RESEARCH ON THE DIGITAL HUMANISTIC PATH OF OVERSEAS DISPLACED ARCHIVES TAKING THE APPLICATION OF THE MISS PLATFORM AS AN EXAMPLE

Rina Su[1] Yumeng Li[2] Xin Yin[3,2] Tao Chen[4]

[1]Sun Yat-sen University Library, Guangzhou 510205
[2]School of Journalism and Communication, Jiangxi Normal University, Jiangxi 330022
[3] School of Information Resource Management, Renming University of China, Beijing 100872
[4]School of Information Management, Sun Yat-sen University, Guangzhou 510205

ABSTRACT

The digitization of displaced archives is of great historical and cultural significance. Through the construction of digital humanistic platforms represented by MISS Platform, and the comprehensive application of IIIF technology, knowledge graph technology, ontology technology, and other popular information technologies. We can find that the digital framework of displaced archives built through the MISS platform can promote the establishment of a standardized cooperation and dialogue mechanism between the archives' authorities and other government departments. At the same time, it can embed the works of archives in the construction of digital government and the economy, promote the exploration of the integration of archives management, data management, and information resource management, and ultimately promote the construction of a digital society. By fostering a new partnership between archives departments and enterprises, think tanks, research institutes, and industry associations, the role of multiple social subjects in the modernization process of the archives governance system and governance capacity will be brought into play. The National Archives Administration has launched a special operation to recover scattered archives overseas, drawing up a list and a recovery action plan for archives lost to overseas institutions and individuals due to war and other reasons. Through the National Archives Administration, the State Administration of Cultural Heritage, the Ministry of Foreign Affairs, the Supreme People's Court, the Supreme People's Procuratorate, and the Ministry of Justice, specific recovery work is carried out by studying and working on international laws.

KEYWORDS

Digital Humanity, Displaced Archive, MISS Platform, International Image Interoperability Framework (IIIF), Linked Data

1. INTRODUCTION

According to historical reasons, many precious archives of our country are scattered overseas, including many rare historical archives. It is estimated that the number exceeds millions of volumes, mainly distributed in Japan, Southeast Asia, Europe, and North America. The historical archives stored overseas are an important witness to the splendid Chinese culture. It is an indispensable part of works for clarifying the development of Chinese civilization. In recent years,

David C. Wyld et al. (Eds): AIAA, DPPR, DSA, ICSS, IOTE, NLPTA, WEST, CACIT, BMLI- 2022
pp. 153-162, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.122312

the international archival community and many countries and regions have increased their attention and support for the collection, collation, publication, and digitization of overseas displaced archives. The International Council on Archives has also devoted decades of effort. The emergence of digital humanities as a new interdisciplinary research paradigm has provided many transformative thoughts and methods for the development of archival science and the full utilization of archives themselves. With the help of the multi-dimensional image intelligence system (MISS platform), this study hopes to provide a brand new research model and resource utilization ecology for these precious archives. Thus, it improves the use-value of overseas displaced archives, helps archives displaced abroad "return home" in a special way, promotes cultural cohesion and centripetal force, strengthens the cultural foundation, and promotes cultural confidence construction.

2. STATEMENT OF PROBLEM

Overseas displaced archives are spread in various forms in a wide range. Their preservation state is uneven, and many of those that have historical significance are handwritten. There are also different characteristics of handwriting. The multi-dimensional image intelligence system (MISS platform) developed by the author's team provides a suitable solution pathto realize the "return" and sharing of these archives within the limited scope of time and space. The MISS platform (http://miss.newwenke.com/sas/) can realize the retrieval, indexing, management, and appreciation of image archives, as well as the programming of the background language SPARQL Editor. It's worth noting that this platform takes "image" as the processing object, so the initial sorting of the archives is crucial, which needs us to store them in the format of "image." Based on the above application principles, this research intends to combine popular digital humanity technology such as Deep Learning, Knowledge Map, and Linked Data in the IIIF (International Image Interoperability Framework) to carry out research on the overseas displaced archives.

3. RESEARCH OBJECTIVES

Using MISS to realize the digital application of overseas displaced archives inspires the digital use of archives, especially as a platform for archives sharing.

