RESEARCH AND PRACTICE ON DOMAIN ONTOLOGY OF ANCIENT CHINESE ARTIFACTS

Ye Yipei
The Palace Museum
Beijing, China
yeyipei@dpm.org.cn
0000-0002-7543-1732

Abstract – Collection information is one of the important resources that has long-term preservation value to museums. The paper proposes a domain ontology centered on ancient Chinese artifacts, aiming to help any organizations or individuals who own ancient Chinese artifacts collection better organize and manage their collection information, furthermore, improve collection search experience so that users can retrieve objects in the collection more easily and efficiently to ensure the availability of information as the digital object. The paper elaborates the creation process of "Ancient Chinese Artifacts Conceptual Reference Model" (ACACRM), summarizing and sharing practical experience for providing reference for museums with similar needs and application scenarios.

Keywords – Museum: Collection Searching: Ontology: Knowledge Graph: The Palace Museum

Conference Topics – Exploring the New Horizons

I. INTRODUCTION

Collection information serves as primary source of descriptive information by recording and interpreting collection through the iconological perspective and associated perspectives in the context of art history, natural history, anthropology, and science. It is one of important information resources of long-term preservation value to the museum. As a fundamental layer underlying the higher level of hierarchies in the Preservation Pyramid, the availability needs to be ensured by reasonable action. If information could not be obtained with availability and accessibility, it is subjectively considered non-existent by the public. Therefore, museums should take action to sort out and bring relevant information to users within the ocean of information, including actions of updating searching algorithm and improving online searching experience.

The majority of museums' collection online searching are simple keyword search based on collection management system or collection information system where collection information is recorded and stored. It finds the exact matched term in the title of an object and uncovering variants, misspelled of the term. It is bound by the limits of not being sufficiently powerful to find associated or broader terms or to express complex queries. The Palace Museum turned to knowledge graph as a methodology for making collection information more accessible and making collection much easier for users to find. For better knowledge organization, the Palace Museum is creating "Ancient Chinese Artifacts Conceptual Reference Model" (ACACRM) to serve as a domain ontology. This paper demonstrates the construction of ACACRM and presents a framework of it.

II. REVIEW OF MUSEUMS' APPROACHES TO KNOWLEDGE ORGANIZATION

Museums are memory institutions for recording, preserving, and disseminating the history of material culture. With the widespread adoption of Web2.0 information technologies in the early 21st century, museum visitors increasingly assume they will have a comprehensive, instantaneous and interactive access to museum collections, which in turn builds an expectation that museums function as knowledge spaces equivalent to libraries. However, unlike library or archival counterparts, museums have a knowledge framework exceedingly heterogeneous, reflecting the fields of art history, natural history, anthropology and science [1]. Therefore, museums need a complex and specialized knowledge organization system to describe the diverse ranges of their objects.

Over the latter half of the 20th century to nowadays, efforts have been made to standardize collection cataloguing within and across domains for collection description, management and information exchange. The efforts include as follows.

A. Iconclass

It is a classification system widely accepted in Netherlands designed for art and iconography for the description and retrieval of subjects represented in images. It contains 28,000 hierarchically ordered definitions divided into ten main categories.

B. Getty Vocabularies

17th International Conference on Digital Preservation
iPRES 2021, Beijing, China.

DOI: 10.1145/nnnnnnn.nnnnnnn

Getty Vocabularies contain structured data related to art, architecture, and material culture for cataloguing and retrieval purposes, published as linked open data so that they are interlinked, accessible and shareable on the semantic web.

C. CIDOC-CRM

It is a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information across diverse and dispersed datasets.

On the basis of the above, it can be concluded that more and more museums are thinking strategically about distributing interoperable data as a part of a broader knowledge web in cultural heritage field.

