IDENTIFYING AND CHARACTERIZING TOPICS IN ENTERPRISE CONTENT MANAGEMENT: A LATENT SEMANTIC ANALYSIS OF VENDOR CASE STUDIES

Andrea Herbst  
*University of Liechtenstein, Vaduz, Liechtenstein*, andrea.herbst@uni.li

Alexander Simons  
*University of Liechtenstein, Vaduz, Liechtenstein*, alexander.simons@uni.li

Jan vom Brocke  
*University of Liechtenstein, Vaduz, Liechtenstein*, jan.vom.brocke@uni.li

Oliver Müller  
*University of Liechtenstein, Vaduz, Liechtenstein*, oliver.mueller@uni.li

Stefan Debortoli  
*University of Liechtenstein, Vaduz, Liechtenstein*, stefan.debortoli@uni.li

See next page for additional authors

Follow this and additional works at: [http://aisel.aisnet.org/ecis2014](http://aisel.aisnet.org/ecis2014)
Presenter Information
Andrea Herbst, Alexander Simons, Jan vom Brocke, Oliver Müller, Stefan Deportoli, and Svitlana Vakulenko
IDENTIFYING AND CHARACTERIZING TOPICS IN ENTERPRISE CONTENT MANAGEMENT: A LATENT SEMANTIC ANALYSIS OF VENDOR CASE STUDIES

Complete Research

Herbst, Andrea, University of Liechtenstein, Fuerst-Franz-Josef-Strasse, 9490 Vaduz, Principality of Liechtenstein, andrea.herbst@uni.li

Simons, Alexander, University of Liechtenstein, Fuerst-Franz-Josef-Strasse, 9490 Vaduz, Principality of Liechtenstein, alexander.simons@uni.li

vom Brocke, Jan, Martin Hilti Chair of Business Process Management, University of Liechtenstein, Fuerst-Franz-Josef-Strasse, 9490 Vaduz, Principality of Liechtenstein, jan.vom.brocke@uni.li

Müller, Oliver, University of Liechtenstein, Fuerst-Franz-Josef-Strasse, 9490 Vaduz, Principality of Liechtenstein, oliver.mueller@uni.li

Debortoli, Stefan, University of Liechtenstein, Fuerst-Franz-Josef-Strasse, 9490 Vaduz, Principality of Liechtenstein, stefan.debortoli@uni.li

Vakulenko, Svitlana, University of Liechtenstein, Fuerst-Franz-Josef-Strasse, 9490 Vaduz, Principality of Liechtenstein, svitlana.vakulenko@uni.li

Abstract

Since the turn of the millennium, the notion of enterprise content management (ECM) has been an umbrella term for all the technological and managerial capabilities required to manage every type of unstructured information over its entire lifecycle. Because the term “ECM” characterizes a very broad approach to digital information management, its precise definition challenges researchers and practitioners alike. The aim of this paper is to shed light on the use of the notion of ECM in practice. Grounded in a latent semantic analysis (LSA) of 1,083 case studies published by ECM vendors, the study identifies and characterizes twelve key topics that dominate the ECM discussion, from capture and scanning to archiving and backup. The results can help information systems (IS) researchers to gain a more profound understanding of the ECM concept, to distinguish it from related approaches to digital information management, and to reconsider and realign their ECM research agendas. Practitioners gain an overview of potentially interesting ECM topics that can trigger and guide their future information-management efforts.

Keywords: Enterprise content management, Document management, Content management, Records management, Latent semantic analysis, Text mining.

1 Introduction

The notion of enterprise content management (ECM) is used to refer to the management of all types of documents and their contents in an enterprise. First used by the Association for Information and
Image Management (AIIM) (Blair, 2004), a non-profit organization for information professionals, the term quickly found its way into information systems (IS) research (Tyrväinen et al., 2006). However, since ECM’s emergence in the early 2000s, researchers and practitioners have had difficulty developing a precise definition of ECM. In IS research, Smith and McKeen (2003) experienced “considerable confusion” around ECM (p. 648), a confusion that has remained prevalent ever since (e.g., Nordheim and Päivärinta, 2006; Tyrväinen et al., 2006; vom Brocke et al., 2011a). The situation is similar in ECM practice: Suzanne Mescan, Vasont Systems’ vice president of marketing, characterized ECM as a buzzword that requires a firm definition (Mescan, 2004), and five years later Dan Hooper, principal for the IT consulting firm Integrated Services, Inc., was still calling for a more elaborated definition of ECM (“ECM: Definition Please!”) (Hooper, 2009). Until today, AIIM has updated and expanded its ECM definition several times, indicating the elusive nature of the concept (Clark, 2014).

