XML and global name access control

Ki-Tat Lam

The author

Ki-Tat Lam is Head of Systems, Hong Kong University of Science & Technology Library, Kowloon, Hong Kong.

Keywords

Access control, Computer languages, Foreign languages

Abstract

This paper discusses why the MARC21-based authority format has failed in a global setting and details the use of XML and its related technologies to achieve global name access control.

Electronic access

The research register for this journal is available at
http://www.emeraldinsight.com/researchregisters

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/1065-075X.htm

Introduction

While there are well-established standards and implementations for conducting authority control at the regional level, efforts to extend such control globally are still far from mature. For instance, the Library of Congress (LC) has established one of the world's largest name authority files using the MARC21 standard. However, this implementation was originally designed for Anglo-American libraries. Problems begin to appear when users worldwide adopted LC's records and standards for non-Latin names. For instance, there have been complaints about the inability to store CJK (Chinese, Japanese, Korean) scripts in LC's authority records because it is very difficult to identify who is who if the record contains only the Latin transliterations.

Currently, the MARC21 format is a widely adopted markup standard for authority metadata. This standard has been very efficient for Latin-based names. However, if it is used for non-Latin names, the following issues may cause difficulties:

- providing more than one established heading in the same record;
- determining the field to be used to store multiple scripts with the same name;
- linking multiple fields of the same name in a record;
- identifying the script that is stored in a field;
- identifying the romanization scheme used in a field; and
- determining which character encoding scheme should be used in the record.

Enormous effort will be required to add original scripts to LC's authority files. However, the most fundamental issue for supporting global authority control lies in a search for a markup standard that fits well with the nature of multi-lingual name data. This article presents a solution using XML and its related technologies to achieve global name access control.

Using multi-scripts in MARC21 name authority record

"金庸", the pen name of the renowned Chinese novelist and journalist "查良鏞", will be used to

OCLC Systems & Services
Volume 18 · Number 2 · 2002 · pp. 88–96
 · ISSN 1065-075X
DOI 10.1108/10650750210430150

OCLC Systems & Services
Volume 18 · Number 2 · 2002 · 88–96

illustrate the multi-script issues in name authority records. Although this example involves only Chinese and Latin scripts, the problems discussed are common to other languages and scripts. The following are the various forms and transliterations of "金庸":

(1) "金庸" – pen name used for his martial arts fiction;
(2) "Jin, Yong" – Pinyin romanization of this pen name;
(3) "Chin, Yung" – Wade-Giles romanization of this pen name;
(4) "Kim-Dung" – Vietnamese romanization of this pen name;
(5) "Kin, Yo" – Japanese romanization of this pen name;
(6) "查良鏞" – name used for his non-fiction publications;
(7) "Zha, Liangyong" – Pinyin romanization of this name;
(8) "Cha, Liang-yung" – Wade-Giles romanization of this name; and
(9) "Cha, Louis" – his English name.

Numbers 1-5 are various transliterations of the pen name "金庸" and numbers 6-8 are for his real name "查良鏞". The following shows the major fields of the authority record found in the OCLC database:

```
001 oca00560270
. . . .
100 1 |aJin, Yong,|d1924-
400 1 |aChin, Yung,|d1924-
400 1 |aKim, Dung,|d1924-
400 1 |aKin, Yō,|d1924-
400 1 |aZha, Liangyong,|d1924-
400 1 |aCha, Liang-yung,|d1924-
400 1 |aCha, Louis,|d1924-
. . . .
```

This record contains many weaknesses:

- It contains Latin scripts only; the Chinese scripts, i.e. 金庸 and 查良鏞, are not recorded.
- "Jin, Yong" is selected as the established heading by the Library of Congress, while other authority control agencies may prefer to use his other names. For example, a Hong Kong library would use the Chinese script 金庸 in its catalog.

