DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications

'China Art in the Museums Overseas': Metadata Aggregation of Chinese Digital Collections Images

Xilong Hou
School of Communication
Qufu Normal University, China
houxilong2008@163.com

Xiaoguang Wang
School of Information Management
Wuhan University, China
wxguang@whu.edu.cn

Hua Jian
School of Information Management
Wuhan University, China
jianh54@whu.edu.cn

Abstract

In the World Wide Web, a very large number of online cultural heritage (CH) resources is made available through digital museums websites. They display collections images and share collection metadata, which makes opportunities for aggregating. 'China Art in the Museums Overseas' platform aggregated China digital collections from overseas institutions. This paper introduces the metadata aggregation workflow of this project, especially a unified data model used to solve the metadata standards heterogeneity. Besides, this platform is not simply a database, but also provides search, images semantic annotation, knowledge graph serves, linked open data interface and so on. Our work is an effort to resolve multi-source data aggregation and it is valuable. It improves the availability and discoverability of digital collections, and spread Chinese culture to the public more widely.

Keywords: Linked Open Data; metadata aggregation; digital collection; cultural heritage

1. Introduction

Cultural heritage is the witness and souvenir of human history, containing inestimable human wisdom. As the only unsuspended one among the four ancient civilizations, Chinese culture plays a vital cultural role around the world. However, for some historical reasons a lot of Chinese cultural heritages drifted out and scattered overseas, making it difficult for scholars and the public to access.

With the application of digital technologies, galleries, libraries, archives and museums (GLAM) institutions published digital collections on the web and made data sets more accessible and interoperable. As more and more institutions bring their data to the Semantic Web level, data integration in different cultural domains and institutions becomes possible. Aggregating rich Chinese digital collections, providing a unified platform with the linked data can effectively expand distribution scales, displaying Chinese cultural heritage intensively, and disseminating Chinese culture. This study puts forward a metadata aggregation framework of digital collections based on linked data, IIIF (International Image Interoperability) and knowledge graph, and the present the implementation of 'China Art in the Museums Overseas ' platform. And this platform enables cross-institution metadata aggregation of Chinese digital collection, and effectively resolves the problems of metadata schemas heterogeneity and the lack of semantic linkage in data integration. Besides, it also can support the digital cultural resources reuse and re-creation. Currently, this platform has aggregated Chinese digital collections from more than 10 overseas museums, including the Metropolitan Museum of Art, the Harvard Art Museum, the Chicago Art of Institute Chicago, the Asian Art Museum of the United States, and the Rijksmuseum and so forth, totaling more than 200,000 digital images and 130,000 metadata records.

2. Workflow of Metadata Aggregation

Driven by the movement of open access and linked open data, GLAM institutions became more open and shared their digital collections resources publicly, making possibilities for data integrating from multi-source collection resources. Thus, in order to aggregating Chinese digital collections in

https://doi.org/10.23106/dcmi.953137810

overseas and provide better data infrastructure and knowledge services, the project adopts the data-center model, that is, it integrates digital images and their metadata from different institutions, builds a unified data center locally, and semantic linked with other linked open datasets. Then, it combines the semantic enrichment and crowdsourcing technology to process and reorganize the collections metadata, and provides data access and analysis serves.

IIIF and schema.org vocabulary have been identified as efficient and lower technical barriers solutions for metadata aggregation in the domain of cultural heritage. (Freire el al.,2020) FAIR data principle provides guidelines to improve the findability, accessibility, interoperability and reuse of digital collections. Guided by the FAIR principle, the data aggregation workflow was divided into four main steps .(1) At first, we designed a unified data model; (2) then, used the data model to harvest, mapping and RDFizating the original data from different data source; (3) after been reprocessed, data were stored in an RDF database. Also, the data were interlinked with DBpedia and Wikidata to improve the availability. (4) finally, the platform provides data interface and service interface. Data are semantically enhanced, computed and analyzed, providing serve through serve interface, such as search & browser, building knowledge graphs and so on, which is useful for humanities researches. Also, you can access the linked open data directly through data interface. Figure 1 shows the workflow for digital collection data aggregation.

