Automatic Document Digitalization: DocDigitice

Ramón J. Rivera Acevedo
Masters in Computer Engineering
Dr. Alfredo Cruz
Electrical and Computer Engineering and Computer Science Department
Polytechnic University of Puerto Rico

Abstract — *In today's technological era, the use of cloud computing and cloud storage has become the norm for many big companies. This is no different for Health Insurance Organizations that are regulated by various corporations such as the Commissioner of Insurance of Puerto Rico. They require extensive documentation of every document that is sent and received. The goal is to design a multi-system application program for the Health Insurance Organization. The primary purpose will be created, facilitate, and improve the process of digitizing different documents of the company. It will offer a secure way to save the documents and at the same time ensuring the quality of these in the system. Employees can upload, monitor, and view the documents required by the company in a simple and safe environment. It will also help the company in audits and increase the overall income of the company in terms of storage, employee time and printing of document copy.*

Key Terms — *ERD, hard copy, Model-View-Control (MVC), soft copy*

INTRODUCTION

The health insurance company, the main objective of the company is selling policies, Individual Retirement Account (IRA), Annuities, and other products. Is one of the biggest Puerto Rican insurance companies that offer services that can help people have a better future. To control and sustain the information, process and money transactions, the department on information and technology gives support to the other departments to improve the process. For many years, the company has been highly dependable of internal servers and manual methods to store documents. These become a problem for the accessibility of the documentation, since is not centralize, and the security to edit or delete a document. This inefficiency has been observed as part of audits or annual revisions, as a regulated company this affects the compliance systems and in the long run, will result in loss of money.

The actual system of contract documentation is completely human dependent. Because of the extensive number of customers and products, documentation is manually managed and requires extensive human effort. To store all the information required of every customer, (forms, personal information, bank account information, social security, among others), it requires a person to: receive the physical documents, scan them, move them to the corresponding folder and file a hard copy. Also, the search of documents requires time, seeing that the only traceability that exists is a manual log that specifies who stores the files and where they stored them. Presently, one of the most critical parts is security. There is no traceability of the process and everyone could alter or delete a document.

To eliminate these problems, the development of a software that will contemplate to comply and be secure, has been created. The program has been written in Structured Query Language (SQL) [1] code, following the CIA Triad (confidentiality, integrity, and availability) to comply and ensure a better quality of the documentation process.

PROBLEM

To create the software, it was necessary to understand the needs of the customer and make sure that the information complies with the laws and regulations established in the facility. Another required that was necessary was the access to confidential documents. For the company, traceability and security of confidential information that is managed daily are extremely important. The slow search and misplacement of customer

documentation were becoming an issue during audits an annual revision. This could have had negative repercussions on the company. It can be sued by the customer for misplacing their information or have penalty fees given by one of the regulatory agencies. For that reason, this project aims were stablished in investigate and develop a program that reduce the time needed to search for information and increase the security of the documentation.

Goals

In order to correct the growing problem that could affect the company with bad documentation practices, loss of documents, data corruption, extra employee labor in documenting and the ever-growing number of documents in the company, a software was designed, created and implemented to help maintain and improve the security and accessibility of documents in a single location.

For the creation of the software, the program was developed with a WinForms (GUI) [2] application in the programing language code of C# [3]. The application starts with a primary privilege access logging module that validate the user's credentials against and grant access to all the primary functions of the application. After the credentials have been validated a level of security is given to the user. This will let the user navigate through different modules to manage the documents. All the information from the user data to the document data are been store in a SQL [1] database. The database as well as the installation components for the software are installed in a virtual server. This is for easier installation and maintenance. This will also help with managing folder accessibility and security to physical files that the program or user might need.

As part of the creation of the software the primary objective were to store all previous documentation and current documentation in a more efficient way. This was done by depositing them in a secured, controlled and encrypted scheme in the SQL Database. Another important capability was to bring better and consistent accessibility to all required users. Finally, this provide the company with an importing monitoring system and maintenance of all document functionalities that will be done by an access log, database log and program log systems that are in the virtual server and database. Some other key aspects are to reduce document damage or loss by a 10%. It also reduces a 25% on labor cost of the contract policy department. (cost = (quantity of time takes to complete task)/(employee work hours) * employee salary). Reduce 50% to 60% in the use paper to generate the hard copy.

