Research on document digitization processing technology

Ruili Zhang[1,*], *Yanming* Yang[1], and *Wenxiu* Wang[1]

[1]Qingdao Campus of Naval Aviation University, No. 2 middle Siliu Road, Qingdao City, Shandong Province, China

Keywords: Document, Digitization processing, PDF format.

Abstract. The digitalization of document information is the development direction of the digitalization of document information management, which involves various technologies such as digitalization technology, picture and text editing, storage format, etc. Through the PDF document loading display, change the page replacement storage, the technical page jump to achieve the PDF document programming control.

1 Introduction

Digitization of documentary information includes digitization of papery document information, digitization of photographic information, and digitization of audio-visual document information, etc. , of which a great deal of work is the digitization of papery document information. The digitization of papery document information includes two different levels, one is the digitization of catalogue of documentary information, the other one is the digitization of the full text of documentary information. The digitization of the full text of documentary information can display the whole content of the management of documentary information comprehensively and systematically. Then, we can use the advantages of computer network to conveniently and quickly access the details of all types of documentary information without revising the document ‘s original ideas. The desktop browsing and information exchange can also be realized by means of LAN, hard disk, reader and other tools, which effectively expands the space of documentary information transmission. Therefore, full-text digitization of documentary information is the direction and the main content of digital management of documentary information. However, due to the high accuracy requirement of technical bulletins and other regulatory documents, the original scanning method can be used to digitalize the documents under the situation that the requirements cannot be achieved completely.

The digitization of documentary information plays an important role in protecting the original documentary information, it involves digital technology, digital degree, digital quality, graphic editing, image storage format, and image rename, etc. The whole process includes several important parts such as scanning, indexing, quality checking and archiving and backup of electronic documentary information. The environmentand tools for

* Corresponding author: zrlhao@163.com

production include scanning system, word recognition system, Photoshop processing system, and Acrobat producing and reading system, etc. Graphically, the process of documentary production is shown in Figure 1.

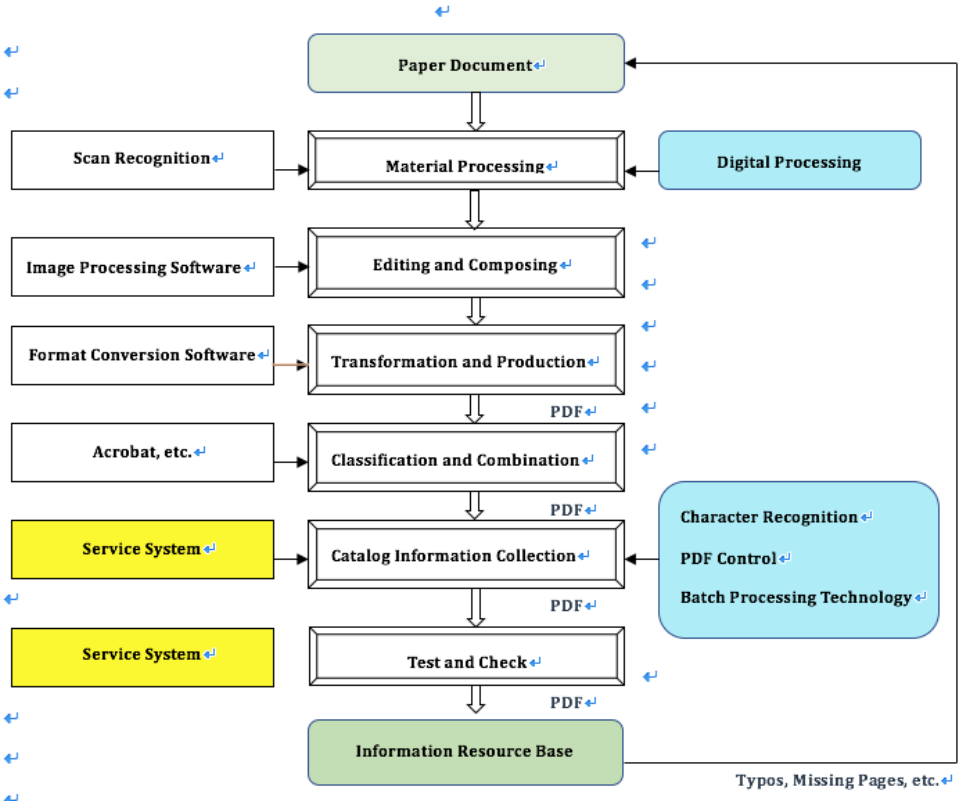

Fig. 1. Digital processing of paper documents.

