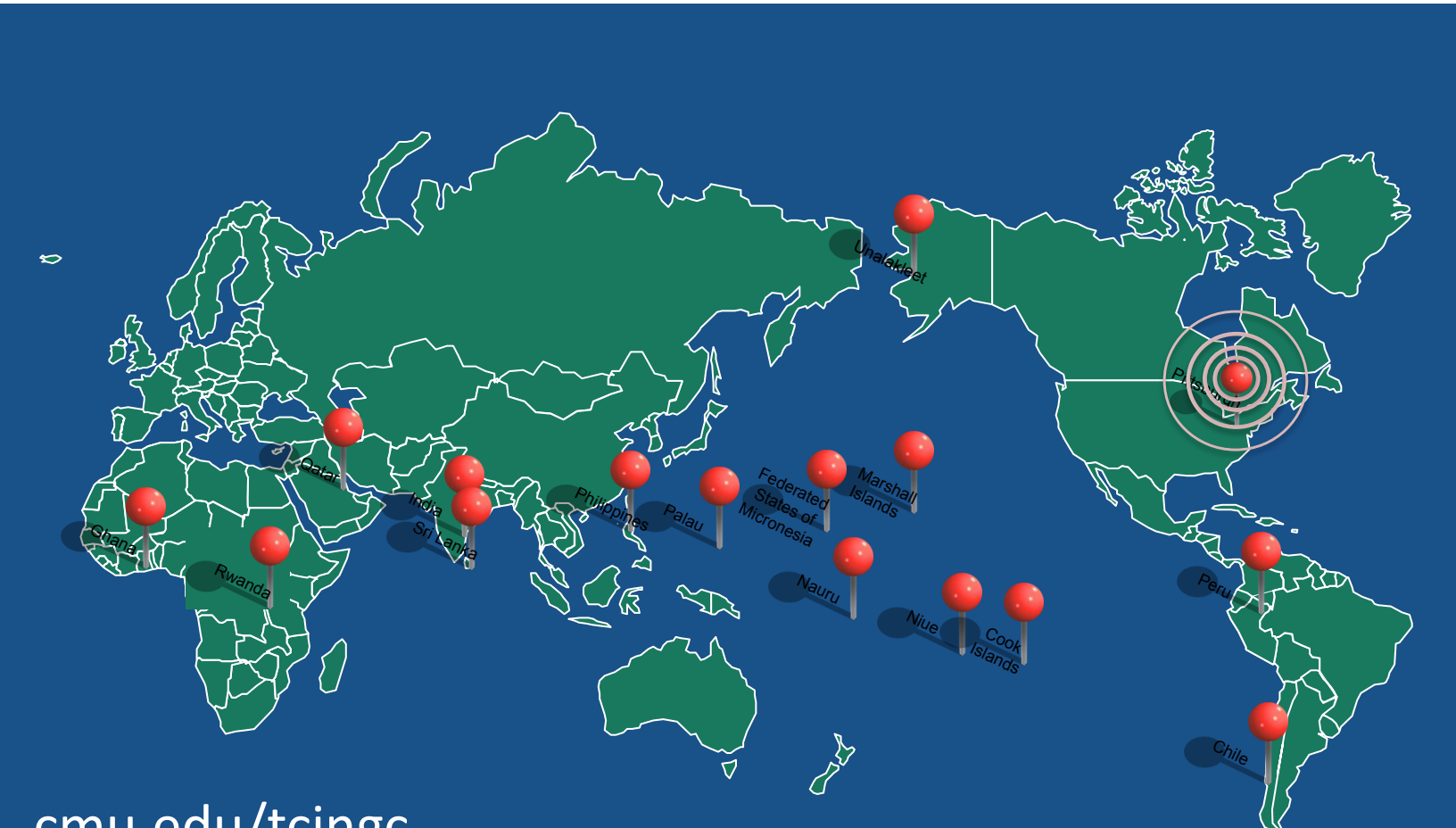
TECHNOLOGY CONSULTING
IN THE GLOBAL COMMUNITY

Final Consulting Report
Palau Financial Institutions Commission

Sunday Zhou

August 2023

Carnegie Mellon University



cmu.edu/tcingc

Palau Financial Institutions Commission
Executive Summary

Student Consultant, Sunday Zhou
Community Partner, Cjay Shiro

I. About the Organization

The Palau Financial Institutions Commission (FIC) is an independent regulatory agency, and a six-member board governs this organization. The FIC was established in 2002 and is located in Koror. FIC has two full-time employees and shares an administrator with FIU. The mission of the organization is as follows:

> *To establish and maintain a financial regulatory and supervisory system for the Republic of Palau, consistent with international standards, which promotes a stable banking and financial sector that facilitates economic growth and development*

FIC has two primary responsibilities, namely bank regulation and the provision of corporations' registry services. The latter responsibility has been recently assigned to the FIC. Presently, the FIC's funding is derived from regulation fees collected from the five banks operating in Palau, amounting to a total annual budget of 1 million dollars. This funding is allocated to support the FIC's day-to-day duties related to bank regulation. Starting in October 2023, the FIC will receive additional funding specifically for the corporations' registry. In general, the organization is adequately funded to support its ongoing operations.