METHODOLOGY & DESIGN

To improve the documentation process and the security a software was created with the language of C# [3] and Visual Studios to create the GUI interface in WinForm model [2]. This software allows to have the documentation centralize and improve the security. To manage the program analysis and creation a waterfall was prepared (figure 1).
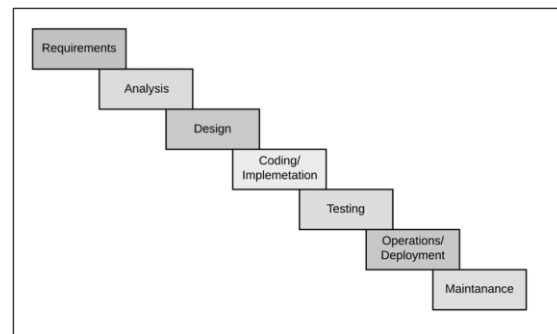

Figure 1
Waterfall Model

The section of methodology and design primary aim was to define the requirements needed to develop an application that will reduce the misplacement of confidential documents in a Health Insurance Organization. A proposal of a software that centralize the data was proposed to management and the stakeholders to resolve the problem presented. The following requirements were established as the key for the software creation:

- Document/Information Storage and Accessibility
- Log Maintenance and Regulation
- Document/Information Security

The waterfall model, sequential design progress methodology, is divided into seven-parts:

- Requirements
- Data Analysis
- Design
- Coding/Implementation
- Testing
- Operation/Deployment
- Maintenance

The waterfall model allows to have a rigid model, each phase has specific deliverables and are completed one at a time having better control in the process.

Database

The database provides us with a better understanding of the relationship that are in the software and the information in the system. The database standard that was use is SQL [1]. SQL is a special-purpose programming language designed to handle data in a relational database management system. A database server is a computer program that provides database services to other programs or computers, as defined by the client-server model. Therefore, the Database Management System (DBMS) [4] that was utilized for the creation of this database was Microsoft SQL Server [5]. SQL Server is the primary database handler in the company now. It is currently running on a remote server for easier access and availability to all systems. In the creation of the database, the process of data normalization [6] was utilized. This is a process of detecting dependencies on the data to improve the integrity of the database. To improve the security the primary information is encrypted utilizing the SQL Server Transparent Data Encryption (TDE) [7] it provides real-time I/O encryption and decryption of data and log file. After implementing the four of the normalization

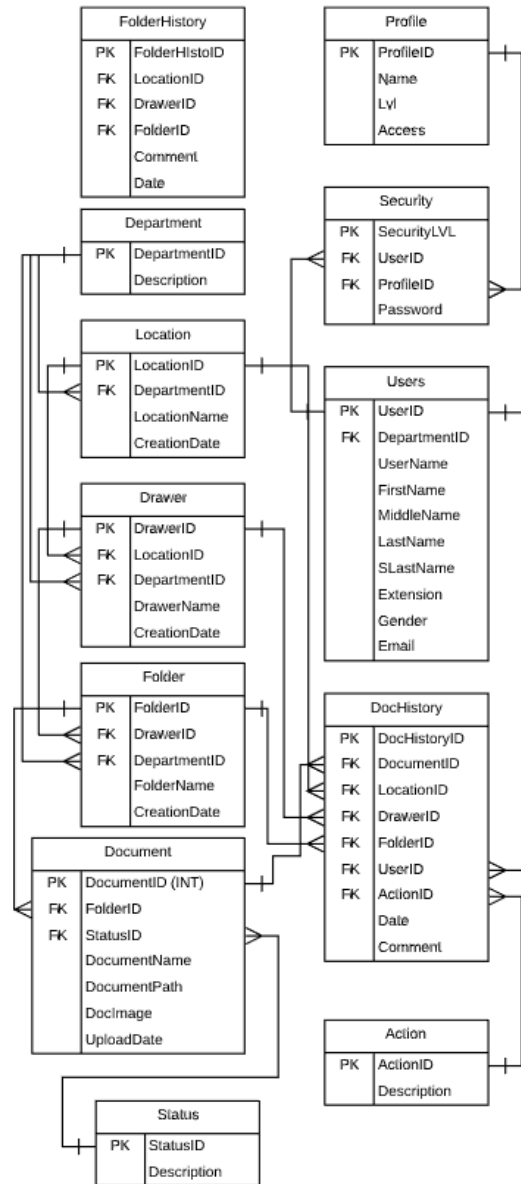principles, the database is composed of 12 tables (figure 2).

