A Semantic Web Approach to Heterogeneous Metadata Integration

Shu-Hsien Liao[1], Hong-Chu Huang[2], and Ya-Ning Chen[1]

[1] Department of Management Sciences and Decision Making, Tamkang University, Taiwan
[2] Department of Information and Library Science, Tamkang University, Taiwan
{Michael,kuanin}@ mail.tku.edu.tw, arthur@sinica.edu.tw

Abstract. Heterogeneous metadata integration is an issue in digital libraries. Mapping is often used for an integrated metadata access, but the implicit knowledge and relations embedded in metadata are ignored. This paper aims to present a semantic web approach to heterogeneous metadata integration of biodiversity repositories. First, implicit knowledge and relations in metadata are extracted out and transformed into a shared ontology with expression of RDF and OWL languages. Next the shared ontology plays an inter-lingua role in harmonizing heterogeneous metadata to achieve an ontology mapping with a unified view. Then the shared ontology is expressed by SWRL for inference query to offer in-depth semantic discovery. Finally four question answering oriented queries are employed to examine the feasibility of the shared ontology for heterogeneous metadata integration.

Keywords: Information integration, heterogeneous metadata, semantic web, digital libraries.

1 Introduction

Many biodiversity heritage institutions have built up digital repositories to curate and manage their collections by digitizing biodiversity materials. Owing to various purposes and functions, digital repositories will adopt different metadata formats as data schemas or elements to describe their collections. Mapping is often-used method to achieve metadata integration. Usually, mapping just focuses on lexical equivalence of metadata elements from source format to target one, but contextual relationships between metadata elements are ignored or lost.

[5] points that data integration refers to combining data in such as way that a homogeneous and uniform view presented to users. [16] also regards that heterogeneous information integration is referred to as information synthesis of different information and data sources across disparate systems on the supply chain. One may draw a generalization that integration of data or information involves two key issues: heterogeneous and distributed. Therefore, the ideal integrated metadata access also needs to solve similar issues.

Semantic web has been proposed as the next generation of web and is composed by three components: ontologies, XML and RDF, and inference rules [3]. Actually ontologies play a core role for semantic web in offering a set of common agreed terminologies

J.-S. Pan, S.-M. Chen, and N.T. Nguyen (Eds.): ICCCI 2010, Part I, LNAI 6421, pp. 205–214, 2010.

and relations to harmonize semantic heterogeneity for distributed web-based databases and systems on Internet. Ontologies can be regarded as inter-lingua either to standardize terminologies or to provide the semantic foundations for translators [15]. Therefore, it becomes an issue how to implement the semantic web into distributed and heterogeneous biodiversity repositories to achieve semantic-oriented metadata integration.

2 Methodology

This article aims to use ontologies as a harmony approach for heterogeneous metadata integration at data schema or element level. Target subjects were selected from Catalogue of Life（hereafter CoF）and Specimens Database of Native Plants in Taiwan (hereafter SDNPT), as case study to illustrate the semantic web driven approach for integration of heterogeneous metadata. The system prototype integrated Protégé 3.4 with JavaBean program language to clarify the related ontology engineering tasks for metadata integration. The rest of this article is organized as follows. First, related work based on literature analysis motivates the contribution of our proposed approach. Next, detail of ontology construction and extended applications for metadata integration is illustrated, in terms of knowledge extraction, and ontology mapping. Moreover, a method is proposed to build a connection from ontology terminologies to mashup metadata for typical relational databases. In addition, four question answering (hereafter QA) oriented queries are deployed to examine the feasibility of proposed approach. Furthermore, the discussion will be presented. Finally, concluding remarks are drawn for future research.

3 Related Work

Up to date, several approaches have been proposed to offer an integrated access to heterogeneous metadata from distributed biodiversity repositories. In terms of data value, the ID-based [10] or name-based [13] linkage approach only uses data value as a pointer to retrieve related metadata for a specific species. This approach relies on an authoritative list of unified identifiers or data value to reconcile the issue of semantic heterogeneity for biodiversity species. To this day, an official authoritative list has not agreed in biodiversity heritage. On the other hand, mapping or crosswalk is another often-used approach to provide a harmony basis for metadata integration according to a specified metadata format, such as Dublin Core（hereafter DC） and Darwin Core. One strand approach employs OAI-PHM with DC as a data aggregation mechanism to harvest and integrate various metadata formats and their elements from distributed digital repositories [2]. The other strand approach combines a networked retrieval protocol（such as Z39.50 and DiGIR） with a unified metadata format（such as EML and Darwin Core） as a federated searching service to offer integrated access to various biodiversity repositories [4][11]. However, mapping is classified as a lexical mapping which is based on lexical form, appearance or meanings [6], and contextual information embedded in metadata formats is excludes out.