However, the FIC does face certain limitations. Firstly, there is a constraint in terms of office space availability, as the current FIC office is fully occupied. Secondly, the employees of the FIC are operating at maximum capacity, as both individuals are consistently occupied with their daily responsibilities.

II. Assist with Scanner, Desktop, Synology Hard Drive Connection Set Up

Cjay was keen on establishing a comprehensive workstation to cater to all facets of their role as a new central registry for corporations. I provided assistance in choosing and authenticating the scanners and ensured that all interconnections among the various components were properly established.

- **Outputs**: A work station for scanning and looking up files. The main components of the working station include:
    - A commercial scanner
    - A Dell computer
    - The Synology hard drive
- **Outcomes**: Tikei has a designated work station for managing all legacy documents.
- **Recommendations:** Be very cautious not to break the connection between the scanner and the desktop, and between the desktop and the Synology hard drive. Do not change the name or the position of the folder `Scanned Folder` on Dell Desktop, and do not remove Dell Desktop from the scanner. Additionally, do not change the name or the position of the folder `CMU Consultant` on Dell Desktop, the Dell Desktop should stay connected to the FIC Wifi all the time, etc.

III. Create a Database with Search, Filter, and Sort Features

FIC expressed a need for a database comprising significant field values gleaned from each company's annual report, enabling a rapid lookup for any specific company. This database is required to be equipped with search, filter, and sorting capabilities.

- **Outputs**: An Excel sheet stored in Synology hard drive with built-in backup capabilities.
- **Outcomes**: The office now has an accessible, maintainable, and robust database that the team is comfortable using.

IV. Develop a Pipeline to Automate Text Extraction and Files Upload

I designed and developed a streamlined pipeline that automates the process of uploading scanned documents onto the Synology hard drive, extracting pertinent values from these documents, and inserting this data into the database. This pipeline also links each row of data to the corresponding scanned document's file location.

- **Outputs**: The pipeline consists of six scripts that handle tasks such as file format conversion, text extraction, file uploading, the creation of new rows in Excel, and system cleanup. Notably, it is designed with robust fault tolerance and error handling mechanisms in place.
- **Outcomes**: The automation process I've developed streamlines the manual tasks of scanning, relocating files, interpreting data, and inputting them into a database. As a result, it has been able to expedite the overall process by a minimum of 60%.
- **Recommendations**: I suggest scanning no more than 10 files at a time, even though the script can handle an unlimited number in one go. Keeping the number small will ensure each file is manageable during the review phase. If you encounter any errors or issues, please refer to the 'Tutorials' section in Appendix A. In addition, after all documents' scanning are completed, feel free to change the format (color, font, font size) of the Excel sheet.

Consulting Partner
Cjay Shiro
commissioner@ropfic.org
Financial Institutions Commission
P.O. Box 10243
Koror, Palau 96940
https://ropfic.org/

About the Consultant
Sunday Zhou
sundayzhou0225@gmail.com

Sunday is a senior in Information Systems and Computer Science. She will begin working as a software engineer for Apple this fall.

Palau Financial Institutions Commission
Final Consulting Report

Student Consultant, Sunday Zhou
Community Partner, Cjay Shiro

I. About the Organization

The Palau Financial Institutions Commission (FIC) is an independent regulatory agency, and a six-member board governs this organization. The FIC was established in 2002 and is located in Koror. FIC has two full-time employees and shares an administrator with FIU. The mission of the organization is as follows:

> *To establish and maintain a financial regulatory and supervisory system for the Republic of Palau, consistent with international standards, which promotes a stable banking and financial sector that facilitates economic growth and development*

FIC has two primary responsibilities, namely bank regulation and the provision of corporations' registry services. The latter responsibility has been recently assigned to the FIC. Presently, the FIC's funding is derived from regulation fees collected from the five banks operating in Palau, amounting to a total annual budget of 1 million dollars. This funding is allocated to support the FIC's day-to-day duties related to bank regulation. Starting in October 2023, the FIC will receive additional funding specifically for the corporations' registry. In general, the organization is adequately funded to support its ongoing operations.

However, the FIC does face certain limitations. Firstly, there is a constraint in terms of office space availability, as the current FIC office is fully occupied. Secondly, the employees of the FIC are operating at maximum capacity, as both individuals are consistently occupied with their daily responsibilities.

