Concept-based Multimedia Information Retrieval System using Ontology Search in Cultural Heritage

Ridwan Andi Kambau
Faculty of Computer Science
University of Indonesia
ridwan.andi@ui.ac.id

Zainal Arifin Hasibuan
Faculty of Computer Science
University of Indonesia
zhasibua@cs.ui.ac.id

(***Abstract***) **-- The richness of Cultural Heritage and Natural History is abundant. Many of the cultural heritage collection in Library, National Archive, and Museum in the form physical object or digital format in a different type of media (text, image, audio and video). One cultural heritage object can have the relationship with other objects in different media format and do not mention query term explicitly. Using the various media format causes problems in search. A monolithic search engine like Google, Bing, Google Image, Youtube, or Findsounds only retrieve one media format. Besides, the search result of the existing search engine is less relevant and incomplete in searching cultural heritage. Several multimedia information retrieval techniques used in building the relationship using ontology like ontology based search, content-based search with ontology and hybrid search with ontology. This paper proposes Concept-based Multimedia Information Retrieval System (MIRS) with ontology using Indonesia's cultural heritage dataset to increase relevance and completeness of the system. Concept-based MIRS using manually built thesauri or by extracting latent word relationship and concept from the Ontology that provides definition and formal structure for describing the implicit and explicit concepts and its relationship in cultural heritage documentation. Ontology-based Semantic similarity measure is defined which measure the semantic relationship between document based on the likeness of their meaning. The search results indicate that the document being retrieved becomes highly relevant, more complete, enrich the keyword and in varying media formats when is compared to existing search engine results such google, bing, google image, youtube and findsounds in specific domain.**

Keywords— Indonesia's Cultural Heritage, Ontology, Concept-based Search, CIDOC-CRM

I. INTRODUCTION

The richness and uniqueness of cultural heritage need to be known to the public more deeply. It is necessary to present complete and relevant cultural heritage information in the process of searching. Cultural heritage objects are available in various form (text, image, video, and audio) and are provided by different memory organization (libraries, archives, and museum) and individuals. The heterogeneous content provision and data format environment create an obstacle for the user to find and related cultural heritage information.

Relevance and completeness are other issues encountered in the search cultural heritage information. Some objects of cultural heritage have the relationship to another object with different media format. In another case, the search result is less relevant as only referring to the explicitly given query and can not relate to other objects that have the same meaning and have a strong relationship with the query.

To overcome the problems several techniques used in building the relationship like Ontology-based Search, Content-based Search with Ontology, and Hybrid search with Ontology. Multimedia Information Retrieval System using Thesaurus or Corpus and Searching with Ontology that describing the explicit and implicit concept and its relation in cultural heritage documentation. Ontology creates concept the relationship not only for cultural heritage objects but also create the relationship between multimedia format and make relevancy ranking based on the relationship of ontology.

Building cultural heritage ontology is the first step in developing the concept-based MIRS for Indonesia's cultural heritage. There are two kinds of ontology that are used, cultural heritage ontology using CIDOC-CRM[1] and media format ontology using the combination of CIDOC-CRM and MPEG-7 standard[2]. The classification and structure of cultural heritage refer to UNESCO classification[3].

Furthermore, designing the concept-based MIRS with ontology using Indonesia's cultural heritage dataset. This system is designed to retrieve any kind of media format and at the same time obtain complete and highly relevant information of Indonesia's cultural heritage from any kind media format of the query. Comparing the system with existing monolithic search engine like Google, Google Image, Youtube, and Findsounds is restricted in the specific domain (Indonesia's cultural heritage).

The rest of paper is organized as follows. In section 2, we show some work related to this paper. Section 3 focus on Indonesia's Cultural Heritage Design with Ontology development. Section 4 is about concept-based search design using Indonesia's cultural heritage domain. Section 5 is the conclusion.

II. Related Work

Designing concept-based multimedia information system with cultural heritage ontology require several stages, starting from describing cultural heritage objects with metadata, collecting and building databases for cultural heritage metadata, integrating various cultural heritage databases on heterogeneous cultural heritage databases and managing multimedia cultural heritage collections using the ontology classification and placing this ontology on the concept-based multimedia information retrieval system. In the next stage designing the concept-based multimedia information retrieval system query to find the multi-format object in the multimedia collection of cultural heritage.

