DocuLex

Content Management Solutions

White Paper:

The Importance of Metadata to Electronic Document Management

by David Bailey

Jan 31, 2013

Doc# 012013-02

Intended Audience

This document is intended for reference by anyone interested in the basic principles of electronic document management.

Scope

This document describes the definition and use of metadata in typical electronic document management solutions, and particular with regard to the management of documents under DocuLex's *WebSearch* electronic document management system.

The Importance of Metadata to Electronic Document Management

Introduction

The concept of *metadata* is one of the most important and fundamental concepts in document management. The term literally means *"data about data".* If you've been working in technology for any time you've undoubtedly seen many examples of metadata – however it is so widespread that it's easy to take for granted. For example, when editing a document in Microsoft WORD™ and clicking on *File* and *Properties* you are presented with a dialog box that allows you to enter information about your document which may not appear explicitly in the document itself. In this case, you are allowed to enter (among other fields) the title of the document, the name of the author, the date published and associated keywords (which may someday be used to search for and retrieve the document).

If you inspect the source of virtually any commercial web page you will see normally-invisible lines near the top that contain metadata about that page (see LISTING 1).

```
<title>About Us</title>
<meta name="robots" content="index,follow" />
<meta name="revisit-after" content="7 days" />
<meta name="keywords" content="EDMS, electronic document management"/>
<meta name="description" content="The world's top EDMS solution." />
```

LISTING 1 – An example of metadata embedded in a web page.

In this case the value of the embedded metadata is quite apparent; it instructs Internet search engines how to navigate the site for indexing, and provides a recommended revisit criteria. Additionally, it contains keywords that are intended to bring Internet visitors to the site when they conduct searches on those search engines.

Some image file formats, such as GIF images, have provisions for embedded metadata that might identify the name of the owner of the intellectual property. Like MSWORD documents, Adobe PDF™ documents provide for the embedding of metadata. In fact, with the exception of regular text files you may be challenged to find a modern file format that does not make provisions for the embedding of some form of metadata. Take it to the next level and of PDF and imagine having the ability to customize the meta data field names, something that directly relates to your business and each particular document.

Custom metadata allows computer users to better manage their files by keeping them more aware of the contents and attributes of those files. Imagine how nice it would be to have your computer run a program that scans millions of files and quickly establishes a visual queue of all relevant documents.

But, beyond all these advantages, imagine keeping all your important documents in an online repository such that those documents can be searched, retrieved and organized based, not only on their contents, but also on their metadata. It's certainly easy enough to imagine wanting to list all documents with a particular meta data value published between two dates – or even documents related to certain keywords (in fact, a search match on keywords is probably a much more interesting match than a match on the full contents of a document).

The Three Roles of Metadata

Metadata can be considered to have three applications within any document management system:

- **Search** – by searching on specific metadata files rather than the full text content of documents, the reported results are almost always delivered much faster and with better accuracy, thus saving valuable time.

- **Display** – it is often very helpful, when viewing a list of documents retrieved through a search, to view the metadata fields for each document along with the original file. With WebSearch the user is able to customize the metadata fields that are displayed with search results.

- **Organization** – It is often helpful to organize the display of documents based on their metadata. For example, if a collection of documents have metadata for client name and contract id, then it is possible for a document management system to allow documents to be presented in a multi-level tree that has clients listed at the top level, and the contract id for each client branching off underneath the branch for the respective client. WebSearch allows an organization to define any number of such trees (known in WebSearch as *File Rooms*), organized into any hierarchy desired by the client *(for more information on the WebSearch File Rooms, see the section "Beyond Searching – WebSearch's File Rooms" below).*

  Additionally, WebSearch allows clients to specify how documents are physically stored in its directory structure by metadata, which can make it practical to access managed content through Windows Explorer if needed (generally not a recommended practice).

Hidden Metadata

Once you start to see the world from a document management perspective you'll start to see potential metadata everywhere. Imagine, for example, importing data from a spreadsheet, in which every column has a header. The column headers become metadata names, and the data underneath them become just more metadata. The rows of data in a database table, coupled with the names of the fields in the database, once again can suddenly start to be viewed as metadata.