Programs

The FIC primarily encompasses two key programs and responsibilities. Firstly, the FIC is entrusted with the crucial task of overseeing, supervising, and regulating Palau's financial sector. This entails evaluating proposals for the establishment of new banks on the island and conducting comprehensive on-site and off-site monitoring of existing banks.

Secondly, the FIC is responsible for the management of the Palau Corporate Registry, which serves as a registry platform mandatory for both local and foreign-owned businesses intending to operate in Palau. The purpose of this registry is to ensure proper identification of the business model, foundation, operations, procedures, and policies of registered entities. Currently under development, the Palau Corporate Registry platform is expected to be deployed for production use by December 2023. At present, all corporate information in Palau is maintained in hard copy format, and the FIC has been assigned the responsibility of managing this legacy information.

Staff

Mr. Cjay Shiro- Executive Commissioner
Mr. Shiro joined the FIC team in March 2020 as a Bank Examiner. He holds an Associate's Degree

in Liberal Arts from Kapiolani Community College and a bachelor's in Business Administration and Marketing from the University of Hawaii. He is currently working on his Master of Science in Finance from the University of Hawaii Manoa. Mr. Shiro has prior experience in commercial insurance and finance. He also has certified training in operational analysis and risk management in anti-money laundering. Mr. Shiro brings additional management accreditation from his ten-plus years of active-duty service in the United States Military. Mr. Shiro conducts onsite examinations and performs off-site analyses of reporting financial institutions from quarterly prudential returns.

Ms. Valorie Tikei Sbal- Office Manager
Ms. Sbal joined the FIC in 2010 as the office manager and in January 2013 became the filing officer for the secured transaction registry for the Republic of Palau. Ms. Sbal is responsible for the maintenance and organization of the FIC's administrative tasks, duties, and documentation.

Technology Infrastructure

Hardware	● 5 Operating Computers ● 4 Operating Laptops ● 4 Printers ● 1 Internal Network ● 4 Mobile Cell Phones
Hardware-Specifications	● Processors – All machines have Core i3-i7 processors ● Ram-8-12 GB ram ● Storage- 4 TB of total storage
Software	● 9- Microsoft Office Licenses in 5PC's and 4 Laptops ● 8-Antivirus software packages in 4 PC's and 4 Laptops ● 8 Security Encryption certified-4PC's/4Laptop's ● 8 Internet Security software installed-4PC's and 4Laptop's
Internet connection	Wifi

Operating System	All machines have Windows 10 Operating System installed.

Technology Management

Cjay assumes primary responsibility for managing the technology infrastructure within the FIC. This entails crucial tasks such as data backup and software installation.

Daily backups are diligently performed from the Network Access Storage (NAS) to a secondary drive, as well as two USB drive passports. One of these passports is securely locked in a fire- and weather-proof safe located within the office premises. The other USB drive is kept off-premise. Before transporting these USB drives outside the building, Cjay dutifully records this information in a logbook maintained by the Office Manager. Upon Cjay's return to the office each morning, the logbook is once again updated with the Office Manager to ensure a comprehensive record of the process. This diligent routine is adhered to consistently, including on weekends.

Since joining the FIC, Cjay has taken the initiative to acquire licensed Microsoft software, ensuring compliance and legality. Furthermore, he is tasked with procuring any necessary hardware required by the organization.

Technology Planning

Technology planning at the FIC follows a relatively informal process. Cjay is allocated a budget of approximately $20,000 annually for IT purposes. In cases where additional funding is required, the FIC has the option to apply for funds from the Palau Government. When there is a need for new technology, Cjay approves the purchase and subsequently informs the Office Manager, Tikei, to initiate a purchase order through the government procurement office.

To ensure a competitive and transparent procurement process, the FIC submits three different quotes for the desired items to the Procurement Office. In instances where the required item is not available locally in Palau, alternative means such as utilizing online services like Amazon are employed to facilitate the procurement process.

Communication

The FIC utilizes various communication methods to interact with external parties. Email and telephone are the primary means of communication for external engagements, while virtual meetings are conducted via Zoom. For internal communication purposes, email is the preferred mode of correspondence. In-person meetings take place at the FIC's physical office when engaging with the Governing Board. Additionally, for more informal group messaging among employees, a Signal group is utilized as it offers a secure communication channel.

Internal file sharing is facilitated through the use of the Synology Network Access Storage (NAS) drive. Staff members are provided with email accounts through webmail (https://www.ropfic.org/webmail), ensuring that staff email communication is encrypted and

secure. When communicating with banks, the bank examiner employs encrypted email communication when interacting with the two non-FDIC-insured banks. For accessing prudential return forms from FDIC-insured banks, a bank portal is utilized through a provided link.