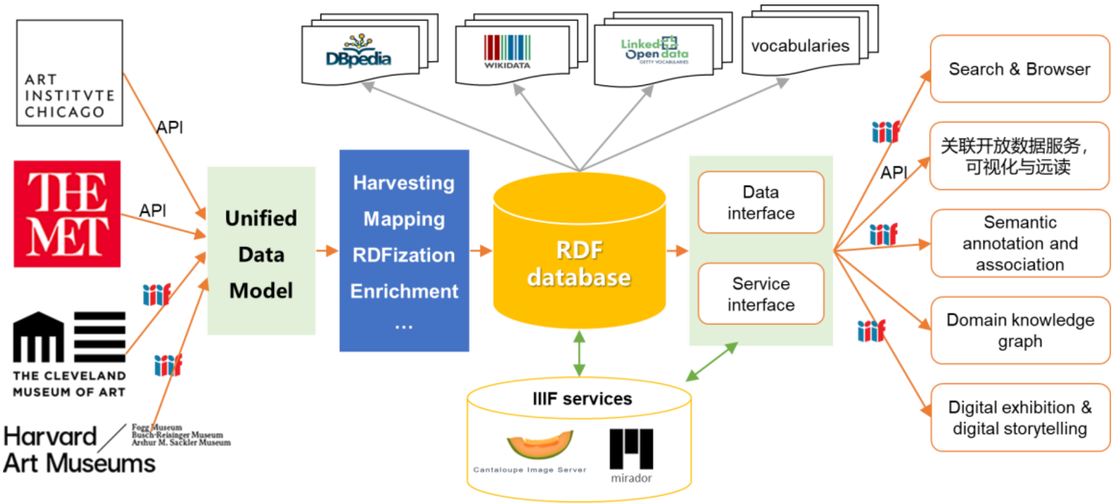


FIG. 1. Workflow for the aggregation of digital collections

2.1. Designing Data Model

The semantic diversity of multi-source and heterogeneous data will adversely affect data storage and management, and it is difficult to achieve semantic interoperability. The metadata standards used by different GLAM institutions are diverse, and so are the data publishing methods and formats (include linked data interfaces, APIs, OAI-PMH, CSV, XML and other formats). Therefore, in order to solve metadata interoperability, we summarized common metadata elements of different institutions and incorporated heterogeneous metadata schemes into a unified data model. The common elements are like labels, classifications, description, creators, etc. Based on the schema.org vocabulary, the data model for image metadata aggregation is designed and constructed, referring to the study of machiya et al. (2020). At the same time, the data model also reuses existing standards and vocabulary, such as SKOS and FOAF.