The technical scheme and research ideas adopted in this study have a certain foresight and have been successfully utilized in Digital Humanity Project like “The newspapers within the Republic of China Era, and cultural heritage”. As for archival resources, the purpose of this study is to carry out the following research:

1.Based on understanding and mastering the current situations and the spreading history of overseas displaced archives, digitize archives of a specific region or a group with special significance as the research object, providing a reference to carry out mass photocopying, digital return, and utilization of overseas displaced archives.

2.The structure of knowledge related to organizing displaced archives is carried out through Ontology and Knowledge Graph using popular Digital Humanity technology; the semantic annotation and deep organization of image archive content are combined with the IIIF framework and Web Annotation model.

3.Using the MISS platform, based on semantic annotation of image archives, relate and combine knowledge (people, place, time, event) related to different data sets (sources), eliminate information silos between different archival resources, to realize the serialization of archives across time and space.

4. RESEARCH DESIGN AND METHODOLOGY

4.1. International Image Interoperability Framework，Iif

International Image Interoperability Framework is a set of standards that define the interoperability framework for digital libraries through a standard set of application programming interfaces (APIs). It provides a unified way to describe, distribute, and access images on the Web. In June 2015, under the guidance of IIIF, the British Library and 29 non-profit institutions such as Oxford University Library and Harvard University carried out image resource storage to ensure the interoperability and accessibility of global image resource storage. Use images as a carrier to promote the unified display and use of online resources such as books, maps, scrolls, manuscripts, music, and literature. The use of IIIF enables image resource storage institutions to break through the limitations of their resources and fully realize the interoperability of image resources with other collection institutions, which greatly improves the research ability of the institutions in the network data environment. Once IIIF was proposed, it quickly became a research hotspot in the field of GLAM (Art Gallery, Library, Archives, and Museum). At present, most major international cultural heritage research institutions have joined the IIIF framework.

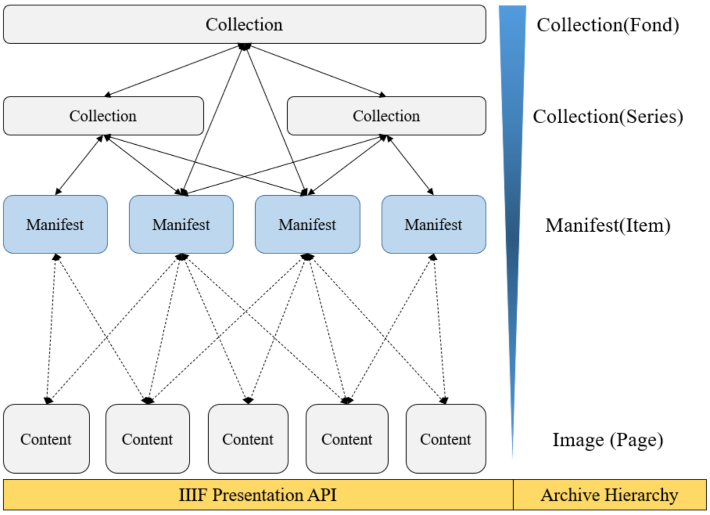

Fig. 1 Organization structure of archival thematic research knowledge

IIIF, as it were, provides an unprecedented new method. It is a set of standards-defined digital library interoperability frameworks that provide a unified method for describing, distributing, and accessing images on the Web through a standard set of application programming interfaces (APIs). This method uses a standardized image request format to share the digital content of images and improves the ability of online research of image resources. Developed through the joint efforts of several institutions, IIIF has quickly been adopted by the wider cultural heritage sector and is receiving more attention in digital humanities construction and research. Dispersal of archival digital resources is where the IIIF framework is put into use. Figure 1 shows the organizational structure diagram of archival image resources using IIIF, and "interaction" becomes the core idea of the whole architecture. 1. "Image interaction": the lowest level is the archival image resources collected by various institutions. These resources are organized into Manifest files through Canvas in the IIIF framework. The archival images contained in the Manifest file are interactive, that is, the archival images of different institutions can be shared and interacted with each other and cross-organized into the Manifest file of archives. 2. "Archive Interaction": Manifest files organized according to different archive images can also interact with each other, thus organizing them into "Archive Collection." 3. "Topic Interaction": Different research topics can also interact with each other and be reorganized into a set of topics at a higher level.