A. Cultural Heritage & UNESCO World Heritage Classification

The cultural heritage is the things inherited from generation to generation and is placed on the heritage institution. Cultural heritage can be tangible like movable cultural heritage (paintings, sculptures, coins, manuscripts), immovable cultural heritage (monuments, archaeological sites, and so on), and underwater cultural heritage (shipwrecks, underwater ruins, and cities) or intangible like oral traditions, performing arts, and rituals [4].

Cultural heritage is composed of abundant information. Heritage information and its related data are not just explicit information; rather this information is a fundamental resource of heritage value and knowledge[5].

Containing various digital formats and is rich in semantic terms is a special characteristic of the cultural heritage. Collection items have their history and are related in many ways to our environment, to the society, and to other collection items. In MIRS term we using the concept to represent the cultural heritage object that has a relation between them. The important resource of cultural heritage is from UNESCO.

UNESCO ratified an international treaty called the Convention Concerning the Protection of World Cultural and Natural Heritage in 1972 for cultural heritage protection and preservation in the world. Based on UNESCO Classification of cultural heritage there are three classifications, cultural heritage, natural history and combining both of them. In cultural heritage is divided into two categories, tangible like clothing, books, monument, building and other artifacts and intangible like languages, social value, tradition, artistic expression and other aspect human activity.

For natural history refers to the elements of biodiversity (including flora and fauna) and geodiversity (including mineralogical, geomorphological, paleontological, etc.). And the combination cultural heritage and natural history have characteristic both of them. [6].

B. Ontology-based Information Retrieval

The used of ontologies to overcome the constraint of keyword-based search. Ontology is the motivation of Semantic Web since the late 90's. At the time traditional semantic web still using Boolean and Vector Space technique. Ontology performance overcomes the traditional semantic search.[7]

Gruber [1993] propose a formal definition of ontology, according to whom *"an ontology is an explicit and format specification of a shared conceptualization"*. Besides Gruber's definition, there is also important definition from Neches at al [1991] *"an ontology defines the basic term and relations comprising the vocabulary of a topic area, as well as the rules for combining terms and relations to define extensions to the vocabulary."* [8]

The ontology can be seen as a set of the term or a set of media and relation between them, showing the concepts that are used in a specific domain[9]. Besides ontology as a representation of knowledge and provides definition and formal structure for describing the implicit and explicit concepts and its relationship, ontology also provides ranking measurement with semantic similarity[10].

Some techniques for searching based ontology [11] like (i) ontology-based search when user takes benefit from ontological data structure and searching for specific information; (ii) content-based visual search with ontology using content features and exploits textual metadata from visual object; and (iii) hybrid search with ontology that combining ontology with visual features from multimedia object. This techniques influence ontology based search of this paper.

C. Concept-based MIRS and Semantic Similarity

Most retrieval or search systems based on text or keyword. Some techniques using words in the text of documents in the corpus; like Latent Semantic Indexing model performs the match based on the concepts. Further, in order to perform concept mapping, Singular Value Decomposition (SVD) is used. Furthermore, keyphrases are an important means of document summarization, clustering and topic search[12]. This is an initial concept based text retrieval, before Explicit Semantic Analysis.

The Concept-based retrieval with Explicit Semantic Analysis (ESA) approach is a new method that enhancing keyword-based text representation with features of concept-based, automatically extracted from massive human knowledge repositories such as Wikipedia or using ontology [13]. Almost same with LSI, ESA is concept-based text retrieval.

In this paper, MIRS is searching technique that is used with Ontology but before it is required to discuss concept-based image retrieval, concept-based video retrieval, and concept-based audio retrieval.

One of the approaches is the concept-based image retrieval is the concept of image training a classifier for detection using tags and variants of Support Vector Machine which allows the use of weight training per sample. Combined with weighting appropriate tag mechanism, sample more relevant that plays a more important role in calibrating the final model concept detector [14].

Another approach is the concept-based video retrieval using a method based on the integration of knowledge-based like ontology and corpus-based semantic word (Wikipedia) similarity measures in order to retrieve video shot for the concept whose annotation is not available for the system[15].

Concept-based audio retrieval technique based on the semantic concept, the audio tracks are mapped into a semantic feature space, where each dimension indicates the strength of

the semantic concept. Audio retrieval is then based on ranking the database tracks by their similarity to the query in the semantic space[16].

