Research on Editions Digitization of Ancient Books Using IIIF and TEI—With Tokyo University “Water Margin” Various Editions as an Example

Yanping Wang
Wuhan University, School of Information Management, 430072 Wuhan, China

Abstract. After more than 30 years of development, the digitization of ancient Chinese books has accumulated a lot of resources. However, the construction of various institutions lacks standards, and it is difficult to exchange and reuse resources and technologies. In order to alleviate the dilemma of version research in the digital construction of Chinese ancient books, this paper uses the literature research method and case study method to analyze the digital practice of the engraved versions of the Water Margin at the University of Tokyo in Japan, discuss the significance of the application of IIIF and TEI, and propose the resources of Chinese ancient books Suggestions for improving digital construction.

1 Introduction

Ancient books are an important material evidence and carrier for the inheritance of Chinese civilization. The further development of modern information technologies such as computer science and big data has provided new ideas for the protection and inheritance of ancient books. After more than 30 years of development, the practice of digitizing domestic ancient books has achieved remarkable success in the digital construction of resources, and gradually developed from "surface data" with simple entry and storage to "deep data" that deeply studies the content and meaning of ancient books, [1] Gradually realizing the orderly development of knowledge organization.

However, the study of ancient book editions is extremely complicated, including the examination and revision of the origin of single books, the identification of rare and inferior editions, original engravings and re-engravings, as well as the printing paper ink colors, font cutting methods, book imprints, layout models, and decorative styles. [2]

2 IIIF and TEI

2.1 Research status of the issue of digitization of ancient books in China

With the further development of modern information technologies such as computer science and big data, various digital technologies are widely used in humanities and social
science research. After more than 30 years of development, the three major categories of ancient books digitization construction in public libraries, academic institutions and ancient book digitization enterprises have constructed the ancient book imaging type, graphic comparison type, electronic ancient book finalization type and ancient book finishing achievement conversion type four Based on categories,[3] there are more than 400 ancient book databases, which have accumulated a lot of resources for the digital protection of ancient books.

However, most of the current ancient book databases in China are mainly based on the content of the literature, and the insufficient utilization of the research results of ancient book edition studies is not conducive to "disciplinary scholarship and examination source".[4] Affected by the disadvantage that standards and specifications cannot be unified, digital resources and technologies of ancient books are difficult to exchange and reuse, which is contrary to the development trend of open data.

2.2 IIIF and digitization of ancient book versions

In June 2015, 29 European and American image resource storage organizations proposed to establish IIIF (International Image Interoperability Framework). IIIF is based on HTTP technology and improves the interoperability and availability of image storage by defining a common set of application programming interface specifications. IIIF takes libraries, archives, and museums that are willing to share digital resources as the main body of application, and has terminal interaction functions to achieve the harvesting and sharing system of image resources between institutions. It does not create a new system, but a set of shared API specifications for the interoperability of digital image libraries.[5] Due to historical reasons, many versions of ancient books are kept by various institutions. The digitization of ancient books is not a discrete project of a single institution, but a systematic project.[6] IIIF has specific functions for image comparison and annotation. Institutions can format existing ancient books' images in IIIF format, integrate them with other IIIF format data items on the same page, and integrate them for re-output. They can also use IIIF's dedicated browser to add expansion slots to compare or annotate images on the page. If the ancient books with the same text content are collected and used by multiple parties, researchers can use the IIIF format to digitize the ancient book images and reassemble them online to provide a complete version of the resource. Sharing technology and research results through IIIF to enable ancient book resources to be used on a unified platform and call each other is an important measure to solve the problem of digitization of ancient book versions.

2.3 TEI and digitization of ancient book versions

In the transcription process of digitized texts of ancient books, the problems of insufficient character set and vertical typesetting are often involved. TEI, the Text Encoding Initiative, was released in May 1994 and used to encode the full text of text in digital form, providing guidelines for the marking of humanities research materials. TEI can not only transcribe text from digitized images, but also record various modifications, annotations, damages, blurs and other traces on the manuscript.[7] The latest version of the P5 guidance document has added detailed instructions on the use of ancient texts, involving many specific reference elements in the catalog. TEI has a tag set function that supports version comparison. The analysis tools represented by Versioning Machine 5.0 and TEI Critical Apparatus Toolbox provide functions such as label sharing and image annotation, which improve the efficiency of version comparison. Applying TEI to the study of digitization of
ancient books, and constructing functions such as version comparison and identification, are of great significance to the study of ancient book edition.