Presently, MARC21 can only tackle part of these problems. It is possible to encode multi-script by adopting the UCS/Unicode specification (Library of Congress, 2000, January). Position 9 in the *Leader* would contain the value "a" to indicate that the metadata is in UTF-8. MARC21 also offers two markup models for multi-script records, known as model A and model B. The following shows how the Chinese script is handled in these two models:

Model A

```
001 oca00560270
. . . .
100 1 |6880-01|aJin, Yong,|d1924-
880 1 |6100-01|a金庸,|d1924-
400 1 |aChin, Yung,|d1924-
400 1 |aKim, Dung,|d1924-
400 1 |aKin, Yō,|d1924-
400 1 |6880-02|aZha, Liangyong,|d1924-
880 1 |6400-02|a查良鏞,|d1924-
400 1 |aCha, Liang-yung,|d1924-
400 1 |aCha, Louis,|d1924-
. . . .
```

Model B

```
001 oca00560270
. . . .
100 1 |aJin, Yong,|d1924-
400 1 |aChin, Yung,|d1924-
400 1 |aKim, Dung,|d1924-
400 1 |aKin, Yō,|d1924-
400 1 |aZha, Liangyong,|d1924-
400 1 |a查良鏞,|d1924-
400 1 |aCha, Liang-yung,|d1924-
400 1 |aCha, Louis,|d1924-
700 1 |a金庸,|d1924-
. . . .
```

By adopting tag 880 and subfield 6 in model A, it is possible to:

- make 金庸 parallel to the established form "Jin, Yong"; and
- maintain a link between 查良鏞 and its Pinyin form "Zha, Liangyong".

MARC21 specifies that the transliteration is stored in the regular tag with its vernacular forms stored in multiple 880 tags. With a slight modification to MARC21 by exchanging the

OCLC Systems & Services
Volume 18 · Number 2 · 2002 · 88–96

regular tag and tag 880, model A will be able to link all transliterations of the same name form:

```
100 1 |6880-01|a金庸,|d1924-
880 1 |6100-01|aJin, Yong,|d1924-
880 1 |6100-01|aChin, Yung,|d1924-
880 1 |6100-01|aKim, Dung,|d1924-
880 1 |6100-01|aKin, Yō,|d1924-
```

In model B, tag 700 (established linking entry) is used to store the Chinese script 金庸, making it an alternate established form. However, model B fails to maintain the parallel relationship between multiple transliterations of the same name. The two fields for 查良鏞 and "Zha, Liangyong" are not linked. Therefore, one can fall back from model A to model B, but not from model B to model A. Although model A allows better linking of the vernacular script with its transliterations, neither library automation vendors nor bibliographic utilities such as OCLC are eager to support it for authority control. None of these models are sufficient for use in the global setting.

Three areas in the MARC21 authority format could benefit from enhancement. Firstly, 1XX should be repeatable, so that instead of depending on 7XX, more than one name can be coded in 1XX as established forms. Secondly, better linking capability between related forms of the name should be developed to avoid using tag 880 with subfield 6. This is because library automation vendors are reluctant to implement tag 880 for authority control, a task that requires substantial effort. Thirdly, there should be a place to hold the multi-lingual attributes of each name, including information about the script, language and romanization scheme. Although effort is needed to identify these attributes, and it is in some cases not an easy task, such information is particularly helpful in differentiating names in a multi-lingual environment.

From MARC to XML

XML (eXtensible Markup Language) is a simplified standard of SGML. It allows plain text data to carry not just layout information, but also semantic structure. Since its official release in early 1998 by the World Wide Web Consortium (W3C), XML has rapidly been adopted and implemented by many industries. For the past few years, many initiatives and projects that make use of XML for library metadata have emerged (e.g. bibliographic, authority, holdings, patron records).

MARC has been widely used as a markup language for library metadata since the1960s. MARC21, UNIMARC and their variations can be considered as MARC's DTD (Document Type Definition) or markup schema. Due to the similarity between markup concepts in XML and MARC, it is feasible to use XML to describe a MARC record. By converting MARC to XML, library applications would be able to exchange library metadata in the emerging XML format and avoid working with the proprietary MARC interchange format, known as ISO2709.