https://doi.org/10.23106/dcmi.953137810

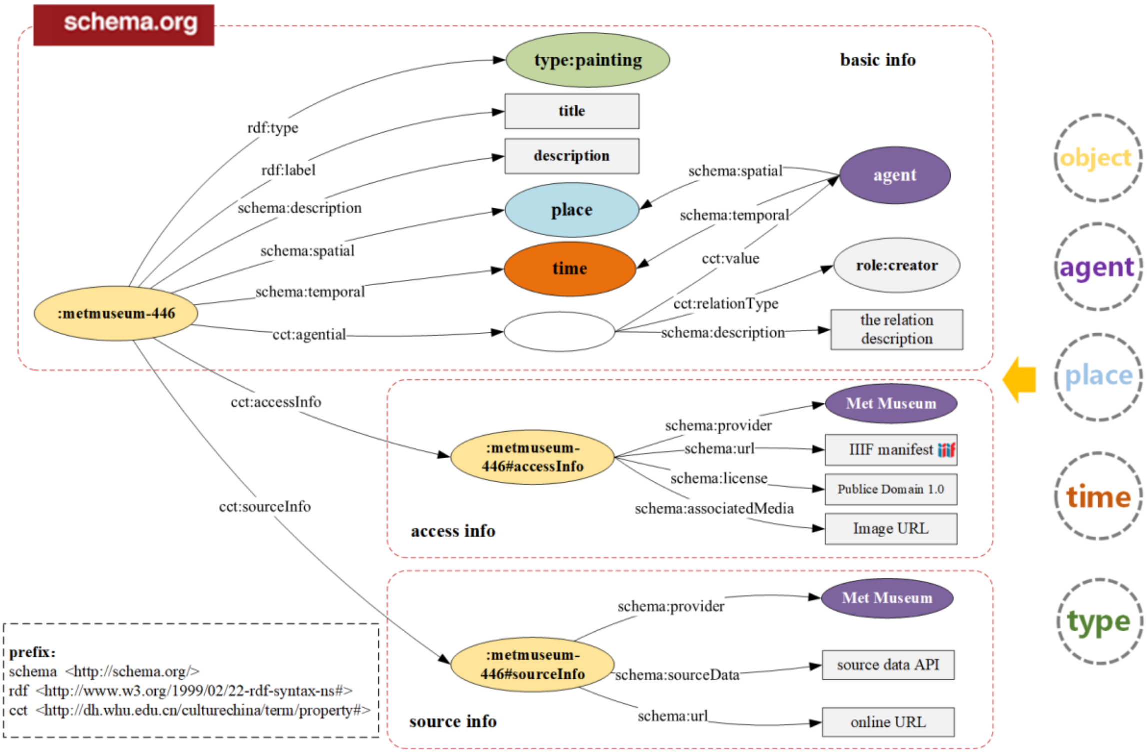


FIG. 2. The unified data model

The data model mainly describes digital collections from three aspects: the basic information, data source information and access information. The basic information records background information such as cultural relic type, name and description, creator, dynasty, excavation time, and material. The source information records the source and provenance of metadata to ensure data traceability and credibility. The access information is about open access to data, including the inked data URL, IIIF access address, and copyright notice. While designing data models, we were focusing on aggregating and organizing cross-institutional image metadata from some core concepts such as cultural relic objects, people and organizations, locations, times or dynasties, types, and themes.

2.2. Metadata Mapping

Data aggregating workflow in cultural heritage domain includes data harvesting, ingestion, mapping, indexing, storing, enriching and publishing linked open data (Siqueira & Martins, 2022). According to the data model designed above, the team incorporates images metadata under different standards into the unified framework. In this process, the data model acts as an ontology, ensuring the semantic integration between different metadata schemas, to solve the semantic interoperability (Alma'aitah *et al.*, 2020). After harvesting and acquiring the digital collection metadata from institutions, a mapping relationship is established between different data formats, types of heterogeneous metadata and data models. Thus, the digital collection image metadata is converted into a structured data format, stored in an RDF database. Providing collection metadata in a standards-compliant, structured format means tools can be used to access, integrate, and manage them. Next, the platform provides machine-readable and sharable data formats through data access interfaces such as SPARQL endpoints to improve the interoperability of metadata resources.

2.3. Semantic enrichment of Digital Collection Images

Semantic enrichment is an effective strategy to improve the data value through semantic technology and was applied as a main strategy to construct smart data in LAM (Zeng & Tan, 2021).

https://doi.org/10.23106/dcmi.953137810

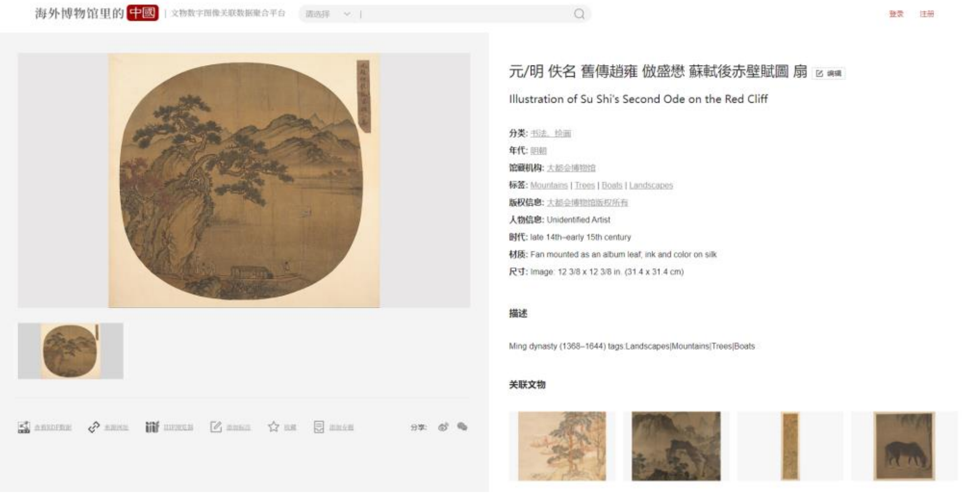
The semantic enhancement of digital collection images mainly includes the semantic enhancement of the images themselves and image metadata.

