The Document Explosion in the World of Big Data – Curriculum Considerations

Michelle (Xiang) Liu
xliu@marymount.edu

Diane Murphy
dmurphy@marymount.edu

Information Technology and Management Science Department
Marymount University
Arlington, Virginia 22207, USA

Abstract

Within the context of "big data", there is an increasing focus on the source of the large volumes of data now stored electronically. The greatest portion of this data is unstructured and comes from a variety of sources in a variety of formats, much of which does not conform to a consistent data model. As business and government organizations become "paperless", the system of "record" in these documents (including text messages, social media postings, tweets, and email messages) becomes more important. The job of "records management" is becoming more significant in the information systems discipline, as businesses and government agencies struggle to control and manage their electronic resources for regulatory, compliance, legal, and business analytics purposes. Records management has risen to be a separate discipline with its own certifications and job classifications. Many of the principles in the discipline were developed in the time of paper records but still apply in this electronic age, and are now considered part of the responsibilities of the information technology worker. As this is a potential large job market, it is time to consider whether electronic document management (in addition to database management) should be included in the college-level preparation of undergraduate students who will join the workplace in the time of "big data" and its exploitation. This paper looks at the field of electronic document (records) management and its insertion in the undergraduate information systems curriculum.

Keywords: big data, records management, electronic documents and records management, information systems curriculum, document management

1. THE BIG DATA CONTEXT

"Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone." (IBM).

This data comes from everywhere: documents created in the course of doing business, posts to social media sites, digital images and videos, emails, web pages, e-commerce transaction records, security cameras and logs, cell phone transactions, and text messages to name a few. As business and government activities become "paperless", much of this data is only available in electronic form and this digital form is the only permanent record of important events (Franks, 2013; Ganz et al., 2007).

Much of the data is "unstructured" or "semi-structured" with no underlying data model or consistent format such as a social media comment. A recent study shows that 95 percent of information in business today is unstructured objects such as messaging logs, call history, usage trends, and weather information (Yu, 2012). This lack of a standard format makes it difficult to fit much of the "big data" into the conventional relational data model, which relies on data items that are consistent in form and format (data type and size). Unstructured data such as text documents is typically stored as a "blob" in a database that limits its usefulness. It can be retrieved and displayed though. In the big data context we are looking for insights above and beyond what traditional, structured data and content analysis techniques can provide. For example, in order to see into the text and find words or phrases that correlate with similar words in other text documents, techniques such as customer sentiment analysis need to be applied. Text analytics techniques such as "word clouds" can be employed to provide visual representations of unstructured data. Furthermore, the techniques that enable people to trace the origins of parts of documents are important as new documents are formed which include data from other sources. Technology exists today for this type of analysis but using them gets complex as the volume and variety of data is exploding.

2. ELECTRONIC DOCUMENTS AND THE ISSUES THEY PRESENT

Where is this unstructured data coming from? Electronic documents are everywhere. The traditional and most common document format include MS-Word "doc (or docx)" or Adobe "pdf". However, many different forms of documents now are commonplace in business- spreadsheets, email messages, SMS messages, Facebook comments, Twitter tweets- are now all components of a big data solution.

It has been reported that one in three business leaders do not trust the information they use to make decisions (IBM). How can you act upon information if you do not trust it? Establishing trust in big data presents a huge challenge as the variety and number of sources grows. While there are many techniques to verify and validate discrete data (e.g., the age of a person can be validated against their birth date), the nature of electronic documents makes them more difficult to validate. Specific issues associated with managing and safeguarding "unstructured" electronic documents include:

1. Version control. The key question is which version of document is counted as the "official version". In a typical business environment, a document may exist in many places as it goes through the draft, review and approval processes. How do we keep track of that workflow leading to the final "document"? Which version is the "system of record"? Is it located in a central repository?

2. Duplication. Even if a document is not revised, it may occur multiple times in many forms, including in someone's email, on someone else's private hard drives, on a variety of corporate servers, in the "cloud", on flash drives, etc. Which documents do we need to archive and store as permanent records and which do we discard? How do we remove known copies? Duplicated data can invalidate results or present bias in a data analysis leading to the lack of trust in the data.

3. E-discovery. When it comes to a legal dispute or a criminal investigation, documents and records are an important part of the discovery process and are the subject of evidence in civil and criminal trials. Amendments to the US Federal Rules of Civil Procedure (U.S. Courts, 2010) codified the requirement to provide electronic documents in the discovery process, using the term electronically stored information (ESI). E- discovery includes not just looking at the content of the documents themselves, but also an analysis of the "metadata" attached to such files (who created them and when, for example).

