

# Document management system

A **document management system (DMS)** is usually a computerized system used to store, share, track and manage files or documents. Some systems include history tracking where a log of the various versions created and modified by different users is recorded. The term has some overlap with the concepts of content management systems. It is often viewed as a component of enterprise content management (ECM) systems and related to digital asset management, document imaging, workflow systems and records management systems.

## History

While many EDM systems store documents in their native file format (Microsoft Word or Excel, PDF), some web-based document management systems are beginning to store content in the form of HTML. These HTML-based document management systems can act as publishing systems or policy management systems.<sup>[1]</sup> Content is captured either by using browser based editors or the importing and conversion of not HTML content. Storing documents as HTML enables a simpler full-text workflow as most search engines deal with HTML natvely. DMS without an HTML storage format is required to extract the text from the proprietary format making the full text search workflow slightly more complicated.

Search capabilities including boolean queries, cluster analysis, and stemming<sup>[2]</sup> have become critical components of DMS as users have grown used to internet searching and spend less time organizing their content.

## Components

Document management systems commonly provide storage, versioning, metadata, security, as well as indexing and retrieval capabilities. Here is a description of these components:

Topic	Description
<b>Metadata</b>	Many document management systems attempt to provide document management functionality directly to other applications, so that users may retrieve existing documents directly from the document management system repository, make changes, and save the changed document back to the repository as a new version, all without leaving the application. Such integration is commonly available for a variety of software tools such as workflow management and content management systems, typically through an application programming interface (API) using open standards such as ODMA, LDAP, WebDAV, and SOAP or RESTful web services. <sup>[4]</sup> <sup>[5]</sup>
<b>Integration</b>	Capture primarily involves accepting and processing images of paper documents from scanners or multifunction printers. Optical character recognition (OCR) software is often used, whether integrated into the hardware or as stand-alone software, in order to convert digital images into machine readable text. Optical mark recognition (OMR) software is sometimes used to extract values of check-boxes or bubbles. Capture may also involve accepting electronic documents and other computer-based files. <sup>[5]</sup>
<b>Capture</b>	Data validation rules can check for document failures, missing signatures, misspelled names, and other issues, recommending real-time correction options before importing data into the DMS. Additional processing in the form of harmonization and data format changes may also be applied as part of data validation. <sup>[7]</sup> <sup>[8]</sup>
<b>Data validation</b>	Indexing tracks electronic documents. Indexing may be as simple as keeping track of unique document identifiers, but often it takes a more complex form, providing classification through the documents' metadata or even through word indexes extracted from the documents' contents. Indexing exists mainly to support information query and retrieval. One area of critical importance for rapid retrieval is the creation of an index topology or scheme. <sup>[9]</sup>
<b>Indexing</b>	Store electronic documents. Storage of the documents often includes management of those same documents, where they are stored, for how long, migration of the documents from one storage media to another (hierarchical storage management) and eventual document destruction.
<b>Storage</b>	Retrieve the electronic documents from the storage. Although the notion of retrieving a particular document is simple, retrieval in the electronic context can be quite complex and powerful. Simple retrieval of individual documents can be supported by allowing the user to specify the unique document identifier, and having the system use the basic index (or a non-indexed query on its data store) to retrieve the document. <sup>[9]</sup> More flexible retrieval allows the user to specify partial search terms involving the document identifier and/or parts of the expected metadata. This would typically return a list of documents which match the user's search terms. Some systems provide the capability to specify a Boolean expression containing multiple keywords or example phrases expected to exist within the documents' contents. The retrieval for this kind of query may be supported by previously built indexes, <sup>[9]</sup> or may perform more time-consuming searches through the documents' contents to return a list of the potentially relevant documents. See also <i>Document retrieval</i> .
<b>Retrieval</b>	A document ready for distribution has to be in a format that cannot be easily altered. An original master copy of the document is usually never used for distribution, rather, an electronic link to the document itself is more common. If a document is to be distributed electronically in a regulatory environment, then additional criteria must be met, including assurances of traceability and versioning, even across other systems. <sup>[10]</sup> This approach applies to both of the systems by which the document is to be inter-exchanged, if the integrity of the document is imperative.
<b>Distribution</b>	

