京都大学防災研究所年報 第 49 号 C 平成 18 年 4 月
Annuals of Disas. Prev. Res. Inst., Kyoto Univ., No. 49 C, 2006

# Building an Integrated Database System of Information on Disaster Hazard, Risk, and Recovery Process – Cross-Media Database (3)

Hironori KAWAKATA*, Paul YOSHITOMI, Go URAKAWA**, Kelly CHAN, Hideki MATSUURA, Kenichi TATSUMI, Takeshi HARA***, Munenari AGUSA, Haruo HAYASHI, and Yoshiaki KAWATA

* Department of Science and Engineering, Ritsumeikan University, Japan (presently)
** Institute of Sustainability Science, Kyoto University, Japan (presently)
***ESRI-Japan, Japan (presently)

Synopsis

Research in adopting digital library initiatives into the XMDB project has begun during the year. This report specifically focuses on these activities: digital repository researches, social bookmarking, and dynamic classification. Some of the technologies that were examined include JHOVE, which provides automated format validation and characterization; DSpace, which is a digital repository system that captures, stores, indexes, preserves, and redistributes research data; Cannotea, which is a free online reference management service; and Dynamic Classification, which is an approach to organize search results by the types of information provided by the organizations themselves.

**Keywords:** database, digital repository, digital library initiative, disaster management

## 1. Introduction

During the 2005-2006 academic year, the XMDB project has captured and indexed a sizable volume of materials related to recent natural disasters: the Niigata flood, the Chuetsu earthquake, and the Indian Ocean tsunami. Data from hurricane Katrina in southern United States were also collected. We have used these test datasets to confirm the design and feasibility of the system. As we solidified the code base of XMDB, we have begun to adept other digital library initiatives into the XMDB project during the year. This report specifically focuses on these activities: digital repository researches, social bookmarking, and dynamic classification. Other related activities, such as researches on controlled vocabulary, automated terms extraction, and classifications, are reported elsewhere.

## 2. JHOVE

JHOVE is the JSTOR/Harvard Object Validation Environment. It is a collaboration between JSTOR and the Harvard University Library[1)] to develop an extensible framework for format validation.

Functions provided by JHOVE can be summarized as: format-specific identification, validation, and characterization of digital objects.

These are explained in more details on the project web site, http://hul.harvard.edu/jhove/.

· Format identification is the process of determining the format to which a digital object conforms; in other words, it answers the question: "I have a digital object; what format is it?"

· Format validation is the process of determining the level of compliance of a digital object to the specification

for its purported format, e.g.: "I have an object purportedly of format F; is it?"

· Format validation conformance is determined at two levels: well-formedness and validity.

* A digital object is well-formed if it meets the purely syntactic requirements for its format.
* An object is valid if it is well-formed and it meets additional semantic-level requirements.

For example, a TIFF object is well-formed if it starts with an 8 byte header followed by a sequence of Image File Directories (IFDs), each composed of a 2 byte entry count and a series of 8 byte tagged entries. The object is valid if it meets certain additional semantic-level rules, such as that an RGB file must have at least three sample values per pixel.

· Format characterization is the process of determining the format-specific significant properties of an object of a given format, e.g.: "I have an object of format F; what are its salient properties?"

The set of characteristics reported by JHOVE about a digital object is known as the object's representation information, a concept introduced by the Open Archival Information System (OAIS) reference model [ISO/IEC 14721]. The standard representation information reported by JHOVE includes: file pathname or URI, last modification date, byte size, format, format version, MIME type, format profiles, and optionally, CRC32, MD5, and SHA-1 checksums [CRC32, MD5, SHA-1]. Additional media type-specific representation information is consistent with the NISO Z39.87 Data Dictionary for digital still images and the draft AES metadata standard for digital audio.

Identification, validation, and characterization actions are frequently necessary during routine operation of digital repositories and for digital preservation activities. These actions are performed by modules. The output from JHOVE is controlled by output handlers. JHOVE uses an extensible plug-in architecture; it can be configured at the time of its invocation to include whatever specific format modules and output handlers that are desired. The initial release of JHOVE includes modules for arbitrary byte streams, ASCII and UTF-8 encoded text, GIF, JPEG2000, and JPEG, and TIFF images, AIFF and WAVE audio, PDF, HTML, and XML; and text and XML output handlers.