In addition, using XML, XSLT (Extensible Stylesheet Language Transformation) and other Web technologies, it is possible to do the following:

- create library metadata once and publish it in different formats;
- view library metadata directly from Web browsers, search engines, and potentially, library automation systems, without conversion;
- convert robustly between XML and MARC formats without data loss; and
- resolve many problems that were inherent with the MARC format.

In 1995, the Library of Congress began to look into the feasibility of using SGML to encode USMARC (a previous version of MARC21) data (Library of Congress, 2001) in the subsequent release of MARC DTDs and conversion software. In 1999, Lane Medical Library at Stanford University released its XMLMARC software, a byproduct of its Medlane Project, for public use (Lane Medical Library, 2002). Based on its in-house developed DTDs for bibliographic and authority formats, this java-based XMLMARC program offers conversion of MARC21 to XML. Many library automation vendors also developed their own DTDs for their MARC data. For example, Innovative Interfaces Inc.

OCLC Systems & Services
Volume 18 · Number 2 · 2002 · 88–96

has its own DTDs to handle various record types.

All these projects are defining their own XML schema (DTDs). However, to facilitate wider acceptance, it is essential that a globally agreed upon XML schema for MARC metadata be developed. This global MARC-XML standard should closely follow the MARC structure. In addition, it should allow for variations, changes and enhancements in the content designation.

Unlike the DTDs developed by LC and the Lane project, which are tightly based on MARC21, the global MARC-XML elements and attributes should be independent of regional content designation standards. To achieve this, Figure 1 shows the proposed namespace.

In the example, <marc> and </marc> mark the beginning and ending of a MARC record. <fd> and </fd> mark the beginning and ending of a MARC tag. <sf> and </sf> mark the beginning and ending of a subfield. Elements fd and sf contain a number of attributes. The name attribute defines the MARC tag name or the subfield code. ind1 and ind2 are the two indicators of the MARC tag. There should be additional attributes for fd and sf, to encode information not found in MARC. For instance, we can have:

- id – to give a unique id to each MARC tag and to allow for linking among MARC tags;
- script – to encode the language, the script and the romanization scheme used in the MARC tag;
- label – to describe the MARC tag, etc.

Attributes for the marc element can also be defined to encode record level information found in the leader and the controlled fields (e.g. the 00X tags in MARC21).

Global name access control in XML – an experiment

With the adoption of XML, it is possible to expand authority control to access control. The idea of access control, as opposed to authority control, is to make obsolete the notion of "authority" so that variant forms of a name are equal in status and users are able to select any of them for searching, retrieval and display.

Based on the proposed MARC-XML schema and the MARC21 specification, the name access control metadata for "金庸" in XML is shown in Figure 2.

Note that all transliterations for the name "金庸" are assigned a value of "100" for their name attribute. And for linking, the id attributes of these fd elements have the same number stem (e.g. 1.1, 1.2, ...). Similar assignments are applied to the name "查良鏞". Also note that the multi-lingual information is stored in the script attribute, in the form of:

```
"script.language.romanization"
```

A feasibility study on the above MARC-XML format was conducted at the Hong Kong University of Science and Technology (HKUST) Library in the summer of 2001. A prototype system for a Global Name Access Control Metadata Repository was developed. The system was built on the XML-based Tamino Server, using the Perl programming language and XSLT stylesheet transformation to generate search results and displays. Figures 3-6 show MARC21 extended pages of the full record, and sample screens of the system: the main page, the search results page, and the public display.

