2005－CH－68（5）
2005／10／28

シソーラスによる複数のメタデータの統合

藤沢 仁子，Frederic Andres

総合研究大学院大学　　国立情報学研究所

(〒101-8430　東京都千代田区一ツ橋 2-1-2, satoko@grad.nii.ac.jp)

本研究では，多面的文化遺産メタデータセットに従ってデータを分類する新しい手法の研究の成果として，文化遺産の分野での意味論的相互運用の管理に関する多面的なデータ分類管理について明らかにした．本研究で提案するオントロジーの概念に基づいた多面的メタデータモデルを用いることにより、複数のメタデータセットを「AAT」の様な垂直方向のシソーラスの分類に関連付けることができる．本研究では，文化遺産コンテンツにおける Dublin Core、VRA、CDWA、MPEG-7 から成るメタデータセットによる実装を行った．本研究で用いた多面的なデータ分類は，米英語による米国の標準的な美術シソーラス「AAT」に基づいている．

Metadata Integration based on Multi-facet Thesaurus

Satoko Fujisawa, Frederic Andres

The Graduate University for Advanced Studies
National Institute of Informatics

Abstract

This paper points out a key issue related to the integration between cultural ontology based metadata management services and content management of digital resources. A multi-facet resource categorization approach is applied to metadata sets, allowing the metadata optimization of the description of cultural resources. It is part of the cultural ontology based metadata management system which supports interoperable semantic management.

1. INTRODUCTION

Following the evolution of cultural heritage archives accessible over Internet, new requirements for semantic understanding of multilingual and multi-disciplinary cultural contents, in the field of the historical silk roads studies, have been pointed out in two major symposiums [Ono 2001[1], Ono 2003[2]]. One key issue is related to the semantic interoperability management in cultural fields. It applies to the application of different vocabularies and terminology used in the description of cultural digital objects for both research and education purpose. A digital resource management system can operate as a repository of digital objects identified using separate metadata records. But more sophisticated digital archives have evolved to process objects in a dynamic way to extract more semantic. Multiple metadata points of view can be added according to the end-users. It is a way to customize the document access.

[1] http://www.nii.ac.jp/dsrtokyo/

[2] http://www.nii.ac.jp/dsrnara/

In the paper, we investigate a new way to categorize resources according to multi-facet cultural metadata sets. We demonstrate that this approach allows metadata optimization of the description of the cultural resources. Multi-facet resource category management is integrated in a cultural ontology based metadata management service part of the Advanced Scientific Portal for International COoperation (ASPICO) platform. This cooperative system aims at providing a web portal service in order to enable international and multi-disciplinary researchers and fellows to cooperate on research about cultural projects (e.g. the historical Silk Roads project, the visual cultural topic maps online project).

In Section 2, we introduce the problem issue and the states of the art in the field of cultural ontology–based metadata management. Then we present the ontology-based metadata platform architecture, its different layers from data collection to semantic management and delivery in Section 3. Section 4 explains theaurus as ontology. Then Section 5 gives an implementation case study related to Silk Roads Contents and discussion. Finally, Section 6 concludes and gives the direction of the future work.

2. STATE OF THE ARTS

An example of the widespread ambiguity in cultural terminology is the misperception that some terms can be represented in different forms and different meanings. For example, three objects (Fig.1), two paintings[b,c] and one statue[a], have the same name "Chakrasamvara Mandala" but different meanings. Moreover, it is difficult to integrate and to organize data provided by multi-cultural multi-disciplinary researchers.

Figure 1. Chakrasamvara Mandala

2.1 Metadata and Categories

Our research works on the categorization of multi-facet resources according to multiple cultural metadata sets is motivated by the recognition that many metadata standards have been created in the past years in the field of culture resources. Such metadata standards reflect any information related to the content, the context, and the structure of resources which supports their effective use, including information which can facilitate their management, their access and their analysis. Also those metadata standards have been developed in parallel by different communities. Content metadata of a cultural resource relates to what the resource contains or is about, and is intrinsic to a specific resource. This category has been developed from

several perspectives, based on who makes or provides the metadata. Dublin Core (DC)[3], VRA[4], CDWA[5], or ECAI[6] metadata enables researches on bibliographical information related to the resource (e.g. author, title, creation date, resource format etc.) from the content producer's perspective. The core set is the fifteen elements defined by DC. Though the metadata of DC are very generic, the descriptions are refined to more specific by other metadata.

