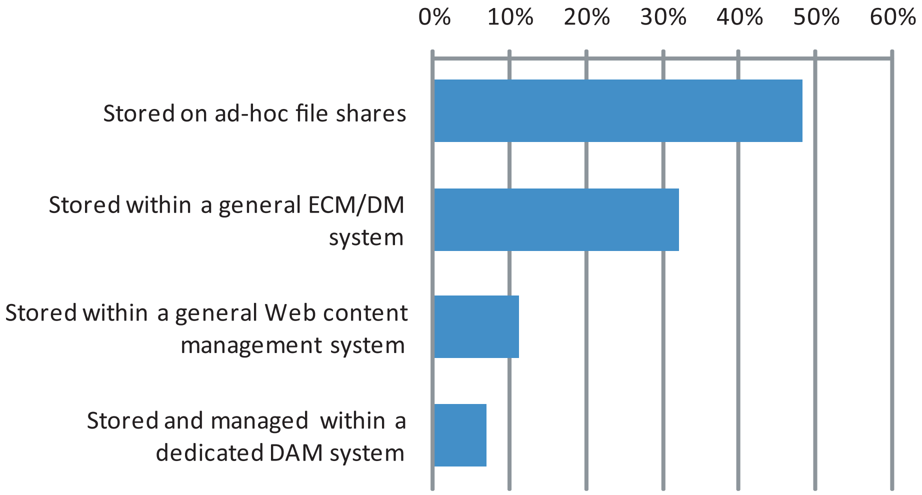
Obviously, some of these applications are specific to certain industry sectors, but we have normalized against this. Highest rewards would come from the ability to link unstructured documents and case notes to structured data held on a finance, HR or CRM system. In a similar area, automated redaction of sensitive information has application in recruitment and in Freedom Of Information requests. Analyzing claims or loan applications for fraudulent behaviour would also likely produce some very strong hard-dollar benefits.

Rich Media and Digital Asset Management

Rich media – pictures, graphics, sound and video files – present particular problems for content management due to their file size, and the inability to search by textual elements within the file content. In addition, if they are also considered to be digital assets, they will come with associated rights-to-use, or rights-of-ownership, which must also be duly managed. Rich media files may also need complex coders, decoders or format transformers to play them out.

70% of our survey respondents "store and manage" photo images and graphics files, 52% have video files and 42% audio files. Management mechanisms vary widely. 48% store them on file shares, 32% in a general ECM/DM system, and just 7% in a dedicated Digital Asset Management (DAM) system.

Figure 12: Which of the following would best describe the way in which you manage your digital assets? (N=252 rich media users)



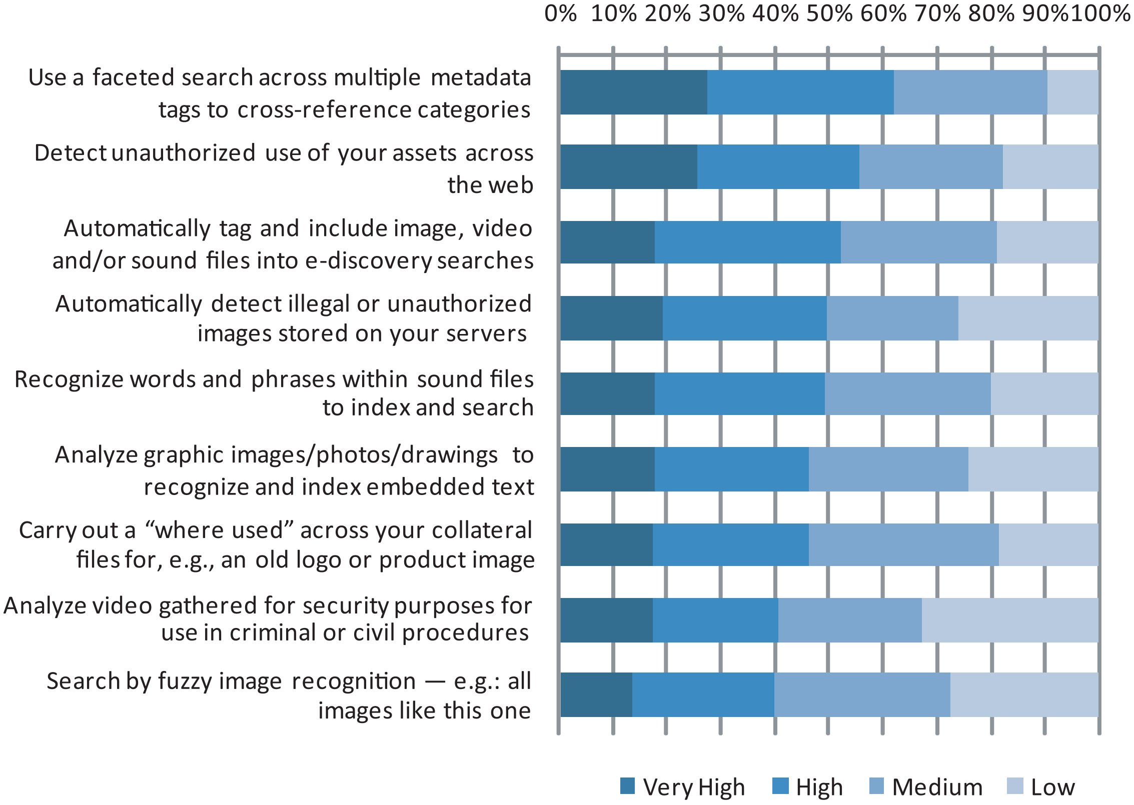
Digital Rights Management (DRM) takes things one step further as regards managing permitted use and providing copy protection. The majority of our respondents felt that they had content that could be protected in this way, but only 10% are using DRM – albeit with a further 14% saying they plan to do so in the next 18 months.

