# Hierarchical (Taxonomical) Approach for Handling External Documents in a Database System

Eliza Mazmee Mazlan, Rohiza Ahmad, Rozana Kasbon

*Abstract*— **This paper presents an approach that can be used to store reference to external documents in a database. The approach, which is hierarchical (taxonomical) in nature, groups each document according to their general to specific attributes. Documents, themselves will be stored in the application server with folder names similar to the tracing of the hierarchical path. Using this method, documents can be easily retrieved by combining the folder names in the hierarchical path. To demonstrate the proposed approach, a database storing maintenance data for storage tanks was developed and discussed in this paper.**

*Index Terms* — **Document Handling, Database System, External Documents, Hierarchical (Taxonomical) Approach**

## I. Introduction

Keeping scattered data into a common pool like a database has been a common practice in most organizations today. With a proper structuring of the database, information can be easily retrieved and produced in the form of reports. However, transforming manually the kept data into a format acceptable for a database is not always straightforward. For example, letters, reports and documents can be considered as source of data. An expert would be required to identify the data in the document so that they can be stored in a database. Failing to identify the important data would make information inaccessible. Hence, documents should be captured as close to its original form as possible. However, there are few issues related to this concept. One of it is regarding the storage space. Since a document can be large in term of size, it makes no sense to directly keep them in the database itself. Even by using electronic devices such as an Optical Character Recognition (OCR) to transform semi-structured text data into a database may result in errors such as in recognizing punctuation marks, alphabetic signs, numerical data and key words [4]. To transform the paper document to electronic format may also bring along certain disadvantages such as those discussed in [5]. The work discussed on electronic document addressing where a paper's citations to all other papers are linked. The disadvantages include the volatility of the document content, the movement of the documents, documents that are deleted and the change of the domain or directory structure where the documents are stored. Hence, we believe, rather than dissecting each portion of the documents for individual data or electronically transforming the paper documents, the documents in their original form should be kept as they are. This will ensure no data is loss from the process of converting the document into data for a database. Hence, we propose a method for storing their reference information in the database. Using the reference, we believe a proper search can be performed in order to find the relevant data.

In this paper the hierarchical approach of storing the reference information in the database is demonstrated using an information system that was developed to store maintenance information of storage tanks. The storage tanks have to undergo assessment exercises as part of its maintenance process. These exercises resulted in huge amount of data being generated as there are more than one locations where the storage tanks are located as well as there are more than one tank for each of the location. In addition, the assessment exercises need to be done in a regular basis. As times goes by the amount of documents used to record the maintenance information of these tanks increased significantly. The information system stored the reports produced in a database and 2 main problems related to its used were identified. First is the amount of storage required to store the data generated and second is the volatility of the stored data and its need for version control.

## II. Hierarchical (Taxonomical) Approach

According to [1] taxonomy is a classification of objects or concepts including hierarchical arrangement in which the objects can be classified into subtypes. In [2] an example of taxonomy is used in the web portal design to describe categories and sub-categories of topics that are found on a Web site. In [3] a taxonomical organization of species is given as an example of the approach. In [6] a document taxonomy or high-level, hierarchical classification for documents and records is used to facilitate the management of recorded

E. M. Mazlan is with the Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Perak, Malaysia (Phone: 6019-3303447; fax: 605-3656180; E-mail: elizamazmee@ petronas.com.my).

R. Ahmad is with the Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Perak, Malaysia. (E-mail: rohiza_ahmad@petronas.com.my).

R. Kasbon is with the Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Perak, Malaysia. (E-mail: rozank@petronas.com.my).

information. In this work a taxonomical approach is used to store document reference information in a database instead of transforming the data from the paper documents into a database. In other words, a database is used to find the document code while preserving the documents in their original form.