Grahlmann et al. (2012, p. 5) defined ECM as “the strategies, processes, methods, systems, and technologies that are necessary for capturing, creating, managing, using, publishing, storing, preserving, and disposing content within and between organizations.” This definition is grounded in a comprehensive review of the IS literature, and it shows that many IS researchers characterize ECM as a broad and integrated approach to digital information management (Päivärinta and Munkvold, 2005). ECM is an integrative concept because it covers all the activities in the content lifecycle (e.g., capture, storage, and retention), all types of semi- and unstructured information (e.g., reports, e-mails, and images), a variety of technologies (e.g., hardware, software, and standards), and all areas in an enterprise (Simons and vom Brocke, 2014). As such, ECM can be considered an umbrella term for many other information management topics like records management, web content management, and electronic document management, each of which typically focuses on specific lifecycle phases, document and content types, and/or business applications (vom Brocke et al., 2010).

The expanding scope of the ECM concept comes with serious threats to its study in research and to its application in practice, as it is difficult to define precisely the boundaries of ECM and to draw a sharp line between ECM and related approaches. From a research perspective, the overarching nature of the ECM concept comes with the risk of studying the same or similar phenomena under different headings (or studying different things under the same label), which makes it difficult to identify and build on related studies, especially in a highly diverse discipline like IS research. From an application perspective, misunderstandings regarding the scopes and objectives of ECM programs can mislead system users into believing that ECM is nothing but old wine in new bottles—an old solution sold as a new one—or into raising expectations that ECM cannot meet in the end. All of these risks could contribute to low levels of acceptance.

As a result, research is challenged to develop a clear understanding and definition of ECM, a solid conceptualization that not only justifies the position of ECM as a distinct field of IS research (Tyrväinen et al., 2006) but also is aligned with the use of ECM in industry. Clearly, the many information-management themes that information professionals discuss under the label of ECM, from the scanning of documents to their retention and deletion, complicate the development of such a conceptualization. Against this background, this paper uncovers the core of the ECM discussion in practice. Grounded in a latent semantic analysis (LSA) of more than 1,000 case studies shared by ECM vendors, we seek to identify and describe seminal ECM topics and to discuss these topics in light of the research themes from the IS literature.

The rest of the paper proceeds as follows. Section 2 characterizes ECM as an important yet elusive concept whose scope is expanding, making it difficult to determine its precise objectives and boundaries. Section 3 explains the procedures for collecting and analyzing the ECM case studies that provide the grounds on which section 4 identifies and characterizes twelve ECM topics by means of a content analysis. Section 5 discusses these results in light of ECM research from the academic discipline of IS. Finally, section 6 highlights the implications and limitations of the study, and section 7 concludes the paper with a short summary.
2 The expanding scope of ECM in practice and research

ECM has become a hot topic for information and knowledge-management professionals from all areas of industry (Päivärinta and Munkvold, 2005). The increasing importance of ECM in information-management practice is reflected in the rapidly growing ECM market. For example, Gartner (2013) estimated that the size of the ECM market was $4.7 billion in 2012, and the Radicati Group (2013), another technology market research firm, forecasted the growth of the ECM market from $5.1 billion in 2013 to $9.3 billion in 2017, an 82 percent increase in only four years. Another market study by TechNavio supports these figures, estimating the ECM market to reach $6.7 billion by 2015 (Infiniti Research, 2012).

While these and many other market analysis reports show that ECM is steadily growing in importance, they also reflect that the ECM market is undergoing rapid changes (Roe, 2013). While the market was in the beginning fragmented and dominated by document management companies, web content management companies, and records management companies, it soon grew and expanded, as did the software products that carried the ECM label. Many ECM vendors extended their software packages with new functionalities like workflow and process management and began to support collaborative content management environments (Cameron, 2011) and to target other (especially unstructured) types of documents and content, such as e-mails, reports, invoices, and audio and video files.

A wave of mergers and acquisitions occurred in the ECM market from the early 2000s on, with big players acquiring smaller niche players in order to expand their software portfolios. For example, Documentum, an ECM platform developed by the company of the same name and now owned by EMC Corporation, acquired between 2001 and 2007 Acartus (archiving), askOnce (search and retrieval), Authentica (digital rights management), Boxcar (content syndication), Bulldog (digital asset management), Captiva Software (image capture and scanning), Document Sciences (output management), eRoom (content-sharing), ProActivity (process analysis and monitoring), TrueArc (records management), and X-Hive (XML database capabilities) (Documentum, n.d.). Among the many other acquisitions on the ECM market, IBM acquired FileNet in 2006, Oracle acquired Stellant in 2006, HP acquired Autonomy in 2011, and Hyland acquired AnyDoc in 2013 (Hyland, 2013; Oracle, 2006; Reuters, 2011; Waters, 2006).

The timeline of mergers and acquisitions in the ECM market shows that ECM vendors continue to absorb technologies to add capabilities to their products, so their ECM suites are increasing in size, making it increasingly difficult for business professionals to determine the scopes of their ECM programs. The expanding nature of ECM makes it an elusive field that is difficult to understand or to distinguish from related fields in the area of information management. The unclear definition of ECM is apparent in a statement from Lubor Ptacek, vice president of strategic marketing at Open Text Corporation, in AIIM’s community blog (Ptacek, 2011):

The best definition we can think of is “anything to do with content.” Even the analyst firms disagree on the ECM subcategories. Is portal part of ECM? No! So why is Web content management (WCM) part of ECM then? Are you saying that digital asset management (DAM) is not part of ECM and yet WCM is? Are you suggesting customers should build their websites without rich media? How about social software? How about search? Ask 10 people and you get 10 different responses.