For the semantic enrichment of digital images, the machine learning technology such as visual object detection and semantic segmentation, as well as crowdsourcing, semantic annotation were used to annotate the objects in images and label them semi-automatically. For example, the deep semantic annotation framework (DSA-CH) proposed by Prof. Wang was effective not only in describing the fine-grained objects and their semantic information in an image, but also in revealing the invisible information of an image such as themes, concepts, cultural background and so on (Wang et al., 2021).

Image metadata records are also enriched by using entity recognition and entity linking with controlled vocabularies. Establishing a semantic link between entities such as persons, places, time, concepts, and types involved in image description information and external linked open datasets. Through the disambiguation of person names, the person entities are entity-aligned with vocabularies such as the Virtual International Authority File (VIAF) and the Getty Union Catalog of Artist Names (ULAN). At the same time, terms such as subject, keywords, and tags of the images are matched and aligned with thesaurus such as the Getty Art and Architecture Thesaurus (AAT).

3. Case Study：China Art in the Museums Overseas

In this section, we will introduce the 'China Art in the Museums Overseas' platform based on the work above. It was implemented for retrieving and displaying cultural heritage images for scholar and the public, providing classification retrieval and keyword retrieval. The platform builds IIIF service, and can uses mirador viewer[1] to display and browse images, and also supports semantic annotation to those images. The detailed information display page of the collections is shown as Figure 3, which not only displays the image metadata information, but also identifies the URLs of the data sources, the IIIF manifest and RDF data in the form of hyperlinks. The linked data browsing page of the collection is as Figure 4.



元/明 佚名 舊傳趙雍 倣盛懋 蘇軾後赤壁賦圖 扇

Illustration of Su Shi's Second Ode on the Red Cliff

分类: 书法、绘画
年代: 明朝
馆藏机构: 大都会博物馆
标签: Mountains | Trees | Boats | Landscapes
版权信息: 大都会博物馆版权所有
人物信息: Unidentified Artist
时代: late 14th–early 15th century
材质: Fan mounted as an album leaf, ink and color on silk
尺寸: Image: 12 3/8 x 12 3/8 in. (31.4 x 31.4 cm)

描述

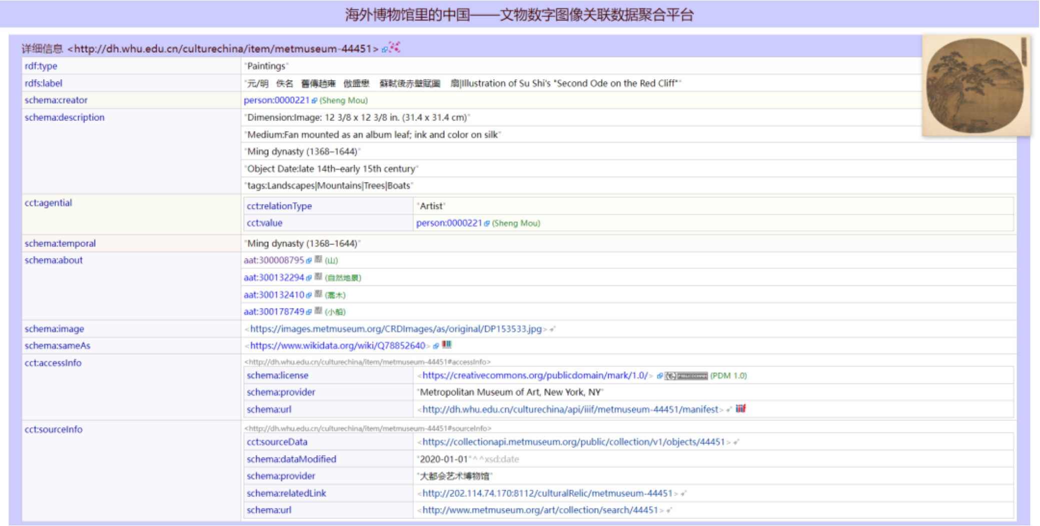
Ming dynasty (1368–1644) tags:Landscapes|Mountains|Trees|Boats