Few studies have focused on how to transform metadata into ontologies in cultural heritage. One study is to illustrate in mapping DC into CIDOC/CRM ontology for

metadata integration [7], the other is to mapping from DC and EAD to CIDOC/CRM [14]. These approaches are based on an existing ontology to transform metadata elements into equivalent semantic terminologies and then select and build up the required ontological concepts and relations. However, the first prerequisite of this transformation lies in that a common ontology has existed and agreed as a domain of discourse to share and exchange knowledge for a specific domain. According to the above discussion, it is worth to explore how a semantic web approach to offer a unified logical view for metadata integration in distributed biodiversity repositories.

4 A Semantic Web Approach to Integrating Metadata

CoF and SDNPT are digital repositories to manage biodiversity information relating to species and specimens respectively. CoF is a typical relational database to record the species information for 50,804 unique species in Taiwan based on Species 2000 metadata format. SDNPT is also a RDB to manage plant specimens for herbarium with 50,027 specimen's records by adoption of the HISPID 3 format. Actually implicit knowledge of CoF and SDNPT is embedded in different metadata formats. Therefore, the approach proposed by this study is to extract implicit knowledge from biodiversity repositories and then transform into machine readable and understandable with standard expression of XML-based RDF, OWL and SWRL languages. The proposed approach is illustrated in detail as follows: knowledge extraction for building a shared ontology, ontology mapping, inference query, and metadata mashup.

4.1 Knowledge Extraction from Metadata Elements for Building Ontologies

Generally metadata is defined as data about data. Furthermore, metadata can be regarded as "a materialization of domain-related knowledge that facilitates the management of data warehouses and helps in achieving good performance"[12]. Therefore one may generalize that metadata is also a kind of knowledge with shared meaning and interpretation for specific user communities. However, metadata is still a human-understandable information with implicit knowledge. Our study has to transform implicit knowledge embedded in metadata into explicit one to build up a shared ontology for harmonizing heterogeneous metadata formats and their elements.

At this stage, we first adopted approach provided by [9] to extract knowledge manually from metadata of CoF and SDNPT repositories as follows: determine the domain and scope of the ontology, consider reusing existing ontologies, enumerate important terms in the ontology, define the classes and the class hierarchy, define the properties of classes, define the facets of the slots, and create instances. Then we used RDF data model to define classes and hierarchy, properties of classes and facets of the slots, to illuminate and re-contextualize the original ontological structure and relations embedded in metadata elements. Lastly, we inputted all classes, properties and their instances into Protégé 3.4 ontology editor to build and validate a shared ontology with expression of XML-based RDF and OWL languages. During this stage, this study is successful in transforming implicit knowledge into explicit one. Moreover, we also build up a shared ontology with a unified logical view. In addition, this study also transforms human-understandable metadata into a machine readable format in RDF and OWL.

4.2 Ontological Mapping

Traditionally, mapping or crosswalks are an imperative task for metadata integration in digital libraries. In practice, crosswalk is a chart or table to represent the semantic mapping of fields of data elements in one element set to fields or data elements in another element set [1]. Once a crosswalk between two metadata formats has completed, an integrated access to heterogeneous metadata of various sources can attain. However, not all metadata formats and elements have been included into existing official crosswalks maintained by authority institutions.

A shared ontology of this study is a set of common terminologies and relations with a unified logical view. It can be used to harmonize metadata from heterogeneous data sources in biodiversity heritage. At this stage, we adopted ontology alignment to generate ontology mapping between the shared ontology and elements of CoF and SDNPT respectively. Protégé and iPromptTab are employed to perform the semi-automatic ontology mapping, because iPromptTab can perform ontology alignment for classes according to both lexical strings and their path-based class hierarchical relations [9]. However, manual revision is still required to complete the final mapping rules. In fact, it reveals that almost elements of CoF and SDNPT can be mapped to the shared ontology, owing to the shared ontology stems from CoF and SDNPT.