The situation is similar from the perspective of the academic discipline of IS. While the relevance of ECM for IS research has been reinforced repeatedly (Simons and vom Brocke, 2014; Tyrväinen et al., 2006; vom Brocke et al., 2011a), a common conceptualization and understanding of ECM remains to be developed. The latest literature reviews confirm that ECM has gained momentum in IS research (e.g., Alalwan and Weistroffer, 2012; Grahmann et al., 2012; Rickenberg et al., 2012; Usman et al., 2009), although academic interest falls far behind the software market. Päivärinta (2014) compared ECM to Hans Christian Andersen’s tale of the ugly duckling:
We can still state that ECM is something of an “ugly duckling” in academia, struggling to find its natural flock of swans with which the field could start to fly on its own. Compared to the market, ECM may have stayed in its egg and in the flocks of established ducks and geese for too long, focusing on the traditional ways of thinking about enterprise systems and software. For example, if we make the blunt assumption that a field’s market size reflects its relevance and compare ECM to, for example, ERP, ECM should deserve at least twice the amount of academic research as it receives today. (p. viii)

The investigation of the ECM concept in IS research largely followed the market, with roots in document management that date back to the 1980s and in web content management from the mid-1990s (Munkvold et al., 2006). In early days the ECM concept was considered a combination of document management and content management (O’Callaghan and Smits, 2005), but other concepts, such as records management, information resource management, and knowledge management, were soon added to discussions in the context of ECM (Munkvold et al., 2006).

To stimulate and guide ECM research from the viewpoint of the IS community, Tyrväinen et al. (2006) proposed that the concept can be studied from four perspectives: content, processes, technologies, and enterprise. The content perspective concerns issues of content identification, semantics, structuring, organization, creation, and use, while the technology perspective examines ECM hardware, software, and standards. The process perspective relates to procedures for ECM system implementation and maintenance and the implementation of content lifecycle activities. Finally, the enterprise perspective refers to the surrounding context of ECM and includes organizational, social, legal, and business content management issues. Thus, Tyrväinen et al.’s (2006) research framework suggests that ECM research covers organizational and technological issues alike, a view that IS researchers have adopted in a wide range of ECM studies that involved various application areas, software systems, and document and content types (e.g., Grahlmann et al., 2012; Rickenberg et al., 2012; Simmons and vom Brocke, 2014).

The goal of this paper is to identify and characterize the topics that are at the core of the ECM discussion in practice. This provides the ground to discuss ECM’s scope and objectives from the viewpoint of IS research.

3 Collection and analysis of ECM case studies

Our identification and characterization of the ECM topics is grounded in a content analysis of case studies shared by ECM vendors. Content analysis is a method of systematic text analysis, traditionally done through manual coding, with the goal of identifying concepts and patterns (Jauch et al., 1980). However, manual content analysis reaches its limits when the number or length of the documents to be analyzed is high (Fisher et al., 2010). Since we collected a large number of ECM cases, we decided to use an automated content analysis.

We used latent semantic analysis (LSA) for content analysis, a text-mining technique that IS researchers have used successfully for a variety of purposes in the recent past (e.g., Evangelopoulos et al., 2012; Indulska et al., 2012; Larsen et al., 2008; Sidorova et al., 2008). LSA detects patterns of word use in texts through statistical analysis (Landauer et al., 1998) and identifies words that co-occur frequently in the same contexts (e.g., in documents, paragraphs, or sentences), which can then be interpreted as topics inherent in a corpus of documents (Sidorova et al., 2008). For example, an LSA that detects that the words “scanner,” “capture,” and “OCR” frequently co-occur in a subset of the document corpus might indicate that the common topic of these documents is “document scanning.”

An LSA-based text-mining study can be subdivided into three phases: data collection and pre-processing, dimensionality reduction, and analysis and interpretation (Figure 1).
Data collection and pre-processing. As the foundation for the content analysis, we collected case studies that vendors or consultancies share online as ECM project reports. Such studies are much shorter than other case studies in research, but they typically include a detailed description of the business problem, the implemented solution, and the benefits achieved. In addition, such case studies usually cover a variety of ECM topics, so they are likely to form a solid basis for the analysis.

We collected the ECM case studies in two ways. In a first step, we searched the websites of all ECM vendors listed in Gartner’s Magic Quadrant for ECM (Gartner, 2012), which resulted in 795 case studies from 16 ECM vendors. (Six of the ECM vendors in Gartner’s Magic Quadrant have not yet published any case studies or did not share their case studies in such a way that we could identify them as ECM-related.) In a second step, we performed a Google search using combinations of the keywords “ECM,” “enterprise content management,” “case studies,” and “case study,” which revealed another 345 case studies from websites like AIIM, IThound, KnowledgeLake, and ImageSource. The final corpus included 1,140 case studies from 25 sources. The majority of the case studies were published between 2004 and 2013, but not all case studies included a publication year. We converted all of the case studies to plain text for further processing.

