ISSN 2278-3091
Volume 8, No.3, May - June 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse01832019.pdf
https://doi.org/10.30534/ijatcse/2019/01832019

Flexible Metadata based Document Management System for Research Organization

Ferose Babu TA[1], Padmakuamr MP[2], Raji S[3], Shaji TG[4], Nibeesh K[5]
[1]Scientist, NPOL, DRDO, India, ferosebabu1@gmail.com
[2,3,4,5] Technical Officer, NPOL, DRDO, India

ABSTRACT

Document management inside an organization is an interesting socio-technical problem. A typical research organization, in addition to various technical documents manages a range of documents meant for publishing, limited circulation, sharing, uncirculated personal reference copies, physical files and dynamic paper files. Flexible metadata structure specific to type of document eases efficient description of document and retrieval. A system integrated with relevant applications having different access levels using databases over intranet can be a very effective document management system by incorporating a combination of structured and unstructured workflow. Technical documents in intranet over a period of time turn into knowledge evolution of a research organization.

Key words: Flexible metadata, Document management, Knowledge evolution**.**

1. INTRODUCTION

An early prediction of the paperless office was made in 1975 [1]. Components of document management system are metadata, integration, capture, indexing, repository, retrieval, distribution, security, work flow, collaboration, versioning and reproduction [2]. Combination of fixed metadata specific for each document class and flexible metadata can retrieve documents with accurate results for pre-defined document classes [3]. Classes of documents specific to a research organization can be classified into technical papers and reports, daily orders and circulars, documents for sharing, routine administrative documents of individuals. Few of these classes of documents are not shared; others include shared through unstructured workflow, published, through structured workflow and physical files. A Document Management System design must be flexible with its goal being a healthy ecology which breeds optimal knowledge transfer [4]. Technical content in a Document Management System (DMS), over a period of time turns into an indexed repository of evolved knowledge. Document Management integrated to workflow becomes a powerful tool for the overall management and improvement of business process [8] .

2. KEY FEATURES

Features of a DMS are metadata, integration, capture, indexing, repository, retrieval, distribution, security, work flow, collaboration, versioning and reproduction. Key features of the proposed system are

2.1 Creation

Creation involves accepting and processing images or paper documents from scanners includes electronic documents and computer-based files. All documents belonging to an entity is combined by appending into a single document in PDF format. It supports all kinds of document types like images, text, video, audio [7]. Each document type having specified format helps in standardizing and predicting storage requirements of repository.

2.2 Flexible Metadata

Metadata is the data about document. When posting a new document, users would often fill in only the mandatory metadata [3]. Each type of document has a set of specific metadata. Every document has a set of mandatory metadata to be filled in. Metadata in the form of {Key – Word} pair binds the word with a key. This puts a limitation to the user/author in efficiently describing the document.

In addition to structured metadata, here the user/author is given an option to describe the document as free style unstructured text without associating it to a key. Combination of structured metadata[11] and freely described unstructured metadata in the form of keyword is proposed to assist user/author in defining the document.

2.3 Search

Search is conducted in two or more cycles. Users start with initial query, depending on the results conditions are added or removed and continue further [3]. Documents can be searched using various attributes and keywords of document. More flexible retrieval allows the user to specify partial search terms involving the document identifier and/or parts of metadata. This would return a list of documents which match the user's search terms. DMS provides the capability to specify a Boolean expression containing multiple keywords.

The retrieval for this kind of query is supported by custom built indexes, to return a list of the potentially relevant documents. Result includes exact match as well as similar matches from common documents.
Combination of structured metadata and freely described unstructured metadata assist user to effectively frame search query.
Search domain include technical documents, published documents, documents shared to the user, limited circulation documents the user has access to, the user's personal files and physical files the user has involved. The search domain exclude documents the user does not have access privilege.

In a search, first the domain is restricted by the type of document, then fixed metadata conditions are applied and finally, flexible metadata based on value without key are applied. Broader search constraints - domain, fixed metadata and flexible metadata are connected by AND. Within each of the constraint, user can specify Boolean operator AND/OR to narrow down or to widen the result set [8].

2.4 Publish

Publishing a document involves the procedures of proofreading, peer or public reviewing, authorizing, and approving etc. Any careless handling may result in the inaccuracy of the document and therefore mislead its authors and readers [4]. A technical document for international/national journal/conference is checked for worthiness with respect to quality and declassification [12] necessities of institution. Those steps ensure prudence and logical thinking**.**

2.5 Work Flow

Metadata Meta Workflow is a complex process enabling documents to move in the predefined order. A structured workflow has a set of well-defined roles through which the document is processed through. A structured workflow is a rule based routing system to streamline document handling procedures. Each of these roles acts on the document. The final role approves the document. This structured workflow is used in technical and published documents.