Business Drivers for Rich Media Analytics

Faceted search has a particular benefit for rich media search due to the heavy reliance on metadata tags as described above, for example finding photos suitable for a holiday brochure by family/couple, beach/villa/pool, Florida/Caribbean, etc. Automated tagging can go along with this to derive the original metadata from the image itself.

Unauthorized use of assets may not just be copyright violation, but could also be unauthorized use of corporate logos. It also extends to licensing and merchandising, either as regards the licensee needing to manage who can do what, or the licensor needing to stay within the restrictions and time frame of their license.

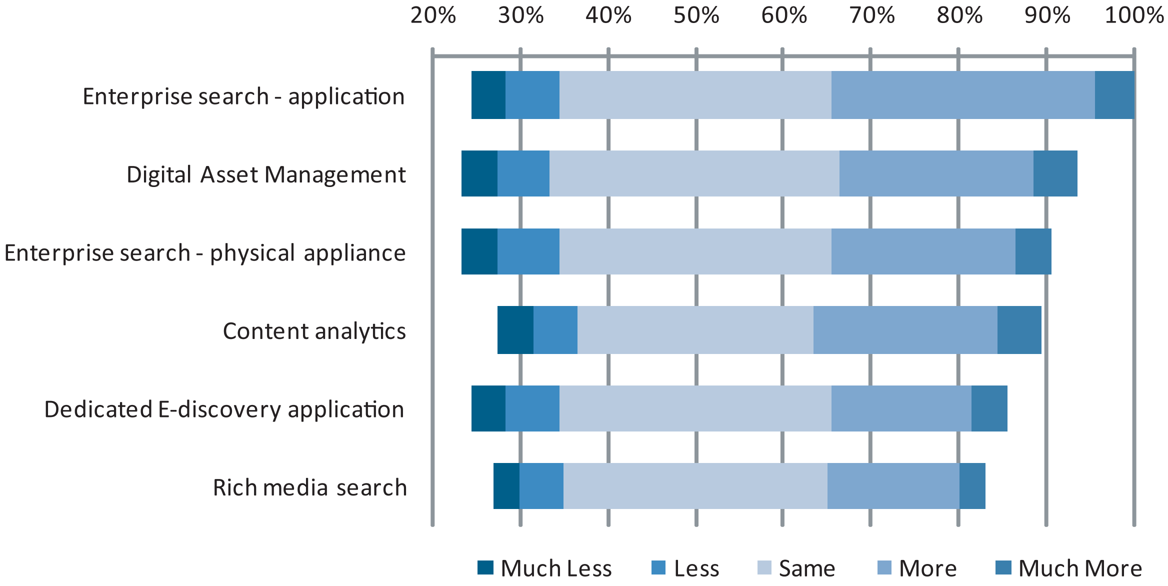
Figure 13: What is or would be the commercial benefit to you if you could:
(N=252 rich media users)



Projected Spend

The most popular spend indicated by our respondents is Enterprise Search, whether as an application or module, or as a physical appliance. However, there is strong interest in Digital Asset Management, and Content Analytics looks to be at the start of a strong expansion, along with rich media search for those who need it.

Figure 14: What are your spending plans for the following product areas in the next 12 months compared to the last 12 months?
(N=398, non-trade. Shorter lines indicate "We don't spend anything on this2)



Conclusion

The benefits of investment in Finance and ERP systems have only come to the fore with the increasing power of Business Intelligence (BI) reporting tools and the insight they provide for business managers. In the same way, the benefits of Content Management systems can be much more heavily leveraged by the use of Content Analytics tools.

