DATA INTEGRITY & TRANSPARENCY

3S International, LLC

September 9th, 2015

BIG DATA has been a public topic for quite some time now. However, data integrity and transparency are rarely touched. If your organization runs an ECM (Enterprise Content Management) application, or your organization is involved in the manufacture, integration, customization and services of ECM products and solutions, you should be aware of potential data integrity and data transparency issues that are revealed in this article.

ECM manages a mixture of structured and unstructured data. ECM system places the data in storage commonly referred to as data repository. Every ECM system has at least one data repository at the backend. A chunk of data is not randomly placed in the repository. It might have relationships with or be dependent upon other chunks of data in the same repository. Or more likely, a chunk of data depends on a program or a component of the system to make it useful and meaningful to the users of the system. Users here refer to human users as well as client computer programs of the system. Data integrity issue may arise if the internal relationships among the data are broken, or a chunk of data is no longer compatible with a program or component of the ECM system that supposed to present the data to the users of the system. When data integrity is breached, users of the system will experience data loss and sometimes more severely data security issues.

Another potential issue to an ECM repository is the data transparency. An ECM system is comprised of many sub-systems and components each may generate/consume data to/from the repository. If a chunk of data generated in the repository by a component cannot be consumed from other components of the system, there arise a data transparency issue. For example, if a component of an ECM system generating data in the repository that is proprietary in the data format that the search component cannot comprehend, that chunk of data is not transparent to the search component and thus the search component is not able to look into that chunk of data for search hits. Data transparency issue in ECM systems prevents annotation data from serving as the first class document indexes, fine grained indexes that are able to point to a specific area in a specific page of a document.

When and how these potential issues might become reality?

Migration from one ECM system to another is surely one scenario for the manifestation of the data integrity issue especially when the source ECM system and the target ECM system are from two difference vendors. There are many commercial ECM systems available in the market and deployed in organizations today, IBM, EMC, Alfresco, Oracle and HP just to name a few ECM vendors each provides one or more ECM systems. When an organization considering migrating from one ECM system to another, the application data generated by the old system might be at risk. Unless you don't care about the data in the old system, a data migration plan

must be orchestrated and implemented prior the transition. However data might be smoothly moved from the old system to the new one, still data integrity in the new system can be broken if the application data generated from a component in the old system is not compatible with the counterpart component in the new system. The counterpart component in the new system is not able to consume the data migrated from the old system due to the incompatibility between the two components. A typical example of such application data is the annotation/redaction data generated from the document viewer that works as a component of the ECM systems. If Viewer A is used in the old system and Viewer B is used in the new system, and the two viewers are not compatible with each other on annotation data formats, then no matter how smooth the data migration from the old system to the new, annotation/redaction data generated from Viewer A is not consumable from Viewer B, thus from the end users perspective, annotation/redaction data is all lost after the system migration. Although Viewer B in the new system can display documents migrated from the old system, annotations/redactions that are associated with the documents will not show up in Viewer B as end users viewing the documents in the new system. If the lost data are redactions that are supposed to cover some sensitive information in areas of the document content, in the new system those sensitive information is wide open to every end user who can display the document from Viewer B, a serious security breach scenario to the organization. To avoid annotation/redaction data loss, vendors came up with solutions such that permanently burn annotations/redactions into the documents before move them into the new system. Obviously, this is not a good solution because it solves a problem by generating a new one. Annotations/redactions are originally separate objects in the repository from the documents that they are associated with, after the migration they are permanently burnt onto the documents. Types of documents are transformed into images. Text annotations become not text searchable in the new system. And furthermore, sensitive information on the documents that are originally covered by redaction objects is forever lost after the migration, physically this time.

Another scenario for the manifestation of the data integrity issue is when organizations switching components from one to another within an ECM system. Componentization for large software systems is the trend. An ECM system comes with many components. As the deepening of standardization and specialization in the software industry, some ECM systems allow switching and even plug-and-play of some components. One of such components is document viewer. Modern ECM systems all support multiple document viewers integrated with them. Some of the ECM vendors even provide guidelines and APIs for 3rd party viewer vendors to integrate their viewers into the target ECM systems. The potential issue for switching document viewers is the risk of losing annotation/redaction data generated by the old viewer. In this scenario, legacy annotation/redaction data is still stored in the repository, however the new viewer, due to data incompatibility cannot read the legacy data, thus from end users perspective all existing annotations/redactions are lost when the documents that they are associated with get displayed in the new viewer. Data integrity of the application data is broken after the switching of document viewers.

Fearing the side effects of annotations/redactions, some ECM vendors went to the extreme that restrict document viewers from generating annotations/redactions, effectively

making document viewers a read-only tool. This approach is a significant drawback for modern document viewers. This approach not only disables the content level document indexing, but also throws away content level collaborations among different users.

Wouldn’t it be nice for everyone that annotation/redaction features can be fully utilized from ECM systems, and data integrity/transparency issues are resolved, and document viewer integrations in ECM systems are made such that deployment of document viewers becomes as easy as plug-and-play?

3Si products are designed and developed with this question in mind!