Note that in Figure 3, the prototype system is capable of converting the MARC-XML

Figure 1 Proposed namespace

```
<?xml version="1.0" encoding="UTF-8"?>
<marc>
  <fd name="245" ind1="1" ind2="4" ....other attributes.... >
    <sf name="a">The ABCs of XML :</sf>
    <sf name="b">the librarian's guide to the eXtensible markup language /</sf>
    <sf name="c">by Norman Desmarais</sf>
  </fd>
  ....more fd elements follow....
</marc>
```

Figure 2 An example of name access contol metadata for "金庸" in XML

```
<?xml version="1.0" encoding="UTF-8" ?>
<marc>
  <fd id="1.1" script="cjk.chinese" name="100" ind1="1" ind2="b" label="NAME
       AUTHOR">
     <sf name="a"金庸,</sf>
     <sf name="d">1924-</sf>
  </fd>
  <fd id="1.2" script="latin.chinese.pinyin" name="100" ind1="1" ind2="b"
     label="NAME AUTHOR">
     <sf name="a">Jin, Yong,</sf><sf name="d">1924-</sf>
  </fd>
  <fd id="1.3" script="latin.chinese.wade-giles" name="100" ind1="1" ind2="b"
     label="NAME AUTHOR">
     <sf name="a">Chin, Yung,</sf><sf name="d">1924-</sf>
  </fd>
  <fd id="1.4" script="latin.japanese.hepburn" name="100" ind1="1" ind2="b"
     label="NAME AUTHOR">
     <sf name="a">Kin, Yō,</sf><sf name="d">1924-</sf>
  </fd>
  <fd id="1.5" script="latin.vietnamese" name="100" ind1="1" ind2="b"
     label="NAME SEE FROM">
     <sf name="a">Kim, Dung,</sf><sf name="d">1924-</sf>
  </fd>
  <fd id="2.1" script="cjk.chinese" name="400" ind1="1" ind2="b" label="NAME SEE
     FROM">
     <sf name="a"查良鏞,</sf><sf name="d">1924-</sf>
  </fd>
  <fd id="2.2" script="latin.chinese.pinyin" name="400" ind1="1" ind2="b"
     label="NAME SEE FROM">
     <sf name="a">Zha, Liangyong,</sf><sf name="d">1924-</sf>
  </fd>
  <fd id="2.3" script="latin.chinese.wade-giles" name="400" ind1="1" ind2="b"
     label="NAME SEE FROM">
     <sf name="a">Cha, Liang-yung,</sf><sf name="d">1924-</sf>
  </fd>
  <fd id="3" name="400" ind1="1" ind2="b" label="NAME SEE FROM">
     <sf name="a">Cha, Louis,</sf><sf name="d">1924-</sf>
  </fd>
  <!—other fd elements ... -->
</marc>
```

metadata to various MARC21 formats, including model A and model B. It should also be possible to convert the MARC-XML format to the UNIMARC format.

It is necessary to extend MARC21 so that the globalization attributes contained in the MARC-XML format can be maintained after the mapping. This can be achieved by enhancing subfield 8 (Field Link Control Subfield) for storing the `id` and the `script` attributes, for example:

```
|81.2\s\latin.Chinese.pinyin
```

"s" (script) is a new field link type that differentiates the multi-lingual attributes of the linked fields. See Figure 6 for the use of subfield 8.

MARC-XML offers other added-value features. It will be possible to implement Barnhart's (1996) idea of reducing data content redundancy in a name access control record by separating and linking data elements such as uniform title and dates associated with the name from the name elements.

With sufficient globalization information encoded in the MARC-XML name metadata, a virtual international authority file as proposed by Tillett (2001) can become a reality. It is feasible to have globally distributed repositories, offering name access control metadata as Web services, for access by various client applications such as OPACs (Online Public Access Catalogs), authority control modules, bookstore catalogs and rights management systems.