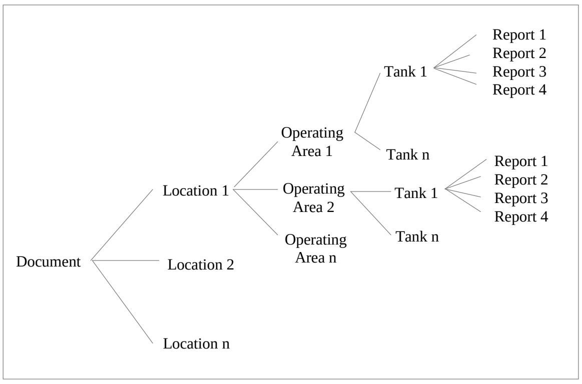

Fig. 1a Taxonomical Design of Document Reference Information

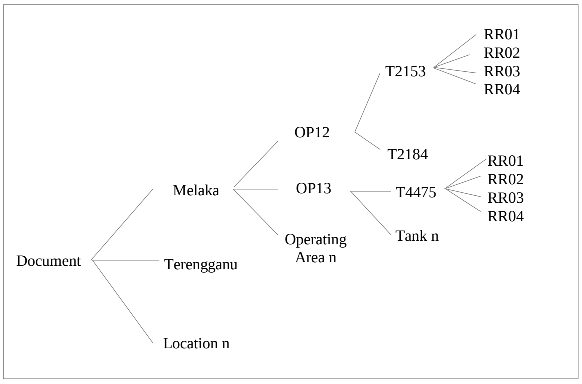

Fig. 1b Sample data of the Document Reference Information

In this work the hierarchical approach is used to store the reference information of oil storage tanks in a database while the documents that are related to the tank are maintained in an application server. The hierarchical (taxonomical) design of the document reference information is as depicted as in Fig. 1a. The design is based on the data that is generated from the assessment exercises of the storage tanks. There is more than one location and for each of the location there are also several operating areas. As a result there can be many storage tanks per location and per location also have many operating areas. The assessment exercises are carried out on a regular basis resulting in generation of many reports per storage tanks. This in turn generated massive volume of documents. The sample data of the reference information is as depicted in Fig. 1b. Basically when a search for a specific document(s) is performed, criteria such as the location, operating area and/or tank number need to be specified.

Fig. 2 depicts the class diagram for the database design where the reference information is stored. The design is based on the hierarchical (taxonomical) approach which is shown from the relationships among the entities. The result of querying the database will produce a document code which is unique for each of the documents produced by the assessment exercises. Table 1 shows the sample data that is stored in the database table. Basically a new report (hence a new Doc_Code) will be generated for each new assessment that is being done for each specific tanks. The Doc_Code is then used to store new document or to retrieve existing document in the application server.

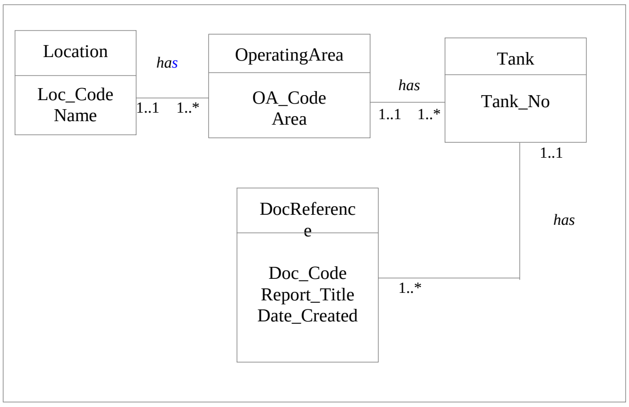

Fig. 2 Class Diagram

Table 1 Sample Data

| **Doc_Code** | **Report_Title** | **Date_Created** | **Loc_Name** | **Loc_Area** | **Tank_No** |
|---|---|---|---|---|---|
| RR01 | Tank_2153 Assessment Report | Jan 2001 | Melaka | OP12 | Tank 2153 |
| RR02 | Tank_2153 Assessment Report | July 2001 | Melaka | OP12 | Tank 2153 |
| RR03 | Tank_2284 Assessment Report | Jan 2001 | Melaka | OP12 | Tank 2284 |

## III. SYSTEM DESIGN

Fig. 3 shows the architectural design of the proposed system. Basically the reference information of the documents is stored in the database server while the documents are stored in the application server.

Fig. 4a, Fig. 4b and Fig. 4c depict the flowcharts of the system. Fig. 4a shows the overall flow of the system. Basically the users have the option either to upload a new assessment report or download existing report. If the option is to download existing reports, users have to choose the location, operating area and finally the tank within the location and the operating area. The flow of the uploading process is as shown in Fig. 4b.