4. Security and privacy. There have been considerable "hacking" activities in today's networked computer environment, with "rogue states" and criminals stealing everything from trade secrets to business plans to individual identities. The more documents available electronically, the more risk of a security breach and the increased possibility of "reputation damage" as such breaches impact the privacy of individuals and businesses.

5. Information leakage: The 2011 CyberSecurity Watch Survey revealed that 46% of the respondents thought that damage caused by malicious insiders was more severe than damage from outsiders (Silowash et al., 2012). Confidential information can be compromised by employees and contractors who have uncontrolled access to important documents. Insider threats are well known since the advent of the WikiLeaks scandal. Wikileaks (wikileaks.org) is an international organization whose sole purpose is to publish confidential information from anonymous sources through its web site, so disclosing such information to the public (Karhula, 2012). How do companies prevent one of their confidential documents from appearing on a website like Wikileaks or being sold to a competitor in the U.S. or overseas?

These are just some of the major electronic document issues that face businesses, government agencies, and other organizations in today's digital world.

3. ELECTRONIC DOCUMENT MANAGEMENT

Document management is becoming increasingly important in the "big data" world, not just for the role of a document as a "record" of a transaction but also for the content that the document contains, in relation to information in other documents and when aggregated in a larger context.

One important consideration is the difference between a document and a record. ISO (International Organization for Standardization) 15489 is the key international standard on records management. It defines records as "information created, received, and maintained as evidence and information by an organization or person, in pursuance of legal obligations or in the transaction of business" (International Organization for Standardization, 2001).

According to the above definition, a record plays the role as "evidence", which implies that it must be complete and unchanged. Furthermore, the definition implies that organizations keep records in order to fulfill "legal obligations" or "transaction of business". A specific document "may or may not meet the definition of a record" (Department of Defense, 2007). ISO 15489 defines a document as "recorded information or Standardization, 2001). This definition is relatively generic compared with the definition of records. It does not specify whether and how the documents need to be kept and managed. Therefore, one major distinction between documents and records lies in that documents may or may not become records. Once a document becomes a record, it must be managed and controlled against change.

In today's digital world, however, the distinction between records and documents becomes vague as documents may be retrieved and then become evidence. In this sense, any document can be considered a record and any piece of its content can be extracted and used in a context different from the original intention of the document.

A document was originally considered to be text on paper and the term records management was used to describe a profession that manages physical documents (Franks, 2013). It was also more frequently associated with the traditional library catalogue and archive management. However, advances in technology (e.g., cloud computing, big data, etc.) and emergence of communication tools (e.g., social media) have completely changed the traditional view of records management as a discipline. Furthermore, how records are created and used in organizations is also fundamentally reshaped as a response to those developments (Bailey, 2013). More and more organizations today are overwhelmed by the volume of electronic documents and records and are searching for more efficient ways to store, manage, and maintain documents and records, and more importantly, to ensure compliance of the records with policies and standards (Franks, 2013).

One of the technical solutions is the implementation of an EDRMS (Electronic Documents and Records Management System), defined as "a system designed to manage electronic content, documents, and records and support four key functions: input (creation/capture); management (content, documents, records); collaboration/process management; and output/delivery" (JISC InfoNet, 2012). An EDRMS includes software tools that can be used to manage all kinds of records and documents and enforce retention and disposition rules and policies (Smallwood, 2012). However, it should be noted that EDRMS is only one piece of the puzzle. A successful document and record management program

requires a seamless integration of different components including the EDRMS, records management policies and procedures, content management techniques, effective information governance strategies, as well as well-trained personnel (Franks, 2013; Smallwood, 2012).

In her book Records and Information Management, Franks summarizes several adverse impacts on organizations which do not have a comprehensive document and record management program in place (Franks, 2013, p33):

- Damage to the organization's reputation;
- High costs for information management and storage;
- Lost files and risk of spoliation;
- Legal discovery penalties or sanctions; and
- Audit and compliance violations.

The above adverse aspects further reinforce the importance of implementing document and record management programs in organizations and the need for personnel to support them effectively.

4. DOCUMENT MANAGEMENT IN THE IT/IS CURRICULUM

Having established the importance of effective electronic document (record) management in the workplace, we next need to consider how to train young professionals to support the implementation and management of such systems.

As mentioned in the prior section, documents in business today can be rendered in a myriad of electronic forms. What poses challenges for managing these documents is the increasingly growing amount of unstructured information and the increasing need to analyze the data inside these documents, sometimes in ways not envisaged when the document was first created. In contrast with a database that stores and manages structured content (numbers in columns and rows), unstructured content is more difficult to capture, classify, maintain, and search than its structured counterpart (Franks, 2013).

