(Xiao and Huai 2017)

Xiao, Hong [肖红], and Yan Huai [槐燕]. 2017. “An Analysis of Quality Checking Problems in the Practice of Digitization of Newspapers of Republic of China” [民国报纸数字化实践中的质检问题探析]. Researches on Library Science [图书馆学研究], no. 07: 61-78+87.

DOI:10.15941/j.cnki.issn1001-0424.2017.07.010

Questions:

Which OCR engine used? Or manually typed?

National Library started digitizing newspapers from the Republic of China in 2014

Image scan of each page, OCR of article headings and titles, headings and titles marked with boxes (for title search)

Quality control: number and naming of documents, the resolution 分辨率, size, and bit depth 位深 of scanned images and other technical parameters, and the recording of OCR character recognition the content of the title in XML text file.

Data processing

Acceptance data

Check the number of files, file names, image quality

qualified

Sampling the quality of the image XML

qualified

Finished product data

Figure 1: Quality inspection process for digital data of newspapers in the Republic of China

After the digital processing of the newspaper is completed, the acceptance data is generated, and all data is completely checked. unqualified data is returned to processing workflow for modification, qualified data enters the random inspection抽检workflow. Random inspection refers to the sampling inspection of data according to a certain proportion, requiring detailed page-by-page or word-by-word in-depth inspection 精细检查of the data obtained. The unqualified data is returned to the processing link, and the qualified data is the finished product data, which can provide readers with services.

The digital data quality inspection of newspapers in the Republic of China adopted a combination of general inspection通查and random inspection抽查. The general check is to use a computer program计算机程序 to quickly check the regular parts 有规律的部分of the data, mainly including file naming文件命名, image size图像尺寸, resolution分辨率, compression压缩情况, and regular label content in the XML text dataＸＭＬ文本数据中有规律的标签内容

Random inspection is to conduct sampling inspection of data on a proportional basis. The sampling inspection ratio is determined by the quality of the data. When there are many problems in the initial stage of the project, 100% full coverage is generally used; after the project is stabilized and various problems have been exposed and resolved, the sampling inspection ratio is generally 30% to 100%. The spot-checked data needs to manually check the clarity, skewness 清晰度、歪斜度of the image and all the label content 所有标签内容of the XML text data.

|  |  |
| --- | --- |
| Label name | Label meaning |
| Record identification number | A unique identification number of a newspaper, taken from the first 9 digits of the MARC record 097Sa field |
| Title | Newspaper name, 200Sa from MARC records |
| Publication date | 8 digits, 4 digits year + 2 digits + 2 digits of day, published on February 12, 1936 as 1936 0212 |
| Volume | The volume or issue number of each issue of the newspaper, such as the 72nd, it is 72 |
| Edition | The corresponding edition of the XML file |
| Title coordinates | The position information of the title of the article, recorded in the form of pixels |
| Article coordinates | The position information of the article title, recorded in the form of pixels |
| Transition information | The content of the article on this edition is not finished, please follow the content on the edition later |
| Author | Article author |
| Attached picture group | attached picture included in the title |

Table 1 Main contents of XML file

Sampling inspection work is mainly to check the content of each label of the XML file and the corresponding part of the corresponding newspaper image word by word, including the identification number纸记录标识号, name名称, publication date出版日期, volume卷期, edition版次, title 标题and title position of the newspaper篇目位置的置标, and the title text recognition results标题文字识别结果, etc.

1.4 Inspection tool

The content of the XML file involves information such as column, title, author and other recognized text and the labeling of the scope of the content. The content of these texts is not convenient to check 不方便检查in the XML file, especially the marking data is a series of coordinate numbers representing the pixel points, which cannot be intuitively judged for its correctness无法直观判断其正确性. In order to improve the accuracy of inspection and work efficiency, the project team developed software for the comparison and acceptance of XML and images ＸＭＬ与图像的对比验收. The main functions of the inspection tool are as follows:

* Be able to judge判断 the correctness 正确性of the XML grammar, if it does not conform to the grammatical rules, it will prompt "load failure 加载失败".
* Judge whether the necessary label 必备标签in the XML file exists, and give a prompt if it does not exist.
* Check the legitimacy 合法性 of the "Publication Date出版日期" to avoid situations like " ５月４０日 ".
* Check the number of digits in the title篇目号的位数. If the number does not match 3 characters, a prompt will be given.
* Check that "Title标题" and "Article Coordinates篇目坐标" are not empty, otherwise an error will be displayed.
* For the correct XML file, display the corresponding 对应image and extract the corresponding content in the XML to facilitate manual proofreading in the next step.

When the XML file is correct, the software interface is as shown in Figure 5. Area区域 1 is the version and title number currently checked, area 2 is the text recognition and marking information part of the corresponding article in XML, and area 3 shows the images corresponding to the XML file. The two large boxes contain the position range of the current examination article on the image.

mapping of the column栏目, title标题, author作者, and article篇目coordinates 坐标 in area 3 and area 2. The software provides an intuitive visual inspection method, which greatly improves the speed and accuracy of inspection

Common problems include wrong record identification number, wrong newspaper name, wrong publication date or not corresponding to the storage folder, wrong version naming or the version number of the XML record does not correspond to the actual image, the issue of the command range of the column, between different titles The boundary and labeling errors, title labeling errors, various text recognition errors, etc

2 Information problems at the newspaper level报纸层面

2.1 Record identification number

2.2 The name of the newspaper  
2.3 Date of publication  
2.4 Page版次  
2.5 Typographical errors

3 Layout issues

3.1 Column issues  
3.2 Title setting  
3.3 Title Marking

4 OCR character recognition error

OCR text recognition includes column, title and author

On the basis of the correct division of each article in the layout, that is, the scope and position of the column, title and author are accurate, the text content is recognized, and the incomplete and unrecognizable words are replaced with „ ⚌ “[&#x268C;].

4.1 More and fewer words  
4.2 Symbol 符号error   
4.3 Ambiguous word 模糊字errors  
4.4 Recognized as similar characters or other characters