OCLC Systems & Services
Volume 18 · Number 2 · 2002 · 88–96

Figure 3 Main page of the metadata repository system

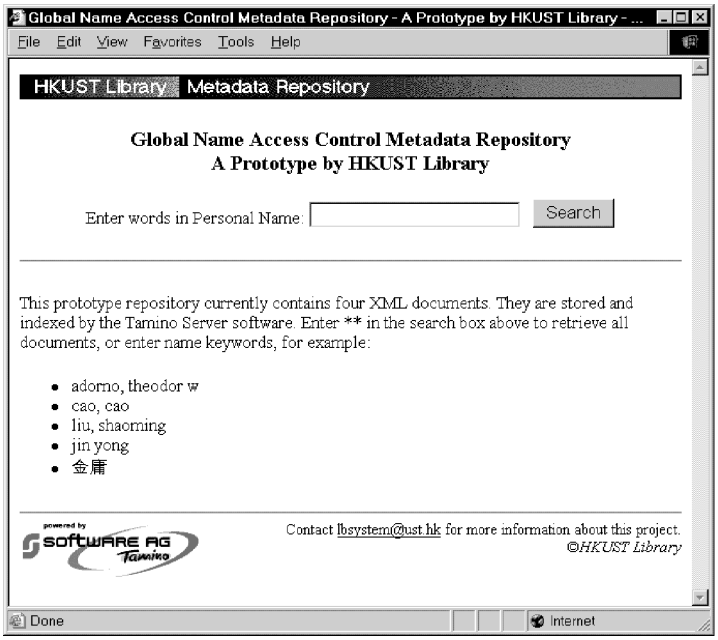
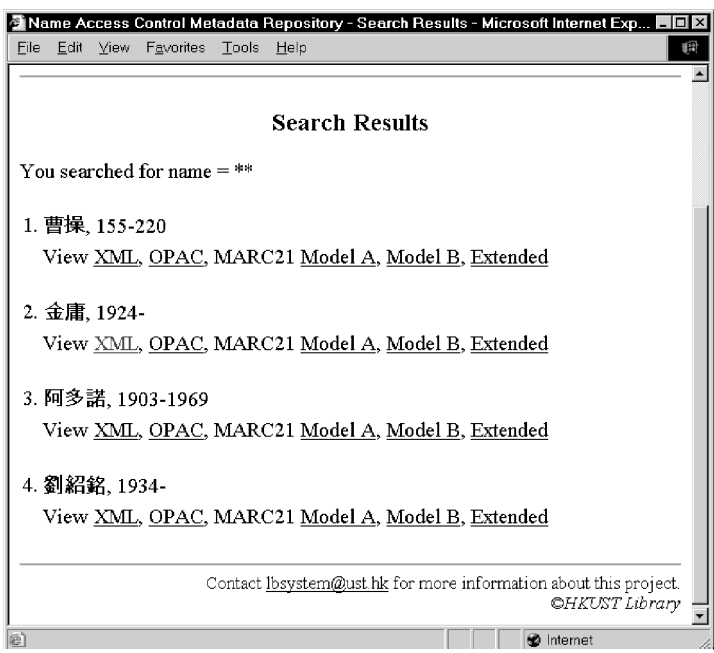


Figure 4 Search results page of the metadata repository system



SOAP communication with the repository

SOAP (Simple Object Access Protocol) (W3C, 2001) is an emerging standard of the W3C. It is a simple protocol for information exchange in a distributed environment. Using this standard, the prototype name access control metadata repository can be published as a Web service. A Web service is a collection of functions that are published on the Web for use by other programs (Glass, 2000).

To verify the SOAP communication concept, the prototype repository was enhanced as a SOAP node. Another SOAP application, pretending to be an INNOPAC WebPAC interface that communicates with the repository for author search assistance, was also developed. Figure 7 shows the SOAP messaging flow of this setup.

We chose to use the SOAP technology instead of the Z39.50 standard in this project. Although Z39.50 is a well-established information retrieval standard for library applications, it was designed before the World Wide Web era. Because of the steep learning curve for Z39.50, developers need to invest extra effort in order to deploy and support Z39.50 systems. On the contrary, SOAP uses the two common standards XML and HTTP for messaging and transport. Because it introduces no new technology and is simple to follow, SOAP can quickly be deployed for distributed applications requiring remote procedure calls. Driven by the advantages of these Web technologies, a group of Z39.50 implementers has begun to study the integration of XML, HTTP and SOAP to the Z39.50 standard (ZNG Initiative, 2001).