3 The digital project of "Water Margin" by the University of Tokyo

3.1 Project introduction

The digital project of "Water Margin" originated from the support of two digital projects of the University of Tokyo, namely the Uehiro Project for the Asian Research Library and the Utokyo Digital Archives Project. The project of the Shanghai Library of Asia Library is based on the Chinese books and epigraphs collected by the University of Tokyo Library and the East Asia Research Center. It is committed to publishing high-resolution images of its materials on Flickr and other photo sharing sites. The Digital Archives Project of the University of Tokyo is based on the resources obtained by the Shanghai-Asia Library Project, and aims to promote the sharing and utilization of academic resources in schools. It releases shared digital image data based on international standards such as IIIF and TEI, building the "University of Tokyo File Sharing Server".

3.2 Implementation of data association between IIIF and TEI

3.2.1 IIIF and illustration research

The source of the digital image resource of "Water Margin" is Uehiro Project for the Asian Research Library. In order to improve the interoperability of images, the researchers will open the digital image resources on Flickr, choose the open source content management system "Omeka" that is suitable for the IIIF standard, and disclose the hidden resources again. Its open source code is conducive to third-party reuse. Secondly, through the plug-in "IIIF Toolkit with Mirador" of "Omeka Classic", the project built a research environment where users can freely add metadata and annotations.

On this basis, for the illustration problem in the ancient book version, the "IIIF Curation Viewer" developed by CODH (the Center for Open Data in the Humanities) was used to build an illustration database. The plug-in can automatically segment the rectangular part of the image that meets the IIIF standard, create a character image set and tag metadata information. Ancient books are published for profit, with different styles and forms of illustrations according to the time of publication and the region of publication. Through the illustration database, not only can the comparison of illustrations between different versions be realized, but also the collection function of the image database with people as the main body can be realized.

3.2.2 Comparison between TEI and text

At present, the process of generating text data from image data on most platforms is basically outsourced to OCR technology companies, but ancient books are more difficult to process complex shaped characters, and are often affected by typesetting and have a high error rate. However, due to the current limitations of digital technology, the recognition of nuanced texts still relies on researchers with professional version knowledge. The researchers refer to the "different reading" method proposed by Kitamoto Asahi [8].
computer discriminates and emphasizes the subtle differences in the images, and improves efficiency through the separation of man and machine. Taking the two editions of "Water Margin" in the library attached to the University of Tokyo as the research object, using TEI to support the label set function of version comparison, and the two versions were compared with the school-based text in Versioning Machine 5.0 and TEI Critical Apparatus Toolbox. Apparatus entry is used to display the text difference, and reading tag is used to display the reading difference, so that the text difference between different versions can be intuitively displayed.

3.2.3 Implementation of graphic and text association based on IIIF and TEI

IIIF and TEI are not independent. The researchers realized the association data between the text content based on TEI and the digital images of the original ancient books based on IIIF. Specifically shown in Figure 1. First, add the facs attribute of the surfaceGrp tag to the URL (Uniform Resource Identifier) of the IIIF manifest of the image set. Secondly, add the URL and n attributes of the graphic tag to the image URL and Canvas ID of each image. Finally, the association between the single-line text of "Water Margin" and the corresponding image area in IIIF is selected according to the zone label. That is, when the text data on the left is clicked, the image corresponding to the text part in the image on the right is highlighted.

![Figure 1. Examples of IIIF and TEI related marks.](image)

4 Conclusion

The development of the digital age has brought development opportunities to the protection and development of ancient books, but also faced multiple tests. Ancient books are an important carrier of our civilization. As an important subject in the study and management of ancient books, ancient book edition studies apply modern information technology to its research, and provide tools for the comparison, identification and sharing of ancient book editions. Building a database of Chinese ancient book versions based on IIIF and TEI
standards is conducive to the integration of China's digital construction with international standards and enhancing international influence.

References