Concept-based text retrieval, concept-based image retrieval, concept-based video retrieval, and concept-based audio retrieval are monolithic concept-based multimedia retrieval. In this paper the design is unified concept-based MIRS, it means the system can retrieve any kind of media format document from any kind of query format and using monolithic concept-based MIRS as the main references to design the Concept-based MIRS.

III. Ontology Design

Designing ontology in this paper is divided in two main ontology. The first is media format ontology and the second Indonesia's cultural heritage ontology. Dataset in this paper using Indonesia's cultural heritage and for media format is using text, image, video, and audio.

Media format ontology helps the user to retrieve relation object in different media format. For example, the query using the text format and then the search result or retrieved document in text, image, video, and audio format.

In order to enable multimedia format using in the concept-based MIRS, the ontology follows the standard of Concept Reference Model where a variety of the cultural heritage contents forms in the multimedia format that must be defined relation format between the object. The multimedia relation between the different format of the objects based on CIDOC-CRM and MPEG-7 standard [2]. Visual Feature, Audio Feature and HTML(Doc) are classes of the Cultural Heritage object and has_visual_feature, has_audio_feature, is_documented_in are object properties or relations between objects. For image and audio only using Visual Feature class. The ontograph in Fig.1. describe the class entity and their relation between different media format.

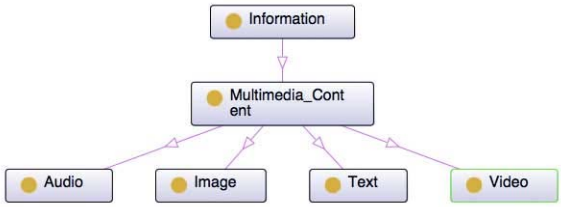

Fig.1. OntoGraf of Multimedia Content (Media Format Relation)

Designing Indonesia's cultural heritage ontology requires three stages, first collecting cultural heritage object and describing the object as metadata or digital representation. The second stage is integrating cultural heritage database from any resources (library, archive, museum, the ministry of education and culture etc) to create one heterogeneous cultural heritage database or we call here "national cultural heritage database". The third stage, using CIDOC CRM to design Ontology for describing the implicit and explicit concepts and relationships used in Indonesia's cultural heritage. (Fig.2)

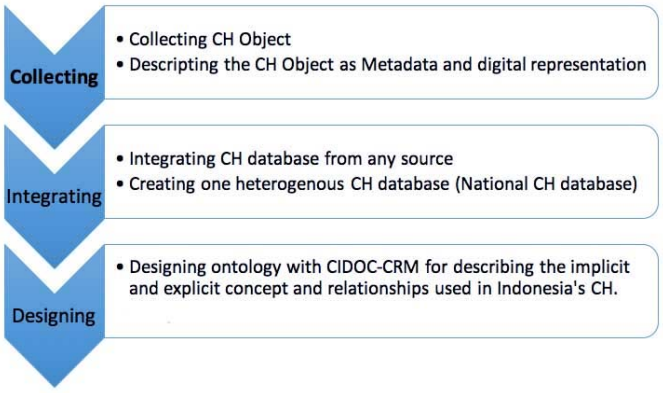

Fig. 2. Three Stages of Ontology-based Indonesia's Cultural Heritage Design

We define five superclasses referring to EDM [17]: Object, Event, Actor, Place and Time. These five class serve as contextual classes containing information on who, when, where and what. Besides that we using CIDOC-CRM as data model and WHC of UNESCO to classify the object.

Each superclass consists subclasses. For instance, the superclass 'Actor' has two subclasses called Person and Group; And superclass 'Object' that referring to WHC of UNESCO consist of subclasses Cultural, Natural and Mixed. Cultural means cultural heritage object, Natural means natural heritage object and Mixed means object that has characteristic of cultural and natural heritage.