Most contemporary undergraduate IT/IS programs include at least one required course on relational databases and the processing of structured data. The role of "big data" in the IS curriculum is currently under discussion (Topi, 2013), the focus being to manage big data technologies. However, there appears to be a lack of courses offered on managing, retrieving processing, and maintaining unstructured data (or documents containing such data) in undergraduate IT/IS programs. The ensuing question that we as IT/IS educators are facing currently is whether it is the right timing to incorporate electronic document management topics into the existing curriculum. And if so, how do we justify such a decision? Where does it fit into the curriculum? Is it a separate discipline or part of the database realm? In order to answer those questions, we apply an existing model as the framework to analyze the scenario.

The "holistic" model was proposed to help IT/IS educators make a valid decision as for "when" to incorporate new technology topics into the curriculum and the "tactical model" was developed for "how" to insert new courses into the existing curriculum (Liu & Murphy, 2012). In the model, several "forces" (i.e., factors) were integrated as a foundation for making "when" decision. These factors are summarized in Appendix: Figure 1.

We primarily examined three of these factors (i.e., impetus for the new topic, technology certification status, and avoiding curriculum bloating) to inform our decision on when and where to place electronic document management in the curriculum.

Impetus for the new topic

In our model, a new topic is considered to be a higher priority if it is recommended by industry or a curriculum advisory board.

As Franks pointed out in her book Records and Information Management, "… records professional must be a specialist when it comes to records management but a generalist when it comes to understanding the core business responsibilities of the organization and possessing the skills and abilities to interact with professionals from other domains, including legal, compliance, business units, information technology, and security/risk management." (Franks, 2013, p289). Most of these general topics are already covered by other courses in the IT/IS curriculum. In our institution, all IT students take courses in project management, computer security, and general business. Students who take the information systems (IS) specialty also take business law and organizational management. The missing

element is therefore electronic document record) management.

Although the document (record) management field was underrepresented in the past (US Office of Personnel Management, 1979; US Office of Personnel Management, 1965 (revised 2005)), employers are starting to recognize the important value of well-trained records management professionals to the business and organizations (Franks, 2013). This is also echoed by OPM in the memorandum Managing Government Records which states that the goal of establishing the records management occupational series is "to elevate records management roles, responsibilities, and skill sets for agency records officers and other records professionals (Executive Office of the President, 2012, p.6)."

Our institution is in the Washington DC metropolitan area and so government jobs are significant possibilities for our students. We felt that this job market was important for our students and could justify adding a course to the curriculum.

The technology certification status

The availability of a certification in the technology by a reputable organization is also considered as an important factor in our holistic model. (Liu & Murphy, 2012). The following are some of the current certification available in documents and records management and its related fields by reputable organizations:

- AIIM ERM: The Association for Information and Image Management (AIIM) is a global, non-profit organization that provides independent research, education and certification programs to information professionals. AIIM offers the Electronic Records Management (ERM) certificates. The certificates include two tracks: ERM Practitioner and ERM Specialist. The former covers major concepts and technologies for ERM while the latter focuses on implementing ERM.

- CompTIA CDIA+: CompTIA is the leading provider of vendor-neutral certifications and is well known globally. CDIA+ (Certified Document Imaging Architect) "covers major areas in technologies and best practices used to plan, design, and specify a digital imaging and content management systems.

- ARMA International ERM: ARMA International (formerly the Association of Records Management and Administrators) is a not- for-profit professional association and the authority on governing information as a strategic asset. ERM covers electronic document systems, related standards, and legal requirements.

Based on these technology certifications we concluded that the field was mature and that our students would be more marketable in the regional job market because of taking this course.

Avoiding Curriculum Bloating

The core curriculum for our B.S. in Information Technology degree is very full and designed to be cohesive and compliant with accreditation requirements. We looked at our specialty areas as having the most potential for application of document management skills and identified two areas: the information systems (IS) specialty (focusing on business) and the health information technology (HIT) specialty (focusing on healthcare). The course could be added to both of these areas: HIT was not a problem as this was a new curriculum. For IS, we decided we could move a decision analysis class to the core curriculum (as an option for the second quantitative course) and so could add the electronic document management course without "bloating the curriculum".

In conclusion, we decided that it was timely to introduce a course on electronic document (record) management into the curriculum and that it would serve students well in today's competitive job market.

