

# An Analysis of the Problems and Countermeasures in the Digital Work of Archives

Qun Li<sup>1</sup> and Liying Cui<sup>2</sup>

Archives of Jilin Agricultural University, Changchun, 130118

**Keywords:** Archives; Digitalize the whole file texts; Countermeasures

**Abstract.** The achievement of the input of the text of file digitization mainly adopts the way of scanning paper documents and files into a digital form [1]. This paper puts forward the corresponding countermeasures and suggestions on some problems existing in the current digital work of the archives.

The key of the digitization of Internet file information is to digitalize the whole file texts. Only the full text of the file is published on the Internet, can the practical function of the archives information network be realized. At present, the full text of the file on the Internet is at the experimental stage, there is still a far way to go from its practical application. The following is a preliminary discussion on the suggestions on the digitization of the full text of the file and the establishment of its full library.

## The Existing Problems in the Digitization of the Full Text of the File at Present

**Standard Specification for Construction is Lagging Behind.** At present, there are no uniform norms in the fields, focus, identification, storage format, query, exchange, online transmission and management of the digitization of the full text of the file. It is difficult to ensure the safety and efficient use of digital archives for archives at all levels are of their own ways [1].

**Some Archives Lack Scientific Assessment and Analysis of Digital Work.** Some hastily launched and blindly advanced [2]. Some scan all the files regardless of user needs and the formation of a large number of spam results in unnecessary waste. As the file digitization of all the archives is in accordance with the principle of from near to far and not identify on the file selection in the digital process, there are many garbage data in the established database of the full text. And then the digitization of many files of high utilization rate and in urgent need for digitization is delayed, which will inevitably seriously affect the development and utilization of collection files.

**Investment is Huge, and it is Difficult to Achieve the Digitization of all the Files.** If all the collection files are fully digitalized, significant capital investment will be required[3]. Take 10 integrated archives in Changchun city for example. There are 500,000 volumes all together. If there are 120 pages each volume then there will be about 60 million pages. The fund will be 12 million yuan for 0.2 yuan per page. It is difficult to solve and achieve on the basis of the current financial situation. If the work is done by the archives, the procedures will mainly include the identification, file transferring, unpacking, scanning, checking processing pictures, naming, uploading, binding, and returning the file. Even if all can be scanned with a high-speed scanner for 1000 pages per day, without increasing the collection of circumstances, it takes about 10 years to complete. Moreover, there are a large number of early archives cannot be scanned with a high-speed scanner and it can be imagined how heavy the workload is. The digitization of the full text of the file is an arduous task of archiving work. Only reasonable planning, scientific assessment and analysis can ensure that the digitization of files is improved and used safely and efficiently.

## The Countermeasures of the Digitalization of Collection Files

**Develop Unified Digital Standards and Specifications.** Standards and specifications are guarantees of digital security and efficient use and also the traffic rules on the files information highway. A unified standard and norm should be developed as soon as possible to further standardize the procedures of digital work and prevent blindness. Relevant standards and norms like

"Digital Standard for Paper Files", "Digital Standard for Photo Files", "Digitized Standard for Recording and Recording Files", "Digitized Operation Specification for Digitized Files", "Digital File Network Utilization and Management Specification", "Digital File Management and Privacy Requirements " should be developed.

**Identify the Collection Files and Determine the Range of Digital Areas.** For most archives, it is not necessary and impossible to digitize all the files. The digitization of the files should be carried out on the basis of the selection. As is pointed out by Professor Wang Jian in "Discussion on the Optimized Mode of File Digitization", selection is a compound word of identification, differentiation, and preference. Identification is intended to determine whether the files still have value for preservation. Differentiation is to determine whether it is necessary to be digitized. Preference is intended to determine whether it is included by digitalization. The author believes that selection should be based on the principle of application underlying first, not repeating digital files, value first, and protection first to determine the scope of the digital file and the scope of digital priority[4].