They can provide true research tools for those tasked with applications as diverse as fraud prevention and medical research. Equally diverse are digital forensics for crime detection and sentiment analytics to measure the “feelings of the crowd” towards a brand or product. Some of these tools analyze text usage, others seek patterns in video or sound, whilst others link together structured databases of transactional detail with unstructured case files of documents and forms.

What they all have in common is the ability to answer complex queries and expose difficult to find content by sophisticated analysis techniques that draw heavily from conventional BI reporting tools. As we have seen in this report, it is early days for these products, but the potential business benefits are wide and varied.

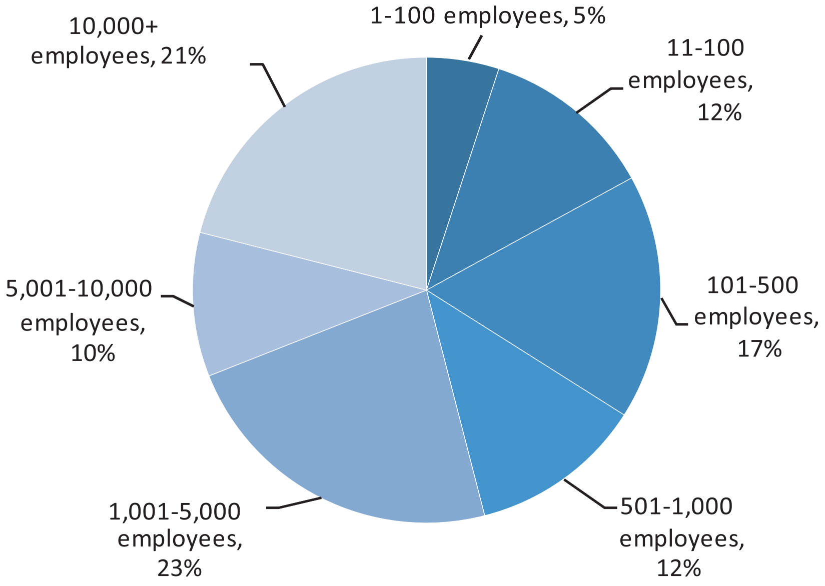
Appendix 1: Survey Demographics

Survey Background

The survey was taken by 527 individual members of the AIIM community between February 9th and February 26th, 2010 using a Web-based tool. Invitations to take the survey were sent via email to a selection of the 65,000 AIIM community members

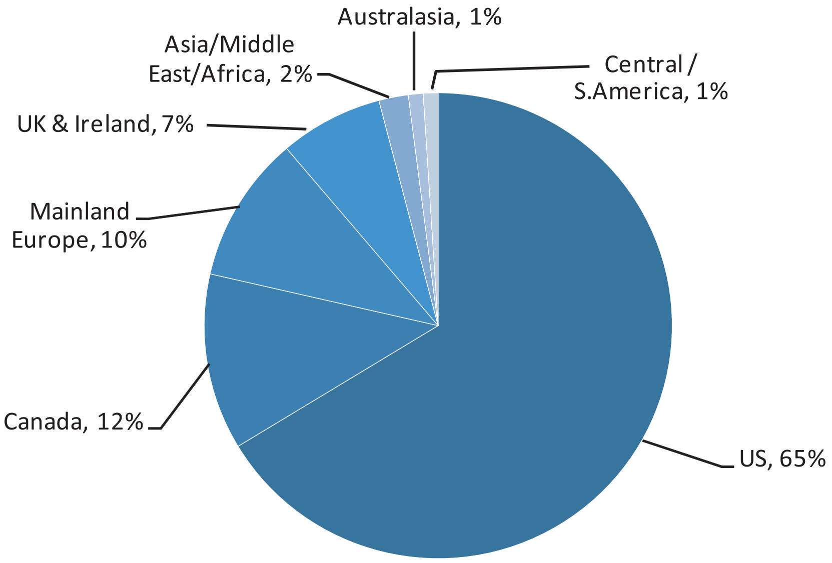
Organizational Size

Survey respondents represented organizations of all sizes. Larger organizations over 5,000 employees represented 31%, with mid-sized organizations of 500 to 5,000 employees at 35%. Small-to-mid sized - 1 to 500 employees - were 34%. Organizations of less than 10 employees were included in the report as they may well include researchers and consultants.