4.2. Linked Data

Linked data has become a key technology in digital humanities research in recent years, especially the integration of heterogeneous resources from multiple sources in interdisciplinary research. Notes that associated data and data association are not synonymous. All associated data can be regarded as data association. Linked Data is a lightweight implementation of the Semantic Web. It is not new data, but a new form of data presentation. Linked data is generally considered only if it conforms to the four principles of Linked Data outlined by Tim Berners-Lee in 2006.

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. Provide useful information about the thing when its URI is dereferenced, using the standards (RDF, SPARQL).
4. Include links to other related URIs in the exposed data to improve the discovery of other related information on the Web.

The main benefits of using Linked Data are:

1. More convenient data access: Your data can be accessed immediately in a machine-processable way through persistent URLs. In terms of sharing data, data from the persistent URI (http://www.mycompany.com/branches/1) is more efficient than in the data warehouse data (for example, the isolation of excel files and even private commercial database server)
2. The schema is more flexible: Your linked data does not follow a particular schema, there are no database tables, just a bunch of declarations, and you can add more statements whenever you want. Or let another data provider supplement the data by adding more statements.
3. More standard query language: Using popular or standardized vocabularies increases data interoperability and allows queries to be performed across multiple data repositories.
4. Data interconnection is more intelligent: the main function of linked data is to link the data with external resources. By linking to your data, you can enrich your data by sharing any other valuable information on the network.

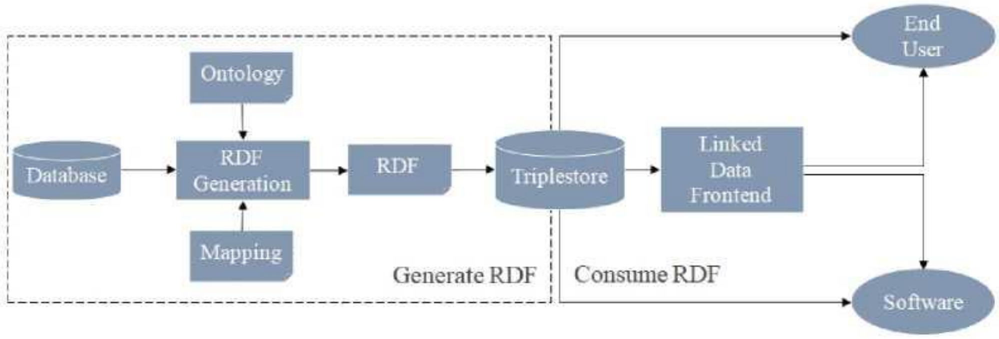

Fig. 2 Process framework of the linked data system

Fig. 2 shows the typical process framework of the linked data system, which is mainly divided into two parts: the generation and consumption of RDF data. GenerateRDF mainly aims at the processing process of RDF data, and RDF conversion is mainly carried out on archive data existing in the Database. During the conversion, the corresponding ontology should be designed, and the program coding should be carried out to map the ontology and archive field information in the Database. The resulting RDF data is recommended to be stored in the Triple Store. Currently, the threshold for structuring data RDF is getting lower and lower, and many mature tools (D2R, Open Refine, and R2RML) can easily convert it. During the Consume RDF phase, RDF data in the Triple Store can be directly called using an interface, such as publishing the data in the Triple Store using the SPARQL Endpoint. Of course, we can also develop some front-end interactive pages with a better interactive experience.