JHOVE currently supports the following digital formats; these do not cover all of the resource types available in XMDB.

JHOVE supported formats are:

· AIFF: Audio Interchange File Format
· ASCII: ASCII-encoded text
· BYTESTREAM: Arbitrary byte streams
· GIF: Graphics Exchange Format
· HTML: Hypertext Markup Language
· JPEG: Joint Photographic Experts Group raster images
· JPEG2000: JPEG2000
· PDF: Adobe Portable Document Format
· TIFF: Tagged Image File Format raster images
· UTF8: UTF-8 encoded text
· WAVE: Audio for Windows
· XML: Extensible Markup Language

XMDB was designed as a format-neutral database of research resources; however, there are no validations in the system. Additionally, format characterization is manually performed and inputted. As the mission of XMDB has been broadened from that of a catalogue or portal, automated format validation and characterization become increasingly important parts to support that mission. Facilities provided by JHOVE enable a significant increase in our data ingest capacity without a large dedicated support staff.

## 3. DSpace

DSpace is a digital repository system that captures, stores, indexes, preserves, and redistributes an organization's research data. The system was jointly developed by MIT Libraries and Hewlett-Packard Labs.[2] Not unlike XMDB, DSpace accepts all forms of digital materials including text, images, video, and audio files. Possible content includes the following:

· Articles and preprints
· Technical reports
· Working papers
· Conference papers
· E-theses
· Datasets: statistical, geospatial, matlab, etc.
· Images: visual, scientific, etc.
· Audio files
· Video files
· Learning objects
· Reformatted digital library collections.

DSpace currently uses a qualified version of the Dublin Core schema based on the Dublin Core Libraries Working Group Application Profile (LAP). DSpace uses

the LAP as a starting point for its application of Dublin Core, borrowing most of the qualifiers from it and adapting others to fit. Some qualifiers were also added to suit DSpace needs. The DSpace meta-data schema is available at http://dspace.org/technology/metadata.html. There is an active project within DSpace to integrate JHOVE into the repository.

While XMDB also implements the Dublin Core schema in its meta-data, the database design in XMDB is considerably more elaborated than that of DSpace. The team will evaluate the suitability of the DSpace schema in terms of harmonization with the existing XMDB design. One of the most reasons for this review is support for the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH). Making collections within XMDB available for harvesting significantly improves the accessibility and usability of the holdings in DRS to libraries and researchers in DPRI and even beyond the university.

DSpace runs on the UNIX operating system; presently it cannot be directly integrated into XMDB. However, DSpace is available in open-source, thus many of it functions, especially those of storage and indexing, can be integrated or selected ported to support XMDB at Kyoto.

## 4. Cannotea

Social bookmarking is saving bookmarks to a publicly accessible web site and assigning tags to entries. Users of these publicly accessible social bookmarking web sites can search for resources by keyword (tags), persons (taggers), and classification schemes that other users have created and saved. Instead of looking into general social bookmarking tools by del.icio.us c/o Yahoo! Inc.[3] – http://del.icio.us, we decided to investigate two more academically focused sites: Connotea by Nature Publishing Group's New Technology team [4] and CiteULike by R. Cameron [5] (http://www.citeulike.org).

The notion of social bookmarking and of tagging represents a shift away from formal taxonomies towards folksonomies and common tags for resources. Lomas (2005)[6] said "Tagging information resources with keywords has the potential to change how we store and find information. It may become less important to know and remember where information was found and more important to know how to retrieve it using a framework created by and shared with peers and colleagues."

Connotea is a free online reference management service. It allows you to save links to all your favourite articles, references, websites and other online resources with one click. Connotea is also a social bookmarking tool, so you can view other people's collections to discover new, interesting content. Connotea is different from general social bookmarking tool such as del.icio.us; it is specifically designed for scientists. Some of these features are summarized on http://www.connotea.org/faq

· Firstly, Connotea recognizes many scientific journals and websites, and when you save an item to your library, it can automatically pull off the bibliographic information, such as author and journal. Using a general bookmarking service, you would have to save all this information as separate tags, which would quickly get unwieldy. This automatic collection by Connotea makes your record of the item much richer than a simple URL. Additionally, development is ongoing, and you will soon be able to search your Connotea libraries by bibliographic information as well, enhancing Connotea as a reference management tool on top of its use as simple social bookmarking service.