Perhaps the greatest example of hidden metadata is the information in the heading of an email The '*from'* and '*to'* addresses, as well as the '*subject'* and '*date received'* of email messages, would all make very fine metadata for search and retrieval.

Calculated Metadata

A good document management system will not only recognize and index all the metadata contained in files, it will also have the ability to generate some of its own. For example, it could add a field indicating the date and time the document was entered into the system (which could then be used by document retention algorithms), or the size of the file. The "document type" is a common field utilized by EDMS systems that is not a common native file meta data type. The "document type" field is particularly useful in a document management system since it becomes an organizations primary electronic repository for all company documents. Due to this particularly specialty, the document type description is important to users when seeking out a particular type of a document, such as "invoice" , "applications" , "purchase order" or "contract".

Imagine that email messages are being indexed into a document management system – maybe by some automated process that catalogs every email message passing through a company's email domain. The algorithm could be configured to flag, via metadata, any emails that contain certain designated 'watch words' or the names of competitors or competing products (or virtually any other criteria).

When a user electronically signs a document, that signature is almost certainly calculated and stored as metadata as part of the document.

The addition of calculated metadata greatly increase the value an organization realizes from incorporating an electronic document management system *(EDMS)* into its infrastructure..

Associating Metadata with Files at Import Time

It is not necessary that metadata be associated with content at the time it is created, or even through use of the software that created the content. It is typical that EDMS systems provide tools and/or support protocols for the addition of new content to the system. It is not uncommon for these interfaces to the EDMS to allow for the association of metadata with content as it is being added. The means provided can range from providing a user interface for manual data entry to the parsing of XML and/or other file formats that associate metadata with the content to be imported. Additionally, the EDMS may automatically associate calculated metadata *(such as the current data and time)* as the content is being added.

Beyond Searching – WebSearch's *File Rooms*

DocuLex's WebSearch EDMS not only allows documents to be searched on specific metadata fields, but also allows documents to automatically become organized by metadata into a hierarchical structure that resembles the nested folder structure generally used to organize files on a computer. With WebSearch the organization's administrator can specify any number of *file rooms*, each of which is presented as an expandable tree with branches associated with values of metadata fields. For example, the accounting department of an organization might list the names of vendors, with branches underneath each vendor listing the transaction date and another branch listing the type of documents. Selecting a level branch or nested folder produces a list of all documents that have metadata associating them with that level. The WebSearch administrator can configure any number of main branches to the file room tree, each with its own criteria for nested metadata relationships. Each department or document collection can have its own secure document hierarchy.

Other WebSearch Metadata-related Features

In addition to the file room features mentioned above, WebSearch provides numerous other features that allow the EDMS user to leverage content metadata. As documents are imported into WebSearch they are processed by its internal *File Gateway*, which is responsible for enforcing the consistency and integrity of all files under management by WebSearch. The file gateway normalizes and categorizes each metadata field presented to it. The normalization provides the assurance that variations in spacing and capitalization do not result in the addition of unintended new metadata fields. The categorization identifies if a metadata field contains date and/or numeric data, and if so defaults it to that data type for purposes of searching (for example, numeric and data fields can be searched by range). The administrator may reassign the data type of metadata fields at any time.

A simple use of normalization and meta data continuity is using a procedure as simple as a meta data list. Users, for example, would choose pre-entered meta data such as a vendor name spelled “Johnson and Johnson” and not “J&J”, “Johnson & Johnson”, “Johnson’s” or simply “Johnson”.

Conclusion

Metadata plays a key role in any electronic document management system, and understanding how an EDMS processes and makes use of metadata should figure significantly into the selection of an EDMS solution and the planning of any EDMS deployment. Even if existing content, that will be managed currently, has no associated metadata, it can be added as documents are added to the EDMS (or even afterwards in many cases) system. I trust that your organization will spend the time necessary to create a well thought-out meta data policy for all file that meet your legal description of a document.