As a Chinese museum whose collection are mainly ancient Chinese artifacts, the Palace Museum finds that the thesaurus and data resources released by European and American institutions are not entirely applicable for its collection as they are primarily rooted in religious subject matters or representational matters in the western art and history. Large amounts of knowledge related to Christianity, the Bible, mythology and ancient western history cannot be used. Concepts related to Chinese culture are organized in deep hierarchical structure and coarse-grained, lacking necessary content. Moreover, Chinese perspective on systems thinking as a cognitive mode for knowledge organization is somehow different. For instance, terms for cataloguing calligraphy was arranged in the layers of Art & Architecture Thesaurus as below, which seems to be an unpredictable layer in the hierarchical structure:

Objects Facet

…Components (hierarchy name)

……Components (objects parts)

……<components by specific context>

……information form components

……<script and type forms>

……scripts (writing)

……<scripts by forms>

……Chinese scripts

……seal clerical script

……cursive script (Chinese scripts)

……running script

……seal script

For better fitting Chinese usage scenario specifically for ancient Chinese artifacts, it is inadvisable for domestic museums to consider knowledge organization to be a mere replica by translating from English. The knowledge organization should be constructed in Chinese socio-historical context. Some relevant explorations have been made as follows.

D. Knowledge Organization in Chinese Socio-historical Context

1. Chinese Iconography Thesaurus

Chinese Iconography Thesaurus (CIT) of Victoria and Albert Museum is the first thesaurus of Chinese iconography, serving as a valuable tool for indexing and accessing images. It is of significance to create a classification scheme rooted in the specificity of Chinese visual culture and provide a conceptual framework. However, iconography as a study focusing on the content of images, CIT is designed for visual art including paintings, prints, and photographs, lacking in vocabularies of areas including classification, materials and techniques.

2. Faceted Thesaurus of Chinese Cultural Heritage

Faceted Thesaurus of Chinese Cultural Heritage (FTCCH) was carried out by the University of Science and Technology Beijing. It is the first thesaurus that focus on the Chinese cultural object concepts and terms, arranged in five facets in its draft including the materials, style and periods, physical attributes, activities, and objects [2]. On the basis of traditional classification methods and the Getty Art & Architecture Thesaurus (AAT), it shows a strong adherence to AAT's hierarchy design, concepts and terms, which may be considered deviant and difficult to apply in practice in some Chinese museum specialists' point of view.

3. Explorations of Domestic Museums

Relevant explorations include Dunhuang Mural Thesaurus Ontology jointly produced by Dunhuang Academy and Centre for Digital Humanities of Wuhan University, Dong Qichang Digital Humanities Exhibition System by the Shanghai Museum, Knowledge Graph Construction based on 100 selected objects launched by the Shanxi Museum and Tianjin University. These applications to specific scenarios explored in-depth knowledge representation of a particular subject, useful but scattered in different knowledge domains. A top-down approach for knowledge sharing and collection information exchanging is in need.

The Palace Museum, under the circumstances, conducted a research on the ontology of ancient Chinese artifacts, serving the dual purposes of upgrading collection searching experience as well as integration of diverse and dispersed collection information sources by a common and extensible semantic framework. In 2020, the project applied a research subject "Concept Reference Model of ancient Chinese artifacts applied in semantic search" of Ministry of Culture and Tourism of the People's Republic of China. The outcomes include a domain ontology (ACACRM), a series of controlled vocabularies and a knowledge graph all available for free.

III. THEORETICAL GROUNDING OF ACACRM

A. Knowledge Organization and Knowledge Representation

With the purpose of making information available, several methodologies, systems and tools are developed in Knowledge Organization (KO) for classification, indexing and searching. In particular, museum collections are typically classified by object type and indexed by properties such as title, creator and materials. Indexing by title and creator is straightforward but indexing by material or subject is otherwise, as it requires an analysis of collection by experts and precise terms taken from a controlled vocabulary. Take the object "Pale-yellow Glaze Bowl with Polychrome Enamel Design of Orchid and Rock" as the search target. Queries such as "yellow glaze", "polychrome enamel", "ceramics", "bowl" are supported in KO. Expanded queries such as "low temperature glaze", "floral design" are not, as shown in Fig.1.