4.3. Ontology

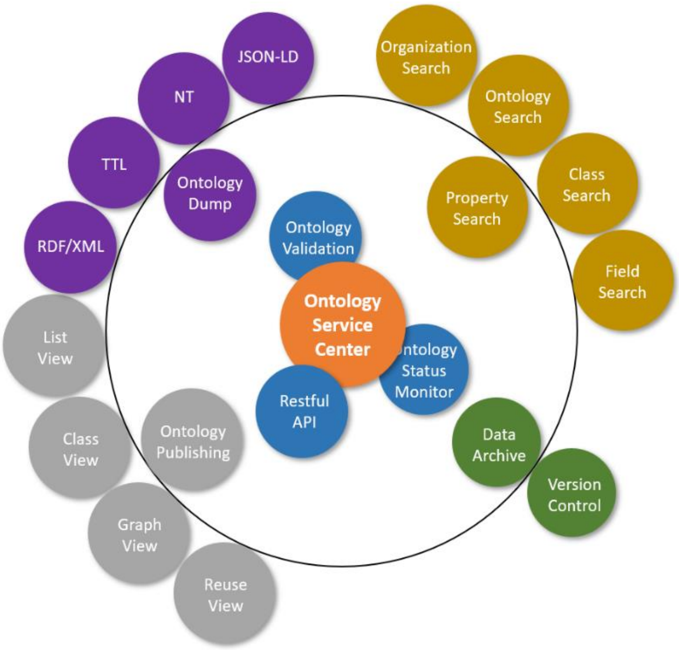

Fig. 3 Functional architecture of Ontology Service Center

The construction of the knowledge graph of overseas displaced archives mainly involves ontology construction, RDF structuring of data, and resource association among different data sources. Firstly, ontology is used to organize the owners, units, semantic-tagged content, and other data of the scattered archives. At present, there are more than 700 native word lists commonly used in LOV (Linked Open Vocabularies). In addition, the Ontology Service Center built by the research group also contains a large number of ontologies related to the digital humanities field. In addition, the Graphic Library in China has released lists of common words in the field of digital humanities, such as people, places, and times.

The ontology service center (OSC) framework, shown in Figure 3, composes of four main components: Ontology Visualization, Property Search, Version Control, Ontology Reuse, Ontology Validation, Ontology Publishing, and Ontology Status Monitor.

1.Data Dump.
OSC provides four data dumps, RDF/XML, TTL, RDF/JSON, and NT, which will be generated automatically real-time of ontology (RDF) model that is a set of Statements of this ontology.
2.Data Publishing/View.
The system supports four kinds of ontology view methods, Class View (C.V.), List View (L.V.), Graph View (G.V.), and Reuse View (R.V.).
-Class View. This view uses a tree structure to display the ontology, which makes it easy to

understand the hierarchical relationship between classes in the ontology.
-List View. This view presents the entire vocabulary on one page with an ordered list of classes and properties, followed by more detailed information panels further down the document.
-Graph View. This view visually displays the ontology with WebVOWL, a web application for the interactive visualization of ontologies. The visualization is automatically generated from the ontology graph.

-Reuse View. This view is achieved through the WebVOWL Editor application, which is designed to serve the skills and needs of domain experts with limited knowledge of ontology modeling.

1.Data Archive.

The archived ontology is saved as a file on disk, and the file name will contain the version number of the archive. In other words, only the latest ontology will be stored in the RDF store. The system will provide the results of the comparison between the archived version and the latest one. If necessary, the archived ontology can be rolled back and restored to the previous version in the ontology graph.

2.Data Search.

OSC enables searching for vocabulary terms (class, object property, data property), term comments, catalogs, namespace prefixes, and contributors.
In addition, you can perform ontology state monitoring, ontology validation, and RESTful interface APIs.

5. Empirical Case Analysis

Based on the core technology and framework, we construct a multi-dimensional image intelligent system (MISS) platform. The platform supports online sorting, publishing, reuse, semantic annotation, and other functions of image resources in various formats, and now supports online interaction of large-scale image resources, providing a solid technical fortress for the reuse of cultural heritage.