Figure 2
ERD Diagram

- **Documents:** This table will contain the data of the uploaded documents, also this includes the Image field store in a Varbinary(max) field.
- **DocHostory and Folder History:** These two will be log tables in the database. they will be used for monitoring, security, and audit logs.
- **Profile:** It has a list of available levels of security that the software will be program to

detect. A profile level must be assigning to each user in the system.

- **Security:** In this table the real level of security is stablish, the union between the user, department and profile will be achieved in this table. As well as the UserID and Password for each user.
- **User, Department:** This saves the users information, as well as the department that will be link to most table for user access purposes.
- **Rest of the tables:** Will contain Location, Drawer, Folder, Action and Status will be for structure of the software and navigation.

Software Design

To comply with the documentation process requirements, a flowchart that show how the company will interact with the application was created (figure 3).
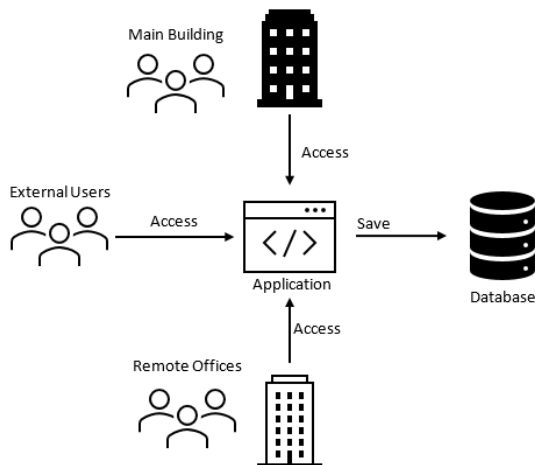

Figure 3
Program Flow Chart

In the application, program flow flowchart, interested entities can see how the users will interact with the application and that everything will be saved to the database. The application will be the center document storage of the documentation system and what will tie everything together in the company. These requirements are based on the end user's needs and law and regulations requirements. Once it was possible to decide the design, the coding/implementation phase was started.

Here the requirements are established of what the software's internal validation will be. It was also established what will be the environment that will be used (Microsoft Visual Studio) and other key requirements that are needed.

In the application internal flowchart (figure 4), it can be seen how internal functions will work in the application, explained in more detail in the Results section.
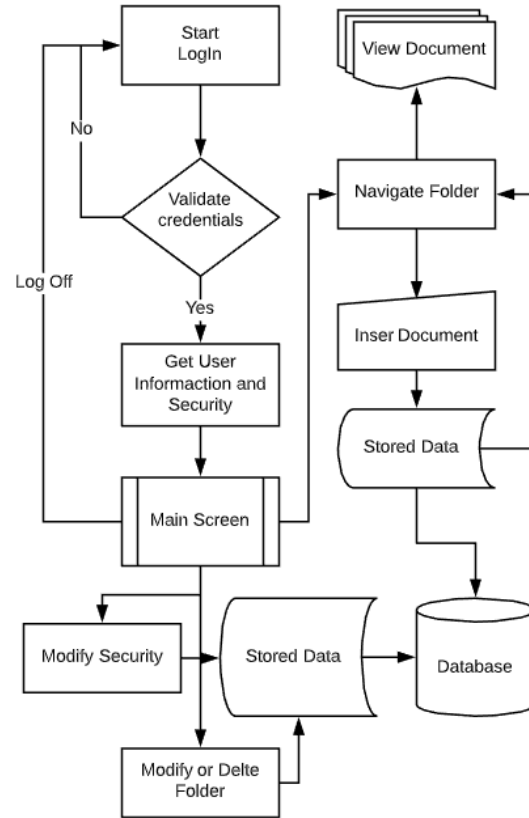

Figure 4
Internal Program Flow Chart

RESULTS

The Results section was started once all the necessary design implementation were completed. Here the software begins to be created, tested, and implemented for use from employees. This is also the final product that the user utilizes, with all its functions and part. It includes all primary parts that will compose the software and as well as key elements that are present in the software. For the software to be presented as a completed software

the three primary objectives determine by management had to be completed as well.

