ECM Data Security & Recovery

A White Paper

3S International, LLC.

December 11, 2013

The ECM Market

ECM (Enterprise Content Management) is the strategies, methods and tools used to capture, manage, access, deliver, search, store and preserver content and documents related to organizational process throughout the content lifecycle from creation to disposition. ECM covers the management of information within the entire scope of an enterprise whether that information is in the form of paper document, an electronic file, a database print stream, or even an email. ECM is an umbrella term covering document management (DM), web content management (WCM), search, collaboration, records management (RM), digital asset management (DAM), workflow management (WfM), case management, capturing and scanning. ECM aims to make the management of corporate information easier through simplifying storage, security, version control, process routing, and retention. The benefits to an organization include improved efficiency, better control, and reduced cost.

According to Gartner, the estimated ECM market is worth approximately $5 billion in 2013, with a close to 10 percent compound annual growth rate. Major players in the ECM market include IBM, EMC, Microsoft, Oracle, HP, OpenText, Alfresco. Consumers of the ECM products range from banking industry, to financial, insurance, medical, healthcare, manufacturing, energy, education and public services. As we move towards the age of BIG DATA, more and more organizations around the globe will face the challenge of increasing amount of digital data.

There are two types of digital data that are under the management of ECM products: structured and unstructured. Content viewers are at the core of the management of unstructured data. Content viewers provide interface between human and the unstructured data which requires human intelligence to process, comprehend and utilize the digital content. Content viewers allow users display, print, manipulate, modify and annotate/markup various digital contents such as pictures, business documents, audio/video files and streams, engineering drawings, medical images etc. There are many dozens of file formats. Some content viewers are standalone applications such as Microsoft Office, Adobe Reader and Adobe Acrobat. Some are browser plug-ins built on top of various browser plug-in technologies. Content viewing is essential to every one and all ECM platforms and systems. An ECM system without a viewer is like giving a blind people a large number of documents to manage. There are many content viewers and content viewing technologies, some of them are home grown from the ECM vendors, and some are from 3rd party vendors. Popular content viewers that are integrated with ECM platforms/systems are Adeptol, Adobe, EMC AXViewer, IGC Brava, Daeja ViewOne Pro, DocVerse, Kyte Viewer, Microsoft Office, MST, Open Office, Snowbound, Spcier etc. There are new viewer vendors coming into the content viewing market with new viewer products and technologies.

An ECM system and a content viewer are normally considered 2 different applications. An ECM system and a content viewer works together to provide content viewing features to end users through integrations. With the content viewing seamlessly integrated with the rest of the content management systems, content viewers that can be embedded in browser pages are becoming more popular than standalone viewer applications.

Many content viewers, during the executions in the production environment automatically generate data that are stored in the repositories of the ECM platform/system. Because of lack of standards, data that are generated by content viewers are proprietary to viewer vendors. Viewer generated data are not compatible with the data that are generated by other viewers. Although the data files are still accessible and visible across the ECM platform/system, the content of the data files can only be consumed by the viewer that generates them. In some extreme cases, the content of the data files are binary, only the viewer that generates them can decipher them. Enterprise data that are generated by content viewers are the topic of this white paper.

Challenges & Opportunities

Data incompatibility among different viewers causes the following issues for the ECM platforms/systems:

- ***Data Loss***. Many feature rich content viewers support annotations on documents and images as a method of collaboration among different users. Annotation data is generated by content viewers that each has its own data format. Annotation data is stored in the repository of an ECM platform/system, separately from the document or image they are associated with. When content viewers are switched from A to B, permanently or only at runtime, viewer B is not able to display the annotations generated by viewer A. Although annotation data are still stored in the ECM repository, from the end user perspective, annotation data are lost. An end user is not able to see annotations or markups that he/she or other users have placed on the same document. Collaboration is broken after the switch of viewers.
- ***Data Security***. Redaction is a feature of many content viewers that allow users to annotate or markup an area of the underlying document that contains sensitive information that the organization wants to protect. Redactions are associated with the annotation securities that

organizations try to enforce: only authorized users can see certain parts of a document in the system. Redaction is a special type of annotations. When annotation data loss arises, data security is broken. Now every one and all users of the ECM platform/system can see the entire document disregarding the security policy. Protected areas of a document are wide open to all users as long as they can see the document.