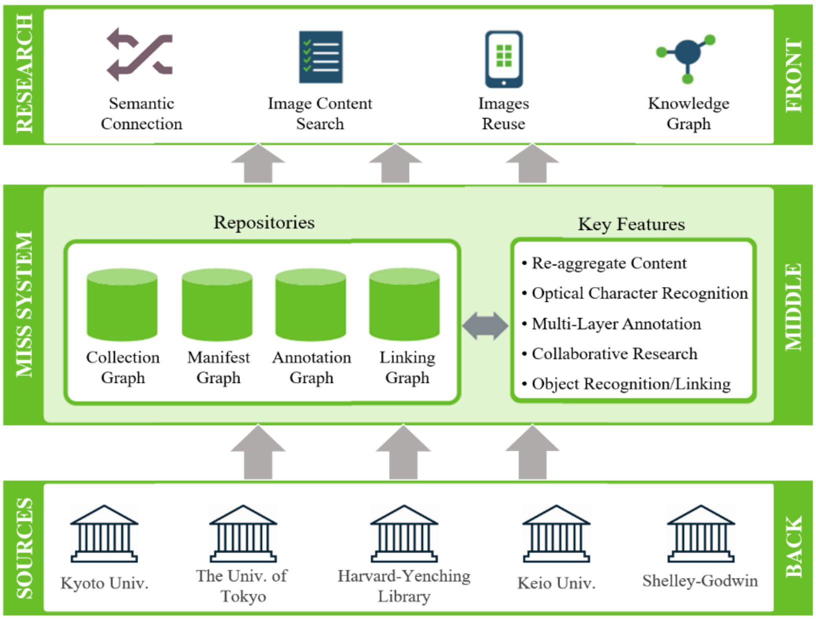

Fig. 4MISS solution for displaced archives

Figure 4 shows the solution for the integration of displaced archives using the MISS platform, which is mainly divided into three layers: front-end application, mid-level integration, and back-end resource import. At the bottom are archival resources from different institutions, which should ideally be published and shared under the standards required by the IIIF framework. The top layer is the specific archival research direction, mainly including Semantic Connection, Image Content Search, Archive Images Reuse, and Knowledge Graph. The middle layer provides the core functions and storage structure for the MISS platform. In the MISS platform, one can perform content re-aggregate, optical character recognition, multi-layer annotation, collaborative research, and object recognition operations. When using the MISS platform to integrate and organize archives, storage will be carried out according to the structure required by the IIIF framework, mainly using Collection Graph, Manifest Graph, Annotation Graph, and Linking Graph.

Next, some innovative ideas for the MISS platform in the utilization of archive resources are described below.

1.“One-click” lower technical threshold

MISS platform provides a “one-click” process for publishing and reusing archive images. Users can not only upload private archive images and generate IIIF resources but also import internet archive manifest resources. Through “One-click”, external resources can be reorganized on the platform. In addition, when reusing the archive image, the original image address will be inherited to the new manifest and form an image “gene chain.”

2.Multi-model annotation for “close reading”

According to the characteristics of Overseas Displaced Archives image resources, we propose a three-layer annotation model: image layer, object layer, and semantics layer. This model can enrich the content and cluster and link the objects in the images, which is convenient for users to quickly obtain and understand the deep meaning of images.

3.The symbiosis between images and text

OCR separated from images will lack the soul of Chinese characters. Therefore, the concept of real-time OCR is proposed, and third-party APIs are transferred to perform real-time OCR and manual proofreading of the text on the image. The generated OCR text can also be used as a machine learning corpus to improve theOCR accuracy. The concept of real-time OCR reduces the complexity of the conversion process of images and texts and breaks them into parts, which can better assist researchers in humanities.

4.Infrastructure construction and transmission of Overseas Displaced Archives

IIIF includes images from different countries and institutions and forms the “Image of Web.” As more and more organizations join IIIF-C, the images become more and more valuable. We believe that the IIIF will become the basic framework for the entire country and even the global archive organizations and further form the ecosystem. In the information technology age, openness and integration can be creative.

Here, taking as an example the collection archives of Keio University Library, Harvard-Yenching Library, Kyoto University Library, and Chester Beatty. Online integration of the resources of these institutions is made in the MISS platform.

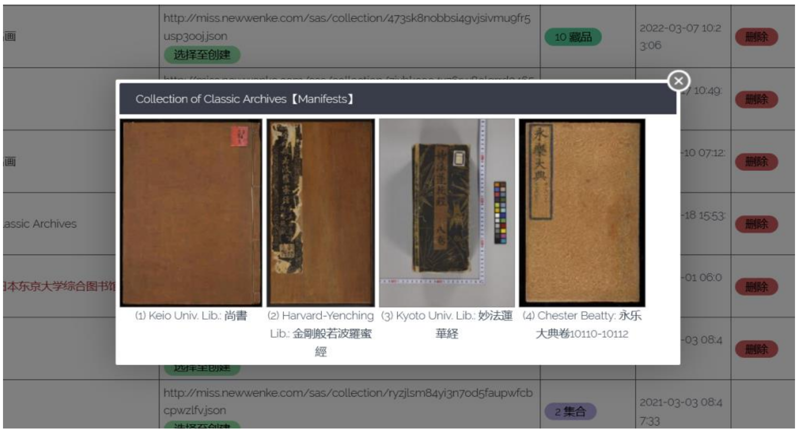
http://miss.newwenke.com/sas/collection/473sk8nobbsi4gvjsivmugfr5usp3ooj.json
选择至创建
10 藏品
2022-03-07 10:23:06
删除
10:49
-10 07:12
assic Archives
-18 15:53
日本东京大学综合图书馆
-01 06:0
-03 08:4
选择至创建
http://miss.newwenke.com/sas/collection/ryzjlsm84yi3n7od5faupwfcbcpwzlfv.json
2 集合
2021-03-03 08:47:33
Collection of Classic Archives【Manifests】
(1) Keio Univ. Lib.: 尚書
(2) Harvard-Yenching Lib.: 金剛般若波羅蜜經
(3) Kyoto Univ. Lib.: 妙法蓮華経
(4) Chester Beatty: 永乐大典卷10110-10112