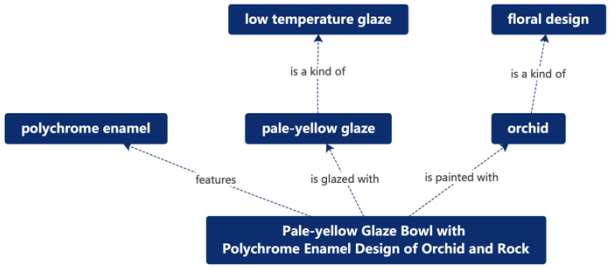

Figure 1 Relevant Knowledge of the Search Target

KO is limited in expressivity as it fails in situations when users do not know what terms should be used or how to describe the object directly [3]. A more expressive technique is needed to support queries like that, which is Knowledge Representation (KR).

As a field of research in artificial intelligence, KR is a mean by which mankind express things about the world, the medium of expression and communication in which human tell the machine (and perhaps one another) about the world [4]. A number of approaches have been proposed and semantic network is of them. It was invented to address the growing need for a KR framework, represented by graph structure G=(V,E). Knowledge can be described in natural language by semantic network. In 2001, Tim Berners-Lee coined the term "Semantic Web" aiming to "provide a common framework that allows data to be shared and reused across application, enterprise, and community boundaries [5]." Grounded on the hierarchical layers illustrated in the Semantic Web Stack, it is possible to construct a massive knowledge base in semantic web. Moreover, semantic web makes semantic search possible. Semantic search goes beyond the text and knows the underlying meaning of users' queries. It is very efficient when users misspell or do not say it correctly for what they want to search for. And semantic search performs better when confronting with ambiguity of lacking context.

B. Ontology and Thesaurus

Ontology as a framework within a specified domain of interest provides formal specifications and definitions of concepts. Thesaurus with selected list of words and phrases being used to tag collections aim to organize knowledge for subsequent retrieval. The combination of two not only set statement on how terms can be connected in semantic relations but also allow different vocabularies to be mapped onto the ontology. According to Project 05CTQ001 supported by the National Social Science Fund of China, integrating thesaurus with ontology is theoretically feasible by studying the definitions of thesaurus comparatively with those of ontology [6].

IV. THE CREATION OF ACACRM

A. Combining Two Approaches for Ontology Building

The knowledge scope of the museum industry is narrower and deeper compared with general knowledge. Intelligent achievements are mainly textual context. Knowledge extraction from unstructured sources requires a model which needs a large amount of training samples and manual labour for persistent error correction. Thousands of descriptive note texts may not produce a satisfactory result. Therefore, the Palace Museum explores manual ontology building based on experts' explicit and tacit domain knowledge by integrating top-down and bottom-up approaches.

B. The Domain and the Scope

Ancient Chinese Artifacts Conceptual Reference Model (ACACRM) is a domain ontology extensible for concepts and information relevant to ancient Chinese artifacts, providing terminology reference, supporting knowledge transformation and merging, knowledge sharing and reuse among different systems for any organizations and individuals who have ancient Chinese artifacts as collection.

The scope of model should firstly be started with the research of categories for description of collection. Two cataloguing guidelines were used as basis: The Palace Museum Collection Cataloguing Code and Categories for the Description of Works of Art. The combination of the two could cover relevant fields of cataloguing data as wide as possible, making ontology consistent to cataloguing terms.

1. The Palace Museum Collection Cataloguing Code

It offers guidance in the creation of collection cataloguing for collections of all kinds. The Palace Museum has the vast holdings of sacrificial vessels and ancient jade artifacts from the earliest dynasties of Chinese history; ceramics comprehensively reflecting the continuous development of Chinese ceramics throughout the past 8,000 years; paintings and calligraphy dating to as early as the seventh century; and antiquities of the imperial collections of all kinds. The representativeness and comprehensiveness of the collection make cataloguing standard of the Palace Museum could become a peer reference.