A document can be distributed as an unstructured workflow by selecting a distribution list and by assigning or not assigning an order to it. In cases where order is assigned, after each members' approval, the document will be automatically routed to the next member in order. If no order is assigned, all members can see the document simultaneously.

2.6 Security

Metadata Security is viewed as combination of authenticity and authorization. Authenticity by means of login-password grants access to a legitimate user. Level of authority or privilege to functionalities for a valid user is determined by pre-defined roles and access control list.

2.7 Notify

Notification agent notifies concerned actors about an activity to be executed in a workflow. Notifications are sent to specific roles by means of automatically generated SMS message to notify relevant role thereby avoiding delay in each stage. Status of progress is notified to the creator of the document at each stage in the flow.

2.8 Publicity

Publicity agents announce arrival of published documents. Technical documents are intimated to user based on the interest of user. All new shared documents are alerted to the respective users. Published documents are highlighted as circular, order, announcement etc in the home page.

2.9 Weeding out

Technical documents, published documents can be marked as weeded out by final approving authority in the structured work. These type of documents marked as weeded out are removed from search and listing but are never really weeded out.
Limited circulation and shared documents can be weeded out by owner of the document. These documents are actually deleted from system along with metadata, member list marked for sharing.
Personal files, physical files, dynamic paper files, confidential/classified are never weeded out from the system as these are more official in nature.

3. TYPES OF DOCUMENTS

An organization handles few types of documents like circulars, notes, fax, letters, email, various types of bills and forms. In addition to these a research organization has to manage various types of technical documents intended for internal and external use. Types of documents are broadly classified as:

3.1 Technical Documents

Technical documents include papers, research report, project proposals, manual etc. This type of document moves on a structured workflow wherein author initiates with keywords as structured and unstructured metadata. A council does proofreading and authorizing. A domain expert reviews the document for quality and de-classifies if any classified content exist. Domain expert forwards to chairman of council who makes the final approval for publishing. Metadata can also be reviewed by the members who act upon the workflow. Approved documents with structured and unstructured

metadata are kept in the repository. This type of document follows structured workflow.

3.2 Published Documents

Published documents include circulars, daily orders. This kind of document moves on a structured workflow with review and approval. All approved documents will be published.

3.3 Limited Circulation

Limited circulation documents are technical or non-technical documents to be distributed to few members. The user can select members to whom the document is to be shared along with the order of circulation. Each of the members has to forward the document to the next member. The document moves across the limited members in the predefined order**.**

3.4 Shared Documents

Shared documents are technical or non-technical documents to be distributed to few members. The user can select members to whom the document is to be shared. All the shared members can view and remark the document at the same time.

3.5 Personal Documents

Personal documents related to official maters like bill settlement, where the documents are required at a future time for reference by the individual and dealing clerk. These types of documents are not shared or circulated but are uploaded by respective clerk and marked to the individual concerned.

3.5 Physical Files

In case of physical file, an entry is made in the system with title and keywords with, in/out date/time at each division, metadata required to trace at a future point of time. OUT date/time indicates when the file is moved out of source division and IN date/time indicates when the file is moved into destination division. Entry for a file is marked to one member at a time. Movements are tracked in order to avoid loss and to determine delay in each stage. Physical file with requisition and set of documents move along with the entry made in DMS. The file may be returned with a fresh DMS entry IN as target division and OUT as source division. Each stage is entered fresh and can take any path. This type of document follows unstructured work flow with member deciding at each stage where the file should move next.

3.6 Confidential/Classified/Secret files

Confidential documents are scanned, classified documents are scanned and encrypted with symmetric key whereas secret documents is dealt as physical files, and never scanned. DMS will have minimal metadata with entries and follows limited circulation with members and order of circulation defined by owner. Owner has an option to mark confidential document to follow a path or all members can see in any order. If marked as following a path, the document will be routed to next member without knowing who it is. The document follows unstructured work flow but predefined by owner.

3.6 Dynamic Paper Documents

This type of document adds in pages/files over a period of time. Initial processing starts with basic metadata in the system with a set of documents merged into a pdf file. This set of document is complete given the stage of processing. After initial processing, few more documents are attached to the basic entry at a later time. Partial processing may repeat depending on number of steps in the activity. Each step adds metadata signifying the step along with set of documents. Final step of processing seals the document and its metadata over a period ranging from few weeks to few years. These files are compressed and/or encrypted with final step considering the nature of content.