An LSA on raw text data may not produce meaningful results, as natural language texts are often too “noisy” to be analyzed automatically (Jones and Martin, 1997), so we cleaned and transformed the original text data. Some of the documents we collected could not be classified as case studies, so we excluded them, along with non-English reports and duplicates. The pre-processing of the documents reduced the overall corpus size from 1,140 to 1,083. To reduce the vocabulary used in the document collection (Manning et al., 2008), we first performed a word transformation called lemmatization (Evangelopoulos et al., 2012), which replaces forms of a word with its common dictionary form (e.g., replacing the term “documents” with “document” and “shared” with “share”). Then we removed common English stop words (e.g., “and,” “then”) and words that appeared in less than three percent of the documents, leaving 3,455 words. Two of this paper’s authors manually and independently analyzed these words to exclude additional uninformative terms and retain only those terms that would lead to the identification of ECM topics in the form of a so-called “go-list” (Evangelopoulos et al., 2012). Tyrväinen et al.’s (2006) four perspectives of the ECM research framework—content, processes, technologies, and enterprise—provided a suitable lens for selecting these terms.

After a pretest that revealed criteria that flawed the results (e.g., industry, software brands), we decided to use only words that characterized (1) types of documents and content (content perspective), (2) lifecycle activities and implementation approaches (process perspective), (3) software, hardware, and infrastructure (technology perspective), and (4) scopes of implementation, including business objectives and challenges (enterprise perspective). This framework led to the creation of two go-lists. The two lists showed an inter-rater reliability of 90.5 percent, which indicates a high level of agreement and exceeds the commonly applied threshold of 70 percent (Harris, 2001). The researchers discussed differences in the two go-lists until they reached consensus and a consolidated go-list of 660 terms.

Next, based on the final go-list, we created a term-document matrix that contained the number of times each term appeared in each document (Coussément and Van den Poel, 2008). We weighted the terms according to their uniqueness or informativeness by applying a Term Frequency-Inverse Document Frequency (TF-IDF) (Salton et al., 1975) weighting transformation to the term-document matrix, which is the most commonly used transformation approach (Evangelopoulos et al., 2012). This weighting scheme promotes terms that are rare in the overall corpus but that occur frequently in the
same documents (e.g., “disaster recovery”) and devalues terms that are frequent across the whole corpus (e.g., “content,” “document,” and “management”) (Manning et al., 2008).

**Dimensionality reduction.** Next, we transformed the term-document matrix using a statistical technique called singular value decomposition (SVD) to reduce the number of dimensions. SVD identifies groups of highly correlated words (i.e., words that co-occur together in documents) and highly correlated documents (i.e., documents that contain similar words) to produce a set of so-called latent semantic factors, each consisting of a list of highly associated terms and highly associated documents. Used together, the lists describe specific patterns of word use that can be interpreted as topics (Sidorova et al., 2008).

The researcher must define the number of factors to be extracted. Evangelopoulos et al. (2012) suggest testing multiple solutions, so we tested different solutions, reaching from three to fifty factors. After reviewing the results, we identified the twelve-factor solution as the most appropriate one, as the use of additional factors did not reveal additional meaningful ECM topics, and solutions with fewer factors lacked some relevant ECM topics.

**Analysis and interpretation.** In the analysis and interpretation phase, we used common statistical procedures (e.g., factor rotation, loading threshold selection) to interpret the extracted word-use patterns (Evangelopoulos et al., 2012). This step involved examining the terms and documents that were associated with a latent semantic factor and then labeling that factor. Two authors of this paper interpreted and labeled the factors independently, and then discussed minor differences in interpretation and labeling until they reached consensus. We also calculated some descriptive statistics, such as the number of documents for each factor and the cross-loadings between topics and documents.

### 4 Topics in ECM

The content analysis of the ECM case studies uncovered the twelve topics listed in Table 1. The LSA algorithm determined the order in which the topics appear in the table. The table also shows the labels we assigned to these topics and ten high-loading descriptive terms that we jointly selected from the list of the 25 highest-loading terms for each topic (at least six of them were among the top ten).

<table>
<thead>
<tr>
<th>ID</th>
<th>Topics</th>
<th>High-loading descriptive terms (excerpt)¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>Records management</td>
<td>Record, paper, record management, capture, scanning, space, storage, compliance, retention, paper document</td>
</tr>
<tr>
<td>F2</td>
<td>Invoice management</td>
<td>Invoice, vendor invoice, account payable, invoice processing, approval, accounting, processing, invoice approval, process invoice, paper invoice</td>
</tr>
<tr>
<td>F3</td>
<td>ECM as a service</td>
<td>Cloud, virtual, mobile, communication, exchange, data center, video, collaboration, file, scalability</td>
</tr>
<tr>
<td>F4</td>
<td>Data analytics and decision support</td>
<td>Analytics, data, case management, decision, content, improve business, improve quality, performance, decision making, processing</td>
</tr>
<tr>
<td>F5</td>
<td>Information exchange and collaboration</td>
<td>Portal, intranet, communication, knowledge, collaboration, exchange, search result, made available, directory, available</td>
</tr>
<tr>
<td>F6</td>
<td>Case management</td>
<td>Case management, collaboration, record management, record, case document, digital, sharing, paperless, archive, documentation</td>
</tr>
<tr>
<td>F7</td>
<td>Document and file management</td>
<td>File, electronic file, folder, file management, file document, metadata, electronic, document management, search, disk</td>
</tr>
<tr>
<td>F8</td>
<td>Enterprise search</td>
<td>Search, enterprise search, search result, searching, metadata, fast, search capability, find, search engine, finding</td>
</tr>
</tbody>
</table>