2. Categories for the Description of Works of Art

It represents common practice and advises best practice for cataloguing, based on surveys and consensus building with user community, providing a set of guidelines and cataloguing rules for the description of art, architecture, and other cultural works. The process of terms enumeration and preliminary classification is illustrated in Table 1. Terms classified as a candidate class will wait for further dealt with.

TABLE I
Terms Pre-treatment (sample)

The Palace Museum Manual of Collection Cataloging					
Term	Type	Ontology	Domain	Range	In ACACRM
Name	Free text	Data Property	any	xsd:string	
Classification	Controlled list	Class			
Creator	Controlled list	Object Property	Collection	People	was written by was painted by
Creation Date		Object Property	Collection	Time: Period/ Reign/ Common Era	was created in

On the other hand, extracting user interests from search queries will give a hint about extending the scope of ontology. The analysis process to find out users' interests is shown below.

1) Pulled users' search query history record.

2) Exported an excel file with search query text list in descending order of frequency.

3) Set the minimum baseline of frequency and conceptualized search query text which are above the minimum baseline.

4) Chose a proper word or phrase to refer to each concept set as a candidate term for class in ontology, such as motif, pattern, color, etc.

5) Created a chart (Fig.2) displaying concept set on the horizontal axis. It is shown that 30% of users submitted search queries of object/work type, 25% of visual design, which means that these concept sets should be enriched with an extensive vocabulary in the knowledge graph.

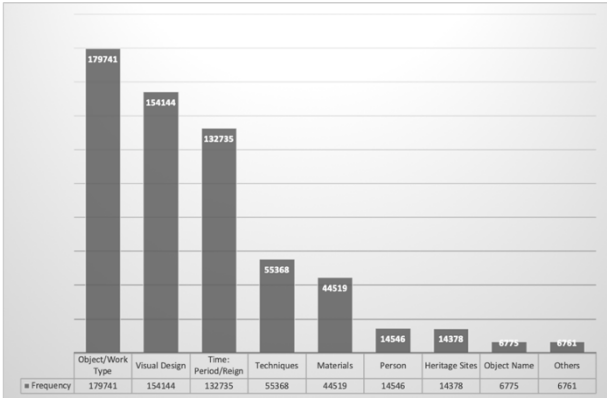

Figure 2 A Chart Showing Users' Interest

6) Took record of special search queries including colloquial expressions which were rarely seen in collection cataloguing.

Candidate terms for class can be further expanded by asking questions as follows: does the motif depicting human activity illustrate a historical event or refer to an allusion in literature and other culture contexts; do individuals have any social roles, such as artistic identities, court officers and administrators; is the object archaeological related to a site or an architecture which can be further related to a geographic location or an administrative unit. Sequence questions will lead to various fields that may help explore the domain of ontology.

So far, candidate terms for class are discrete waiting to be grouped and aggregated into a higher level. For example, motif, pattern and color are visual design elements for enhancing aesthetic appeal, hence making "visual design" a top-level concept set. Grouped terms could be very helpful for thesaurus design.

C. From Thesaurus to Ontology

The process of ACACRM construction is conducted using a hybrid top-down and bottom-up ontology building approaches. Top-down process starts with the most general concepts and subsequent specialization of the concepts by referring to some authoritative and comprehensive classification thesaurus not necessarily limited to museum domain. On the other hand, manual or semi-manual extract terms from authoritative publications and text materials that are reviewed by experts. Enumerating terms as much as possible without considering their semantic overlap will lead to words grouping in semantic clusters. Words will be properly grouped by the repetition of seme within semantic field. A Word with the most specific meaning will be an instance of ontology and its hypernym will be its class. Convergence of top-down and bottom-up needs a series of middle ontological classes to complete the hierarchy.