4. DATA BASE DESIGN

DMS mainly consists of six entities and normal user entity shared from administrative database. The main entities are applications, roles, privileged users, documents, keywords, distributions. Normal user data is shared from administrative data. Entities are designed with essential database concepts like primary key, foreign key, and unique key [9]. Minimal required indexes are built to optimize search without compromising inserts. Indexes are periodically rebuilt to optimize the tree. Garbage collection of junk initiations in workflows is periodically cleaned based on incompleteness in metadata fields and date of creation.

5. USER CHARACTERISTICS

DMS being an intranet, web based, multi user environment, has three broader levels of users. Finer workflow users fall into the second level. They play various roles to keep the integrity of the system. They are broadly classified as:

5.1 Admin

Admin manages all supporting tables that assist in upload of documents. Admin can integrate a new type of document; define its members and rules for workflow.

5.2 Manager

Managers are members having specific roles in structured workflow. They have specific actions to be carried out at any

stage of a type of document. Structured metadata for each type of document is defined by this user. Managers doesn't include unstructured workflow members are defined for a specific document and can be done by normal users.

5.3 Normal

Normal user can search the repository and retrieve documents. There is no restriction on search and retrieval of technical documents, published documents. The domain of search in cases of limited circulation, confidential and shared documents is limited to documents marked to the member. Personal files can only be accessed by the member and respective clerk. Physical files and classified/secret files can be accessed only while in transit by a member as physical file is manually moved between these members.

6. CONCLUSION

Flexible metadata based document management system integrated with structured/ unstructured workflow and structured/unstructured metadata can serve as one of the best document management method for limited user on intranet.
As the system manages almost all kinds of documents in a typical organization, it saves on environment and cost in terms of paper. Technical content in the indexed repository over a period of time evolves into a knowledge base of the research organization.

7. FUTURE WORK

Dublin core metadata along with ontology can be tried out in terms of document definition, retrieval performance and accuracy. Generation of metadata from OCR can assist user in data entry.

ACKNOWLEDGEMENT

The authors wish to acknowledge Shri. S Vijayan Pillai, Director, NPOL for according permission to carry out this work, Shri. Prince Joseph, Group Director (Embedded Systems) for extended support, Shri. Ananda Manoharan Y, Group Director (IT) for motivation and taking the twinge in reviewing, generosity of several colleagues who read manuscript drafts and made suggestions.

REFERENCES

1. **Paperless Office.**, (http://en.wikipedia.org/wiki/ Paperless_office).
2. **Document Management System**, (http://en.wikipedia/wiki/Document_management_system).
3. Mirjana A. Andric, ***Using Metadata for Information Retrieval in Document Management Systems*** , Eurocon 2005, IEEE, Belgrade, pp. 1093-1096, (2005).
4. M. Ginsburg, ***Intranet Document Management Systems as Knowledge Ecologies***, International Conferences on System Sciences - 2000, IEEE, Hawaii, (2000).
5. H. Baban, S. Mokhtar, **Online *Document Management Systems for Accademic Institutes***, International Conferences on Information Management, Innovation Management and Industrial Engineering - 2010, IEEE, pp. 315-319, (2010). https://doi.org/10.1109/ICIII.2010.555
6. Conferences on Information Management, Innovation Management and Industrial Engineering - 2010, IEEE, pp. 315-319, (2010).
7. D. Owczarec, J. Wojciechowski, J. Murlewski, ***Electronic Document Management System*** , MIXDES 2006 Dept. of Microelectronics & Computer Science, Technical University of Lodz, pp. 791-792, (2006). https://doi.org/10.1109/MIXDES.2006.1706695
8. R. Summers, J.J.L. Chelsom & et. el, ***Document Management: An Intranet Approach*** , International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, Amsterdam, pp. 1236-1237, (1996).
9. L. Aversano, G. Canfora, A. D. Lucoa, P. Gallucci, ***Integrating Document and Workflow Management Systems*** , IEEE, pp. 328-329, (2001).
10. Z. Deng-Hong, L. Ziao-Hong, ***Database Technology in Network Document Management System***, IEEE, (2010).
11. Danish Ahamad1 , MD Mobin Akhtar2 , Shabi Alam Hameed , **A Review and Analysis of Big Data and MapReduce**, International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.1, January – February 2019. https://doi.org/10.30534/ijatcse/2019/01812019
12. M.Tawarish1 , Dr. K. Satyanarayana , **An enabling technique analysis in Data Mining for Stock Market trend by Approaching Genetic Algorithm,** International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.1, January – February-2019. https://doi.org/10.30534/ijatcse/2019/06812019