Knowledge-Base System for Lake Basin Management

Tatsuki Sekino¹*, Masahisa Nakamura², Thomas Ballatore³ and Victor Muhandiki⁴

¹Research Institute for Humanity and Nature, 457-4 Motoyama, Kamigamo, Kita-ku Kyoto 603-8047 Japan
²Research Center for Sustainability and Environment, Shiga University, 2-5-1 Hiratsu, Otsu, Shiga 520-0862 Japan
³Daiichi University, College of Pharmaceutical Sciences, 22-1 Tamagawa, Fukuoka, 815-8511, Japan
⁴Department of Environmental Systems, College of Science and Engineering, Ritsumeikan University, 1-1-1 Noji Higashi, Kusatsu Shiga 525-8577 Japan

*Correspondence author: Email: sekino@chikyu.ac.jp

ABSTRACT

LAKES (Learning Acceleration and Knowledge Enhancement System) is a knowledge base system to accumulate and search information required in lake basin management. In the present study, documents described using XML were applied to the system and tried to solve problems in the previous version of the system. The LAKES Schema which can be applied to various types of documents was developed to define document structure. Since the schema adopts structure of XHTML, authors can easily describe text, tables and figures using well known HTML syntax. Test documents according to the schema and experimental system for the test documents were constructed to examine the schema. The results demonstrated that the LAKES Schema and the system for the schema were appropriate and realizable.

Keywords: database, documentation, retrieval, interdisciplinary research, information science, XML, archive, LBMI

INTRODUCTION

In lake basin management, decision making is conducted using information collected from various academic fields (Ballatore et al. 2005). However, technical terms and specific concepts in each field prevent the decision makers from using information in unfamiliar fields. Mass of the information as well as the barriers between academic fields is a potential problem in lake basin management. It is not easy to find required information from the huge amount of accumulated information which is increasing with complexity.

LAKES (Learning Acceleration and Knowledge Enhancement System) has been developed to solve these problems in lake basin management (Sekino and Nakamura 2006). The function of sentence search to find sentences containing given search words from accumulated documents is a feature of LAKES. Users can read directly required information (i.e. description in the documents) without scrolling to find the description, and instantly grasp the information as a sentence. This function completely differs from conventional system to find a document containing given search words and from KWIC (keyword In Context) to display limited words containing the search words.

Another feature of the LAKES is a thesaurus function showing relationship between technical terms. This thesaurus is a list of pairs of words that co-occur in the same sentence in accumulated documents. This function realized that the users use technical terms in unfamiliar academic fields as search words, because they can know the meaning and usage of the terms from the relationship between technical terms (i.e., relationship between nutrients and water bloom) recorded in the thesaurus.

The LAKES was examined in the LBMI Project (Toward a Lake Basin Management Initiative: Sharing Experience and Early Lessons in GEF and Non-GEF Lake Management) which was conducted from March 2002 to May 2005 (Nakamura 2005), and 120 of documents relating with the project were introduced into the system. 58,021 sentences were extracted form the documents and 119,900 relationships between technical terms were recorded in the thesaurus. The project members used LAKES not only to search required information, but also to examine differences about specific issues among lakes.

Some problems in LAKES were found from the experimental usage in the project. The users can not easily know the described position of the found sentence in the document, and can not understand the context. There are some cases that they have to download the whole of the document to confirm the position. Database structure of the system also caused problems. Since sentence search was executed based on thesaurus table in the database, the users have to select search word among words recorded in the thesaurus. The problem that the users can not introduce their own documents to the system is partly because of the complicated structure of the database. There are technical problems in sentence extraction process as well. The system often extracted part of a table as a sentence by mistake.
In the present study, we tried to describe documents introduced into the LAKES using XML (Extensible Markup Language) to solve those problems. Schema design, test data construction and fundamental design of the system were completed at the present time. In this paper, we introduce the schema design (LAKES Schema) for the new version of LAKES and discuss application of the system to lake basin management.

Outline of XML and Related Technology

Applied to the LAKES XML is well known as a meta-language to define markup language for structured documents, and standardized by W3C (World Wide Web Consortium) which is the organization that administers technological specifications relating with the internet (W3C 1996). XHTML which is an updated version of HTML (HyperText Markup Language) is an application of XML (W3C 2000). XML is applied to describe data or documents in various fields as well as web pages. GML (Geography Markup Language) used to describe geographic information data (OGC 2005), EML (Ecological Metadata Language) used to describe ecological data (KNB 2001) and OMF (Observation Markup Format) to describe meteorological data (US Navy 2007) are fine examples of XML applications.