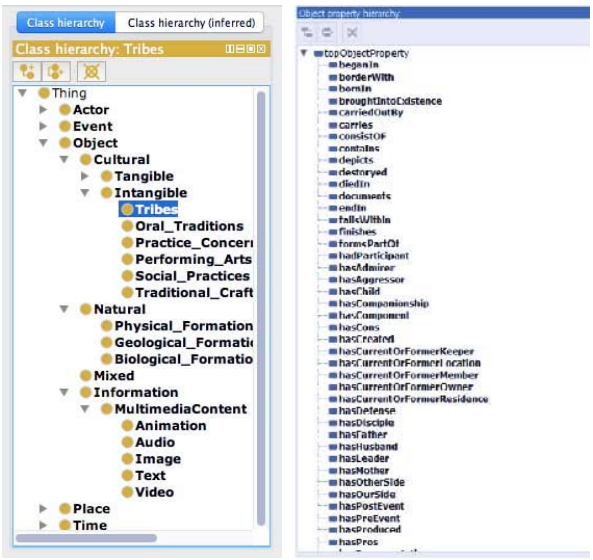

Fig. 3. Indonesia's CH Class and Property by Protege

Fig.3. shows the five superclasses with their relation or object property that is created using Protégé Application. While using Protégé to design Indonesia's cultural heritage, there are three steps to construct an ontology from Netches et.al; (1) identification of the basic term and their mutual relations; (2)

Agreeing on the rules that arrange them; (3) definition of terms and relations among concepts.

Assuming the user try to find information about 'Tana Toraja'. The concept of Tana Toraja has many relationships with another cultural heritage object like 'Rambu Solo' event is the funeral ceremony, 'Tongkonan House' object is the traditional house of Tana Toraja. Tana Toraja also related with another funeral ceremony in Bali is called Ngaben. Fig.4. show a part of Indonesia's cultural heritage with "Tana Toraja" concept and its relation.

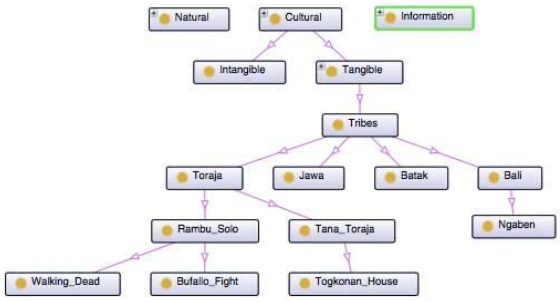

Fig.4. OntoGraf of Indonesia's Cultural Heritage (in part).

IV. Concept-based MIRS Design

The design of Concept-based MIRS using Ontology is the solution of inaccurate and incomplete cultural heritage searching and find cultural heritage objects with the multimedia format at one time searching.

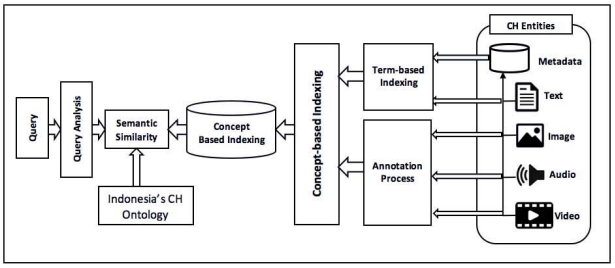

Fig. 5. Concept-based MIRS Design for Indonesia's Cultural Heritage Domain

In the context of using Ontology, concept-based MIRS define a Cultural Heritage objects as a concept and there is the relation between the concepts. In Fig.5. shows concept-based MIRS that can be used to find information about Indonesia's cultural heritage. Starting with cultural heritage entities in multimedia format in the collection. Metadata and text using term-based indexing because of the data in textual format. Multimedia data (image, audio, and video) use annotation process to give information in the object. Information from term-based indexing and annotation process is processed in concept-based indexing to change the information become the concept and put them into Concept-based Index database. In the query side, system process the query into the query analysis and then performing match process in the semantic similarity. The result of the concept-based search is cultural heritage object as a concept and their relation based on the object properties that have been designed.

V. Implementation of Concept-based MIRS using Ontology Search in Indonesia's Cultural Heritage

Implementation of concept-based MIRS using ontology search with Indonesia's cultural heritage dataset that provides 500 data text, 500 data image, 500 data video, and 500 data audio. This data is processed in concept-based MIRS with media format ontology and Indonesia's cultural heritage ontology.

A. Concept-based MIRS and Semantic Similarity

The implementation of searching cultural heritage objects, start with concept-based indexing task in Indonesia's cultural heritage storage that store cultural heritage entities in text, image, video, and audio format. Text format is indexed using term-based indexing and image, video, and audio using annotation process (automatically or manual) to represent the concepts. The concepts are stored in concept-based index database and waiting for Semantic Similarity process with ontology match with the query. In semantic similarity is performed relation matching and rank the concept based on how strong the relationship with the existing concepts.