Security

The login screen (figure 5) is created as the main security filter.

Figure 5
Login screen

This page is the start of the software and with out the correct credentials the user will not be permitted access into the system. To validate the user input credentials the program generates a connection SqlConnection class string. (Code Example: *SqlConnection sqlCon = new SqlConnection (ConnectionString)*)

This provides the software with access to the database. After this Queries can be perform on the data on the database. The query that is utilize in this part of the code (figure 6) will return with a true or false depending if the user credentials are found.

```
SELECT CASE WHEN COUNT(*) > 0 THEN 1
                  ELSE 0 END
FROM dbo.Security
INNER JOIN dbo.Users
      ON Security.UserID = Users.UserID
WHERE FirstName +
            RIGHT('0000' + CONVERT(VARCHAR,
            Security.UserID), 4) = @username
      AND Password like @password
```
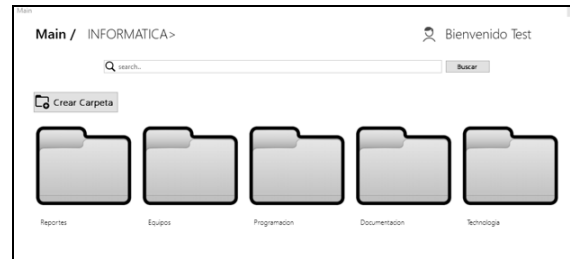
Figure 6
Login Query

For access or error in credentials to the application users must go to their designated supervisor and/or the security personal to allow access. This screen will evaluate the user's credentials compose of a User ID (name of user composes of four-digit number and a security password that is reset every 90 days). If the users have forgotten their credentials, the system will generate a ticket for assistant in the Information and Technology department to assistant. No single user will be able to create their own credential and only certain users will be able to modify other user credentials.

Document Storage and Accessibility

Following the structure of the flowchart, after successfully logging in, the user will be presented with the main screen (figure 7). This view will depend on the security access given to the user.

Figure 7
Main Folder screen

Not all user will have access to the same folders. The search tool helps the user to find documents that are indexed in the system, faster than searching manual documents. Depending on the security level, the user can create or edit folders in the main screen and other screens.

This screen is structure with three primary tables, they are three primary tables that are utilized to show the User the folder they are at. This hierarchy is structure in the fallowing order: Location, Drawer, and Folder. An example of the Location table (table 1) shows the folder name, date created, department that belongs to as well as the level of security that is required to access this table. All the other tables (Drawer and Folder) have a similar level of information each having a Department and an LVL.

First, the software prepares a SQL Store Procedure that provide us with the folders that the user can see. (Query Example: *EXEC dbo.[SP_FIND_FOLDER] @TypeFolder = @Parameter1, @NumUsuario = @Parameter2, @DepUsuario = @Parameter3,@locationID = @Parameter4, @drawerID = @Parameter5*) Base on the FolderType it will select from the previous table the necessary information. After the information has been found, the program will execute the *resetInformation*() Function this clean all the variables utilize by the software, this will prevent any errors from happening.

From here the user can navigate to all the location where documents are stored. All the information uploaded and edited is indexed in the database. Once you get to the last folder you will access the documents that are stored in the designated folder (figure 8).

Figure 8
Main Document screen

It provides an icon of the type of image with the designated name in the top of each image. This is the only location where you can insert an image, the image inserted are uploaded to the designated folder. There is also a search provided here, and navigation buttons in the bottom.

The image data can be seen stored in table 2, with in hexadecimal date located in the imagen field. A Select statement Query is utilized to get the information from the database. (Query Example: *SELECT DocumentID, TypeID, DateUploaded, UploadUser, StatusID, FolderID, Image, DocumentName FROM dbo.Document WHERE StatusID < 99 and Image IS NOT NULL AND FolderID = "+Folders.folder + " AND LVL <= " + LogIn.Userlvl*)

Figure 9 shows the Create Folder Screen as well as the database when a new folder is created.

Figure 9
Create Folder screen

Here the user can only assign a name to the new folder just like when a folder is created. This folder will be created in the same location where you are as set default. But if the user has the access it can create folders in other directories.