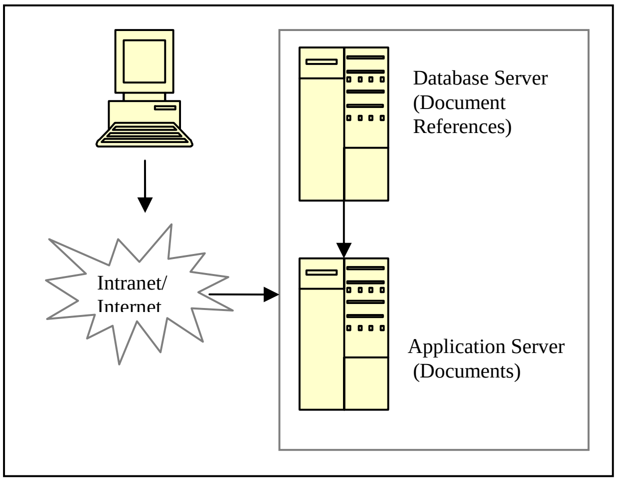
Database Server (Document References)
Intranet/ Internet
Application Server (Documents)

Fig. 3 System Architecture

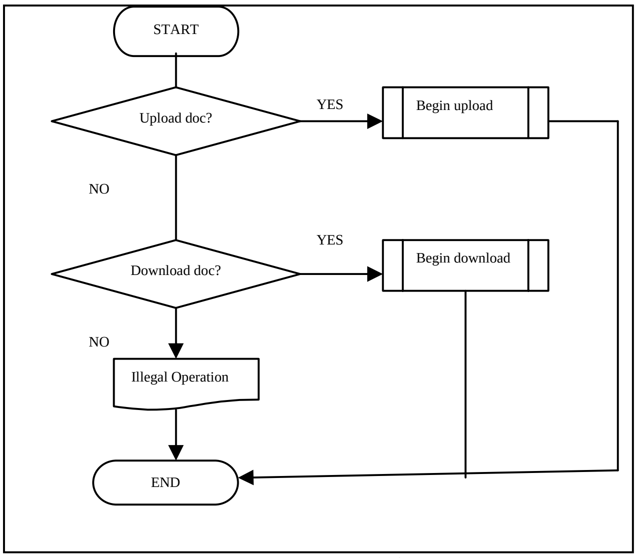

Fig. 4a Overall system flow

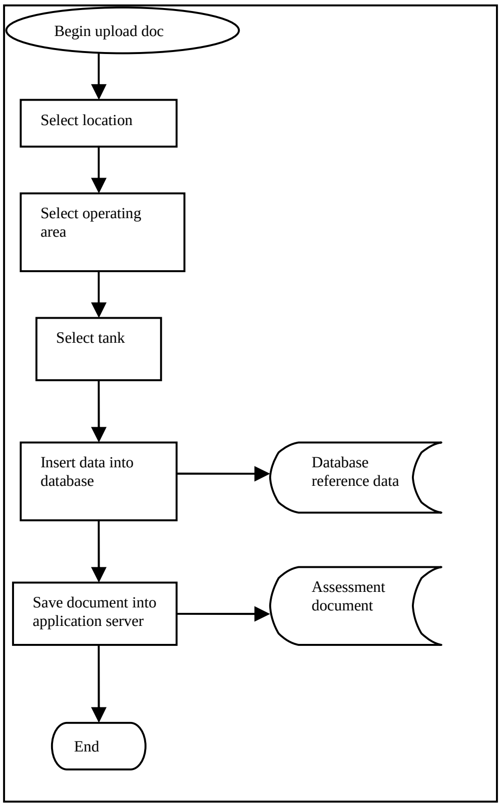

Fig. 4b System flow for uploading new report

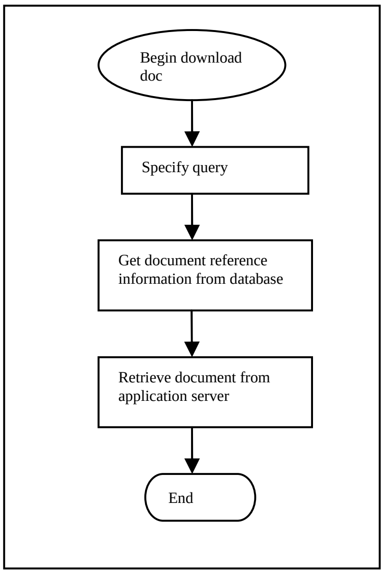
Begin download doc
Specify query
Get document reference information from database
Retrieve document from application server
End