¹ Arranged in ascending order according to their loadings
Table 1. Topics in ECM

Records management was identified as the first ECM topic, with high-loading descriptive terms like “record,” “record management,” “space,” “storage,” and “compliance.” From the content perspective, records are evidence of transactions, so this stream of the ECM discussion focuses primarily on the storage and retention of business-critical and transactional documents with the goal of making business operations more transparent and accountable. As such, this ECM topic also incorporates the variety of standards and directives for reporting and documentation that today’s companies have to consider in digital information management, as suggested by many terms other than “compliance” that loaded on this factor (e.g., “regulatory,” “law,” “track,” “audit”). The scope of the records-management topic is broad—enterprise-wide and beyond—and includes various lifecycle activities, particularly the scanning of paper documents (e.g., “capture,” “scanning,” “paper document”). Objectives of the records-management topic refer to process improvement, most notably to the availability of and access to records (e.g., “access,” “retrieval,” “retrieve,” “available”).

Invoice management is an ECM topic that focuses on a specific type of document, the digitization and processing of invoices. Among the descriptive terms that loaded high on this factor were “invoice,” “vendor invoice,” and “paper invoice.” Companies use ECM to streamline accounts-payable operations, to convert paper invoices into digital formats, and to automate and make more transparent invoice-related processes (e.g., “transparency,” “transparent,” “workflow,” “automation,” “cycle time”). The invoicing topic indicates that the capture and processing of invoice data has become important for the ECM industry. However, from a process perspective, invoice management extends beyond the initial scanning process to include activities like invoice approval and distribution, which find their expression in descriptive terms like “invoice processing,” “processing,” “process invoice,” “approval,” and “invoice approval.” From an enterprise perspective, the topic is especially applicable to the accounting and finance functions, as it is reflected in terms like “account payable,” “accounting,” and “finance.”

The emergence of the topic ECM as a service from the data analysis confirms that ECM applications are increasingly delivered on-demand, so cloud-specific terms like “cloud,” “scalability,” “virtual,” and “data center” loaded high on this factor. From a process perspective, the cloud-computing topic refers to issues of collaboration in particular (e.g., “communication,” “collaboration,” “exchange”), and it is most often used to refer to the management of documents and files (e.g., “file”), though more specific formats like “videos” are also discussed in the context of cloud computing. The data suggest that the cloud-computing topic is not applicable only to specific industries or business functions, which suggests a broad scope of application from an enterprise perspective. Companies also typically want to use ECM as a service that supports information access through the use of mobile devices (“mobile”).

Data analytics and decision support had high-loading terms like “data,” “decision,” “decision making,” “processing,” and “analytics.” The major objective of this topic is to improve decision-making

---

F9 Workflow and process management  
ISO, BPM, process management, business process, report, invoice, workflow, approval, form, automation

F10 Web content management  
Website, web, portal, web content, marketing, digital, publishing, content, intranet, communication

F11 Archiving and backup  
Archive, archiving, backup, disk, protection, data, recovery, disaster, disaster recovery, retention

F12 Capture and scanning  
Scanner, image, recognition, capture, scanning, OCR, PDF, digital, imaging, format
and business in general (e.g., “improve business,” “improve quality,” and “performance”). While this topic seems to pivot on the management of structured data, most of the ECM cases in data-analytics refer to the analysis and processing of unstructured digital assets, such as the protocols of customer interactions, blog entries, and external information from the Web. Still, the analysis of semi-structured information like case-related patient records (“case management”) is also discussed under this ECM topic. The process perspective is not so prevalent, and there were no high-loading terms related to the technology or enterprise perspectives.

**Information exchange and collaboration** refers to the “exchange” of “knowledge,” so the process perspective dominates this topic, in particularly in terms of “collaboration” and “communication.” Companies invest in information exchange and collaboration to ease information access and establish a comprehensive source of information (“available”). Accordingly, from a technology perspective, the use of intranets and enterprise portals is especially relevant (e.g., “portal,” “intranet,” “directory”), although information exchange can also cross organizational boundaries. Mechanisms for storage and retrieval are at the core of the ECM-collaboration discussion (e.g., “search result,” “made available”). Accordingly, from the viewpoint of information types, collaboration in ECM focuses on the explicit dimension of knowledge, that is, knowledge codified in digital documents.

