Facilitating Resource Utilization in Union Catalog Systems

Yung-Teng Tsai and I-Chia Chang

Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan

{alextsai,eiga}@iis.sinica.edu.tw

1 Introduction

Numerous papers have addressed building Union Catalog (UC) technologies, such as metadata schema, data exchange (e.g., OAI-PMH, Z39.50), and resource classification. In contrast to these approaches, in this article, we focus on UC technologies that leverage resource utilization across digital resources. We take the repositories of National Digital Archives Program (NDAP) in Taiwan as the project content to identify the major resource utilization issues. Our solution methodologies include resource unification, information query, information navigation, and unencoded character handling.

2 Background and Issues

The objective of the National Digital Archives Program [2], which is sponsored by the National Science Council (NSC) of Taiwan, is to promote and coordinate the digitization and preservation of content in leading museums, archives, universities, research institutes, and other content holders in Taiwan. Currently, NDAP's digital collections comprise over two million records covering eleven thematic groups. To enhance information sharing among these repositories, we have built a UC prototype to evaluate various building technologies, including OAI-PMH and Dublin Core (DC). However, this consolidation process has raised the following challenging issues:

Resource Unification: NDAP repositories are heterogeneous collections of diverse metadata principles, inconsistent coding systems, as well as assorted and distributed database storage systems. Unifying these heterogeneous repositories is therefore fundamental.

User Needs: UC users range from academic researchers to the general public. Hence, in addition to cross-database and cross-domain search features for researchers, browsing mechanisms are also needed to help general users navigate resources with ease.

Information Exploration: Utilization of NDAP's resources is enhanced by an easy-to-use interface. Access technologies, such as spatial and temporal information browsing, are being developed so that users can view information from different perspectives. Information visibility functions like content classification for resource navigation and data grouping for query results are also being investigated to enhance data clarity.

Unencoded Character Problem: This problem impacts on NDAP resource utilization in the areas of. character encoding, text input, font generation, and display. As the UC unifies all NDAP repositories, all unencoded characters are aggregated; thus the problem is magnified. A total solution for the unencoded Chinese character problem is therefore required.

3 Methodologies & Results

We propose the following methodologies to facilitate resource utilization in the UC.

Resource Unification: To transform and integrate NDAP's heterogeneous repositories into a uniform resource, we use an extended DC scheme as the unified metadata framework. Two approaches (OAI-PMH and XML file import) have been developed for this purpose.

Information Query and Navigation: By incorporating the Apache Lucene Search Engine, we have implemented full-text query and advanced DC query to consolidate NDAP resource utilization. In addition, four information classifications (participant organizations, content themes, spatial regions, and temporal domains) have been designed to provide comprehensive and intuitive information navigation of the unified NDAP resources. Spatial and temporal browsing are realized through the spatial and temporal data format conversion modules provided by the Academia Sinica Computing Center.

Unencoded Character Handling: The approach developed by the Chinese Document Processing laboratory [1] is used to manage NDAP's unencoded characters in the areas of data representation, storage, retrieval, display, and distribution. This resolves the problem of unencoded Chinese characters by applying a glyph expression model and glyph structure database to manage the characters. A set of tools is built into the data model to support character encoding, font generation, text display and input, and document dissemination functions.

A UC application [3] that incorporates the proposed methodologies has been released to the public. The UC adopts a web-based UI method for exploring information. Furthermore, it categorizes navigated content and query results to enhance data visibility.

Acknowledgements

This work was partially supported by NSC grants 94-2422-H-001-006, 94-2422-H-001-007, 94-2422-H-001-008, and 94-2422-H-001-004. The authors would like to thank Dr. L. F. Chien and various professional parties in the NDAP Research & Development of Technology Division for their contributions to this paper.

References

[1] Chinese Document Processing (CDP) Lab, Institute of Information Science, Academia Sinica. http://www.sinica.edu.tw/~cdp/

[2] National Digital Archives Program. http://www.ndap.org.tw/index_en.php

[3] Union Catalog of National Digital Archives Program. http://catalog.ndap.org.tw/dacs4/System/