For example, the VRA standard adds metadata regarding visual source and material/techniques, CDWA enables to describe works of art regarding the ownership, history involving, current location and creator. ECAI is an extension of DC metadata regarding timemap, spatial and temporal contents.

From the perspective of the provider that runs services, typical added-value metadata such as MPEG-7[7] describes information needed for retrieval, e.g., various formats under which a cultural resource is available, or similarity to improve search precision with, e.g., feature descriptions of (fractions of) the contents (e.g., a gold sculpture featuring in a video, parts of some mantra in an audio stream etc.).

Contextual metadata indicates the who, what, why, where, how aspects associated with a resource's creation and is extrinsic to a specific resource. CIDOC/CRM [8] and the ABC metadata[5] set are part of this category. Structural metadata relates to the formal set of associations within or among individual resources and can be intrinsic or extrinsic. It is represented by the entity-relation model integrating information at varying degrees of detail. It can be the Object-oriented "Domain ontology" which formalizes the semantics needed to describe objects and relationships in the cultural heritage context.

2.2 Metadata and Crosswalk

At the same time, many of these standards also have commonalities. Crosswalks[11] are used to "translate" between different metadata element sets. The elements (or fields) in one metadata set are correlated with the elements of another metadata set that have the same or similar meanings. We apply Metadata crosswalks for correlation among the metadata sets in order to facilitate semantic interoperability and to effectively convert data from one metadata standard to another.

The core set is the fifteen elements defined by Dublin Core, hence all the equivalent elements of other sets are linked with the corresponding elements of Dublin Core. Though the classification by Dublin Core is very general, it becomes more and more specific as the Crosswalk goes through.

2.3 Ambiguity issue

An example is that an object can be annotated by the various metadata sets. Figure 2 is an example annotated by the Dublin Core, VRA, CDWA. Classifications of the same object are varied as each standard categorizes the object by their own classification. For example, Location of VRA means the place of creation, but it is defined by Creation-Place and Current Location Repository number defines the current

---

[3] http://dublincore.org/

[4] http://www.vraweb.org/vracore3.htm

[5] http://www.getty.edu/research/conducting_research/standards/cdwa/

[6] http://www.ecai.org/

[7] www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

[8] http://cidoc.ics.forth.gr/scope.html

place of resource in CDWA. Another example, regarding "metal" and "brass", though CDWA defines them as Material & Thecnics, classificatin by VRA is more specialised such as *Material.Medium* is "metal" and *Material.Support* is "brass".

Figure 2. Annotation by multiple metadatas

It involves problem in terms of interoperability, as each institute applies different metadeta sets to classify their cocntents.

Though there are many cultural metadata sets, no single metadata model exists for sufficiently managing interdisciplinary and interorganization cultural contents. Hence, the multi-facet resource category management and the mapping are highly required for the efficient data retrieval.

As there is no simple one-to-one mapping corresponding to each other, the mapping is complex. We propose a new approach which maps multiple cultural metadata sets to cultural-dependant thesaurus avoiding the overlapping of attributes of different metadata sets. First, it enables the interoperability between multiple metadata sets according to one resource category tree. Second, it enables to provide an ontology-based metadata management which is language dependant and cultural dependant.

3. SYSTEM ARCHITECTURE

Our system is an ontology-based metadata management system in order to establish a multi-lingual systematic inventory of the digital cultural contents. It is part of the "myscoper" prototype which intends to provide a platform available over internet to researchers, investigators, and those who are interested in sharing digital archives of cultural heritage contents such as historical silk roads.