Information Management

Important information managed by FIC:

1. Bank regulation information
2. Local corporations' registry information

Everything related to the first type of information was taken care of by the 2019 TCinGC consultant and 2020 TCinGC consultant.

Currently, all corporates' business information submitted before December 2023 is in hard-copy form. The physical forms and folders are stored in an office that is 50-minute drive away from the FIC office. Whenever a new request is received to review the information, an FIC employee has to drive all the way for the file, bring it back to the FIC office, and drive to put back the file after it's used.

Business Systems

The office manager uses Quickbooks pro for bookkeeping within the FIC office.

II. Provide Assistance for Establishing Connection between scanner, desktop, and Synology hard drive

Motivation

Cjay was looking to establish a comprehensive workstation tailored to accommodate all aspects of their new role as the corporate registrar. The key tasks encompass document scanning, executing scripts for value extraction, uploading scanned documents to the hard drive, and dealing with future inquiries from corporations. The workstation is intended to cater to these requirements; facilitating any task pertinent to the new role, offering easy access to related documents, and enabling the creation of authorization assignments.

The purpose of this workstation isn't limited to scanning legacy documents. Once the new corporate registry platform is officially implemented in December, the workstation will also manage tasks related to this new platform.

Upon discussion with Cjay, I concurred with the concept of inaugurating a dedicated workstation, given the magnitude and escalating importance of their new role at FIC. I recognized the significance of a smooth interconnection between the scanner, computer, and the Synology hard drive, which boasts substantial storage capacity for document archival. This connection forms the fundamental hardware infrastructure for file transfer, script execution, and file upload operations.

Process

Evaluate the volume of existing legacy documents: This is a crucial step in determining the necessary storage space on the hard drive, the scanning speed, and the computational power required for the computer handling the processing tasks. I carried out an upper-bound estimation by tallying the number of companies, the years of annual reports per company, the number of pages in each annual report, and the storage space required to accommodate a PDF file of a certain page number.

The number of companies	Years of annual reports	Page count of each annual report	Space needed to store a 30-page pdf
350	6	30	2MB

The number of companies	Page count of other documents	Space needed to store a 100-page pdf
350	100	4MB

Total storage space:

350 * 6 * 2 + 350 * 4 = 5600 MB = 5.6 GB

Evaluate the adequacy of current physical devices: The current scanning capabilities of the office scanners were inadequate when juxtaposed with the volume of legacy documents estimated in the initial phase. The scanning process was considerably sluggish, averaging roughly 5 seconds for each page. Additionally, there was no available computer within the office setup to efficiently handle the new scanning and processing responsibilities. Consequently, Cjay and I have resolved to invest in new hardware, particularly a commercial-grade scanner and a state-of-the-art computer.

Exploring local electronics stores in Palau: Cjay and I embarked on an exploration of five electronics stores (Master's electronics, WCTC, two Surangles, and Globus), seeking the most advanced scanner and computer obtainable on the island. When we discovered potential devices, we solicited quotations and specifications from the respective stores for later comparison. It became clear to me that beyond speed, the readability of text and image clarity after scanning were paramount. As such, we made subsequent visits to Globus and WCTC to assess the scan quality of various commercial scanners. We requested store owners to perform demonstration scans of text faintly inscribed on paper. After a thorough evaluation of all factors, we chose a Canon C38261 commercial scanner from Globus and a 12th Gen Intel Core i7 Dell computer equipped with 16 GB RAM.

Installation and device interconnection: The newly procured devices needed to be connected to the office's internet network and set up to transfer files and communicate seamlessly. The primary requirement is for the scanner to transfer scanned files to the computer and for the computer to upload these files to the Synology hard drive. Working alongside the installation technician, and with ChatGPT's support, I was able to gather the IP addresses of the different devices, set up the new scanner, register the new computer with the printer, and create a shortcut for the Synology hard drive on the Dell desktop.

Outputs

Components

1. Dell computer; 12th Gen Intel Core i7, 16 GB RAM.
2. Canon; ImageRUNNER, ADVANCE DX, C38261.

Connections

1. Dell computer is recognized as “Dell Desktop” on Canon scanner.
2. Synology hard drive is recognized as “192.168.1.200” short-cut icon on Dell Desktop, and Synology folder for storing scanned legacy documents is marked as “CMU Consultant” on Dell Desktop.

Upon configuring the devices and establishing the necessary connections, the entire workflow encompassing document scanning via the scanner, processing on the Dell computer, and uploading to the Synology hard drive can be automated using Python code.

Outcomes

Scanning

- Speed

The table below states the scanning speed of the old office scanner and that of the new commercial scanner.