For example, assuming the user search for 'Tana Toraja', a famous place in South Sulawesi, Indonesia using query text. Fig.6. is user interface design for the query input. The query field is not only for text, but also can use the image, video, or audio format.

Fig.6. User Interface Query

B. Media Format Ontology

The design of Concept-based MIRS allows any kind of media format is used as the query and the search result also generate information in the various form. Media format ontology arranges and manages the situation that enables relation between media format concept is related. Fig.7. shows media format relation in Indonesia's cultural heritage. When the searching of 'Tana Toraja' with query text, the retrieved document is related not only in text format but also image, video, and audio format.

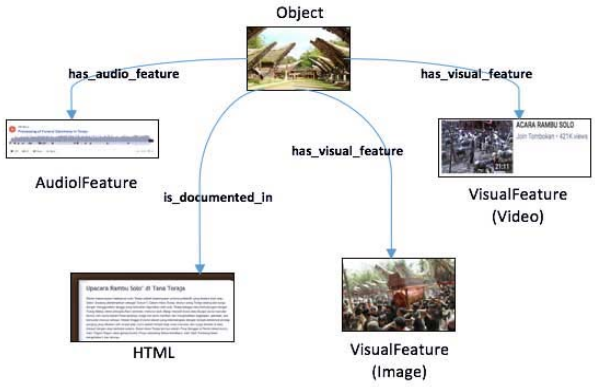

Fig. 7. Different Format Media Relation in Indonesia's Cultural Heritage Ontology

C. *Indonesia's Cultural Heritage Ontology*

The Fig.8. Shows how a single piece of data or the concept of Tana Toraja can be semantically enriched by the implementation of Indonesia's cultural heritage ontology. To begin with, Tana Toraja as a place is identified by traditional Tongkonan house, and the leader in the region will provide access to anyone who visiting Tana Toraja. Rambu Solo funeral ceremony as an event took place at Tana Toraja and funeral ceremony in Toraja was influenced by the funeral ceremony in Bali that called Ngaben.

The relationship could be extracted from the heterogeneous database based on Indonesia's cultural heritage Ontology design, where each relationship is specified by assigning properties.

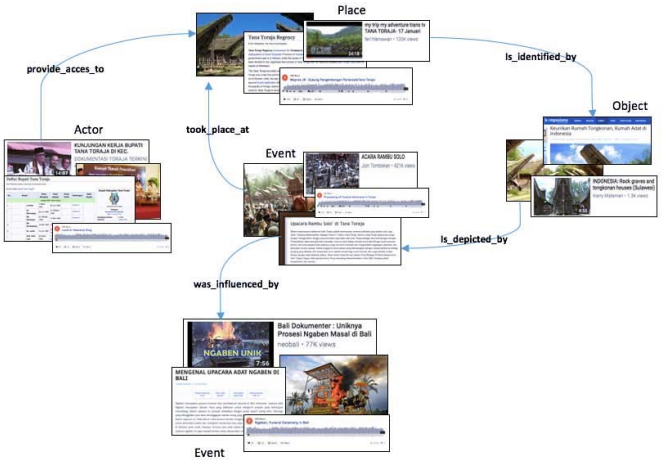

Figure 8. Indonesia's Cultural Heritage Ontology in the concept of "Tana Toraja"

D. *Semantic Similarity using Concept Weight*

Measuring the rank list in the retrieved documents with Semantic Similarity measure using Concept Weight. Concept weight is determined using term frequency and semantic distance. Cosine similarity using concept weight measure is applied to find similarity between different documents. According to the similarity score, the document is clustered. Concept weight based clustering increase the accuracy rate of the document[10].

Another rank technique that combined with concept weight is semantic weight using semantic relationship extraction and the result is the weight properties[18]. The combination of these two techniques will result in an accurate ranking list.

VI. Comparison With Existing Search Engine in Indonesia Cultural Heritage Domain

The ontology will enrich the result of the search based on the design of entity and property of the cultural heritage object (See. Fig.9 and Fig.10). If we make the comparison with the existing search engine, in this case, we using google with the same query ('tana toraja'), our concept-based search with ontology is complete and highly relevant with any kind of media (text, image, audio, and video). Google only shows the result in text and image format included the map. The text, image, and map on the display depend on the query of 'tana toraja' explicitly it means every document has 'tana toraja' word in their data. (See Fig.9.)

