Accounting and Corporate Management (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acccm.2023.051220
ISSN 2523-5788 Vol. 5 Num. 12

Application of Python Automation Tool Combined with OCR Technology in "Financial Checkup" of Universities

Shumin Xia

Audit Office, Wenzhou Medical University, Wenzhou, Zhejiang, 325000, China
975291800@qq.com

Keywords: Python, OCR, Financial Checkup

Abstract: In this case study, auditors conducted research on the financial reimbursement process of a university and identified insufficient risk controls in the finance department of University W using process testing and risk assessment. Based on the needs of "financial checkup" and the objective condition of digitized financial vouchers in the university, auditors combined information technology audit tools and big data auditing concepts. They integrated Python automation tools with OCR technology and connected to an intelligent cloud OCR financial document image recognition interface to automate the process of reading files, recognizing invoice images, collecting data, and writing them into Excel. Subsequently, through data analysis, audit suspicions were identified for further verification, thereby improving audit efficiency and accuracy.

1. Introduction

According to the "Fourteenth Five-Year Plan for the Development of National Audit Work" issued by the National Audit Office, it is required to adhere to the use of technology in audit work, fully implement the requirements of General Secretary strengthening technological capabilities in audits, enhance innovation in audit techniques, and make full use of modern information technology to improve the quality and efficiency of audits. Internal audits in universities focus on their primary responsibilities and carry out audits of the economic responsibilities of the leadership, conducting regular 'economic check-ups.' These 'economic check-ups' are based on economic supervision and focus on the authenticity, legality, and effectiveness of financial matters. This paper presents a case study that thoroughly investigates the financial reimbursement process, introduces information technology audit tools, improves the efficiency of internal audits' 'economic check-ups' in universities, and enhances the accuracy in identifying audit concerns.

2. Overview and Risk Analysis

2.1. Risk Points and Solutions for Electronic Invoice Reimbursement

With the widespread adoption of electronic invoices in 2020, extending from value-added tax general invoices to special invoices, internal management risks related to financial reimbursement in universities have emerged. Liu Li (2022) identified risks such as duplicate reimbursement and

verification risks in electronic invoice reimbursement. Additionally, there is an issue of information asymmetry in the correction of electronic invoices, as the recipient may not be aware of the correction, leading to the reimbursement being canceled after it has been processed, making it difficult for the recipient to detect and recover the funds. The study also proposes internal control suggestions for managing these risks in enterprises and institutions: firstly, establishing an electronic invoice archive database to avoid duplicate reimbursement; small enterprises can create an electronic invoice Excel ledger, while large enterprises with multiple financial auditors should establish an invoice management system for automated verification. Secondly, addressing the issue of information asymmetry through process control by regularly batch verifying previously reimbursed electronic invoices and comparing them with the corrected value-added tax invoices[1].

2.2. Risk Analysis of W University's Financial Reimbursement Process

W University adopts the Fudan Tianyi Wingsoft Financial Information Portal System for financial reimbursement. Reimbursement personnel enter reimbursement details into the financial reimbursement system, submitting reimbursement forms online with information such as financial project codes, reimbursement summaries, reimbursement amounts, payment recipients, and corresponding payment amounts. The online submission primarily serves the functions of budget and payment controls for project funding.

When submitting reimbursement materials offline, original receipts, printed electronic invoices, and relevant attachments are attached. The printed electronic invoices require the reimbursement personnel to handwrite a statement of "commitment not to duplicate reimbursement." The university's financial department plans to collect reimbursement materials at the delivery counter and distribute them to the respective counter's financial auditors for review and reimbursement. The review process does not include logging the electronic invoice numbers in a ledger, and the control measures for checking the duplication of invoice reimbursement are limited to the subjective commitment of the person handling the reimbursement form. Consequently, there is still a risk of information asymmetry due to the cancellation of invoices after reimbursement. The specific process and risk points are depicted in Figure 1.