Old Scanner Canon imageCLASS MF644Cdw	New Scanner imageRUNNER ADVANCE DX C3826i
14 ipm	80 ipm

The new scanner is more than five times faster than the old scanner. The new scanner significantly speeds up the process.

- Legibility

The new scanner exhibits superior capabilities, such as the ability to detect and scan pencil marks, stamps, and documents printed with light ink, which the previous scanner failed to recognize. Furthermore, it has the added functionality to scan color documents. As a result, the quality of the scans has improved significantly, a vital enhancement for the subsequent OCR text recognition stage.

Dell computer

By dedicating all of the new computer's resources (CPU, RAM, storage, etc.) to these specific tasks, there is no need for context switching to perform unrelated tasks. This enhances the processing speed for running scripts I have written. Furthermore, it prevents any potential confusion between FIC's varied responsibilities. It's clear that all the documents on this computer are critical for FIC's role as a corporate registry.

Recommendations

The primary goal is to ensure that the connections among the devices are sustained. Consequently, the FIC office must maintain the integrity of all IP addresses and the names of all files, folders, and devices. However, Cjay has a trusted IT technician who is well-acquainted with the office's entire network and device setup. Therefore, I am confident that any connection issues that may arise can be promptly and efficiently resolved.

III. Create a Database with Search, Filter, and Sort Features

Motivation

They currently arrange all physical files in alphabetical order. However, it's common to have over 30 corporations falling under the same initials, making it quite cumbersome to locate documents for a specific corporation. The ultimate goal of this digitization project is to create a database with search, filter, and sort capabilities. This will enable FIC employees to retrieve specific documents

easily once all the information is input. The database will be used primarily when a corporation requests to review its information, which happens once or twice per month. Consequently, the database does not need to support high-volume queries. Apart from the search, filter, and sort functionalities, no additional features are necessary, as all extracted information will be reviewed and corrected during the data entry stage.

Process

The process of selecting the most appropriate database involves extensive research and communication with various parties along the way.

Initially, I considered **Azure SQL Database** and **Azure Cloud**. FIC handles highly confidential corporate information, and security is one of Azure's strongest assurances. Additionally, Azure boasts high availability rates of 99.9%, according to their official website. The service provides a wealth of customizable options, such as the number of cores for computation, storage space, varied pricing plans, authorization, and more. Azure is also flexible, allowing for modifications at any time. For instance, we could effortlessly scale up or down the storage space and computation resources of the database whenever needed. Another significant advantage of Azure is its suite of development tools. Azure Data Studio presents a clean interface and incorporates built-in search, filter, and sort features for each query. I test-registered an Azure account, configured a database, created data, ran various queries, explored **Azure Data Studio**, tested its search feature, and all seemed well until I attempted to register an account on behalf of FIC. Unfortunately, Azure's accepted billing addresses for individual and corporate accounts do not include Palau. Although there are workarounds, such as government offices in Palau using a US address, this would make FIC dependent on an external entity for the duration of database usage. I didn't find this entirely sustainable or worth the hassle.

I solicited advice from John, CEO of Paradigm Applications, who is in charge of building the new registry platform for FIC. He suggested SQL Express due to its free nature, thus eliminating the billing address concerns. However, **SQL Express** has a storage limit of 10 GB. Considering the scanned files of 350 companies, along with the storage needed for configurations, this limit might be proved insufficient. Upon expressing my concerns to John, he mentioned that a common practice when dealing with large files in a database is to store file pointers, with the files themselves stored in separate file storage.

Subsequently, I created a local SQL Express database for testing purposes, setting specific columns for extracted values and a final column for file pointers of scanned files uploaded to the hard drive. In addition to Azure Data Studio, I also experimented with **SQL Server Management Studio** (SSMS).

During this testing process, an intriguing thought crossed my mind. The query features I was scripting and planning to instruct Cjay and Tikei to use are basic functions in Excel—a tool with which they are already extremely familiar. I quickly consulted **ChatGPT** about the benefits of using SQL Express versus Excel. As evident from the chart below, almost all advantages of SQL Express are not relevant to our use case, while all benefits of Excel are fitting.

SQL Express

Advantage	**Applicable or not and reason**
Handling large volume of data	**Not applicable.** FIC, at most, has 3000 rows of data.
Concurrent access	**Not applicable.** There is only one commercial scanner, so this bottleneck defines that there is only one person scanning files to the database at one time. Given the low corporation request frequency, concurrent access is also not needed

	for the future.
Query capability	**Not applicable.** FIC is only interested in very basic features like search, filter, and sort. Complex queries are not needed.
Security	**Applicable.** SQL Server Express has robust security features that can help protect data, including user-level permissions and encryption. However, similar features also exist in Synology hard drive. By storing excel databases in hard drive, FIC also has strong security guarantee and authorization management capability.
Automation and scripting	**Not applicable.** The rest of the pipeline is written in Python.