Take the hierarchical structure of "Tools" in the "Object/Work Type" thesaurus as an example. As a general concept, subsequent specialization of "Tools" was made a reference of Classified Chinese Thesaurus K875 "Objects for Various Purposes". Top-down construction method was applied to Level1 to Level4, shown in Table 2. On the other hand, relevant collections were reviewed, and terms were extracted from their title to be the bottom term in the hierarchy, such as Lei and Si in Level6. Certain joint terms which was conceptually generalized from the bottom terms were needed to make them a whole hierarchy.

TABLE 2
Partial Hierarchy of "Tools"

L1	L2	L3	L4	L5	L6	L7
Tool						
…….Tools of production						
……Prehistoric hunting and gathering tools						
……Tools of agriculture, forestry, livestock farming and fishing						
……Tools of agriculture						
……Tools for preparing land						
……Lei						

.....Si
.....Tools for irrigating land
.....Jie Gao
.....Lu Lu
.....Harvest tools
.....Reaping tools
.....Sickle
.....Threshing tools
.....Liu Zhou
.....Sowing tools
.....Weeding tools
.....Tools of forestry
.....Tools of livestock farming
.....Tools of fishing
.....Tools in handicrafts

Each word needs to be created a registration form. Words with a revised form will be added in the thesaurus. The registration form is set on the basis of the Dublin Core. The template is shown in Figure 3.

Figure 3 Word Registration Form (sample)

D. *Reuse Cross-language Object Properties*

Object properties (OP) are restriction added onto classes to describe the relationships between concepts. For better readability and conciseness, Domain-ObjectProperty-Range should be adhering to the syntax and grammatically correct in natural language. Mandarin Chinese is generally taken to be an SVO (stands for "Subject-Verb-Object"), which very closely matches English word order. Therefore, Domain is mapped to S (the agent which acts), ObjectProperty to V, and Range to O (the patient which is acted upon). Take an OP "depicts a historical event" as an example. It has "Storytelling drawing" specified its domain and "Event" as its range. It states that "a storytelling drawing depicts a historical event of (event)."

ACACRM reused partial OPs of Getty Vocabulary Program ontology (GVP ontology). Cross-language reuse is not blind acceptance. The following text will share practice and experience from the Palace Museum.

1) Textually identical OPs were additional remarked to be easily distinguished according to their domains and ranges.

For example, aat2211_produce, aat2425_produces, aat2428_produces are respectively remarked with Chinese characters to show their restrictions on OPs (Fig. 4). Otherwise, textually identical OP may increase difficulty in subsequent writing cypher queries based on relationship. Textual difference helps narrow down relationship types.

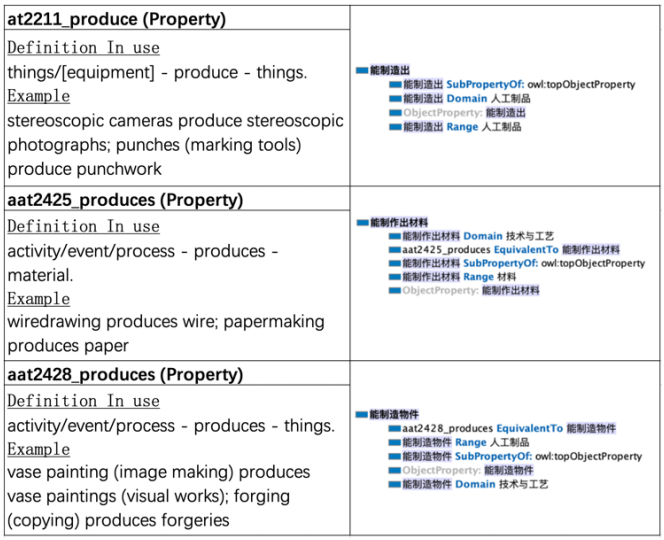

Figure 4 Samples of Reuse Cross-language OPs

2) Redefinition of the domain and range of an OP may be needed. Since ACACRM shares a different class structure with any other ontologies, it is necessary to transfer domains and ranges to classes with similar description.