4.3 Inference Query

Basically our shared ontology is a RDF data model of triples（subject, predicate, and object） with unambiguous associative relations and assertions, and stored in an XML-based RDF/OWL format. It can be extended as a knowledge representation basis to allow computer to meaningfully process the knowledge and provide semantic conclusions from our shared ontology and retrieve corresponsive metadata for answering imposed queries. Thus it can be further utilized to develop a set of semantic units of description logic（hereafter DL） such as IF-THEN rules, to draw inference query from various digital biodiversity repositories. For instance, two RDF triple statements, such as “Species has－product－Specimen” and “Specimen－is_collected_by－Collector”, can formulate an IF-THEN rule like “IF a specific plant Species has Specimen and Specimen is collected by Collector, THEN it means that Collector had collected this Species.” The SWRL syntax can be expressed as follows: product?（?x, ?y） ^ is_collected_by（?y, ?z） →hasCollector（?x, ?z）. Therefore, one can identify and combine two semantic RDF triples and statements as a basic IF-THEN rule for inference query. At this stage, we employed Protégé and SWRLTab software to manifest the deployment of SWRL language for OWL-DL based IF-THEN rules（see Fig. 1).

4.4 Mashup Metadata from Digital Repositories

How to retrieve corresponsive biodiversity metadata from CoF and SDNPT is still a problem in this study. Basically, the CoF and SDNPT are two typical relational databases, neither RDF nor networked retrieval protocol based. The proposed approach is to develop a query agent as a connection from ontological query results to mashup corresponsive metadata of CoF or SDNPT. The component of query syntax of CoF

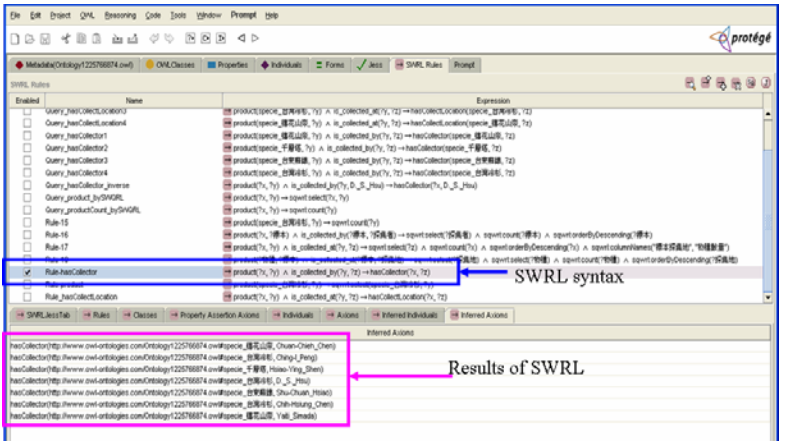

Fig. 1. An instance for SWRL query syntax and result

and SDNPT was analyzed into two parts: URL location as well as query field and string. The first part without underline is to connect the location of specified repository, and the second one with underline is to retrieve metadata records from repository based on either species' or specimen's name. Therefore several specific query syntaxes are generalized to mashup metadata as follows:

- SDNPT – http://db1n.sinica.edu.tw/textdb/hast/hast_label.php?_op=?species_m.speciesE:EngSpeciesName (query field is English species' name)
- CoF – http://taibif.org.tw/taibif_search/species_Detail.php?sc=Engspeciesname (query field is English species' name)

Second, we used JavaBean as program language to extend the function of Protégé for retrieving metadata from various digital repositories. Based on mashup connection, users can retrieve metadata from Protégé to CoF or SDNPT for reviewing detail of species or specimen records directly, no matter users select the specific term by browsing or querying ontologies. On the other hand, users can also query terms in a SPARQL syntax to access the corresponsive metadata from CoF or SDNPT respectively. Moreover, users would further use either SWRL or SQWRL language to perform OWL-DL based inference query to retrieve species' or specimen's metadata.

5 Question Answering

The design of ontological capable applications can facilitate the integrated metadata access and semantic query answering to a subject of interest from various biodiversity repositories. By means of QA-based query examples, we illustrate how biodiversity researchers not only query at various levels of ontological granularity, but also make semantically constrained queries.