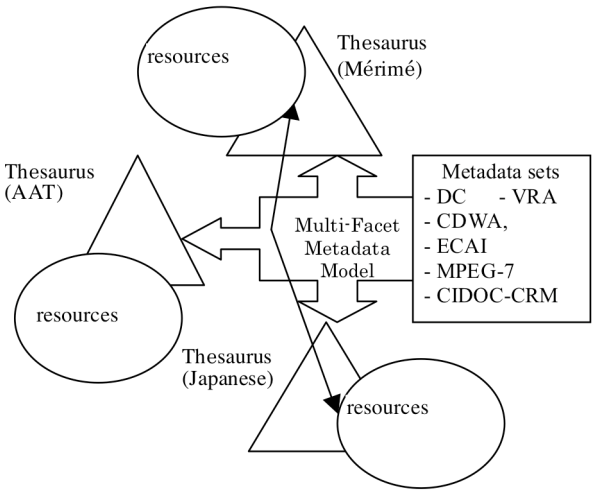

Figure 3. Ontology-based Metadata Management

As it is shown in Figure 3, a central structure called Multi-Facet Metadata Model connects the mapping between the metadata sets and mono-lingual thesaurus in such a way that no language is used as bridge language.

Thesauruses are both language and cultural dependant. Therefore each language is mapped with its related thesaurus following an interlingua representation approach. One key issue is the lack of thesaurus for some target languages such as Farshi or Japanese.

4. THESAURUS AS ONTOLOGY

Ontology is systematic classification of the objects and description of those relationships such as lexical or structural definition, or the definition of the relationship between lexicon and structure. As thesaurus is definition of the lexical relationship in hierarchical structure, thesaurus is one of ontology.

4.1 Art & Architecture Thesaurus

The Art and Architecture Thesaurus(AAT) , developed by The Getty Research Institute, is a cultural dependant structured vocabulary that can be used to improve access to information relating to fine art, architecture, decorative arts, archival materials, and material culture. It contains more than 133,000 terms, descriptions, bibliographic citations, and other information about concept. Concepts can have multiple higher concepts such as multiple-inheritance, and top-level concepts are architecture, materials, styles and periods, types of people, activities, physical attributes and links to associated concepts.

As the terms of AAT are not combined in advance, the users are free to combine the terms and to make professional expression. It enables to express an object from multiple point of view. For example, Figure 4 is a cartographic photo of Korea. It shows the two viewpoint of classification, Facet 1 proposes a path to a digital image and Facet 2 classifies the resource as photomap.

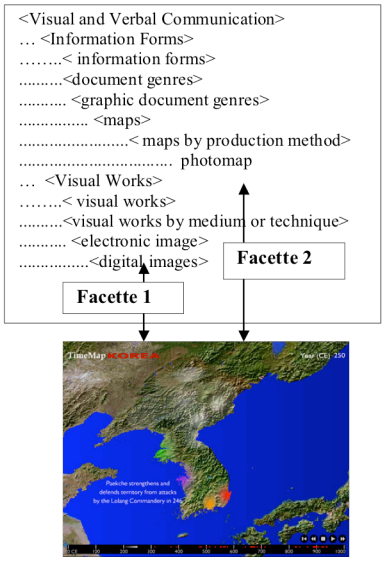

Figure 4. Digital photomap of Korea

Information systems is to provide some supports to enable any users to understand contents. A classification of cultural contents corresponding to a hierarchy of terms or controlled vocabularies are used for categorization and search digital resources in the databases of digital and visual archives. The categories help the users to navigate through digital data collections. the users can easily sort out the interesting subjects by grouping the data by understandable categories

5. CASE STUDY

5.1 Prototyping using Protégé 2000

We developed the ontology-based metadata management in OWL using Protégé 2000 . It is a knowledge-base-editing environment that allows the user to construct a domain ontology and to annotate the ontology. This editor supports the construction in a frame-like fashion with classes and slots.

The structure used for metadata management is composed by the two main classes; *Metadata type* class defining multi-metadata set and metadata class for the AAT thesauric hierarchy. The *Metadata type* class defines hierarchically the attributes of metadata sets and composed of the sub-clasees according to the following classification; Administrative, Contents,Descriptive, Physical, Conservation, Products, Technique and Usage.

We use *AAT_thesaurus* class for projecting the description of metadata sets in the sub-classes accoding to the type of resource. The projection will be detailed in the following section.