Fig .4 Collection of Classic Archives Preview

Firstly, import the Manifest URI address of the archival resources of the four institutions into the MISS platform during integration.

Secondly, after importing, the resources of each institution can be presented independently. That is, a new Manifest URI address can be generated in the MISS platform.

Then, select the Manifest of resources of these institutions in MISS to generate a new Collection.

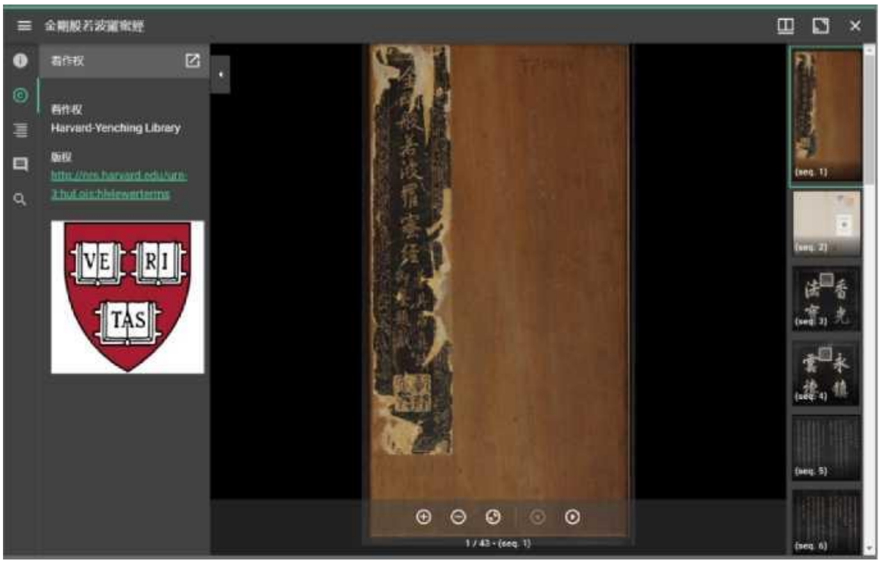

Fig. 6 Full text of the Vajra Prajnaparama Sutra

Figure 5 shows the preview screen for the new collection, where you can see the archives from the four sources. Note that the left and right operations in the MISS platform do not download digital images of online files. Figure 6 shows the screenshot of browsing the full text of Harvard-Yenching Library archives (Vajra Prajna Paramita Sutra) using Mirador 3.

After importing archive resources into the MISS platform, OCR recognition, semantic annotation and association can be carried out on the platform. Further research on the knowledge graph can even be carried out.

6. SUMMARY AND PROSPECT

In conclusion, according to the literature research, a large number of all the archival resources in China have been lost in museums and other institutions in Japan, the United Kingdom, the United States, and other countries due to wars. The cost for redeeming them is exorbitant, which is negative for China's national dignity and the dignity and value of the archives themselves. Therefore, the redeeming need to base on the strategic thinking of great power diplomacy. Effectively safeguarding the ownership and right of the recoursing of China's lost overseas archives, the integrity of China's national memory, and the national attribution of archives is imperative. According to the need of building cultural confidence, it is urgent to search for the scattered archives lost overseas due to war and other reasons at this stage, to maintain the dignity and integrity of Chinese culture. According to the promotion of the status of the archival discipline, the promotion of the discourse power of the discipline, and the need for the development of the discipline, China's archival discipline can explore the unique value of the discipline and improve the status of the discipline by carrying out the successful operation of recovering overseas displaced archives. To solve the discussed archives above, we propose four proposals：

First, promote the establishment of a normal cooperating and dialogue mechanism between archives authorities and related government agencies, embed archives work into the overall framework of digital government construction, digital economy development, and digital society construction, and explore the integrated development of archives management, data governance, and information resource management.