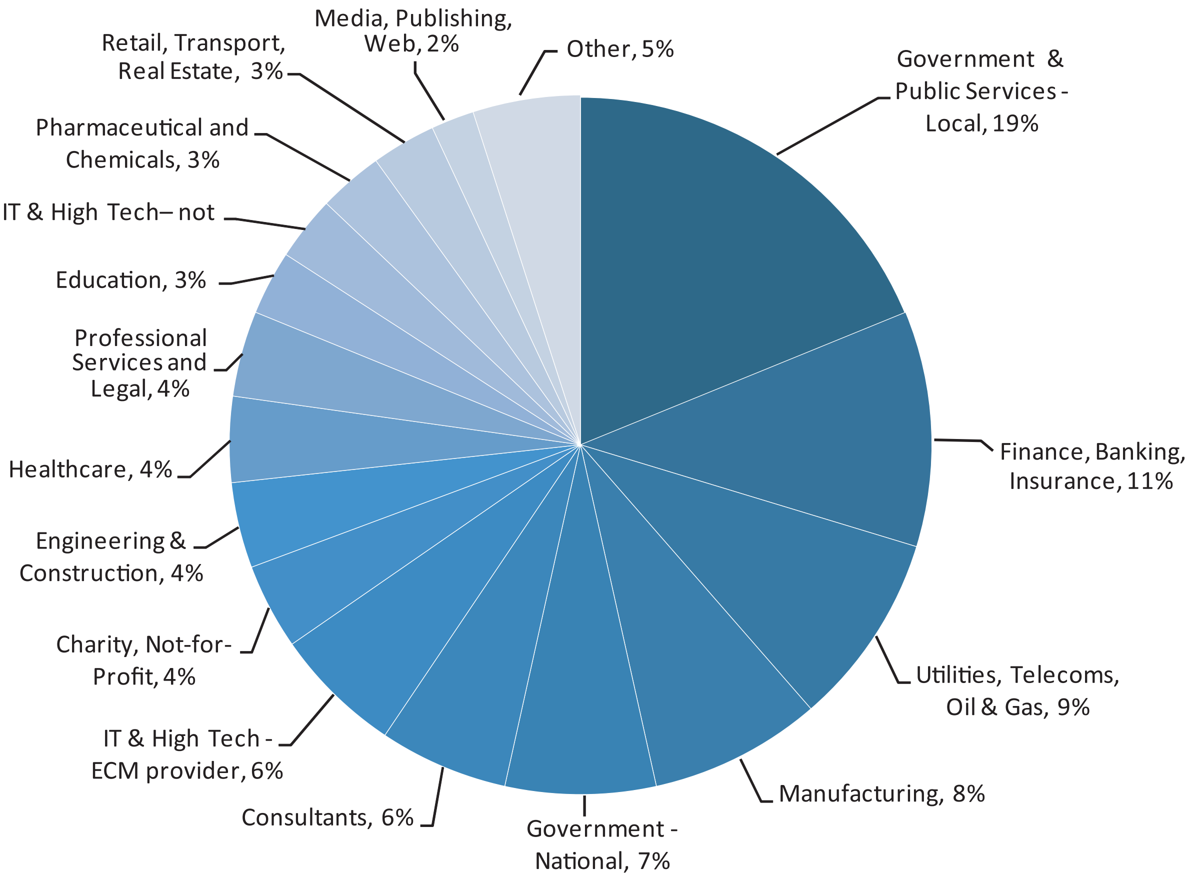
Geography

77% of the participants were based in North America, with most of the remainder from Europe.



Industry Sector

There is a higher participation from Finance, Banking and Insurance and Utilities than is usual for the AIIM demographic, indicating a higher usage of analytics in this sector. Local and National Government made up 26%. The remaining sectors are evenly split. To avoid bias, suppliers of ECM have been removed from most of the report, but consultants have been included as they are likely to be valid users of the tools.



Appendix 2: Glossary

Glossary

Automated redaction: Search, matching and blanking across scanned images or electronic documents of personal details such as social security numbers, phone numbers, names, monetary amounts, etc.

Content Analytics: A range of search and reporting technologies which can provide similar levels of business intelligence and strategic value across unstructured data to that conventionally associated with structured data reporting.

Content assessment: Trawling of stored documents and content to measure relevancy, currency, or frequency of access as an indication of the need to keep, or more particularly, migrate content to another system.

Content de-duplication: Exact or near-exact match of content stored within the same or different systems, albeit with different metadata as to who stored it and when. Scoring system allows automatic deletion of duplicates, saving space and reducing potential errors.

Copyright detection: Web search to match unauthorized use of copyrighted images, sound files, etc, and to serve notice of infringement. Similar tools may also be used to detect out-of-date logos or product imagery across marketing collateral and company websites.