2 Document processing method

2.1 Scan of document

(1) Scan method. The corresponding scanner or professional scanner is selected for scan based on different size of documents. Large-scale document can be processed by large-scale digital platform. For example, we can use A3 Microtek LP1700 scanner or apply image mosaic after small-scale scanning. Flat-panel scanning should be adapted for poor quality paper and the document that is too soft, too thin or too thick. Then, the document with better paper can be processed by high-speed scanner to improve work efficiency, so we can use a FS2500 scanner with A4 format.

(2) Scan color. In general, scan colorincludes binary black and white, gray, color and so on. This project adapts gray scanning. First, pages that are printed in black and white, with clear handwriting and no illustrations can be scanned in black-and-white mode. Second, pages printed with black and white, poor clarity of handwriting or illustrations, and pages that are multi-color can be scanned in gray mode. In addition, pages that have red heads, seals, black-and-white photos, colorized photos or color illustrations can be scanned in colorized mode according to various demands.

(3) Resolution of scan. In principle, the selection of the parameters of scanning resolution depends on the clarity and integrity of images the efficiency of image utilization after scanning. It is generally recommended to select greater than or equal to 100 DPI while scanning documents with black-and-white binary, gray and color modes. Undercertain circumstances, the resolution can be improved appropriately when the characteris small, dense and poor clarified. Documentary information requiring OCR recognition is recommended to select scanning resolution greater than or equal to 200 DPI. In general, we should try our best to achieve the goal that a lifetime use by one scan from the perspective of protecting the original document. It requires higher storage space of the systemif higher scanning resolution, such as 300 DPI or more is applied. From the perspective of commonly documentary information online search and application, 150 DPI can basically meet users' requirements. The relatively smaller storage leads to the relatively faster speed of image browsing. Therefore, the system capacity should be weighed comprehensively according to the practical application requirements and the established system capacity before scanning. This project mainly selects 300 DPI.

(4)Scan registration. To be more specific, it is important to fill in the registration form carefully, register the number of pages scanned, check whether the actual number of pages scanned for each document is consistent with the number of pages filled in when the documentary information is sorted out, and indicate the specific reasons and solutions when inconsistent.

To sum up, the key point of paper documents scan is to reduce the byte size of the final document as much as possible on the premise of grasping the quality (reading effect and character recognition rate).

2.2 PDF processing and merge

The merging function of Adobe Acrobat is used to merge scanned pages such as cover, description, catalogue and text into a single file in sequence. Among them, it is necessary to tailor and synthesize someillustrations with exceeded pagesby Photoshop tools first.

2.3 PDF format conversion and treatment

Once the installation of Adobe Acrobat is completed, the menu bar of Adobe Acrobat will be shown in the menus of Word, Excel and other software. This is the seamless combination of Adobe Acrobat and those applications. Thus, documents created by these applications can be easily converted into PDF format, and the related parameters of conversion can be controlled and adjusted. When it comes to the original locomotive working file in word format, we can use the Adobe Acrobat menu embedded in word to convert, which is simple and fast.

However, it is practically found that in PDF documents directly converted by word, some original floating graphics may be misplaced. At that time, the original format and location can be maintained by using "Wendiantong PDF GoldV9. 51" conversion.

In addition, for some special fonts used in word, such as FangzhengXiaobiao Song simplified style, the font cannot be embedded in the conversion could result in that the conversion cannot be displayed properly due to copyright restrictions. Apparently, this need to be carefully checked and converted into other fonts to ensure the integrity of the document and its quality requirements.

3 Related technological research

3.1 Dynamic loading and embedding technology

So far, all types of PDF document browsers are mature and full-featured to large extent. Adobe Reader and Cajviewer can be can be selected as two common browsers. The reader can be embedded in the information system, and the user-defined control and seamless connection can be achieved in order to make the operation more targeted and simple.

3.2 PDF script control technology

PDF document is an open standard for global electronic document distribution. It is also a structured document format and supports various compression methods. When converting and distributing documents in PDF format, every security inspector can accurately view and print the original document created on any platform including logo, picture, color, etc. , without any additional format conversion.

In addition to the commonly used Adobe Acrobat editing software, there are also CAJ full-text browser, which is a special full-text format reader for www. chinaqking. com. Like superstar reader, CAJ browser is also an electronic book reader (caj reader, CAJ full-text browser). CAJ browser supports CAJ, NH, KDH and PDF format reading of www. chinaqking. com. What's more, CAJ full-text browser can cooperate with the original text on the Internet or read the full text downloaded, and its printing effect is consistent with the original version.

(1) PDF Programming Control

The loading and displaying of PDF documents, the replacement and storage of changing pages, and the skipping of technical pages all need to be controlled by programming of PDF documents. Both Acrobat and CAJviewer ofTongfangHowNet provide powerful development packages with perfect operation interfaces. , so all kinds of functions from document creation to security control can be realized via various APIs. For instance, open, close and jump operations are the most frequently used functions for document browse.