**Case management** refers to the management of case-related documents and records in settings like healthcare, social work, nursing, and law. The high-loading descriptive terms for this topic are “case management,” “record management,” “record,” “documentation,” and “case document.” Case management focuses on specific application contexts and industries like healthcare, public administration, and government agencies, which typically have to deal with case-related documents for specific persons or issues. This ECM topic refers to many lifecycle activities, including capture and scanning and migrating toward the paperless office (e.g., “paperless,” “digital”), content sharing and collaboration (e.g., “collaboration,” “sharing”), and archiving for regulatory-compliance purposes (e.g., “archive”). Case management objectives typically concern efficiency considerations like the streamlining and automation of case-related processes (e.g., “efficient,” “efficiency,” “business process,” “work process”).

Probably the most common and established approach to ECM is **document and file management**. This topic focuses on documents, a general type of recorded information, or even more generally, it focuses on the management of files. The related descriptive terms are “file,” “electronic file,” “file management,” “file document,” and “document management.” Storage and retrieval based on meaningful metadata are the dominant lifecycle phases, and they facilitate an efficient information search, which is supported by terms like “metadata,” “folder,” “disk,” “access file,” and “search.” However, the ECM case studies also discuss digitization of documents as a precondition for the establishment of electronic file management (e.g., “digital,” “electronic,” and “paperwork”). Accordingly, the business challenges that drive companies to invest in file-management systems relate to finding information, moving files around, and losing files, so the guiding objective is to establish an integrated database for efficient information access (e.g., “access file,” “control”).

ECM search functionality is even more prevalent in a topic we labeled **enterprise search**, which is another topic that focuses on a specific lifecycle phase. Some of the many high-loading descriptive terms that support this conceptualization are “search,” “enterprise search,” “search result,” “searching,” “metadata,” “fast,” “finding,” “find,” “search engine,” and “search capability.” The major ECM objective discussed under this topic is the improvement of search functionality in order to quickly deliver more accurate search results by, for example, combining information from several sources (e.g., “accurate,” “relevant information,” “integration”). Other terms that loaded on this factor like “intranet” and “website” suggest that the scope of the enterprise-search topic is broad, addressing internal and external stakeholders. The topic does not refer to specific types of information, as indicated by broad terms like “knowledge,” “record,” “file,” and “content.”

**Workflow and process management** deals with the automation and improvement of processes that involve the creation and use of documents and content, another important topic for ECM profession-
als. Process-related terms like “ISO,” “BPM,” “process management,” “business process,” “workflow,” and “automation” support this view. Common objectives are the reduction of turn-around times and tracking of workflow status (e.g., “tracking,” “monitoring,” “faster,” and “processing time”). The workflow topic puts a holistic focus on the content lifecycle, as indicated by many activity-related terms like “capture,” “approval,” “review,” “monitor,” and “track.” Types of content that are frequently discussed in the context of workflow and process management are “reports,” “forms,” “invoices,” and “contracts.” “Finance” and “customer service” also appeared as high-loading context-related terms, while most of the technology-related terms refer to workflow support, such as “system workflow” and “workflow software.”

Another established ECM topic is web content management, which typically refers to the publication of content on the web, as indicated by terms such as “web,” “website,” “content,” “web content,” and “digital.” However, web content management also refers to the publication and management of content on the intranet for, for example, internal communication purposes (e.g., “intranet,” “portal,” “communication”). The dominant lifecycle phase is “publishing,” though it is also important to maintain and update content so it is consistent and current across the enterprise. Web content is usually public, so it represents a company’s brand and its products and services. Accordingly, from an enterprise perspective, the topic is especially relevant for the marketing function (“marketing”). Many, if not all, industries, including media and entertainment, resale, IT services, government, healthcare, and transportation, as the ECM cases suggest, do web content management to some extent.

Archiving and backup is another topic that concerns virtually every company, as it deals with the preservation and retention of documents and content. The goal of archiving and backup is to implement a centralized approach to the long-term preservation and protection of data (“data,” “protection,” “archive,” and “archiving”) in order to ensure business continuity. Business-continuity initiatives often make use of disk-based backup solutions (e.g., “business continuity,” “disk”) to enable quick data recovery in case of a disaster like water or fire damage (e.g., “backup,” “recovery,” “disaster,” and “disaster recovery”). Therefore, two lifecycle phases, “store” and “retain,” are at the core of this topic. With many different types of documents and records, independent of type and format, issues of archiving and backup are discussed in industries like manufacturing, public administration, government, education, insurance, and media and entertainment.

The final topic we encountered in our data is capture and scanning, which focuses on the conversion of paper documents into digital formats. The corresponding high-loading terms are “image,” “imaging,” “scanning,” “digital,” and “capture.” The central technical components are scanning devices and recognition technologies like optical character recognition (OCR) and intelligent character recognition (ICR) (e.g., “scanner,” “recognition,” and “OCR”). The descriptive terms “PDF” and “format” reflect that digitized documents can be stored in various formats. The portable document format (PDF) is the most common, as it allows content searches in combination with recognition technologies like OCR, and at the same time keeps the original form and content of the documents preserved. Capture and scanning is an important topic for many industries—it is prevalent even in areas like art, history, and science—and concerns many types of documents, perhaps most notably forms and invoices.