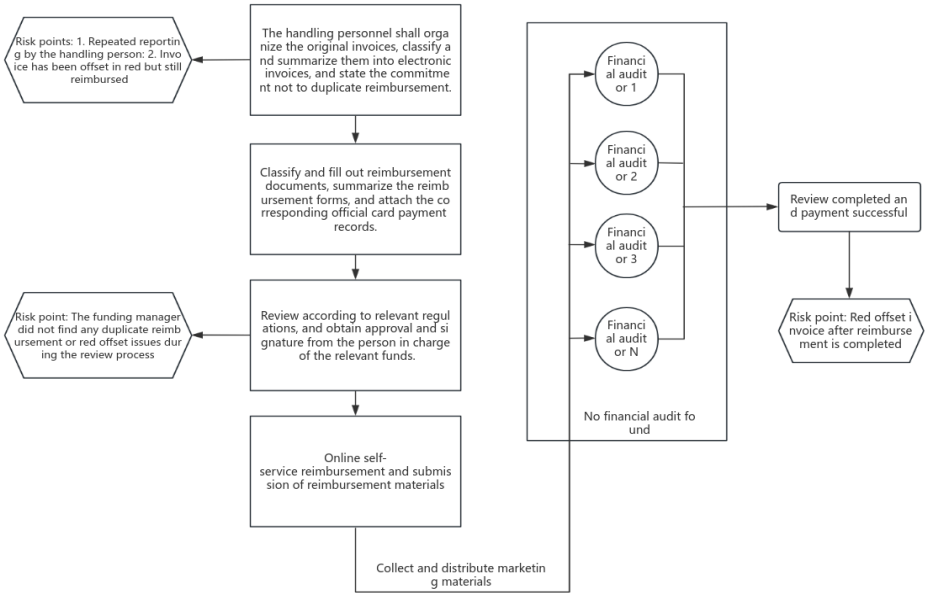

Figure 1: Process framework for financial reimbursement risk

The university's audit department still relies on sampling checks of financial vouchers and reviewing paper-based financial vouchers during internal audits, making it difficult to identify duplicate reimbursements or post-reimbursement cancellation information. The auditing challenge lies in obtaining a complete set of invoice numbers.

2.3. Accounting Archive Imaging Management

In 2022, W University's financial department purchased financial voucher imaging services. Starting from 2022, financial vouchers are scanned and electronically archived, allowing authorized personnel to apply for self-download. The imaging of financial vouchers provides a foundation for auditors to use information technology tools to read the complete set of financial invoice image information and convert it into structured data.

3. Introduction of Information Technology Tools

3.1. Application of Python in Internal Auditing

Python is a widely used programming language in recent years for office automation, with a rich set of libraries to accomplish tasks and is the preferred language for machine learning. Identifying automation tasks, selecting appropriate libraries, and achieving automation goals in office work.

There are also mature cases of Python application in internal auditing. Yuan Lihua (2020) utilized Python data analysis techniques for auditing public vehicle refueling, Zhang Xiaozheng (2021) conducted two direct policy audits based on Python, and He Yazhe (2021) used Python's random forest algorithm for auditing human resources in the power grid enterprise. These cases demonstrate that Python possesses features such as rapid data organization, convenient data analysis with clear results display, scalability, and reproducibility in auditing[2].

However, the aforementioned cases have already prepared relatively mature structured audit data for the introduction of Python, facilitating the direct application of Python data analysis tools. In the university's auditing, there is no readily available financial reimbursement data.

3.2. Application of OCR Technology in Internal Auditing

OCR (Optical Character Recognition) is a mature text recognition technology. When discussing the application of OCR technology in internal auditing, Wang Li (2019) proposed the use of OCR technology to perform image recognition on complete invoices, converting them into structured Excel electronic data. Then, by setting risk points and warning rules, data correlation and filtering can be performed. The study proposed four scenarios for the application of OCR technology that can be applied in internal audits at universities, namely verifying the implementation of the central "eight-point regulations" in official receptions, checking high-frequency transactions and abnormal purchases, verifying consecutive invoice numbers, and verifying proxy invoicing.

Based on the premise of imaging school financial reimbursement vouchers, the introduction of OCR technology enables the full extraction of financial ticket information and conversion into structured data suitable for audit applications, making it possible to achieve full coverage in financial audits.

3.3. Application of Artificial Intelligence Cloud Services

ChatGPT, a chatbot program driven by artificial intelligence technology, was released in November 2022. It is a natural language processing tool trained through human conversations and

can perform tasks such as writing emails, computer programming, and writing papers. Expanding the mindset of internal auditing and making good use of artificial intelligence services is also a trend. Mature artificial intelligence cloud services available on the internet now provide artificial intelligence, big data, and cloud computing services. On the cloud server, there are artificial intelligence-related financial ticket OCR services available, capable of recognizing value-added tax invoices, train tickets, taxi receipts, etc. These services can essentially meet the needs of internal audits in universities[3].