Other the search result comparison with existing multimedia search engine like Bing, Google Image, Youtube and Findsounds show the concept-based MIRS with Indonesia's cultural heritage ontology is outperform.

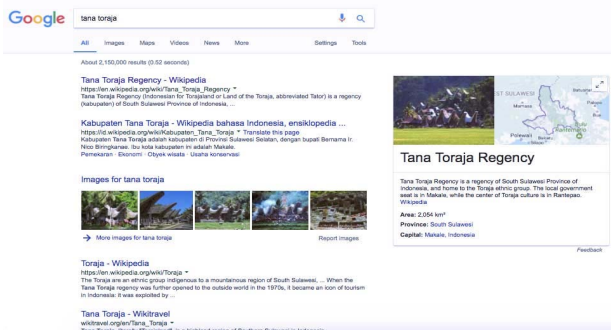

Fig.9. The search result of Google with query 'tana toraja'

Our user interfaces result of the concept-based search with Indonesia's cultural heritage ontology show information with any kind of data (text, image, audio and video) and enrich the query with relevance information that related as the concept. In figure 10. We see tana toraja as the region in Indonesia, tana toraja that has the funeral ceremony, tana toraja that has the traditional house is called Tongkonan and tana toraja related to another area that has a similar funeral ceremony like Ngaben in Bali. (See Fig.10)

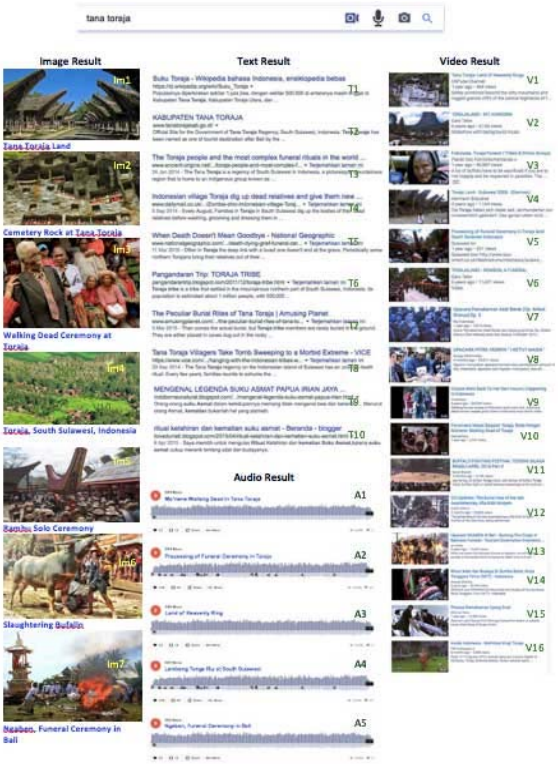

Fig.10. User Interface Design of Concept-based Search with Indonesia's cultural heritage Ontology using query 'tana toraja'

Here, the comparison between Concept-based MIRS with the existing multimedia search engine (Google, Bing, Google Image, Youtube, and Findsounds) with using "Tana Toraja" query in Indonesia's cultural heritage domain. Concept-based MIRS with 'tana toraja' query can retrieve more object with conceptual relation (*took_place_at, is_depict_by, provide_acces_to, is_identified_by, and was influenced by).* Google and another search engine can not retrieve the object with complete relation like the proposed system. (See. Table.3)

TABLE 1. Comparing the result Concept-based MIRS vs Existing Multimedia Search Engine with the query of tana toraja.

Concept Relation / IR System	MEDIA	Relations				
		took_place_at	Is_depict_by	provide_acces_to	Is_identified_by	was_influenced_by
Concept-based MIRS	Text, Image, Audio & Video	Yes	Yes	Yes	Yes	Yes
Google	Text, Image	Yes	Yes	No	No	No
Bing	Text, Image	Yes	Yes	No	No	No
Google Image	Image	Yes	Yes	No	Yes	No
Youtube	Video	Yes	Yes	No	Yes	No
Findsounds	Sound	No	No	No	No	No

VII. CONCLUSION

In order to retrieve complete, highly relevant and multi-format cultural heritage, it required concept-based MIRS with Ontology. Two kinds of Ontology that are used in this work, media format ontology and Indonesia's cultural heritage Ontology.