Excel

Advantage	**Applicable or not and reason**
Ease of use	**Applicable.** Excel's interface is intuitive and user-friendly. It is also a tool that both Cjay and Tikei are very familiar with.
Portability	**Applicable.** Excel files are widely used and can be opened on virtually any device. They are easy to share via email and all other communication channels. It can be directly stored in Synology hard drive.
Built in search, filter, and sort capabilities	**Applicable.** Excel has built in powerful features. FIC can now search, sort, filter for anything without changing or running SQL queries.
Flexible and customizable	Applicable. Excel spreadsheets can be easily customized for various layouts and formats. This includes color coding, conditional formatting, merging cells, and more.

After evaluating the comparison above and discussing with Cjay about their familiarity and comfortability with **Excel**, I chose to use an Excel spreadsheet as the database for storing all extracted key values and file pointers for the scanned files stored on the Synology hard drive. A significant advantage with this choice is that not only can the scanned files be stored on the Synology hard drive, but the database itself can also be housed externally rather than on the Dell computer.

The Synology hard drive comes with automatic backup, encryption, and authorization management features, which are highly beneficial for our use case. I evaluated Excel's performance considering our data volume, and the results confirmed that Excel can handle the current data load (350 * 6 = 2100 rows of data) without issues. Furthermore, the search, filter, and sort results appear instantly, ensuring efficient operations.

Outcomes

The outcome is quite straightforward: an Excel spreadsheet housed on the Synology hard drive. This solution is simple, flexible, sustainable, and sufficient. This spreadsheet offers a clean and user-friendly interface with easily navigable search, filter, and sort features. In addition, it comes with the advantage of automatic backup.

Recommendations

It's essential to regularly check that the Synology backup is functioning without issues. In addition, I recommend archiving all physical forms instead of discarding them before another backup layer is set up. If the billing address issues can be resolved, it might be worth considering moving files stored in Synology to cloud storage as an additional backup guarantee. Cloud storage providers typically maintain servers in various clusters across multiple countries, and usually offer five or six backup layers. With such measures in place, it should then be safe to completely discard the physical documents.

IV. Develop a Pipeline to Automate Text Extraction and Files Upload

Motivation

Cjay emphasized from the outset that they want a searchable database, with field values derived from each company's annual report. The bulk of my project involved designing and developing an easy-to-use pipeline for scanning documents, extracting key values, uploading them to the Synology hard drive, and inserting new rows of data into the Excel database.

The pipeline needs to be flexible and customizable. It has to extract key values from targeted files while also allowing FIC administrators the ability to input additional information or rectify any populated data or files.

Importantly, the pipeline has to be user-friendly. Tikei, who will primarily be in charge of scanning all documents, has limited technical experience. Consequently, my aim was to design a pipeline that requires minimal technical knowledge from the operator, automating most of the process to streamline tasks.

Process

The pipeline has three objectives:
1. Upload the scanned documents to Synology hard drive
2. Extract key values from scanned documents
3. Insert extracted values along with a generated file pointer to Excel.

The trio of objectives has been realized through the execution of six scripts detailed in the Outcomes section.

Of the three objectives, the extraction of key values from scanned documents posed the most significant challenge. My research on various OCR technologies led me to select Google Tesseract, a widely recognized and well-maintained tool. Tesseract's Python-compatible library, pytesseract, offers a clean and well-designed interface. Given that pytesseract processes PNG files rather than PDFs, a format conversion step was necessary, as the scanned documents were originally in PDF format.

I employed several strategies to enhance text recognition accuracy.

In an attempt to improve recognition accuracy, I tested the Open AI API and ChatGPT API. My experimentation with various prompts revealed a three-part prompt – a request to correct spelling and spacing issues, an example of raw detected text and corrected text, and another raw text needing correction – consistently yielded the best results. While these tools improved the overall spelling accuracy of the report, they were less effective in correcting the key fields of particular

interest to my community partner. Given the cost associated with each request, I ultimately decided against using these tools.