This case combines Python as an automation language, OCR, and intelligent cloud technology to batch-read the imaged financial documents of the audited units, forming standardized and structured audit data, and obtaining complete invoice numbers and other information. This is beneficial for achieving full coverage of internal audits in examining financial matters, transitioning from sampling audits to comprehensive audits, conducting audit data analysis, and capturing audit concerns.

4. Case Analysis

4.1. Voucher Structure Analysis

During the audit, financial vouchers are downloaded in batches. The imaged financial vouchers are stored in a certain structured format. Under a folder named after the voucher number, there are: PDF-format cover pages of the vouchers, named after the voucher number, such as "2022 11O 786" (see Figure 2); individual image files in JPG format, named as "Attachment000*". These images include invoice pictures as well as scanned pictures of non-standard voucher attachments, such as reimbursement approval forms.

Figure 2: Credential number named folder

4.2. Program Objective Design

The program design needs to achieve the following objectives: (1) Automatically and sequentially open the folders where the invoices are located and implement looping functionality. (2) Determine the file format of the folder; if it is ".JPG," read the image information. (3) Based on the read information, determine whether the image is a value-added tax electronic ordinary invoice; if it is, read the invoice information. (4) Write the key information, such as the storage path of the read invoice image, invoice number, invoice date, amount, and seller's name, into Excel. Based on this, an electronic document about invoice information is formed for repetitive analysis and recognition. Audit personnel can easily verify corresponding doubts by locating the corresponding folder through the storage path of the invoice image[4] .

4.3. Code Analysis

4.3.1. Introducing the Baidu Intelligent Cloud AipOcr tool.

Introducing the Baidu Intelligent Cloud AipOcr tool. Register for a corresponding account and input the APP_ID, API_KEY, and SECRET_KEY into the program. The code is as follows:

```
from aip import AipOcr
APP_ID = ' '
API_KEY = ' '
SECRET_KEY = ' '
client = AipOcr(APP_ID, API_KEY, SECRET_KEY)
```

4.3.2. Introduction of Third-Party Extension Libraries

Introduction of Third-Party Extension Libraries, Using Tools such as CV2, math, numpy for Invoice Image Analysis and Processing, and Converting them into Structured Data. The code is as follows:

```
import cv2 # import dlib
import math
import numpy as np    # Import the numpy library
import os
from PIL import Image
import pandas as pd
import os
```

4.3.3. Setting the Local configure.

Setting the Local Image Path for Machine Reading, Required Formats for Reading, Data Export Path, and Format. The code is as follows:

```
folder_path = r"C:\Users\97529\Desktop\Invoice recognition tool\test"
jpg_files = []
for root, dirs, files in os.walk(folder_path):
    for file in files:
        # if file.endswith('.jpg') or file.endswith('.JPG'):
        if file.endswith('.JPG'):
            jpg_files.append(os.path.join(root, file))
id_list = jpg_files
count = 0
newary = []
for ids in id_list:
    img = open(ids, 'rb').read()
    res = client.vatInvoice(img)
    # print(res)
    try:
        words_result = res['words_result']
        words_result_num = res['words_result_num']
        print(str(ids)+' '+'Recognition completion')
        Commodity_num = len(words_result['CommodityName'])
        for i in range(Commodity_num):
            try: #There are no missing values
                newary.append({
'Invoice picture Specifies the path for storing the invoice picture':ids,
'Invoice number':words_result['InvoiceNum'],
'Invoice date':words_result['InvoiceDate'],
'Buyer's name':words_result['PurchaserName'],
```

```
'Purchaser taxpayer identification number':words_result['PurchaserRegisterNum'],
'amount':words_result['TotalAmount'],
'Seller's name':words_result['SellerName'],
'Seller's taxpayer identification number':words_result['SellerRegisterNum'],
'remark':words_result['Remarks']})
                    except IndexError as e: #A missing value exists
                        pass
            count += 1
        except:
            pass
newsdf=pd.DataFrame(newary)
newsdf.to_excel('./test/VAT invoice information'+'.xlsx')
import pandas as pd
import time
time.sleep(2)
df = pd.read_excel('./test/VAT invoice information.xlsx')
df = df.drop(df.columns[0], axis=1)
df.to_excel('./test/VAT invoice information.xlsx', index=False)
```

According to the context of the paper and correct English grammar, the translation of the provided section is as follows:

4.3.4. Data Collection Results.

By looping through the folders of financial vouchers, reading invoice images, and writing corresponding fields into an Excel file, Based on the obtained data, further data analysis can be performed using Excel tools.