关联文物

FIG. 3. The detailed information page of the collection

[1] https://projectmirador.org/

https://doi.org/10.23106/dcmi.953137810



海外博物馆里的中国——文物数字图像关联数据聚合平台

详细信息 <http://dh.whu.edu.cn/culturechina/item/metmuseum-44451>

rdf:type		"Paintings"
rdfs:label		"元/明 佚名 舊傳趙雍 仿盛懋 蘇軾後赤壁賦圖 扇|Illustration of Su Shi's "Second Ode on the Red Cliff""
schema:creator		person:0000221 (Sheng Mou)
schema:description		"Dimension:Image: 12 3/8 x 12 3/8 in. (31.4 x 31.4 cm)"
		"Medium:Fan mounted as an album leaf; ink and color on silk"
		"Ming dynasty (1368–1644)"
		"Object Date:late 14th–early 15th century"
		"tags:Landscapes|Mountains|Trees|Boats"
cct:agential	cct:relationType	"Artist"
	cct:value	person:0000221 (Sheng Mou)
schema:temporal		"Ming dynasty (1368–1644)"
schema:about		aat:300008795 (山)
		aat:300132294 (自然地景)
		aat:300132410 (喬木)
		aat:300178749 (小船)
schema:image		<https://images.metmuseum.org/CRDImages/as/original/DP153533.jpg>
schema:sameAs		<https://www.wikidata.org/wiki/Q78852640>
cct:accessInfo		<http://dh.whu.edu.cn/culturechina/item/metmuseum-44451#accessInfo>
	schema:license	<https://creativecommons.org/publicdomain/mark/1.0/> (PDM 1.0)
	schema:provider	"Metropolitan Museum of Art, New York, NY"
	schema:url	<http://dh.whu.edu.cn/culturechina/api/iiif/metmuseum-44451/manifest> iiif
cct:sourceInfo		<http://dh.whu.edu.cn/culturechina/item/metmuseum-44451#sourceInfo>
	cct:sourceData	<https://collectionapi.metmuseum.org/public/collection/v1/objects/44451>
	schema:dataModified	"2020-01-01"^^xsd:date
	schema:provider	"大都会艺术博物馆"
	schema:relatedLink	<http://202.114.74.170:8112/culturalRelic/metmuseum-44451>
	schema:url	<http://www.metmuseum.org/art/collection/search/44451>

FIG. 4. Publishing the linked data of the digital collection

The core function of the platform is to realize the integration and linking of digital collection across institutions, and make it convenient to obtain all collection data related to a specific subject or person. For example, Figure 5 shows all the masterpieces of the painter, "Yun Shouping". Users can access and discover digital images and metadata from different institutions such as the Metropolitan Museum of Art, the Princeton University Art Museum and other institutions with a simple search action. The platform apples Ontodia[2] to visualize, navigate and explore linked data. Semantic linkages are established by topics, keywords, time, etc.