5.2 Case Study and Discussions

The case study have an objective to construct the first version of the multi-facet classification associated with the definition of ontologybased metadata. The discussion is based on three points: (1) metadata specialization by ontology, (2) the similarity of attributes and (3) the attributes categorization by subject.

5.2.1 Metadata Specialization by Ontology

Controlling equivalence between metadata attributes is an important point. An attribute can belong to several metadata sets and has the same significance; in other case, the attribute will be a specialization of another attribute of another metadata set.

R1: in case that an attribute belongs to two metadata descriptions and the domains of value are equivalent, the rule of attribute usage is defined as follow;

$$D_1.a \equiv D_2.a \quad \rightarrow \quad D_1.a$$

Let's see an example of the attributes of "Type", Table 1 shows definition and examples of "Type", one of the core elements, defined by Dublin Core, VRA and CDWA.

metadata	definition	example
Dublin Core: Type	The nature or genre of content of the resource.	collection/image/dataset /software/physicalobject /interactive image
VRA: Type	Identifies specific type of Work or Image described in the record.	print /sculpture painting/pottery/ furniture/photograph
CDWA: Object/ Work - Type	The kind of component that is part of the object, work, or group of objects described.	panels/folios/paintings/ drawings/videos/saucer bowl/cup/coin/mask/ installation/statue/

Table 1. "Type" of metadata sets

Comparing those three, *DC.type* is classification regarding the type of data for all the resource while *VRA.type* and *CDWA.type* cover more more specific domain of the image works and the domain of art object. Hence, *DC.title* is applied, instead of *VRA.title* or *CDWA.title_or_name*. *VRA.title* and *CDWA.title_or_name* can be said as the identic properties.

R2: in case, an attribute is specialization of another

$$D_1.b \xrightarrow{specialisation} D_2.b \quad \rightarrow D_1.D_2.b$$

$$D_1.c \xrightarrow{specialisation} D_3.c \text{ or } D_2.c \xrightarrow{specialisation} D_3.c \quad \rightarrow D_1.D_2.D_3.c$$

Example: *CDWA.context* is a specialization of *VRA.location*

R3: a case of the attribute in the only one metadata set

$$D_3.d \quad \rightarrow \quad D_1.D_2.D_3.d$$

$$D_n.e \quad \rightarrow \quad D_1.D_2.D_3....D_n.e$$

For example, « *style_period group_movement* » is unique and only exists in CDWA, the class hierarcy of this attribute is *CD.VRA.CDWA.style_period_group_movement* The naming rule is always mutual. The application of the rule enables to provide a structure of ontology-based metadata by creating relations between the metadata and the attributes of the metadata.

5.2.1 Similarity of Metadata Attributes

As showed in Section 2, the metadata standards for cultural resources have been developed in parallel by different communities from various perspectives. The bridge of Getty in form of «Crosswalk» provides horizontal correspondence between metadata attributes. The research enables the first optimization that the attributes are mapped in the AAT hierarchy avoiding overlapping of mapping the similar

attributes. The metadata attributes involves the level of details. Figure 8 shows an example of an specialization of categories by the theme according to the specific requirement of community as extension of AAT.(example: hierarchy about the mandalas). As we mentioned, the ontology-based metadata is realized by mapping the metadata sets into the thesaurus classification.

However, there are lots of metadata standards but a universal standard does not exist. It is impossible to create a unique universal metadata set covering all the cultural resources because various communities provide them.

The path in the AAT hierarchy enables to

classify the attributes of each metadata set in the function of applicability.

5.2.1 Similarity of Metadata Attributes

The third advantage of our approach is the Subject-depend Categorization applied under the AAT hierarchies.

According to Appendix, the attributes of mandala are added at the leaf node. 'Mandala' is not metadata standard but one of the major subjects in the cultural art domain. That is, the more specific subject-depend categorizations are added under the AAT hierarchies by metadata standards.