5 Discussion

The content analysis of the collected case studies revealed twelve ECM key topics that considerably differ in their nature and partly overlap also. While some refer to specific application areas (e.g., case management), others refer to specific ECM system components and functionality (e.g., enterprise search and capture and scanning). We identified and characterized the topics using the four perspectives on ECM research proposed by Tyrväinen et al. (2006) as a theoretical lens.

From a content perspective, one ECM topic focuses on the management of a specific document type (invoice management), another focuses on the management of various semi-structured or unstructured digital assets (data analytics and decision support), and others deal with more generic forms of infor-
mation (records management, web content management, and document and file management). From a process perspective, two ECM topics focus on specific lifecycle phases (capture and scanning, archiving and backup), and at least one topic provides a holistic perspective on the implementation of the content lifecycle (workflow and process management). We identified only one topic that focuses on the technology perspective of ECM, ECM as a service, but two others can be located at the intersection of the process and technology perspectives: While the enterprise-search and information-exchange topics both focus on storage and retrieval processes, the ECM infrastructures that are required to support collaborative content management environments (e.g., intranet, portals, and search mechanisms) play an equally important role (similar to the capture-and-scanning and archiving-and-backup topics). From the enterprise perspective, case management is the only topic that focuses on a specific application area, even though other topics, such as invoice management, also dominate specific organizational contexts (accounting and finance).

Our results confirm that ECM is a comprehensive and modern approach to digital information management (Päivärinta and Munkvold, 2005) that covers many types of content, processes, technologies, and enterprise contexts. However, we were not able to identify a single ECM topic that addresses all four ECM perspectives at the same time. Records management, for example, covers many types of records as evidence of transactions and various lifecycle phases like scanning and retention (Svärd, 2014), and it is usually implemented at an enterprise-wide level with the help of various IT infrastructures. However, records management focuses on records, so many other types of documents are outside its scope, and it does not cover all of the phases in the content lifecycle (e.g., content reuse for publication purposes). Similarly, topics like capture and scanning and archiving and backup refer to various content types and industries, but they focus only on selected lifecycle phases and technical components.

These examples show that the topics we have identified, taken together, represent what many have defined as “ECM;” however, seen individually, they all represent older concepts with much narrower scopes. Concepts like records management, electronic document/file management, and (Web) content management all have long histories in IS research, so the question that emerges is what makes them distinct from an ECM viewpoint. Is ECM nothing but a merging of long established ideas and concepts (Laumer et al., 2013; vom Brocke et al., 2011a)?

Clearly, some ECM topics show which application contexts (e.g., accounting), content types (e.g., invoices), processes (e.g., scanning and retention), and technologies (e.g., ECM as a service) dominate the discussion of ECM in practice. Still, these topics are nothing new in information management research. However, a closer look at the case study data suggests that ECM does deserve to be a distinct field of IS research.

The emergence of ECM in practice goes hand in hand with the increasing digitization with which companies must cope. Today’s companies create great quantities of digital information, and information management has become difficult and complex, as evidenced by the many business challenges we found in the case studies concerning storage and retrieval, regulatory compliance, content publication, and so on. While these challenges are not new, they have a new quality today, which we believe justifies the study of the ECM concept from the viewpoint of IS research.

Companies have always had to comply with reporting obligations and standards, for example, but new directives, together with globalization and increasing numbers of legally binding documents, have made such compliance an increasingly complex task. For example, some product-related documents have to be retained for years even after a product has been withdrawn from sale, and the retention periods differ widely among countries (vom Brocke et al., 2011b). Likewise, storage and retrieval processes have long been the focus of information-management research and practice, but the enterprise-wide scope of ECM requires companies to deal with new standardization challenges like the development of corporate taxonomies and the definition and use of metadata (Munkvold et al., 2006). Similarly, content publication poses new challenges, as the increasing number of outlets in which
content is reused makes it difficult for companies to keep their information products consistent and up-to-date (Rockley and Cooper, 2012).

These and related challenges require companies to pool their strengths and to integrate previously isolated approaches to document and content management at an enterprise-wide level, an endeavor that many currently call “ECM.” Munkvold et al. (2006) put it this way:

> While most of the issues related to ECM initiatives can be traced back to established research areas when studied individually, the ECM concept integrates these issues in a new manner […]. The rationale of ECM resides in the global collaboration needs of an organization’s employees, customers, and partners through digital information content. (p. 95)

Concluding, from the viewpoint of the four research perspectives, the enterprise-wide scope in particular justifies research on ECM, along with the increasing amount and types of digital assets from the content perspective. However, the driving force of all these challenges is the innovation at the technological level that makes new workflows and business processes possible (or necessary) at the process level. Therefore, it is ECM’s integrative nature that makes it a distinct research field.