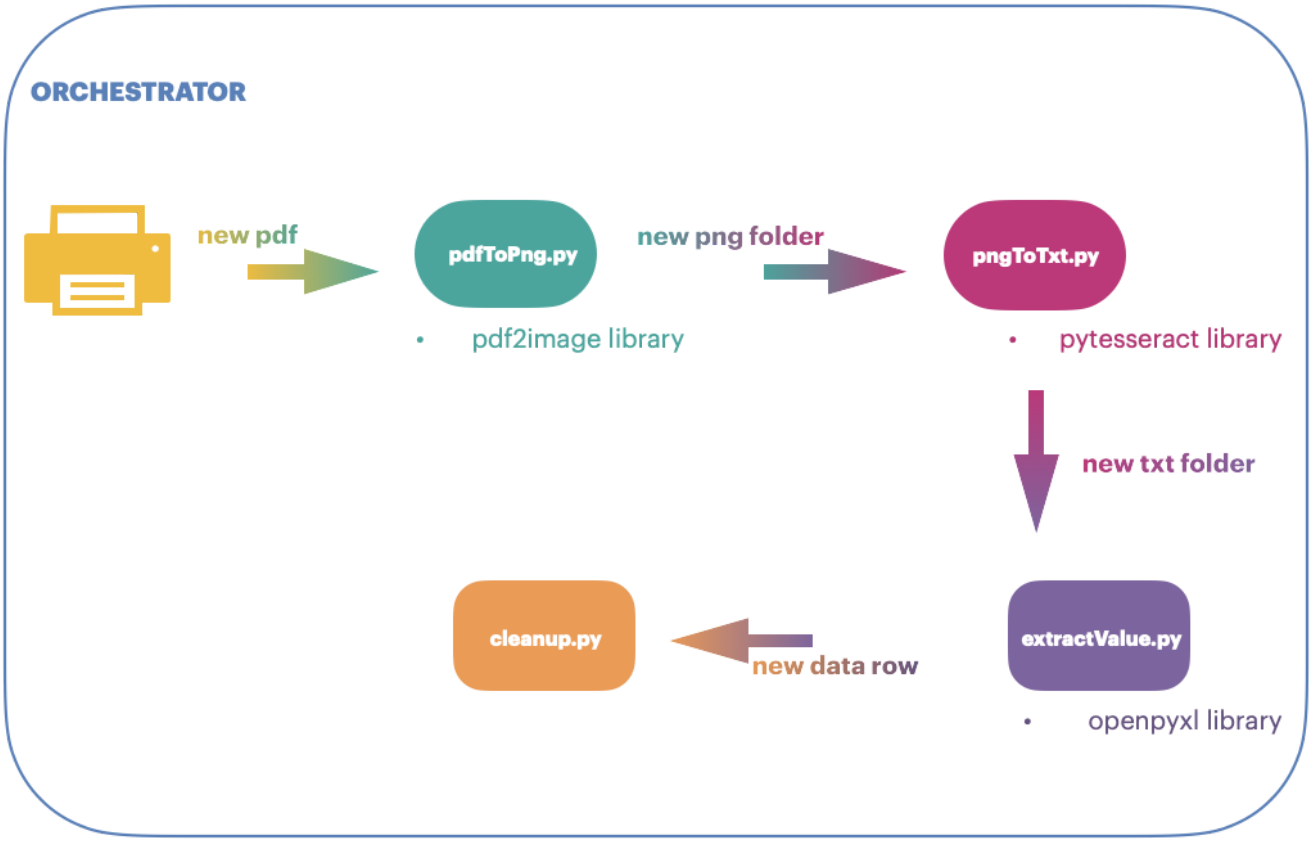
Furthermore, I discovered that changing the orientation of the scanned paper significantly enhanced recognition accuracy. By applying the OCR technology to correctly oriented papers, I was able to improve recognition accuracy by at least 50%. I found out that this is due to the underlying algorithms and models being trained primarily on well-structured, clear, and correctly-oriented data. Therefore, skewed or rotated documents may not be recognized as accurately because they deviate from the training data's norms.

Given that annual reports generally adhere to a key-value format, I looped through lines to identify keys and extract their corresponding values. The main challenge arose from the variability in annual report formats issued by FIC in different years, along with the modifications that some companies made to the base FIC layout. Moreover, variability in the font and font size used by companies sometimes led to errors in value detection. I designed my script to accommodate these variations. The strategy involves identifying all potential formats for each field and then having the script loop through each possibility until it found a match, at which point it would parse the field value based on that format. For fields with short names that often appear on a single line, I parsed by space, while for fields that typically had longer responses, I parsed by line.

Upon extracting the values, my script transferred the file from the local Dell Desktop to the hard drive, generated a new row of data with a file pointer, and pushed this data to the Excel database on the hard drive. It then allowed FIC staff responsible for scanning documents to closely review the data row for potential undetected or incorrect values. Rather than manually searching for the original scanned files, they could access the relevant file directly through the file pointer in the Excel sheet.

Since Cjay and Tikei lack extensive software experience, I prioritized ease-of-use in the design of the program. I created a separate script to orchestrate the entire process, managing all temporary files generated, providing these files for sequential tasks, and cleaning up intermediate temporary files. As such, they only need to double-click on the orchestration file to initiate and complete the process for each scanning round. The orchestrator ensures that the folder status is reset to default after each round.