4.3.5. Analysis of Audit Results.

(1) Discovery of duplicate reimbursements in the college. After the audit pointed it out, the college recovered the duplicate reimbursements from the suppliers. Further analysis revealed that there were two college staff members who were in contact with the supplier, and there was a cross-business situation. Both parties did not communicate clearly before the reimbursement, and the review process did not carefully verify, resulting in duplicate reimbursements. The audit made recommendations to optimize the internal control process for financial reimbursement.

(2) Discovery of four consecutive invoices issued by the same supplier on the same day, with a total amount of 110,000 yuan. These four expenses corresponded to four activities and were reimbursed separately. Audit personnel conducted further verification of the four expenses.

4.3.6. Further Application of Audit Results.

Based on the obtained structured data of the college's complete invoice information, as shown in Table 1, the economic activities of the audited entity are essentially understood. Audit personnel further identify potential issues based on the results of data analysis. In conjunction with the above audit results, audit personnel further consider the audit issues that the data analysis results may indicate.

Table 1: Data analysis results and audit doubts

Data Analysis Results	Audit Concerns
1. Duplicate invoice numbers	Reimbursement duplication
2. Consecutive invoice numbers	Split payments, split invoicing, split reimbursement to avoid approvals
3. Payments to the same supplier exceed the limit	Split payments to avoid bidding, centralized decision-making, etc.
4. Missing or incomplete invoice information	Reimbursement documents do not meet standards
5. Purchases of tobacco, alcohol, etc., which are prohibited by the central "Eight Regulations".	Violation of the spirit of the central "Eight Regulations"
6. Single invoice amounts from the same vendor below the limit	Avoiding approvals, questioning the authenticity of financial transactions
7. Post-reimbursement invoice cancellations found by comparing data exported from the national tax platform	Returns and refunds after reimbursement, misappropriation of public funds
……	……

5. Conclusion

Audit personnel conducted research on the financial reimbursement process of universities to analyze potential risks and audit challenges. Additionally, during the review of electronic vouchers, a pattern was identified in the voucher files. Information technology tools were used to convert non-standard image content into structured audit data that is easier for audit personnel to analyze. During the internal "financial check-up" process, a research-oriented audit mindset and big data audit approach were employed. By reviewing literature and studying relevant audit cases, the combination of Python automation tools and OCR technology was introduced. The Baidu Intelligent Cloud OCR financial document image-text recognition interface was utilized to automatically read folders, recognize invoice images, collect data, and write them into Excel.

The audit tool that combines Python automation and OCR has the capability to automatically read financial vouchers, thereby improving audit efficiency. It also possesses the characteristics of being replicable and reusable, making it applicable not only for this particular internal audit project but also for future projects, with necessary modifications according to audit requirements[5].

University auditors should possess a research-oriented audit mindset and learn from excellent big data and information technology audit cases, incorporating internal audit practices specific to universities. Auditors can apply this information technology tool on a small scale during the internal "financial check-up." However, when using it on a larger scale, there are still challenges to address, such as data recognition accuracy, tool effectiveness, and data security to prevent leaks. Further improvements are needed in these aspects.

References

[1] Liu, L. Research on the Problems and Countermeasures of Electronic Invoice Management and Control. China Collective Economy, 2022(33), 129-131.

[2] Zhang, X. Research on Audit Algorithms for Two Direct Extension Policies Based on Python. China Internal Audit, 2021(10), 48-55.

[3] He, Y., & Cai, Z. Research on the Application of Random Forest Algorithm Based on Python in Human Resources Audit of Power Grid Enterprises. China Internal Audit, 2021(08), 44-50.

[4] Yuan, L. Application of Python Data Analysis Techniques in Audit of Fueling for Government Vehicles. China Internal Audit, 2020(10), 58-61.

[5] Wang, L., & Ye, J. Innovation and Implementation of Audit Technology Based on OCR. China Internal Audit, 2019(04), 44-47.