6 **Implications and limitations**

Grounded in a content analysis of more than 1,000 case studies shared by ECM vendors, we identified and characterized twelve key ECM topics. From a research perspective, the results confirm that ECM can be perceived as an enterprise-wide approach that integrates several information-management issues at the content, process, and technology levels. The results also indicate that researchers should be precise about the boundaries of their ECM studies or risk studying the same or similar phenomena under different headings or exploring different topics under the same label.

Our results show that many of the ECM topics are closely related, so we hope that this overview helps researchers to identify related works in the broader context of ECM. The results might also contribute to establishing ECM as a distinct research area in the IS community, leading to a more profound understanding and research agenda. We identified some of the business challenges that currently drive companies to invest in ECM, including regulatory compliance and storage and retrieval, on which researchers can ground concrete research questions. The discussion of these challenges, especially from the viewpoint of the whole enterprise, showed why ECM, despite its elusive and overarching nature, deserves to be a distinct field in information-management practice and research.

Our results are also useful for information-management professionals. Practitioners gain an overview of current topics in ECM, which can raise awareness about current developments in the ECM area. The ECM overview can also help them to position their own information-management activities in the broader ECM context. Our results suggest that ECM provides an integrated perspective on many highly interrelated topics in the field of information management, so ECM is perhaps best understood as an overall information-management strategy, rather than as a single IT project or even a software package.

The study has several limitations. First, while the set of 1,083 case studies we identified is relatively large, there are many other ECM reports that we did not collect. Although we believe that the case studies provided a solid foundation for the identification of the ECM topics, cases studies like the ones we used are usually published for marketing purposes, so the authors’ attitudes about the reported projects are typically positive. The use of overstatements, “marketing language,” and product names can make the texts “noisy,” which complicates their analysis.

Second, our go-list excluded some relevant terms because they flawed the results. For example, one would expect e-mails to be frequently discussed in the context of the topic of “ECM as a service,” yet the term “e-mail” did not appear in the list of terms that loaded high on that factor because it was frequently used in the contact-details section of the case studies, and we had to exclude it. Other examples for ECM-related terms we excluded from the analysis are “print,” which is often used in instruc-
tions about how to print the document, and “bill,” which is not only a document type but also a first name.

Third, our data analysis followed an automated approach for content analysis that is not yet well-established in IS research, though it is increasingly used (e.g., Larsen et al., 2008; Sidorova et al., 2008). We followed Evangelopoulos et al.’s (2012) guidelines, but some issues remain. For example, there are no standards for determining an appropriate number of factors for the analysis, so we tested several solutions to find the most appropriate one. Solutions with more or fewer factors would produce different results. Some of the emerging topics that are currently discussed in the context of ECM are notably absent from our results, such as social content management (Aladwani, 2014), and more established topics that we have expected to emerge from the analysis like digital asset management (Ljungberg, 2005) are also missing.

Fourth, a limitation can be seen in the selection of the terms we used to characterize each topic, which we did manually. While the majority of the selected terms are among the ten highest-loading terms (at least six per factor), other researchers may have chosen other terms to describe the ECM topics. There are also alternative approaches to determining the threshold for document and term loadings, and factor labeling was highly subjective, even though two researchers performed this task independently and concluded with very similar results.

7 Summary

ECM is an elusive concept that is difficult to understand and to distinguish from related approaches in the area of information management. Since ECM emerged in the early 2000s, researchers and practitioners alike have had difficulty determining its scope and objectives. The confusion around ECM remains prevalent, although some IS researchers have set out to explore the nature of ECM in the past few years. Among the pioneers in the IS community are Päivärinta and Munkvold (2005), who collected and analyzed 58 ECM case studies provided by AIIM. Now nine years later, we have done the same, but because the interest in ECM has considerably increased in the meantime, we were able to collect more than 1,000 ECM case studies, so we analyzed them in an automated content analysis.

We identified twelve ECM topics and described these topics using the four perspectives of content, processes, technologies, and enterprise. We found broad topics in digital information management (e.g., document and file management), and more concrete topics that focus on specific application contexts (e.g., case management), lifecycle processes (e.g., capture and scanning), document and content types (e.g., invoice management), and technologies (e.g., ECM as a service).

Our results confirm that ECM can be understood as a broad and integrated approach to digital information management, but most of the individual topics are older concepts that have narrower scopes. Accordingly, one might argue that ECM is nothing more than an umbrella term for long-established concepts in IS research. However, we conclude that ECM deserves its role as a distinct field of IS research because it addresses several contemporary information-management challenges, including regulatory compliance, enterprise search, and content reuse, that can be met only at an enterprise-wide level.

This paper not only reinforces the meaning of ECM from the viewpoint of IS research but also shows that the study of ECM requires IS researchers to define carefully the boundaries of their research. The close relationships between some of the ECM topics identified also suggest that practitioners should not view ECM as a single IT project but from a strategic viewpoint as a comprehensive approach to digital information management. We hope that our results contribute to establishing ECM as a distinct field of IS research and that they support researchers and practitioners in studying and implementing ECM.
References