· Secondly, Connotea also provides the ability to import and export references in RIS format, meaning you can easily use Connotea to work with your existing desktop reference management software.

· Thirdly, Connotea supports standards such as DOIs and OpenURL that are used by academic publishers.

· Fourthly, because Connotea is designed for the scientific community, you will benefit from a More Signal, Less Noise effect. The focus of the content shared through Connotea is considerably tighter than a general bookmarking service, which helps to provide better, more relevant recommendations via the social nature of the service.

## 5. Dynamic Classification

Dynamic classification is an approach to organize search results such that the organization itself provides information about the kinds of information that are represented by the documents in those results.

Dynamic categorization is based on three key premises:

1. An appropriate categorization depends both on the user's query and on the documents returned from the

query.

2. The type of query can provide valuable information about the expected types of categories and about the criteria for assigning documents to those categories.

3. Taxonomic knowledge about terms in the document can enable useful and accurate categorization.

Pratt compared different methods in organizing search results: term weights, controlled vocabulary, relevance ranking, and clustering. Her research indicated an advantage in dynamic classification over relevance ranking, a common search results presentation method. XMDB currently presents search results organized: (1) by resource types, (2) spatially, and (3) temporally. XMDB will continue to enhance its search functions and the presentation of search results.

## 6. Remarks

In conclusion, we will be focusing on data ingest aspects of XMDB during the 2006-2007 academic year; we will proceed with both open-source and commercial solutions. The commercial alternative, *ChronoStar* by Kubota Comps Corporation[7], is not reported here as that evaluation did not commence until February of 2006. A key goal of this coming year is to accumulate a significant volume of research materials in XMDB to critically test, review, and evaluate the usability of the digital library for the research community within the institute and the university. The team has begun to identify suitable "collections" for input into the database beginning the fall of the year. One candidate dataset is the existing *SAIGAI* database maintained at DRS[8]. The simultaneous available of the familiar dataset in two systems gives the team the chance of an objective usability evaluation.


## Acknowledgements

This research was supported by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) 21st Century COE Program for DPRI, Kyoto University (No.14219301, Program Leader: Prof. Yoshiaki Kawata).


## References


DRS, DPRI, Kyoto University: database SAIGAI, http://maple.dpri.kyoto-u.ac.jp/saigai/

Cyprien Lomas (2005): 7 Things You Should Know About Social Bookmarking, EDUCAUSE, http://www.educause.edu/LibraryDetailPage/666?ID=ELI7001.

JSTOR and the Harvard University Library: JHOVE (JSTOR/Harvard Object Validation Environment), http://hul.harvard.edu/jhove/.

Kubota Comps Corporation: ChronoStar, http://www.chronostar.jp/index_e.htm.

MIT Libraries and Hewlett-Packard Labs.: Dspace, http://dspace.org/technology/metadata.html.

del.icio.us c/o Yahoo! Inc.: del.icio.us, http://del.icio.us/.

Nature Publishing Group's New Technology team: Connotea, http://www.connotea.org/.

Richard Cameron: CiteULike, http://www.citeulike.org/.

# 災害ハザード・リスク・復興過程等に関する情報の統合型データベース・システム（クロスメディア・データベース）の構築（３）

川方裕則*・ポール吉富・浦川豪**・Kelly CHAN・松浦秀起・
辰己賢一・原武士***・阿草宗成・林春男・河田惠昭

*現，立命館大学理工学部
**現，京都大学生存基盤科学研究ユニット
***現，ESRIジャパン(株)

## 要旨

本年度はクロスメディアプロジェクトに新しい幾つかの電子図書館イニシアチブを統合するための研究を開始した。本稿はこれらに関する研究，すなわち，デジタルリポジトリー，ソーシャル・ブックマーキング，またダイナミック・クラシフィケイションに関する調査をまとめたものである。技術検討を行ったものには，自動フォーマット検証及び特徴付けの機能を有するJHOVE；研究用データを入手，格納，インデックス化，保存，そして再配信するためのデジタルリポジトリーであるDSpace；無料のオンライン参照管理システムであるCannotea; また，組織が提供する情報のタイプを基に検索結果を分類するダイナミック・クラシフィケイションが含まれる。

**キーワード**: データベース，デジタルリポジトリー，電子図書館イニシアチブ，防災