Once a folder is created, a modify button is presented every time the user hovers over the button. By clicking this button, the same window will be presented, but the options of modify or delete the folder is presented instead of the Create (figure 10).

Figure 10
Modify or Delete Folder screen

Here users can change or fix the name of the folder as well as change the location that the folder is in currently. Any change that is done to a folder as well as a file must be done with a message explaining why this change are been made to this specific folder. Changes done to any folders are store in the FolderHisto Table. Two queries will be utilized to effectively perform this action one SQLQuery Update instruction and a SQLQuery Delete instruction. The Delete instruction can be seen in figure 11. Depending on the level that the users is in the folders is the type of query that will be utilize. If you delete a Location folder that is show as a level 0 folder everything under it will be remove this includes the Drawer folders as well as the Folder folders. This same principle applies to the Drawer level that is 1 and the Folder level that is level 2.


```
if (Folders.nivel == 0)
{
    connection.query = "DELETE dbo.LOCATION  WHER
                        " DELETE dbo.DRAWER WHERE
                        " DELETE F FROM FOLDER F
}
else if (Folders.nivel == 1)
{
    connection.query = "DELETE dbo.DRAWER WHERE D
                        " DELETE dbo.FOLDER WHERE
}
else if (Folders.nivel == 2)
{
    connection.query = "DELETE dbo.FOLDER WHERE F
}

SqlCommand command = new SqlCommand();
```

Figure 11
Delete Folder Query

Audit Trails

By double-clicking any image an image viewer is presented through a plugin call PDFium [8] a document audit trail (figure 12) will be presented with basic information of the image.

Figure 12
Document Audit Trail

The image is presented utilizing a *MemoryStream*, tis will create an array of bits that will help us create a Bitmap of the hexadecimal data that is store in the *varbinary* field in the database. The document Audit trail that is achieve by creating a query, give us the necessary information from the document to the document history table (DocHistory).

(Query Example: *SELECT CONCAT (dbo. Users.FirstName,'', LastName, ' #000', DocHistory. UserID) AS [UploadUser], Description, CONVERT (VARCHAR (10), Fecha (110) AS fecha, Comment FROM dbo.DocHistory INNER JOIN dbo.Users ON Users.UserID = DocHistory.UserID INNER JOIN dbo.Actions ON Actions.ActionID = DocHistory.ActionID WHERE DocumentID = @Parameter1 ORDER BY DocHistoryID)*

This is utilized for user's knowledge of movements or memos in documents from other users. For supervisor or audits it will be utilized to obtained valuable document movement information and see the changes on the documentation through time.

The last security item that is included directly for users is the User Security Page (figure 13). Here security specialist will be permitted to add, modify, and deleted personal in the system, apart from the administrators of the system.
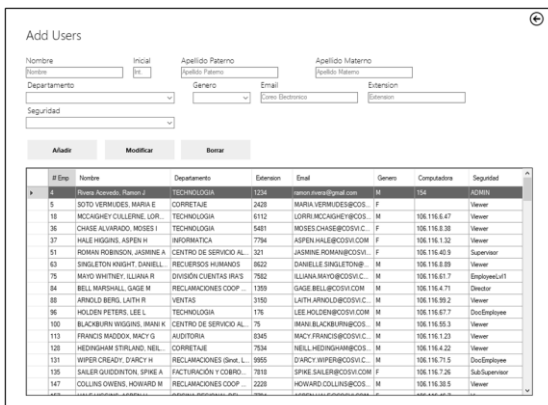

Figure 13
User Security Information

As a safety purpose the user can't delete itself from the system. Also, the system will not provide the tools to change primary user information like employee number, primary name and last name and extension. For these functionalities you will have to call the IT department.

By double clicking the label that includes the name of the user and a welcome message (*Bienvenido, Ramón*) in the top right corner a user information window that all users can see will be presented to them (figure 14). From here the user can Logout of the screen as well as see their current information and their activities. These activities are retrieved by making a query to the database with the UserID.

Figure 14
User Audit Trail

Completion

In the completion of this project there were two key elements that had to be presented to fully complete the project. One was the training manual. A User-Manual and a Technical-Manual was made to give proper training to all employees of the company.

