

Don't Count out this Dinosaur

The role of metadata in today's archiving world

IMERGE Consulting, Inc

Jim Just

November 2008

James. just@imergeconsult.com

As Appeared in Document Media.com

www.imergeconsult.com

The most common definition of metadata is “data that describes data” — not the most useful definition ever written. The ISO Records Management standard definition is more useful in the context of this article and document management in general. ISO 15489 defines metadata as: “Data describing context, content and structure of documents and records and their management through time.”

In the records and document management world, the term “index” is synonymous with metadata values. End users (consumers) of an enterprise content management system (ECM) will typically view only a small part of a document's metadata — the fields that identify the documents from one another as needed for their job function. For example, documents relating to an auto loan would be identified by the index (metadata) fields loan number, document type, loan date, term date and borrower's name. But an ECM administrator would see a much richer set of metadata for the same loan; she would have access to fields such as scan/creation date, file format, image resolution, object unique identifier, storage location, creator user ID and much more.

Various standards have specific metadata requirements that match the business and records requirements of the author; the best known in the US is DOD5015 2, which defines mandatory and optional metadata for each type of object (email, images, pictures, web records, etc.) stored in an ECM. ISO15489, ISO/DIS 23081-1 and the European Model Requirements (MoReq) are other examples of metadata standards. The Dublin Core Metadata Initiative has released a number of recommendations focused on standardizing the structure and vocabularies used across a number of applications. ISO 15836:2003 and ANSI/NISO Z39.85-2001 represent integration of these recommendations into formal standards. Many industry associations have produced standards for the information associated with their documents.

What, then, is the role of metadata in the realm of document archives? There are many, including:

- Access control (security)
- Authenticity of records
- Trustworthy records
- Usability/findability of records
- Fulfill standards requirements
- Reference to the physical location of objects in the storage cloud
- Reliability — data remains unchanged or change is audited

The power of Google, Yahoo Search, Microsoft Live, Verity, Apache Lucene and other content search engines has changed the way we interact with documents and all types of textual content. Visionary thinkers have suggested that only content searching will be needed in the near future. While this vision of the future may come true, it is unlikely that content search alone will replace the need for metadata, that it will instead complement it. However, when establishing your metadata model, it is impossible to consider every eventuality, and the cost to do so would be excessive. Content searching offers a means to complement metadata searching. For example, following the I-35W bridge collapse in Minneapolis last year, many state agencies received open records requests for information about bridges; the Minnesota Pollution Control Agency typically would not relate spills or investigations to bridges. In this case, they needed to identify documents that had any connection or reference to “bridges.” Without content searching, the staff had to manually identify and tag those documents with the metadata term “bridge” to meet the open records requests.

Textual documents, such as Word files, emails and PDFs (textual or OCRd), can be ingested by a content search engine, but scanned documents, pictures, faxes and other non-textual content cannot be ingested — they rely

strictly on metadata values for searching. Even with textual documents, the need for metadata remains when using content searching in order to limit the search to a specific subset of documents, whether by unique identifier (e.g., claim number, loan number, etc.) or range of values (e.g., from date, to date, etc.). The searcher then has a subset of documents that are specific to their need.

Metadata provides certainty to content search results, enhancing the usability of the content and complementing the search engines' power. Another example of the need for metadata that is common with our clients is email management. Email archive systems have tremendous value for mitigating the risks associated with email content during the discovery process, and e-discovery tools have become adept at categorizing documents to minimize the need for costly human analysis. But automated metadata capture is limited to that which is extractable from the email fields such as email sender, email addressee, subject, etc. The email server cannot confidently file the email under a specific topic (e.g., claim number, loan number, document type) — that effort requires human interpretation together with determining whether an email constitutes an official record or not.

Finally, metadata remains crucial to drive-automated workflow processes; workflow typically requires a separate set of metadata that drives the process through its tasks and becomes part of the history of the document. Content search engines complement rather than replace metadata while enhancing the findability of documents, and they are a tremendous aid in e-discovery.

A metadata model should include the data needed to capture the history of a document so that its integrity and trustworthiness can be tracked with certainty. Trustworthy records are core to a sound records management program. The metadata must not only be captured contemporaneously but protected, through access controls, from changes by unauthorized users. Authorized changes, such as correcting erroneous data, must be captured in the audit trail. So, too, the audit trail must capture changes to the access controls to prove chain of custody. When a content management system is used, the system's audit trail should capture all actions related to the documents, including all metadata edits. Obviously, the use of an ECM system greatly simplifies the management of metadata (assuming, of course, the ECM has the necessary controls).

The analysis of business requirements for document management will inevitably include intense discussions on the metadata requirements. With file systems, the creation of non-system metadata is relegated to document properties, file-folder names and object names; this rudimentary metadata capture approach is fraught with problems and very user-unfriendly. An ECM system, on the other hand, uses a database to store the metadata, which provides much greater flexibility and nearly unlimited quantity of values. However, any discussion of metadata must balance the need for data against the effort required by the creators. While there is a tendency to keep the creator's effort minimal (a commendable goal), erring on the side of minimal effort may also compromise the usefulness of the system.

For example, one of my clients scanned legal documents for three years, including both new and existing documents. To keep things "simple," they did not capture the effective dates of the documents, only the document types. Within a few years, some document type search results were pages long, thus, rendering the ECM unusable — the end users had to open every document to find the one they needed. The system was abandoned.

Metadata analysis must consider the requirements of both upstream and downstream creators and consumers of documents. Taking an enterprise view is critical so that all consumers of the metadata and documents are accommodated.

Some quarters will argue for comprehensive metadata so that a broad range of queries can be used to locate content from within the ECM solution; but this approach often duplicates data that is resident in business systems or makes the creator's job too time-consuming, stifling system adoption. Best practice is to determine how the consumers will locate information: Will they use the business system to find a key value first (e.g., loan number, claim number) then retrieve documents? Will business systems be available to feed secondary indices into the capture interface(s), eliminating manual effort create metadata? Will the business system data be available for queries from trusted third parties? Will there be a group of consumers that will only access the ECM through its metadata? Will business system data be available for the entire life cycle of the documents? Each of these questions comes into play when creating a metadata model that will support the full life cycle management of

documents in the content archive.

Metadata remains the crucial component for trustworthy, findable and authentic records, and enterprise content management remains the best solution for managing documents (objects) and their metadata. As always, developing complete requirements will result in a metadata model that meets the requirements of your archiving system.