5. CREATING AND IMPLEMENTING THE ELECTRONIC DOCUMENT MANAGEMENT COURSES

The "tactical" model (Liu & Murphy, 2012) suggests that, in general, there are four insertion approaches for a new technology topic: offer a "special topics" course which students can take as an elective, offer a new course outside of the department/program without "bloating" the IS/ IT curriculum, introduce a new course or make extensive revisions to an existing course, or develop a new specialization in IT program. Considering the factors of timeframe for implantation and the complexity of required approval processes, we originally took the approach of introducing a new course

to the curriculum and created a new course - Electronic Documents and Records Management as an IS/IT elective. This approach involved a formal curriculum review process at our institution.

As digital documents become the system of records in the world of big data, whether medical records or real estate transactions, the need for organizations to manage as well as analyze documents and other electronic records becomes paramount. The new course addresses this growing need, and supports the growing trend of e-discovery in litigation. It was an elective in the BS in IT program and in the IT minor, and a required course in the new BS in Health Information Management program. The broad purpose of the course is to ensure that students understand what is required to establish a sound documentation function within their work area, how electronic documentation systems work and what is needed to keep them compliant with regulations and policies. The course discusses how to go from a paper system to an electronic system and provides students with practical experiences at creating and using document and records management systems. It also discusses what documentation must be in place to support the company's systems and processes. The course objectives are listed below:

- Define electronic document management and records management systems and describe the importance of managing them, as the systems of record, in business and government;
- Explain the business and legal benefits of establishing a comprehensive records retention program and the need for vital records protection and disaster recovery planning;
- Identify factors that help reduce the scope and time for a document search and identify general criteria for indexing systems for effective identification, search and retrieval of records;
- Recommend solutions for common filing problems, for uniquely identifying records and for safeguarding the security and confidentiality of documents, particularly those containing personally identifiable information (PII); and policies, and the regulations for e- discovery in the legal environment; and
- Describe and evaluate document and record management software and how it is used in business, including legal and medical environments;
- Understand the common ground in all regulations, including the International Organization for Standardization (ISO), the HIPAA regulations for electronic medical records, the National Archives regulations and policies, and the regulations for e-discovery in the legal environment; and
- Create or acquire and manage a cost-effective electronic document management system to meet a specific business need government.

We ran this 300-level course as an elective and a pilot conducted in summer 2012 as an online course. There were only six students enrolled but their feedback was positive. Now we are moving to also offer it as a required course for students taking the information systems specialty (the largest subset of our students) beginning in the fall of 2013. The course is running as a hybrid one and has eighteen students. We will closely follow the outcomes (including employment opportunity, learning gains, etc.) from those students.

6. CONCLUSIONS

Big data has in its roots a variety of electronic document formats and, if the results of analytical process on this data are to be "trusted", it is important that the electronic sources of data (structured and unstructured) are managed effectively. It is our belief that as "big data" becomes more prevalent, qualified professionals will be needed to collect, manage, and retain documents in a structured way to meet industry and government requirements. One of these fields is obviously heath care, but there is also a variety of other business needs. While there is a high demand for the "data scientist" (Woods, 2012) to do the analysis and interpretation of the data, there is also a need within the IT profession for the "data custodian" role to manage the data. We believe that an undergraduate course in electronic document (record) management is essential to understanding the role of the data custodian and

making our information systems students more marketable in this fast growing field.

7. REFERENCES

Bailey, S. (2013). Perspective: Realigning the Records Management Covenant in Franks, P. C.'s Records and Information Management (pp. 23-25). Chicago, IL: Neal-Schuman.

Department of Defense. (2007). DoD 5015.02-STD: Electronic Records Management Software Applications Design Criteria Standard Retrieved May 22, 2013, from http://www.dtic.mil/whs/directives/corres/pdf/501502std.pdf

Executive Office of the President. (2012).Memorandum for the Heads of Executive Departments and Agencies and Independent Agencies. National Archives and Records Administration. Retrieved from http://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-18.pdf

Franks, P. C. (2013). Records and Information Management (1st Ed.). Chicago, IL: Neal-Schuman.

Ganz, J. F., Reinsel, D., Chute, C., Schlichting, W., McArthur, J., Minton, S., Manfrediz, A. (2007). The Expanding Digitial Universe: A Forecast of Worldwide Information Growth through 2010. Retrieved from http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf

IBM. Big Data at the Speed of Business, Overview. Retrieved on May 24, 2013, from http://www-01.ibm.com/software/data/bigdata/

International Organization for Standardization. (2001). ISO 15489-1: Information and Documentation-Records Management- Part 1: General. Geneva: ISO.