Specification of related technology are also standardized by W3C (W3C 1996). For example, XML Schema used to define structure of XML documents, XQuery used to search a part of XML documents and XSLT (XML Stylesheet Language Transformations) used to define format transformation of XML document are typical standardized technology relating with the XML. Additionally, XML documents are made in TEXT file type. Therefore, the XML documents are not influenced by software upgrades or discontinuations, while binary files depending on specific software have a risk that they can not read in the future. These features of XML are an advantage in applying the technology into LAKES.

When documents described using XML (XML documents) are introduced into LAKES, these documents have the same structured format according to a common schema. The validation mechanism of XML will ensure that the format of documents is strictly conformed to the schema, even if the documents will be made by some groups individually. The structured format of the documents has a potential to solve the problem in the previous system relating with described position of searched sentence, because the system can obtain information about the described position of the sentences according to the structure of the document. Furthermore, it is expected that these documents described using standardized technology are widely used in other software distributed generally.

LAKES Schema

The LAKES Schema was described according to W3C XML Schema (W3C 2001). In this schema, ‘documents’ element is the root element in a XML document, and contains one or more ‘document’ elements (Fig. 1). Since a document in LAKES was defined as an independent document with authors, a ‘document’ elements contains elements named ‘title’, ‘creators’, ‘contentstable’, ‘keywords’ and ‘references’ to describe information about the document, elements named ‘tablesect’, ‘figuresect’ and ‘appendixsect’ to describe tables and figures, and ‘section’ elements containing main text. A ‘document’ element can also contain child ‘document’ elements. Therefore, a book with chapters written by different authors can be described as a ‘document’ element of the book (i.e. ‘creators’ are book editors) containing child ‘document’ elements of the chapters (i.e. ‘creators’ are chapter authors).

Figure 1: Structure of ‘document’ element in the LAKES Schema. Icon (A) and (B) indicate that following elements sequentially and selectively appears, respectively (icons in Figs. 2 and 3 are the same). Elements belonging to ‘dc:elementsGroup’ are originated from DCML. This figure is automatically generated from the schema using a software XMLSpy (Altova) (Figs. 2 and 3 are the same).
A ‘section’ element contains main text of the document, and is used to describe a block in the document such as a small chapter (Fig. 2). Since a block described in a ‘section’ element is not an independent document, it can not contain ‘creators’, ‘contenstable’ and ‘references’ elements though it can contain ‘title’ elements. A ‘section’ element can contain child ‘section’ elements, and therefore, the complicated structure of chapters in a document can be described using hierarchical structure of ‘section’ elements. Tables and figures are described in elements named ‘tablesect’, ‘figuresect’ and ‘appendixsect’ as an independent block in a document (Fig. 3). These elements can contain related information of the tables and the figures described in elements named ‘title’ and ‘legend’.

The LAKES Schema adopts the structure of XHTML 1.1 to describe text, figures (images) and tables. Therefore, users can describe those using well known syntax of HTML. A ‘document’ element can contain metadata about the document according to DCMI (Dublin Core Metadata Initiative) which is a standard method to describe bibliographical information (DCMI 1995). Information about source and publication of a document can be described in ‘source’ and ‘publication’ elements of DCMI, respectively. When name or location (latitude and longitude) of lakes or basins is described in ‘coverage’ elements of DCMI, users can instantly find information about the lakes or basins.

The Supposed System

We considered the supposed system which accumulates XML documents according to the LAKES Schema. The usage procedure of the supposed system is shown in figure 4. The most notable feature of the system is the function that users can select the search target at the time of entering search words. For example, users can select displayed text blocks as a search result among chapters, paragraphs, and sentences. Additionally, users can find tables as well as text containing given search words on the system. Document structure according to a common schema contributes to realize this function. Another feature of the modified system is the function to display document structure as a tree. Using the structure tree, the users can know the described position of the extracted text blocks as a search result, and grasp its context. It is also possible to display larger text blocks in the document using links in the structure tree.

Figure 2: Structure of ‘section’ element in the LAKES Schema. Elements belonging to ‘html:xhtml.List.class’ and ‘html:xhtml.BlkStruct.class’ are originated from XHTML 1.1.

Figure 3: Structure of ‘tablesect’ element to describe tables in the LAKES Schema. Tables are described using ‘table’ elements of XTHML. In the same way, figures are described using ‘img’ elements of XHTML in ‘figuresect’ elements. Titles and legends relating with the tables are described in ‘title’ and ‘legend’ elements contained in the ‘tablesect’ element, respectively.
Since the LAKES Schema adopted structure of XHTML to describe text, text decoration, layout, figures (images) and tables are exactly represented on a web browser. Therefore, it is expected that the new system will be constructed as a web application in the same way as the previous system. Furthermore, the new system will be able to create PDF files using XSLT which is a technology to convert file format. This function will solve the technical problem in connection between the system and PDF viewers.