Fig. 4c System flow for downloading existing report

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

For the download the user may specify the query by specifying the criteria such as the location, operating area and/or tank number. The reference information of the document comes from the database while the documents are retrieved from the application server. The flow of downloading existing report is as depicted in Fig. 4c.

## IV. Result and Discussion

```
select loc, op, tank,… where docid = …

"http:\\"+loc+"\"+op+"\"+….
```

Fig. 5 Sample codes for database & document retrieval

```
Location: Melaka

Operating Area: OP12
Tank No: T2153                    Tank No: T2284
Report     Date Created           Report     Date Created
RR01       10th Jan 2001          RR01       11th Jan 2001
RR02       10th July 2000         RR02       11th July 2001
RR03       12th Dec 2000          RR03       13th May 2001

Operating Area: OP13
Tank No: T3003                    Tank No: T3004
Report     Date Created           Report     Date Created
RR01       15th Jan 2001          RR01       11th Jan 2001
RR02       15th July 2001         RR02       11th July 2001
RR03       16th Dec 2001          RR03       13th Dec 2001
```

Fig. 6a Sample output for retrieval based on location

```
Operating Area: OP25

Location: Terengganu
Tank No: T4523                    Tank No: 4524
Report     Date Created           Report     Date Created
RR10       15th Jan 2001          RR01       11th May 2002
RR12       15th July 2001         RR02       11th May 2002
RR13       15th Dec 2001          RR03       15th May 2002
```

Fig. 6b Sample output for retrieval based on operating area

Fig. 5 are the sample codes to retrieve the document reference from the database (sql language) and finally to retrieve the document from the application server (php language). Fig. 6a and Fig. 6b show the sample output format of certain queries. Fig. 6a shows the possible result of a query that specifies Melaka as a location. There 2 operating areas for the location in Melaka which are OP12 and OP13. Each operating area has its own tanks. Fig. 6b shows the possible result of a query that specifies operating area OP25. The result shows the location of where the operating area is and the tanks that are located there.

Using hierarchical (taxonomical) approach in handling external documents may provide certain advantages. First, there is no need for the documents to be transformed into a database thus reducing the errors associated with wrong interpretation of document content or errors in recognizing special characters if device such as OCR is used. Secondly, the database size can also be significantly reduced. However, this approach is not suitable if the objects or concepts contained too many sub-categories or sub-types that resulted in a long chain of branches.

## V. Conclusion

Database has been known to be used to store data ranging from text to images. Because of the efficient way of keeping the data and also its reliability to retrieve information has make the use of the database very popular. This paper has demonstrated the use of database to keep only reference information instead of the whole content of documents.

## Acknowledgment

The researchers would like to thank Halimi Ismail, Sharul Niza Ahmad and M Fadly Ghani of PETRONAS R&T Division and Nur Azira Baharuddin, Nurul Afizah Salim of Universiti Teknologi PETRONAS for their contribution during the requirement gathering phase. Also many thanks to the Computer and Information Sciences department of Universiti Teknologi PETRONAS and Universiti Teknologi PETRONAS for their support.

## References

[1] http://www.dictionary.net/taxonomy
Definition of taxonomy

[2] http://searchcio-midmarket.techtarget.com/sDefinition
Definition of taxonomy

[3] http://www.thefreedictionary.com/taxonomy
Definition of taxonomy

[4] Onkov, K. Effect of OCR-errors on the Transformation of Semi-structured Text Data into Relational Database. ACM International Conference Proceeding Series, pp123-124. 2009

[5] Ashman, H. Electronic Document Addressing: Dealing with Changes. ACM Computing Survey, Vol 32, No. 3 pp 201-212, 2000.

[6] Choksy, C. 8 steps to Develop a Taxonomy. Information Management Journal. 2006.