Digital Asset Management (DAM): Content management systems particularly geared up for rich media files such as images and sound which are characterized by large file sizes, proxy representations (low resolution thumbnails or clips) and complex coders, decoders or format transformers.

Digital forensics: Investigative analysis to detect fraud or misbehaviour through word usage, trends, links, photo or sound patterns, particularly with a view to providing evidence for potential legal action.

Digital Rights Management (DRM): Technology that inhibits use (legitimate or otherwise) of digital content that was not desired or foreseen by the content provider or copyright holder, particularly to prevent unauthorized duplication.

E2.0/Social media monitoring: see also Sentiment Analysis. Monitoring of positive or negative comments being made on Facebook, Twitter, etc, in order to head off potential Twitter storms or negative brand impact. For internal E2.0, can be used to monitor staff morale.

E-Discovery tools: Search tools which analyze content for its likely relevancy to litigation by, for example, linking names, time periods and terms used. May also extend to legal hold and partitioning of content for further scrutiny.

E-mail trending: Reporting of email activity to indicate unusual patterns or topics of particular current interest. May also reflect potential faults or incidents, poor service, media coverage or staff sentiment.

Faceted search tools: Ability to sub-divide within search results by a standard set of metadata tags, e.g., as used by shopping websites to sub-set a product search by manufacturer, size, color, price range, etc.

Federated search: Ability to interrogate more than one repository or index from a single search screen. Might include linking internal ECM systems with subscription access to government databases or those of professional bodies.

Image and sound tagging: Pattern recognition search to match and apply additional metadata, e.g., faces, trees, voices, birdsong, to rich media files. May also be part of digital forensics.

Rich media: Generally used term for non- text formats such as photo-images, graphics, video, sound and animation. They cannot be searched by their textual content so must be tagged and/or represented by thumbnails or audio samples.

Sentiment analysis: Analysis of words used in comment or feedback to indicate satisfaction or dissatisfaction with products or services, aggregated to an overall score of satisfaction. Often extended to overall brand response, and monitored for trends or incidents.

Text analytics: Lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining, etc, frequently as part of other processes described here.

Web analytics: Advanced reporting over time of web behaviour including content assessed, paths followed, time on-site, referring sites, access points, search terms, geographic origin and pre-purchasing trends.

UNDERWRITTEN IN PART BY

IBM

IBM ECM provides improved workforce effectiveness by enabling companies to transform their business processes; access and manage all forms of content; secure and control information related to compliance needs, and optimize the infrastructure required to deliver content anywhere at anytime. IBM ECM automates and streamlines all records-based activities, eliminates burdensome end-user participation, enforcing compliance and creating business advantage while reducing the cost of compliance and risk management through the delivery of an integrated, open platform that provides interoperability with the widest selection of IT systems, thereby reducing costs and improving efficiency.

IBM Cognos® Content Analytics gives companies and public sector organizations the necessary tools to access and analyze the volumes of unstructured content found within any organization—in documents, insurance adjuster notes, case worker notes, web pages and more—to gain new business insights. It can:

- **Access** virtually any type of structured, semi-structured or unstructured content found within the enterprise.
- **Spot** issues and important trends that may fall outside of the normal business reporting channels.
- **Supply** new insights to business users across the organization, and help them go from insight to action with confidence.
- **Discover** new insights by automatically identifying and tagging key attributes and entities within unstructured content. Crawl almost any content source and identify key words and phrases.
- **Refine** the analysis with navigation and drill-down capabilities based on identified key attributes, entities and extracted dimensions.
- **Visualize** new insights with advanced visualization that enables exploratory mining and highlights trends, deviations and anomalies for more informed business decisions and actions.
- **Deliver** business insights to other processes and applications – such as IBM ECM (Enterprise Content Management) repositories or CRM (Customer Relationship Management) applications – and integrate content with IBM Cognos 8 BI for additional reporting and analysis

www.ibm.com/software/ecm/compliance

aiim®

Find, Control, and Optimize
Your Information

AIIM (www.aiim.org) is the community that provides education, research, and best practices to help organizations find, control, and optimize their information.

For over 60 years, AIIM has been the leading non-profit organization focused on helping users to understand the challenges associated with managing documents, content, records, and business processes. Today, AIIM is international in scope, independent, implementation-focused, and, as the representative of the entire ECM industry - including users, suppliers, and the channel - acts as the industry's intermediary.

© 2009
AIIM
1100 Wayne Avenue, Suite 1100
Silver Spring, MD 20910
301.587.8202
www.aiim.org