In accordance with the principle of application underlying first and based on the study of borrowing registration and the analysis of application, determine those files of high utilization as the scope of digital priority. Take the archives collection in Jilin Agricultural University as an example. Several types of files of high utilization in recent years like the student status files, student enrollment list, and subject acceptance certificate are arranged in topic order to complete the digitization of the full text of the file.

In accordance with the principle of value first and protection first, prefer the files with special collection value and the precious files or solitary copies which are old, broken, and with vague handwriting to be digitalized. Let's also take Jilin Agricultural University as an example. The historical files of the first 50 and 60 years of the establishment of the school are digitalized and are currently continued to be digitalized in accordance with the principle from near to far.

**Earnestly File the Electronic Documents and Preserve and Converse the Electronic Version of Paper Documents.** At present, the system or department that has realized office automation can carry out the filing and application of electronic document in accordance with the relevant specifications and requirements of the electronic document filing. But for the various non-standard electronic files formed by each unit are referred to as the electronic version of paper documents. They are in the form of electronic documents and their contents are exactly the same with corresponding paper documents, but there are no external features electronic documents such as electronic seals and electronic signatures. Many units just let them go with these electronic documents. The author believes that these electronic files of various departments should be preserved in two formats of text files and pictures. They should be transferred to the archives together with the archived electronic file directory of the second year in order to avoid the vicious circle of restoring to digital files after re-scanning, so that archives are no longer caught in the increasingly amazing scan whirlpool.

### **The Overall Suggestions on the Digitization of the Full Text of the File**

The realization of the digitization of the full text of the file can be divided into two cases:

In the first case, the archived file itself is an electronic file of a text, image or the like. In other words, such archival information is generated in the form of electronic files and it has been digitized before being uploaded. Things need to do are to convert such data into a canonical format, and then deposit them into the file library according to a certain organization form and establish a full-text retrieval system.

In the second case, the archived file is a text or image of a traditional carrier. This type of files can select three data storage modes when converting to a digitized file:

Firstly, scan the file page by page into the image file with required format which usually is PDF format by using the scan entry mode. Upload it to the online version of the file management system through the technical processing. The advantage of this approach is that the original picture of the file can be saved and the system technology is relatively simple. The disadvantage is that a larger

storage space is occupied which is not conducive to improving the speed of online transmission. This method applies to the digitization of image file, the historical manuscripts of higher fidelity requirements, multimedia files and so on[5].

Secondly, store the contents of the file by adopting text form and use a full-text search database as supplement. There are two ways to record the full text of the file. Manual input in text form. Scan the original file into image files and convert them into text file by using OCR. So that the real word-by-word full-text search to the full text of the file can be realized.

Thirdly, The above two kinds of storage patterns can be combined, which is to store the attached scan in text form. The basic way of making is to scan the image file, and then use OCR to convert it to text format, so as to establish a one-to-one relationship between text and image page. Users can conduct full text search on the library database in the text format to view the original files and achieve its certificate authentication function.

In short, we will encounter a variety of difficulties and challenges in the process of digitizing the files. In the face of these problems, we must continue to explore and study new methods, analyze the new situation, solve new problems, and speed up the process of digital file.

## References

- [1] C.L. Yang Ping Lv. Records the full digital work problems and countermeasures, Yunnan archives.2007 5.
- [2] L. Liu. On the Identification of Archives before Digitization. Archives Science Bulletin, the 1st edition, 2007.
- [3] Z. Ou. Reflections on the Digitalization of Archives. China Archives, the 1st edition, 2007.
- [4] J. Wang. Discussion on the Optimization Model of Archives Digitalization. Archives Science Bulletin, the 1st edition, 2007.
- [5] Q.X. Wang, Z.Y Xu. Research on the Construction Principle of Archival Information Database. Archives Science Study, the 2nd edition, 1998.
- [6] R .Wang. Related Issues on the Digitalization of Paper Files. Beijing Archives, the 6th edition, 2006.
- [7] X..C. Wu. Research on Digital Pre - processing of Archives. Archives Science Study, the 2nd edition,, 2006.
- [8] R..X.. Zhang. Management of electronic documents. China Information Review, the 7th edition, 2007.