3) Translation of OPs may change the part of speech.

In English, the predicate always contains a verb or verbs that link with the subject and object, fluently expressed in active or passive voice formula. Some sentences in the active voice can also be expressed in passive voice. In Chinese mandarin, active voice is more commonly used. And passive voice is used with lexical items, usually functional words such as "*bei*" and "*rang*" especially "*bei*" as a passive voice marker. The passive voice sentence is formed by the structure "Subject-*bei*-Object-Verb-AspectParticle", which is not adhering to the structure of Domain-ObjectProperty-Range. Therefore, part-of-speech shift is in need. From the Palace Museum's practical experience, the conversion from verb to complement of the subject which can be a noun, or an adjective is commonly applicable.

For example (Fig. 5), aat2801_conjuncted_with was converted from a verb in passive voice to the subject's noun complement, which stated as "sth's conjunction item is sth."

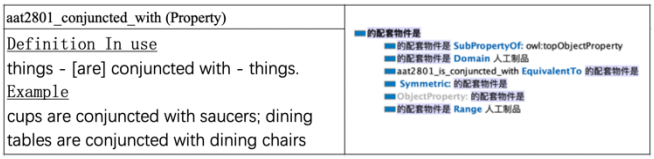

Figure 5 Sample of Change the Cart of Speech

Under the circumstance, ACACRM only reused one of OPs that are inverse functional in GVP ontology, the one whose translation is adhering to SVO sentence structure. The absence of the other one will not have an adverse effect on subsequent retrieval if graph database is used for semantic search. It is not necessary to add duplicate relationships in the opposite direction.

E. Mapping ACACRM in a Graph Database

Although graph databases including Neo4j are often described as schema optional, which means it is not necessary to create a schema up front. But for better and efficient data management of large amounts, an ACACRM-GDB mapping rule is proposed as follows in Table 3.

TABLE 3
ACACRM-GDB Mapping Rule

ACACRM	Graph Database
Class	Label
Object Property	Relationship
Data Property	Property
Instance	Node

Variables such as object name, object ID, measurement including size and weight, are infinite observational units which are considered as the data property in ACACRM, namely property in graph database. Property created with numeric value may be participated in subsequent retrieval. For example, a curator is trying to search collections whose height is less than 50 cm. His/Her Cypher query would be:

```
match (n:collection)
where n.'height'>50
return n.name
```

The cypher query will fail if the numeric value is suffixed with a string of a unit of measurement. Therefore, default measurement units should be created, such as centimeter as the unit of length and gram as the unit of weight.

F. Current Progress and Future Prospects

ACACRM is currently developed in progress and far from a final work. The paper presents the visualization of ACACRM framework to help overview this domain ontology (Fig. 6).

Figure 6 Visualization of ACACRM Framework

Mini-knowledge graph was constructed on the basis of ACACRM framework. The following expectations were achieved by tests of small-scale data.

1) When a user's search query does not correspond to any legal characters in collection cataloging, it is still a successful query.

For example, if users create queries to search collections with "flower/floral design/plant pattern design", the object catalogued with "orchid" will appear in the result.

2) ACACRM shows a high degree of scalability and flexibility of covering users' query language.

It is very helpful for users who are not familiar with relevant vocabularies of ancient Chinese artifacts.

3) It is possible to search collections by the subject matter (or referred to as content).

It proves that it is feasible to upgrade collection searching from the perspective of knowledge organization and knowledge representation. But as the research carrying on, some difficulties emerged and remained unsolved, one of which is the representation of archaeological controversies. Relevant studies and test case are still underway. Since the project is in the implementation stage, this paper summarizes and shares the current practical experience for providing reference for museums with similar needs. ACACRM is expected to help museums drive acquisition, retention, growth and value of their collection information (Fig.7), which as one of the most important digital resources deserves to be given consideration of methods to ensure the permanent accessibility and usability for long-term preservation.