In the second term we must provide a report to management proving that we have completed the expected milestone that were stablish by the software in the beginning. This includes the stablish three criteria:

- Document/Information Storage and Accessibility
- Log Maintenance and Regulation
- Document/Information Security

The document information storage and accessibility were stablished by providing a united database system that employees will store all the information in and giving them access through the application to this data has also given us a 25% boost in productivity.

Thanks to the audit logs that were explain before we completed with the second requirement from the stakeholders by providing a log maintenance and regulation. Supervisors as well as

employees can oversee their movements and document that they have perform giving them extra security in the action that they have done.

Finally, for security, by including the logging mechanism and page, as well as an internal level of security to access document. We have eliminated the fact that users will see the information that they should not be seeing. The removal of the function to delete images from the system only hide it from the users also provides a security measurement that no image can really be delete from the system and lost. This also helps us with the goal that was stablish of 10% lost in document.

We the inclusion of the PDFium or PDF viewer system the need for printing copies of document has been almost completely remove and users can now generate a copy of the document and send it to other employees by email saving the company the 60% in paper lost or misuse.

Maintenance

For effective and seamless support to the application features from Visual studio will be utilize in the deploy procedures. One will be a default directory in the server for installation of the deploy files, and secondly will be automatic update. This will instruct the software to look for updates on every time it starts or on every time it its close and will automatically install any update without the necessity of no personal involve.

Conclusion

Even though the software has not eliminated the physical object or hard copy, the risk of loss or theft has been scientifically lowered. Thanks to the security implemented the supervisor have been able to monitor document entry to the system and see how users utilize this information. Also base on the findings a phase two has been stablish, some future updates and improvements to the security and user functionalities in the system has been added to this phase.

Since the feedback has been positive, the company is moving forward with expanding the

Digitalization Application to the mobile application domain. It has been determined that the expand in the data storage system will become required and stablish a more frequent backup cycle for contingency. Some of the shortcomings where the amount of time left with a gap in the accessibility that will be talking in phase two also required a bit more tuning in the security department in the project and a group level security for easier management and complicated levels of access. Some previous errors with users uploading blank documents and corrupted file.

Some future work that are plan for the next phase, a mobile app in Xamarin [9] and/or web application will be introducing. In the mobile application the utilization of the mobile device camera for scanning documents and uploading images is a strong upgrade. To ensure the upload users integrity a digital signature will be implemented as a watermark to identify the person that upload the document as a visible mark. Other improvements that are been requested by personal is the inclusion of in application reports, document rotation to a Upright portrait mode and document word scanning to help with searching text found in documents (this could include reference number and lots number).

References

[1] M. Knight, "What is Structured Query Language (SQL)?", Dataversity, August 21, 2017. [Online]. Available: https://www.dataversity.net/structured-query-language-sql/

[2] Microsoft, "Windows Forms," March 30, 2017. [Online]. Available: https://docs.microsoft.com/en-us/dotnet/framework/winforms/

[3] C. Sells, Windows Forms Programming in C#, 1st ed. Boston: Addison-Wesley Professional, 2003.

[4] Techopedia, "Database Management System (DBMS)," August 30, 2019. [Online]. Available: https://www.techopedia.com/definition/24361/database-management-systems-dbms

[5] Computer Business Review, "What is SQL Server?", July 12, 2020. [Online]. Available: https://www.cbronline.com/what-is/what-is-sql-server-4914415/

[6] W. Kent, "A simple guide to five normal forms in relational database theory," Communications of the ACM, vol 26(2), pp. 120-125, Feb. 1983. [Online]. Available:
https://www.cs.dartmouth.edu/~cs61/Resources/Papers/CACM%20Kent%20Five%20Normal%20Forms.pdf

[7] Microsoft, "Transparent Data Encryption (TDE)," May 9, 2019. [Online]. Available:
https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/transparent-data-encryption?view=sql-server-ver15

[8] Patagames, Vladivostok. C# PDF Viewer, ver. 4.30.2704. [Online]. Available:
https://pdfium.patagames.com/csharp-pdf-viewer/

[9] Microsoft, "Xamarin." Accessed May 1, 2020. [Online]. Available:
https://dotnet.microsoft.com/apps/xamarin