Examination of the LAKES Schema and the new System

Test documents according to the LAKES Schema were made to examine the schema and the supposed system. Example of a XML document according to the schema is shown in figure 5. In the present study, proceedings of the 11th World Lake Conference (Nairobi) were used for the test document. Files of the proceedings were saved as XML format by Adobe Acrobat 6.0 (Adobe Systems, Inc.), and then required tags for the schema were manually entered to saved file. After the entering of tags, it was confirmed that format of the documents was strictly conformed to the LAKES Schema using validate function in commercial software (Macromedia Dreamweaver 8 (Adobe Systems, Inc.) or XMLSpy (Altova)).

The test documents were sufficiently made according to the LAKES Schema from the proceedings. Hierarchical structures of ‘document’ elements and ‘section’ elements worked well and were applied to describe papers with different structure. Metadata according to DCMI were also useful. Although information about location of lakes or basins (‘coverage’ element) were not added to the documents in the present study, elements originated from DCMI were used to add information about document source (‘source’ element) and conference session (‘description’ element) in the documents. These results of the construction of test documents showed that the LAKES Schema was appropriate to describe XML document relating with lake basin management. Nevertheless, there were some problems in the construction process of test documents. It was revealed that about one hour was required to enter the tags into a 10 pages of document. This implies that method to easily enter these tags has to be prepared. An editor which automatically enters the tags to the area selected by users using GUI (Graphical User Interface) is a method to solve the problem.

Figure 4: The supposed procedure to use the new version of LAKES accumulating XML documents. Lists in rectangle indicate what is displayed for users in each step.
An experimental system with some tentative modules for the new version of LAKES was developed according to the supposed system in the present study. The previous version of LAKES had the database structure that the same database controlled both functions of text search and thesaurus, and this database structure caused some problems. Therefore, database for text search and thesaurus were separated in the experimental system. Although XML Database which is specialized to search XML document is ideal, it is too large for personal use. Considering needs that the LAKES was distributed widely and operated on personal computers, the experimental system was constructed using a relational database for personal use (Microsoft Access 2003 (Microsoft Corp.)). This database was used to administrate index table of elements in documents.

In indexing module test, a part of the test data (407 KB) containing 14 documents was introduced into the experimental system, and then, the index data was made on the database. The number of extracted ‘section’, ‘tablesect’ and ‘figuresect’ elements was 210, 15 and 22, respectively. Required time to make the index data was a few seconds using a personal computer with standard specifications (Intel Pentium M processor 1.10GHz, 760 MB RAM, Microsoft Windows XP SP2), and was practical. In search module test, it is confirmed that the system extracts ‘section’ elements containing given search word and output the search result as a HTML file. This process of search and file output was conducted within few seconds. These result about the experimental system indicated that the new version of LAKES was realizable.

Application and future subject of the LAKES

In the present study, the LAKES Schema for XML documents was developed and examined. The results about the test document and the experimental system demonstrated that the new version of LAKES to manage XML documents was appropriate and realizable to share information required in lake basin management, though there were some problems in the system.

When LAKES is used in various lakes in the world, documents for LAKES are individually made by many researchers and research groups. Therefore, a schema file of the LAKES Schema has to be exhibited on the web for validation process of XML documents. We have to consider administration of the schema as well as distribution of the system. Additionally, a multilingual system will be required for widespread distribution of the system, because LAKES can be applied only for documents in English. Function of sentences extraction has to be developed for more languages.

The documents described according to standardized specification (i.e. XML) bring another potential for LAKES. The system can work as an interface module to search information, and can be

Figure 5: An example for documents according to the LAKES Schema. Prefix es are attached to element names for elements of the LAKES Schema and DCMI
used in other databases such as the World Lakes Database (ILEC 1995). Furthermore, since the LAKES Schema adopts standardized DCMI and XHTML, the document according to the schema will be used in other software that adopts the same standard (ex. Semantic Web (Daconta et al. 2003)).

ACKNOWLEDGEMENT

We are grateful for Ms. M. Tanigawa for her assistance in construction of the test document. This work was supported by Kakenhi (Grant-in-Aid Scientific Research (C)) from Japan Society for the Promotion of Sciences (JSPS) (19510053).

REFERENCES