- ***Data Transparency***. The content viewer generated data is stored in the repository of an ECM platform/system. Due to the incompatibility and proprietary nature of the content of the viewer generated data, they are black box to the other parts or components of the ECM platform/system. For example, the search component of an ECM platform/system would not be able to search into the contents of the data for important textual information even though the data contain user inputs. User data are not transparent to other parts of the ECM platform/system. Some ECM products claim that their search can look into the annotation contents while it really couldn't.

These issues arise when customers of an ECM system switch from one viewer to another, or switch from one ECM platform/system to another ECM platform/system. Although modern content viewers are feature rich and very powerful, switching of viewers is not uncommon among ECM customers with some of them deploy multiple viewers in a single ECM system in order to cover a variety of content viewing needs, and some replace one viewer completely with another after usage in the production for some time. Customers have many different reasons to switch content viewers. It is their decision. Data loss and security issues are certainly important factors that balance the motivations to move away from the existing viewer. However, when customers decide to make the switch, they are willing to pay big price for tools and services to recover the lost data due to the importance of the enterprise data to their businesses. We are seeing more customers making switch of content viewers.

With the consolidation of the ECM market place, switching of EMC products and vendors are on the horizon. Migrations of large amount of enterprise data, structured and unstructured from the repository of one ECM system to the repository of another is inevitable. This includes the data that the content viewers have generated over the years of deployment. Without proper data migrations, precious enterprise data are doomed to be lost after switching.

Aside from data loss and security issues, another concern that may stop customers from switching content viewers is the viewer integrations in an ECM platform/system. Some ECM platforms/systems have only one content viewer integrated into the ECM platform or system. The integrated content viewer is "hard coded" into the hosting system. Switching the content viewer means rewriting portions of the ECM platform/system, thus are not practical. Customers of such ECM platforms/systems do not have choice on content viewers. For ECM platforms/systems that allow selection of content viewers from customers, traditionally, they are integrated into the underlying system directly by viewer vendors. Direct integrations have the potentials of exposing unique and special features from a content viewer to the end users. However the disadvantage of the direct integration approach is obvious. Imagine a personal computer system allows peripheral equipment vendors, such as sound, video, networking, memory and hard drive, to integrate their products directly with the PC system. Without a well-defined middle layer, direct integration of content viewers will bring confusions to the customer and difficulties to the platform vendors. How to identify and decide which party should handle a field issue? When the ECM platform/system comes up with a new use case for content viewing, would every one and all content viewer integrations have to go through rewrite? How to systematically handle the data that each viewer generates without causing data loss, data transparency issues? A Viewer Integration Framework (VIF) is desired between the viewer and the underlying the ECM platform/system. VIF defines the integration interface between a viewer and the underlying system. Borrowing the concept of hardware integrations in personal computers, a VIF with plug-and-play capabilities makes switching content viewer really easy and effortless for customers. For details on the topic of viewer integrations and VIF please refer to another white paper from 3S International.

Our Vision & Solutions

Our vision for the future of ECM is the standardization on the data format among content viewer providers.

Standardization of annotation/redaction data format is essential to the ultimate resolution to the data loss and data transparency issues revealed in this white paper. CMIS (Content Management Interoperability Services) is a step forward towards that direction. It is still far from our vision. CMIS currently only emphasizes on operations (application programming interfaces, RESTful services etc.) and object modeling. It is yet to provide a standard on the data format for annotations and markups. For the standardization on the data format for annotations and markups, representatives from the vendors of content viewers must be included into the board of the technical committee which currently only includes representatives from ECM vendors.

Prior the standardization on annotation data format, which may take long time, there is something that we can do now to help customers to switch content viewers without concerns on data loss and data security issues. Our annotation data recovery (ADR) tool provides a systematic approach for conversions among the annotation data formats from many popular content viewer vendors. Our ADR tool is a plug-in platform where new annotation data format can be added easily. It is also designed to work with multiple data repositories from different ECM vendors. It seamlessly converts annotation content stored in a repository from one format to another. In addition to ADR, we provide SaaS interface from our website that allows the invocations of the data conversion programmatically from remote locations. Finally, combining our annotation conversion libraries with an implementation of VIF in an ECM platform/system, a practical resolution to the data transparency issue can be achieved.

With data conversion tools such as ADR available, we shall see more customers willing to switch viewers since the tool breaks down one of the barriers that stops them from picking the viewer of their choices.