Figure 8 shows some SOAP messages between the INNOPAC and the repository.

It is desirable that a globally agreed-upon profile for these kinds of SOAP request and response messages be formed. Once the specification is in place, distributed authority repositories can be published as compatible Web services, serving MARC metadata in XML format. Similar to the idea of Z39.50, client applications can then communicate with selective repositories worldwide for their name access control.

Conclusion

There are limitations and problems in MARC21 for name access control at the global level. These problems can be resolved with the

Figure 5 OPAC format view of the metadata repository system

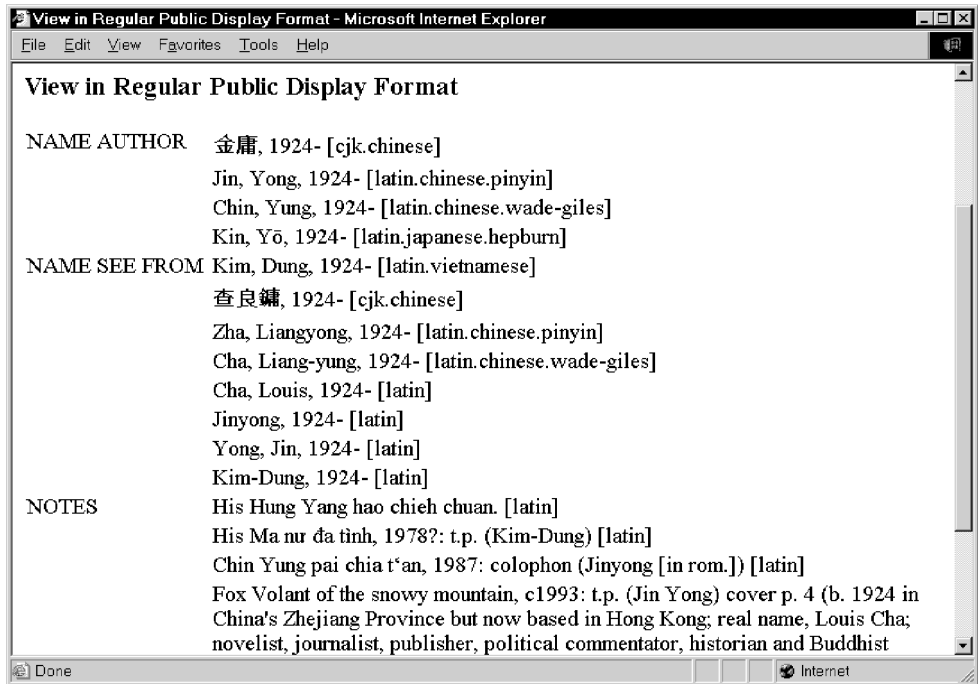


Figure 6 MARC21 extended format view of the metadata repository system

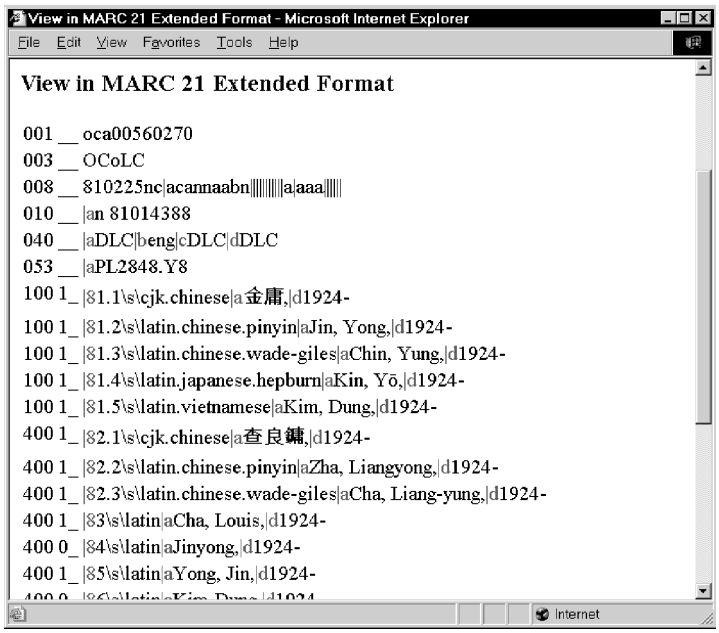
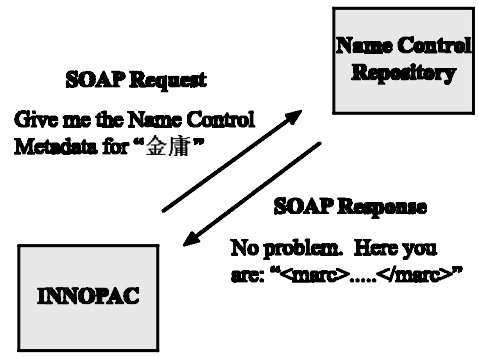