Query 1: Which species has not specimen? There are two approaches may be the answer to this query. At the beginning, it can add zero into datatype attribute of specimen in our shared ontology. However, this approach is not a correct way for answering this query. Thus this study selects the second approach to find answer for this query. Essentially ontologies are assumed to be an open world. This query could convert to prove an assumption to be true or false. Therefore in this study we use SPARQL query to find the answer. The answer is Keteleeria davidiana（台灣油杉）. The syntax and query result of SPARQL are shown Fig. 2.

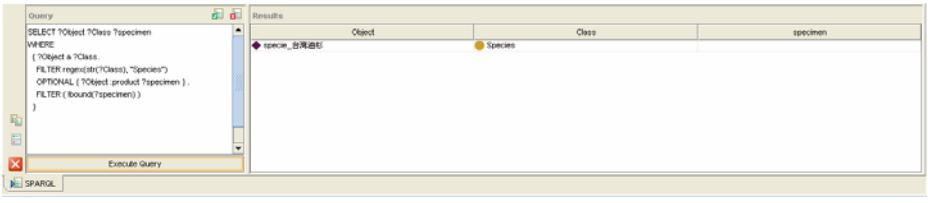

Fig. 2. SPARQL result for query 1 from Protégé

Query 2: Which species are two different species but with synonymous name? This graph pattern for this query is illustrated in Fig. 3. It clarifies the use of OWL differentFrom property as a query restriction to assert that two species（Aralia decaisneana Hance and Aralia bipinnata Blanco) are different but with the same Chinese common name（鵲不踏）.

Query 3: Which species is collected by Ching-I Peng at MIAOLI_HSIEN? This graph pattern for this query is illustrated in Fig. 4. It represents a more sophisticated query that spans over several RDF triples. This RDF graph query is a set of tuples in a sequential order, especially containing values for collector and collecting place of specimen respectively to satisfy three conditions: (i)a species has specimen, (ii) a specimen has been collected at specific place and （iii）a specimen has also been collected by specific person. The answer for this query is Abies Kawakamii（台灣冷杉） species.

Query 4: Which species has specimen, reference literature, scientific name, English common name and Chinese common name simultaneously? This graph pattern for this query is illustrated in Fig. 5. It represents another more sophisticated query that spans a greater portion of our shared ontology. The answer to this query is a set of tuples containing a species, specimen, reference and a complicated relation for name which includes scientific name, English common name and Chinese common name. The answer for this query is Hedychium coronarium Koenig（穗花山奈）with specimens（no. 101527 and 93998）, Flora of Taiwan（vol. 5, p. 717） reference literature, English common name（e.g. white ginger）, and Chinese common name（野薑花）. For this query this study uses SWRL rather than SPARQL, because SPARQL query can not de-duplicate the results.

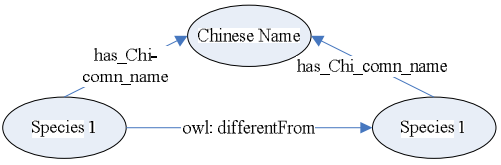

Fig. 3. RDF graph for query 2

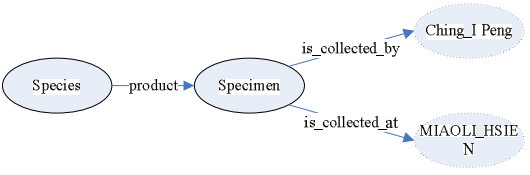

Fig. 4. RDF graph for query 3

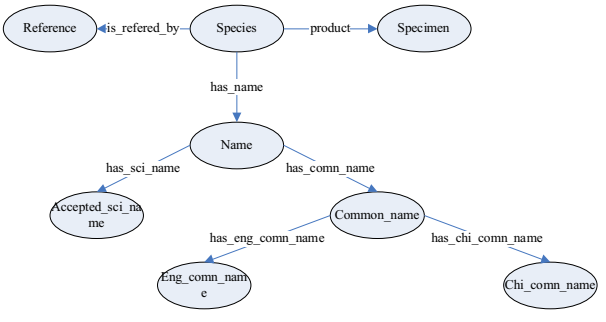

Fig. 5. RDF graph for query 4

6 Discussion

6.1 Transformation from Human-Understandable Metadata into Machine-Understandable Ontology

In fact a common ontology at element level is not available in biodiversity. In this study the proposed approach is to construct a shared ontology by transforming metadata into ontologies, rather than a mapping from metadata to ontologies. Thus this study is first to transform heterogeneous metadata into ontologies. It means that we have to extract and re-contextualize the original ontological concepts and relations embedded in metadata elements. This study has made implicit knowledge in metadata into explicit one. Therefore a human-understandable metadata expression has changed into a machine-readable RDF/OWL format. Thus the proposed approach in this study is a more practical solution than the above [7][14].