JISC InfoNet. (2012). infoKits. Northumbria University on behalf of JISC Advance Retrieved May 23, 2013, from http://www.jiscinfonet.ac.uk/infokits/

Karhula, P. (2012). What is the effect of WikiLeaks for Freedom of Information? International Federation of Library Associations and Institutions, (October 5, 2012). Retrieved from http://www.ifla.org/publications/what-is-the-effect-of-wikileaks-for-freedom-of-information

Liu, M., & Murphy, D. (2012). Tackling an IS Educator's Dilemma: A Holistic Model for "When" And "How" to Incorporate New Technology Courses into the IS/IT Curriculum Paper presented at the Proceedings of the Southern Association for Information Systems Conference, March 23rd-24th, 2012, Atlanta, GA.

Silowash, G., Cappelli, D., Moore, A., Trzeciak, R., Shimeall, T., & Flynn, L. (2012). Common Sense Guide to Mitigating Insider Threats, 4th Edition (CMU/SEI-2012-TR-012). Retrieved from Software Engineering Institute, Carnegie Mellon University website: http://www.sei.cmu.edu/library/abstracts/reports/12tr012.cfm

Smallwood, R. F. (2012). Safeguarding Critical E-Documents: Implementing a Program for Securing Confidential Information Assets. Hoboken, New Jersey: John Wiley & Sons, Inc.

Topi, H. (2013). Where is Big Data in Your Information Systems Curriculum? ACM Inroads, 4(1), 12-13.

U.S. Courts. (2010). Federal Rules of Civil Procedure. Retrieved on May 26, 2013, from http://www.law.cornell.edu/rules/frcp/

US Office of Personnel Management. (1979). Position Classification Standard for Support Services Administration Series. GS-0342. Retrieved from http://www.opm.gov/fedclass/gs0342.pdf

US Office of Personnel Management. (1965 (revised 2005)). Position Classification Standard for Archivist Series. GS-1420. Retrieved from http://www.opm.gov/fedclass/gs1420.pdf.

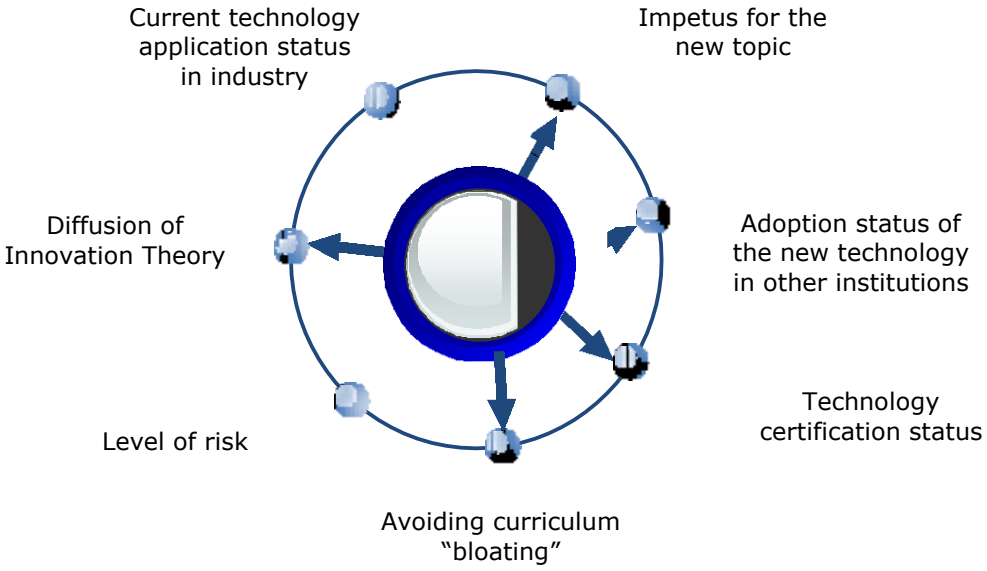
Woods, D. (2012). IBM's Anjul Bhambhri on What is a Data Scientist? Forbes, (February 16, 2012). Retrieved from http://www.forbes.com/sites/danwoods/2012/02/16/ibms-anjul-bhambhri-on-what-is-a-data-scientist/

Yu, E. (2012). Oracle Looks to Clear Air on Big Data. ZDNet, (October 4, 2012). Retrieved from http://www.zdnet.com/oracle-looks-to-clear-air-on-big-data-7000005211/

Editor's Note:

This paper was selected for inclusion in the journal as the ISECON 2013 Best Paper The acceptance rate is typically 1% for this category of paper based on blind reviews from six or more peers including three or more former best papers authors who did not submit a paper in 2013.

Appendix


Current technology application status in industry
Impetus for the new topic
Diffusion of Innovation Theory
Adoption status of the new technology in other institutions
Technology certification status
Level of risk
Avoiding curriculum "bloating"