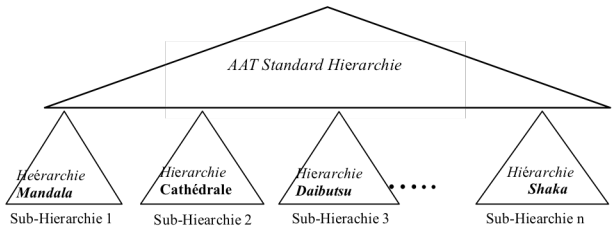

Figure 5. Subject-depend Sub-Hierarchies

The cultural resource involves common subject among many datas such as Mandara, Buddhistic temple, Bodhisattva ect. Those subjects have specific categories as like style or structure of period or groups. Then, classification by subjects plays important role for cultural resource.

The double hierarchies, of AAT classifying by format and contents covering all cultural resource and Sub-hierarchies by specific subjects, enable more precise search.

6. Conclusion and Future work

This report presents a new way to categorize resources according to multi-facet cultural metadata sets. It aims to improve the interoperable semantic management. We also introduced a categorization of digital resource according to an ontology-based metadata hierarchy.

By the Ontology-based Multi-Facet Metadata Model, the multiple metadata sets are mapped not only flatly by the Crosswalk but also vertically by the AAT hierarchies at the same time.

Our work on developing the Multi-facet Resource Category Support is useful in the Cultural Ontology based Metadata Management system, as it allows metadata optimization of the description of the cultural resources.

Furthermore, the multi-facet metadata model of resource provides a unified metadata set. It takes an advantage to link the metadata sets between institutes or communities. It is the first approach enabling to realize a single hierarchical metadata set covering the cultural resource.

We demonstrated an implementation in the case of the Silk Roads involving the metadata sets Dublin Core, VRA, CDWA and MPEG-7.

The categorization of multi-facet resources is defined by only the AAT, a monolingual thesaurus of American English. The system is extensible to multi-lingual management, as we have been developing multi-facet metadata model and a thesauric classification independently.

Then, the next step is to merge the contextual metadata set such as CIDOC-CRM into the system, and to model the multilingual categorization of multi-facet resources

ACKNOWLEDGMENTS

We like to thanks to Prof. Kinji Ono for his guidance and help for the research. Furthermore, we like to thanks NII for this research support.

REFERENCES

[1]Andres F. and Ono K., ASPICO-DSR Advanced Portal for International Cooperation on Digital Silk Roads, Proceedings of nara Symposium for Digital Silk Roads, pp. 335-346, 2003, ISBN 4-86049-024-X

[2] Godard J., Andres F. and Ono K., ASPICO: Advanced Scientific Portal For International Cooperation on Digital Cultural Content, International Journal "Information Theories & Applications" Vol.11, 2004

[3] J.Hunter, "Combining the CIDOC/CRM and MPEG-7 to Describe Multimedia in Museums", Museums on the Web 2002, Boston, April 2002

[4] Ono K., Overview of the Digital Silk Roads Project, Proceedings of Nara Symposium for Digital Silk Roads, pp. 63-70, 2003, ISBN 4-86049-024-X
[5] Lagoze C. and Hunter. J., The ABC Ontology and Model, In Journal of Digital Information, Volume 2 Issue 2 Article No. 77, 2001-11-06
[6] Shabajee P., Miller L., and Dingley A. Adding Value to Large Multimedia Collections Through Annotation Technologies and Tools:Serving Communities of Interest, in proceedings of Museums and the Web 2002, Boston, USA
[7] Weinstein P. C., Ontology-Based Metadata: Transforming the MARC Legacy, in the proceedings of Digital Libraries 98, Pittsburgh, USA, p254-263
[8] Hunter J., "Working Towards MetaUtopia - A Survey of Current Metadata Research", Library Trends, Organizing the Internet, Edited by Andrew Torok, 52(2), Fall 2003
[9] Colyer G., Ishii K., Hunter J., "Mapping between Dublin Core and JPX (JPEG 2000) Metadata", ISO/IEC JTC1/SC29/WG1 N2736, 15 October, 2002
[10] Hunter J., "A Proposal for the Integration of Dublin Core and MPEG-7", ISO/IEC JTC1/SC29/WG11 M6500, 54th MPEG Meeting, La Baule, October 2000
[11] Godby C. J., Young J. A., Childress E., " Repository of Metadata Crosswalks", in D-Lib Magazine, December 2004, Volume 10 Number 12.

APPENDIX: Inherited Metadata Hierarchy