Second, foster a new partnership between archives departments and enterprises, think tanks, research institutes, industry associations, and other social organization, letting multiple social subjects play a role in the modernization process of the archives governance system and governance capacity.

Third, the National Archives Administration led a special operation to recover overseas displaced archives, drawing up a list and a recovery action plan for archives lost to overseas institutions and individuals due to war and other reasons.

Forth, carries out specific recovery work through the study and work of international law by the National Archives Administration, in conjunction with the National Cultural Heritage Administration, the Ministry of Foreign Affairs, the Supreme People's Court, the Supreme People's Procuratorate, and the Ministry of Justice.

ACKNOWLEDGMENTS

This research is granted financial support from the National Social Science Fund of China (19BTQ024).

REFERENCES

[1] Feng, H.-L., Jia, X.-S., Li, Y.-Z.(2018).Historical development and international trend of the disposing of displaced archives.Archives Science Bulletin.(04):4-9.

[2] Qu, C.-M.(2020).The "Emotional Turn" of Foreign Archives Studies.Archives Science Study.(04):128-134.

[3] Abood, P. (2011).Archive of the Displaced.Journal of the Association for the Study of Australian Literature : JASAL.11(1):2.

[4] Lowry, J.(2017).Displaced Archives. New York:Routledge,2.

[5] Munir K. and Anjum M.-S. (2018). The use of ontologies for effective knowledge modeling and information retrieval. Applied Computing and Informatics, 14(2):116-126.

[6] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol, 25: 1251-1255.

[7] Chen, T., Zhang, Y.-J., Liu, W. and Zhu, Q.-H. (2019). Several Specifications and Recommendations for the Publication of Linked Data. Journal of Library Science in China, 45(1): 34-46.

[8] Feitosa, D., Dermeval, D., Avila, T., Bittencourt, I.I., Loscio, B.F. and Isotani, S. (2018). A systematic review on the use of best practices for publishing linked data. Online Information Review, 42(1): 107-123.

[9] Garijo D. and Poveda, M. (2020). Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web. Applications and Practices in Ontology Design, Extraction, and Reasoning, IOS Press, Netherlands.

[10] Chen,T., Liu,W., Sun,X., Zhu,Q.-H., Zhao,Y.-X.(2021).A New Mode of Cultural Heritage Application under the Effect of IIIF and AI. Journal of Library Science in China,47(02): 67-78.pdf.

[11] Xia, C.-J.,Wang,L.-H.,,Liu, W.(2021). Shanghai memory as a digital humanities platform to rebuild the history of the city. Digital Scholarship in the Humanities, (3) URL.

[12] Chen,T.,Shan, R.-R.,Zhang ,Y.-J., Sun,X., Xu,X.(2021).Research on the Construction of Semantic Support Platform for Digital Humanities Research -- Taking ECNU-DHRS Platform as an Example. Library Journal:1-12.http://kns.cnki.net/kcms/detail/31.1108.G2.20200628.1836.002.html.(accesse07January 2021)

[13] Chen, T., Shan ,R.-R., Li,H.(2020)Research on Semantic Annotation of Image Resources in Digital Humanitie.Journal of Library and Information Technology in Agriculture,32(09): 6-14.pdf.

[14] Li, H., Chen, T.,Shan ,R.-R.(2019)Conversation across Time and Space: Construction of Calligraphy and Painting Memory Chain Based on IIIF-IIP Semantic Annotation Platform . Journal of Library and Information Science in Agriculture, 32(09):15-21.PDF.

[15] Liu, W., Hu, X.-J.,Qian, G.-F.,et al. (2012)RDA and Linked Data. Journal of Library Science in China, (1): 34-42.pdf.

[16] Liu, W. (2011)Linked Data: Concept, Technology and Application Prospect. Journal of University Libraries, 29(2):5-12.