Second, it is not always feasible in heterogeneous situations for diverse user communities or domains to agree on using either the same authoritative identifiers [10] or

names [13], or a specific metadata format [2][4][11]. Therefore this study provides a new and more flexibly customized approach to build up the shared ontology from metadata elements, and enrich ontological relations between elements as a domain of discourse for any communities and domains. Furthermore, the proposed approach in this study also transforms machine-readable metadata into a machine-understandable ontology in a SWRL language that can be furthering processed and inferred by semantic web software.

6.2 Manifestation of Semantic Web Technologies on Heterogeneous Metadata Integration

In this study we build up a shared ontology as a harmony mechanism to integrate metadata from heterogeneous digital biodiversity repositories at conceptual level. With addition of ontologies to metadata integration, our contribution can be drawn as follows. First, it is distinctive from simply physical or virtual data aggregation based on the same metadata element set without semantic relations [2][4][11]. The shared ontology proposed by this study is a manifestation of knowledge representation to represent associative relations of class and property. Many relations of our shared ontology are not expressed straightforward in digital repositories. It can include relations into query indexing and inference query, in addition to metadata elements. Second, the shared ontology can further support semantic query formulation across various levels of granularity, to discover any relation between two or more objects for answering in-depth questions as same as our demonstrated queries. This provides a data mining approach to discover relations between two or more resources in biodiversity. Thus, the use of associative relations among objects is an advantage of the use of ontology mapping over typical metadata mapping. Therefore semantic web driven approach is a conceptual crosswalk for heterogeneous metadata integration, rather than an element mapping without semantic relations and logic axioms.

Finally, each RDF triple can be regarded as a unit of DL to formulate into a series of IF-THEN inference rules. As shown as our demonstrated queries, the proposed approach is successful in generating IF-THEN rules compliant with SPARQL or SWRL/SQWRL languages to achieve inference query. On the other hand, biodiversity domain also needs negation, e.g. species has not specimen as same as illustrated in our query 1. The ability of using negation as failure plays an important role in QA, especially for knowledge discovery. Biodiversity researchers would like to query with the assumption that all the knowledge is available at certain point to discover new research issues or provide insight into research trends. However, “closed world” inference such as QA for negation is usable in biodiversity. Therefore, apart from representation of ontological hierarchy and relations, it can formulate inference from the shared ontology based on semantic expression of knowledge representation.

7 Conclusions

This study is successful in implementing the technology of semantic web on metadata integration for distributed digital repositories in biodiversity. First, this approach proposed is distinctive from most current studies in building up a shared ontology from bottom up, instead of adopting existing ontologies or metadata formats and elements.

In this study we also manifest how to transform metadata from implicit knowledge into explicit one. It means that this study changes metadata from human-understandable biodiversity formats and elements into a machine-understandable specification of explicit knowledge in an ontological expression. Next, in this study we employ the shared ontology to develop a set of conceptual mapping rules with logical relations and axioms. Based on conceptual mapping rules, one can integrate heterogeneous metadata from digital repositories, instead of a pure schema or element mapping table. Third, this proposed semantic web driven approach also uses the shared ontology as a knowledge representation mechanism. Thus it is allowed to include logical relations into query indexing and perform OWL-DL inference query, in order to find any possible relations among objects for in-depth QA in biodiversity heritage.

Although the approach proposed for metadata integration is semantic web driven, our target subjects are still belonging to typical relational databases without any RDFization. Furthermore, the proposed agent in this study is a tentative solution to mashup metadata from relational databases. Therefore, the RDFized normalization is needed to transform these proprietary relational databases as qualified SPARQL Endpoints for providing RDF compliant query in a distributed online environment. Moreover, the biodiversity heritage still needs to develop a common agreed ontology at data schema or element level for knowledge sharing and discovery in the long term, because this study reflects partial requirements of institutions and their digital collections.