Upon completion of the pipeline, I conducted stretch testing by running the pipeline against 30 randomly picked documents from an additional six boxes provided by Tikei. This testing surfaced several parsing issues, leading me to refine my program to handle these edge cases.



At the conclusion of my testing phase, my program had developed the capacity to extract the majority of information across different years and from various companies.

Outcomes

The project outcome consists of six scripts:

1. **pdfToPng.py**: Converts a PDF file into a collection of PNG files, with each PNG file representing a single page of the PDF. This creates a temporary folder for each scanned document in the "pngFolder" subfolder.

2. **pngToTxt.py**: Converts a collection of PNG files into a corresponding collection of TXT files, with each TXT file representing a single PNG file. This creates a temporary folder for each scanned document in the "txtFolder" subfolder.

3. **correctSpellingChatGPT.py**: Sends requests to the ChatGPT server for spelling correction. Sends one request per TXT file and updates the TXT files with the corrected text returned by the

ChatGPT server. This script is currently not in use for reasons explained earlier, but may be beneficial if comprehensive text extraction from the annual reports is needed in the future.

4. **extractValues.py**: Extracts key values useful to my community partner for searching the corresponding annual report of any company. The key values include:

- Organization Name
- Tax Identification Number
- Year
- Names, mailing addresses, and citizenship of directors
- Number and expiration date of the Foreign Investment Permit and/or Foreign Investment Approval Certification of the organization
- File link

"Company Shareholders" field for manual input due to inconsistent formatting across companies. This script also generates a new row of data and uploads it to the Excel spreadsheet stored on the Synology hard drive.

5. **cleanup.py**: Clears any intermediate temporary files and transfers the scanned documents to the Synology hard drive's "Legacy Scan" folder under the CMU Consultant directory.

6. **ORCHESTRATION.bat**: Orchestrates all other scripts to run in sequence. Double-click the script to start running the pipeline. Each run extracts values from all files in the Legacy/PDF FOLDER, creates a new row of data for each file, and extracts values accordingly.

These scripts streamline the process for my community partner, saving them the time and effort required to 1) manually transfer scanned documents to the Synology hard drive, 2) manually input all five fields of data, and 3) manually link each file to each row of data. The process only requires human intervention during the scrutiny phase. It is estimated to expedite the process by approximately 60% on average, once my community partner becomes more adept at managing the pipeline and can handle scanning and running the scripts synchronously.

Recommendations

Follow the video and text tutorials in the Appendix closely for using the pipeline. Please don't change the names and locations of the following files and folders:

On Dell Desktop:

- CMU Consultant
  - Legacy Scan
  - Legacy.xlsx

- Legacy
    - ANNUAL REPORTS
    - OTHERS
    - RECOVERY
    - Release-23.05.0-0
    - development
- Scanned Folder

For troubleshooting and Q&A for the pipeline, also see Appendix.

About the Consultant

Sunday is a recent graduate in Information Systems with a double major in Computer Science at Carnegie Mellon University. She is interested in distributed systems, computer vision, and machine learning. She will be working as a software engineer at Apple in the fall.

Appendix A. Tutorials for Palau FIC Legacy Pipeline

Video tutorial demonstrating the process of scanning 5 units of files:

https://drive.google.com/drive/folders/1AuxciIbMGL1TpefvXlcFXZRvUno8tuam?usp=sharing

Text tutorial demonstrating the process of scanning 5 units of files:

1. Prepare x number of annual reports, text facing you orientation, scan in one by one, separately.
2. Drag all 5 scanned pdf files from Scanned Folder on Desktop to Legacy/ANNUAL REPORTS/PDF FOLDER
3. Within Legacy/ANNUAL REPORTS, double click ORCHESTRATE_ANNUAL_REPORTS
4. Wait for the command prompt to disappear
5. Head to CMU Consultant/Legacy, check the file link works properly, check and correct any errors and undetected fields

Similar process holds for scanning other documents.

If any error occurs during this process, check out the Recovery section below.

Recovery:

Scripts stop running but no new rows are added in excel?

1. Head to RECOVERY, double click ORCHESTRATE_RECOVERY
2. Repeat the previous scanning process and check the new status of excel

If you still have issues, feel free to contact Sunday Zhou via email.

Where scripts and copies are stored:

Desktop of Dell Computer: Desktop/Legacy/development
Synology Hard Drive CMU Consult folder: sunday/Scripts
Github: shared with Cjay

Appendix B. Contacts of CMU Consultant Sunday Zhou

Email: sundayzhou0225@gmail.com
Cell: +1 412-439-6279