Media format ontology enhances the capability of Concept-based MIRS using multimedia query and retrieving the document in multimedia format. Indonesia's cultural heritage ontology increase retrieved document relevance and completeness and enrich the keyword based on the likeness of their meaning.

The expectation of this paper is the user will be able to recognize all aspect of Indonesia's cultural heritage as meaningful. With conceptual relation define by this ontology model, the user will be able to take away not only rich meaning from cultural heritage information but also a seamless information experience.

REFERENCES

[1] C. T. Aalberg, D. Balzer, C. Bekiari, L. Boudouri, N. Crofts, Ø. Eide, T. Gill, G. Goerz, M. Hagedorn, G. Hiebel, J. Inkari, D. Iorizzo, J. Kotipelto, S. Krause, K. H. Lampe, J. Lindenthal, M. Nyman, P. Riva, L. Rold, R. Smiraglia, R. Stein, M. Stiff, and M. Žumer, "Definition of the CIDOC Conceptual Reference Model Documentation Standards Group V6," 2015.

[2] J. Hunter, "Combining the CIDOC CRM and MPEG-7 to Describe Multimedia in Museums," *Museums Web 2002*, no. 1-88625-27–4, 2002.

[3] UNESCO, "World Heritage Centre," *UNESCO*, 2014. [Online]. Available: http://whc.unesco.org/en/about. [Accessed: 26-May-2017].

[4] M. Doerr, "Handbook on Ontologies," pp. 463–486, 2009.

[5] S. Kim, "Towards a Semantic Data Infrastructure for Heterogeneous Cultural Heritage Data," 2015.

[6] A. Rubhasy, "Management and Retrieval of Cultural Heritage Multimedia Collection Using Ontology," pp. 2–6.

[7] D. Vallet, M. Fernández, and P. Castells, "An Ontology-Based Information Retrieval Model," *Semant. Web Res. Appl.*, pp. 455–470, 2005.

[8] A. M. Rinaldi, "An ontology-driven approach for semantic information retrieval on the Web," *ACM Trans Internet Technol*, vol. 9, no. 3, pp. 1–24, 2009.

[9] G. Tan, T. Hao, and Z. Zhong, "A Knowledge Modeling Framework for Intangible Cultural Heritage Based on Ontology," 2009.

[10] S. A. Elavarasi and K. Menaga, "Ontology Based Semantic Similarity Measure Using Concept Weighting," pp. 15–20, 2014.

[11] C. Doulaverakis, Y. Kompatsiaris, and M. G. Strintzis, "Ontology-based access to multimedia cultural heritage collections-The REACH project," *Comput. as a Tool, 2005. EUROCON 2005. Int. Conf.*, vol. 1, pp. 151–154, 2005.

[12] R. Rodrigues and K. Asnani, "Concept Based Search Using LSI and Automatic Keyphrase Extraction," *2010 3rd International Conference on Emerging Trends in Engineering and Technology*. pp. 573–577, 2010.

[13] O. Egozi, S. Markovitch, and E. Gabrilovich, "Concept-based information retrieval using explicit semantic analysis," *ACM Trans. Inf. ...*, vol. 0, no. 0, pp. 1–38, 2011.

[14] V. Papapanagiotou, C. Diou, and A. Delopoulos, "Improving Concept-Based Image Retrieval with Training Weights Computed from Tags (accepted, awaiting publishing)," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 12, no. 2, 2015.

[15] S. Memar, L. S. Affendey, N. Mustapha, and M. Ektefa, "Concept-based video retrieval model based on the combination of semantic similarity measures," *2013 13th Int. Conf. Intellient Syst. Des. Appl.*, pp. 64–68, 2013.

[16] L. Barrington, A. Chan, D. Turnbull, and G. Lanckriet, "Audio Information Retrieval Using Semantic Similarity," *Tracks A J. Artist. Writings*, vol. 2, pp. 2–5, 2007.

[17] A. I. (Ed), "Europana Data Model Primer, Europana Technical Document," 2013.

[18] J. Lee, J. K. Min, A. Oh, and C. W. Chung, "Effective ranking and search techniques for Web resources considering semantic relationships," *Inf. Process. Manag.*, vol. 50, no. 1, pp. 132–155, 2014.