References

1. Baca, M.: Practical Issues in Applying Metadata Schemas and Controlled Vocabularies to Cultural Heritage Information. Cataloging & Classification 36(3/4), 47–55 (2003)
2. Barros, E.G., Laender, A.H.F., Gonçalves, M.A., Barbosa, F.A.R.: A Digital Library Environment for Integrating, Disseminating and Exploring Ecological Data. Ecological Informatics 3(4-5), 295–308 (2008)
3. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 284(5), 34–43 (2001)
4. Best, B.D., Halpin, P.N., Fujioka, E., Read, A.J., Qian, S.S., Hazen, L.J., Schick, R.R.: Geospatial Web Services within Scientific Workflow: Predicting Marine Mammal Habitats in a Dynamic Environment. Ecological Informatics 2(3), 210–223 (2007)
5. Hakimpour, F., Geppert, A.: Resolving Semantic Heterogeneity in Schema Integration: an Ontology Based Approach. In: FOIS 2001 proceedings of the international conference on formal ontology in information systems, pp. 297–308. ACM Press, New York (2001)
6. ISO/IEC JTC 1/SC32 WG2: Information Technology: Semantic Metadata Mapping Procedure: ISO/IEC WD 20943-5 (2008), http://metadata-standards.org/metadata-stds/Document-library/Documents-by-number/WG2-N1201-N1250/WG2-N1217-WD_SMMP_20081119.pdf
7. Kakali, C., Lourdi, I., Stasinopoulou, T., Bountouri, L., Papatheodorou, C., Doerr, M., Gergatsoulis, M.: Integrating Dublin Core Metadata for Cultural Heritage Collections Using Ontologies". In: Sutton, S., Chaudhry, A., Khoo, C. (eds.) Proceedings of the 2007 International Conference on Dublin Core and Metadata Applications, DCMI, Singapore, pp. 128–140 (2007), http://dcpapers.dublincore.org/ojs/pubs/article/view/877/873

8. Noy, F.N., McGuinness, D.L.: Ontology Development 101: a Guide to Creating Your First Ontology (2001), http://protege.stanford.edu/publications/ontology_development/ontology101.pdf
9. Noy, F.N., Musen, M.A.: Anchor-PROMPT: Using Non-local Context for Semantic Matching. In: Workshop on Ontologies and Iinformation Sharing at IJCAI (2001), http://www.dit.unitn.it/~accord/RelatedWork/Matching/noy.pdf
10. Page, R.: A Taxonomic Search Engine: Federating Taxonomic Databases Using Web Service. BMC Bioinformatics 6(48), 1–8 (2005)
11. Peterson, A.T., Vieglais, D.A., Sigüenza, A.G.N., Silva, M.: A Global Distributed Biodiversity Information Network: Building the World Museum. The Bulletin of The British Ornithologicsts' Club 123A, 186–196 (2003)
12. Ralaivao, J.-C., Darmont, J.: Knowledge and Metadata Integration for Warehousing Complex Data (2007), http://hal.archives-ouvertes.fr/docs/00/32/06/61/PDF/ista07-ralaivao-darmont.pdf
13. Sarkar, I.N.: Biodiversity Informatics: Organizing and Linking Information across the Spectrum of Life. Briefings in Bioinformatics 8(5), 347–357 (2007)
14. Stasinopoulou, T., Bountouri, L., Kakali, C., Lourdi, I., Papatheodorou, C., Doerr, M., Gergatsoulis, M.: Ontology-based Metadata Integration in the Cultural Heritage Domain. In: Goh, D.H.-L., Cao, T.H., Sølvberg, I.T., Rasmussen, E. (eds.) ICADL 2007. LNCS, vol. 4822, pp. 165–175. Springer, Heidelberg (2007)
15. Uschold, M., Gruninger, M.: Ontologies: Principles, Methods and Applications. Knowledge Engineering Review 11(2), 93–136 (1996)
16. Xu, Q., He, F., Qiu, R.G.: Heterogeneous Information Integration for Supply Chain. In: Proceedings of 2005 IEEE International Conference on Systems, Man and Cybernetics. pp. 97–105. IEEE Press, New Jersey (2005)