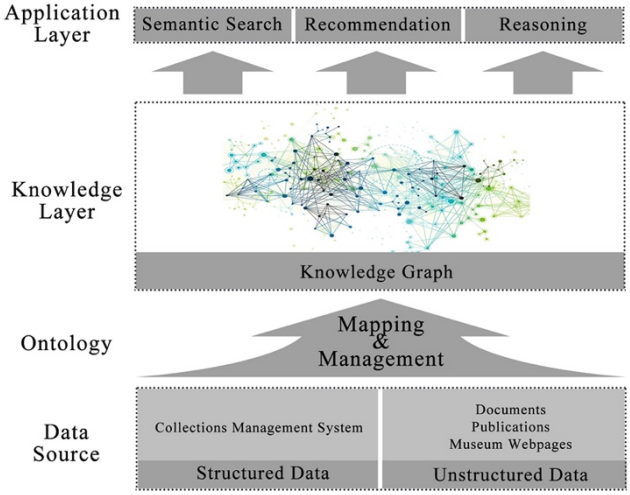

Figure 7 Context Diagram of ACACRM

The long-term preservation is a subject of ongoing development and knowledge graph construction requires extensive and unremitting efforts. The museum industry, in particular, needs manual revision to ensure the accuracy of knowledge. The Palace Museum always strives to have its activities informed to the public and shares its outcomes and experience with the world. Under the framework of ACACRM, a co-construction knowledge organization platform is a promising approach to encourage more volunteers to participate in knowledge graph construction. The Palace Museum welcomes any organizations or individuals to participate in knowledge co-construction and

knowledge sharing, creating a sustainable development and environmentally friendly ecosystem together.

V. CONCLUSION

The museum in the digital age is an information provider and a knowledge transmitter. Boasting a rich collection, museums are paying more attention to processing, storage and usage of their collection information resource. In the context of DIKW pyramid, the essence of knowledge organization is to deal with collection information. Knowledge organization acts as a bridge between the public need to information and museum information resources. It helps break the barrier of inconvenience that museums could not provide subject indexing or personalized semantic retrieval. It is a tool for processing information resources, enabling the public to efficiently, comprehensively and accurately search what they want, thereby enhancing the availability of collection information and maintaining the long-term preservation value of collection information as digital objects.

In addition, information infrastructure construction for integrated information services is also one of the development directions for museums under the trend of cultural resource integration and sharing. Knowledge organization can promote information exchange and integration between different information sources, so that local information sources can be processed into consistent, global, Internet resources. The integration of heterogeneous digital resources dispersed in physical space can be promoted. A one-stop access to digital resources of museums makes museums truly a public knowledge space for the whole society.

ACKNOWLEDGEMENT

The author would like to thank all members of Data Management & Services Team at the Palace Museum, as well as all the individuals for general supervision, administrative support, technical support for ACACRM project, and individuals proofreading the paper of Chinese version.

REFERENCES

[1] Melissa Gill, “Knowledge Organization within the Museum Domain: Introduction,” *Knowledge Organization*, Vol. 44(7), pp.469-471, 2017.

[2] Mingyu Huang, “Methodology for the Faceted Thesaurus of Chinese Cultural Heritage: An Initial Discussion,” *Journal of Library and Information Science*, Vol. 43(1), pp.171-190, April 2017.

[3] Fausto Giunchiglia, Biswanath Dutta, Vincenzo Maltese, “From Knowledge Organization to Knowledge Representation,” *Knowledge Organization*, Vol. 41(1), pp. 44-56, January 2014.

[4] Randall Davis, Howard Shrobe, Peter Szolovits, “What is a Knowledge Representation?” *AI Magazine*, Vol. 14(1), pp. 17-33, 1993.

[5] What is the Semantic Web? W3C. https://www.w3.org/2001/sw/

[6] Zeng Xinhong, “OntoThesaurus (Chinese-Thesaurus-Ontology) - An Integration of Thesaurus and Ontology,” *Information Science and Archives Science*, Vol.1, pp.34-43, 2009.