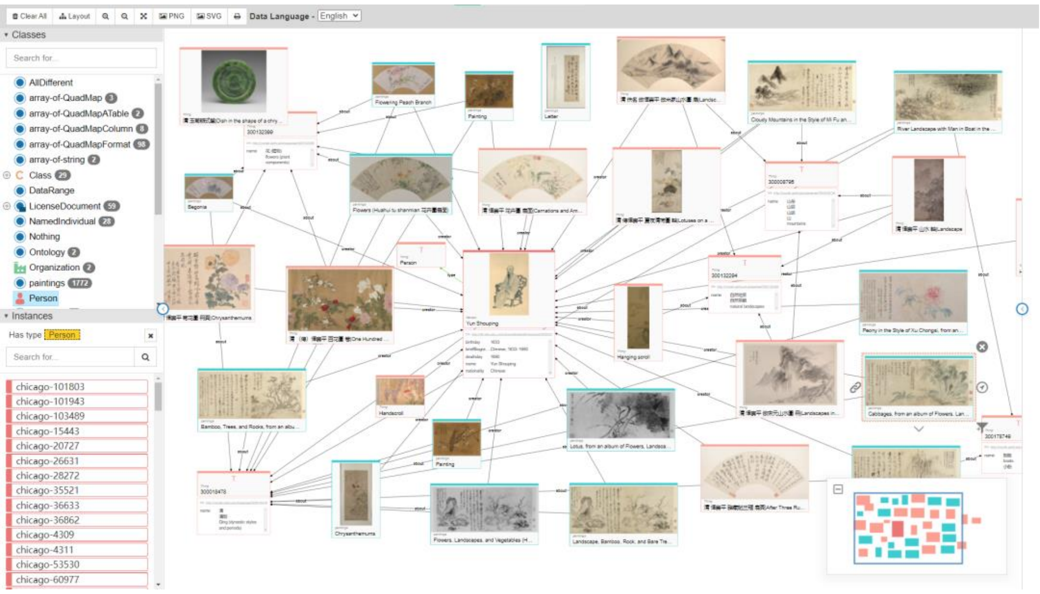


FIG. 5. Visual Browsing of linked data

The platform is developing innovative applications such as online curation, digital narrative, and visualization of distance reading to achieve in-depth interpretation of digital collection resource, as well as knowledge mining and discovery. At the same time, the neural network is used for object detection and semantic segmentation of cultural heritage images, and a fine-grained cultural heritage image knowledge base is under construction as well.

[2] https://github.com/metaphacts/ontodia

https://doi.org/10.23106/dcmi.953137810

4. Conclusion

This paper presents preliminary results of methods and practices for metadata aggregation of digital collections. Our primary goal is to find a solution to aggregate digital collections related to Chinese culture published by overseas GLAM institutions. We put forward a unified data model and aggregation workflow of heterogeneous digital collections, and uses technologies and methods such as linked data, IIIF and semantic enrichment to build the website, 'China Art in the Museums Overseas'. In the future, we will further strengthen cooperation with cultural institutions, enrich the integration of digital collection resource, and promote innovative applications of the digital cultural resources.

Acknowledgements

This work was supported by national social science foundation of China (21CTQ041, 21&ZD334). We thank the reviewers for their comments, which helped to improve the article.

References

Alma'aitah, W. Z. A., Talib, A. Z., & Osman, M. A. (2020). Opportunities and challenges in enhancing access to metadata of cultural heritage collections: a survey. Artificial Intelligence Review, 53(5), 3621-3646.

Freire, N., Robson, G., Howard, J. B., Manguinhas, H., & Isaac, A. (2020). Cultural heritage metadata aggregation using web technologies: IIIF, Sitemaps and Schema.org. International Journal on Digital Libraries, 21(1), 19-30.

Machiya, D., Okuda, T., and Kanzaki, M. (2020). Japan Search RDF Schema: a dual-layered approach to describe items from heterogeneous data sources. In International Conference on Dublin Core and Metadata Applications, 2020, 21-25.

Siqueira, J., & Martins, D. L. (2022). Workflow models for aggregating cultural heritage data on the web: A systematic literature review. Journal of the Association for Information Science and Technology, 73(2), 204-224.

Wang, X., Song, N., Liu, X., & Xu, L. (2021). Data modeling and evaluation of deep semantic annotation for cultural heritage images. Journal of Documentation, 77(4), 906-925.

Zeng Marcia Lei & Tan Xu. (2021). Semantic Enrichment of Data-Interpreting the New Trend of LAM Data in Supporting Digital Humanities. Digital Humanities Research, 1(01),65-86.

https://doi.org/10.23106/dcmi.953137810