(2) Control of Virtual Buttons

For assembled PDF files, since multiple files are stored in the same PDF file, they need to be printed or extracted separately when used. According to the function of PDF control itself, the start and end pages are required to be input. For the convenience of user operation, we simulate the keyboard operation by programming to realize the automatic filling of page numbers, which requires that we can identify the status of dialog box and the position of interactive objects, and then simulate the key operation.

3.3 Character recognition technology

For PDF files generated by WORD, the character selection tool of PDF reader can be used to select and extract the characterdirectly. However, if PDF documents are generated by scanning paper files, they cannot be extracted directly, then OCR,optical character recognition technology is adapted to export the text.

(1) CharacterRecognition of CAJviewer Reader

It is relatively convenient to use CAJviewer reader in character recognition, and OCR technology of Tsinghua Mandarin Tong is used in CAJviewer reader. However, the manual operation steps provided are complicated, and the requirements for massive text extraction are obviously less humanized. Therefore, the internal code control should be considered,

such as file title input, which can be quickly extracted and filled into the corresponding text box by the following code.

(2) CharacterRecognition of Adobe Reader

In Adobe Acrobat, wecan operate the menu "Document"→ "OCR "→"Using OCR to Recognize Text" →"Selection Tool" to select text to copy after recognition.

Adobe Reader does not have its own OCR function, but it can be operated by operating the menu "File" → "Print", setting the desired page, and selecting "Microsoft Office Document Image Writer" for the printer's name. It is a virtual printer installed in the computer with Microsoft Office 2003. Besides, it prints a PDF document into a file with the suffix "mdi" and opens it automatically. In the "mdi" file, the menu "tools"→ "using OCR to recognize text" are operated in sequence, then we can directly select text in the page, and complete copy operations.

Moreover, we can export text to a Word document if we use the operation menu "Tools"→ "Use OCR to recognize text" and "Tools"→ "Send text to word".

The recognition rate of OCR technology depends on the scanning accuracy when creating a PDF document, and those documents with illegiblehandwriting will not recognize too many words correctly.

To copy the illustrations in the PDF document, we can open the PDF with Adobe Reader or Adobe Acrobat, and use the "Select Tool" to select the illustration and then press the "Ctrl" + "C" buttons to copy the illustration to the clipboard. Furthermore, all images in the document can be exported to one image file simultaneously by operating the menu "Advanced" → "Document Processing" → "Export All Images" in Adobe Acrobat.

4 Conclusion

In summary, the digitalization of a document can be monitored by the operation steps of the document processing, its process and the interactive interface, whilethe intelligence of the documentcan be realizedby the object and process monitoring technology. In addition, the operation control of the mouse and the keyboard can be implemented to realize the pipeline operation or batch. According to automatic collection through program design and database record, repeated manual work can be reduced and work efficiency can be improvedsubstantially, thus achieving intelligent processing. The main technologies are as follows.

First, the combination of OLE automation and VBA technology enables the recognition of the status of the word document and the reliable transmission of operational commands when the typesetting system and the word document are separated from each other, thereby achieving intelligence and fast and flexible format control.

Second, the utilizationof dynamic loading and embedding technology solves the problem of integrated design of file retrieval and use, also realizes seamless connection and self-renewal of multi-applications. Specifically, the embedded design method can realize the seamless connection between the retrieved directory list and the document browsing window, so the display interface can be more clear and concise, and the operation is more simple and intuitive.

Third, the idea of special search vocabulary should be proposed and the application of word segmentation technology should be used to realize the fuzzy search of keyword sentences. What's more, "secondary search", which means that multiple repetitive search in the previous results based on new conditions should be permitted, not only the practical search become more flexible and convenient, but also the flexibility and effectiveness of information queries can be enhanced.

References

1. Liu Zhutao. Development of aeronautical meteorology equipment management system based on Delphi [J]. Computers and Telecommunications. 2017 (04) 41-43
2. Wu Mengqi. Government Service system based on E-government [D]. Jilin University. 2016
3. Xiaowei. Cong. Realization and management and control application of digital security grade mark of electronic document[J]. Security Science and technology. 2013(03).
4. Jie Shao. Application and process of optically variable picture[D]. 2015.
5. Jin Gao. Application of gis digital processing and storage technology[J]. 2006
6. Hui Liang. Application of gis digital surveying and Mapping Technology in engineering survey[J]. Introduction to architectural research. 2019(8)
7. Xianhai Tan. Research and implementation of data conversion from unstructured to structured data[D]. 2013.
8. Wenbin Li. Discussion on digital image processing and compression technology. Digital technology and application[J]. 2012(6) .