Figure 7 SOAP request and response



help of XML and related technologies. The proposed MARC-XML schema can crosswalk with MARC21 and any other MARC content designation versions, such as UNIMARC. With the use of SOAP, XML-based name access control metadata can be easily published as Web services. If major authority files, such as the Library of Congress name authority file, are converted to the proposed XML format and are enabled as SOAP nodes, a true global name access control environment will be achieved.

To make the above scenario happen, a global effort should be established to specify two standards, namely the XML schema for the MARC format and the SOAP messaging profile for bibliographic information retrieval.

OCLC Systems & Services
Volume 18 · Number 2 · 2002 · 88–96

Figure 8 SOAP messages between the INNOPAC and the repository

```
Request:
<env:Envelope xmlns:env="http://www.w3.org/2001/09/envelope">
  <env:Body>
    <m:SearchName xmlns:m="http://library.ust.hk/ac">
      <searchquery>
        <query>金庸</query>
      </searchquery>
    </m:SearchName>
  </env:Body>
</env:Envelope>

Response:
<env:Envelope xmlns:env="http://www.w3.org/2001/09/envelope">
  <env:Body>
    <m:SearchNameResponse xmlns:m="http://library.ust.hk/ac">
      <searchresults>
        <marc> <!– the MARC-XML metadata for 金庸 –> </marc>
      </searchresults>
    </m:SearchNameResponse>
  </env:Body>
</env:Envelope>
```

References

Barnhart, L. (1996), "Access control records: prospects and challenges", *Authority Control in the 21st Century: An Invitational Conference, March 31-April 1, 1996*, available at: www.oclc.org/oclc/man/authconf/barnhart.htm (accessed 29 January, 2002).

Glass, G. (2000), "The Web services (r)evolution, part 1: applying Web services to applications", available at: www-106.ibm.com/developerworks/webservices/library/ws-peer1.html (accessed 29 January, 2002).

Lane Medical Library, Stanford University Medical Center (2002), "Medlane XMLMARC", last updated 7 January, available at: xmlmarc.stanford.edu/ (accessed: 29 January, 2002).

Library of Congress (2000), "MARC 21 specifications for record structure, character sets, and exchange media. Character sets: part 2. UCS/Unicode environment", January, available: http://lcweb.loc.gov/marc/specifications/speccharucs.html (accessed: 29 January, 2002).

Library of Congress (2001), "MARC SGML and XML", last updated March, available at: lcweb.loc.gov/marc/marcsgml.html (accessed 29 January, 2002).

Tillett, B.B. (2001), "A virtual international authority file", *67th IFLA Council and General Conference, August 16-25, 2001*, available at: www.ifla.org/IV/ifla67/papers/094-152ae.pdf (accessed 29 January, 2002).

W3C (2001), SOAP Version 1.2 working drafts, available through the links in: http://www.w3.org/TR/2001/WD-soap12-part0-20011217/, December 17 (accessed 29 January, 2002).

ZNG Initiative (2001), "Z39.50 next generation", July, available at: www.loc.gov/z3950/agency/zng.html (accessed 29 January, 2002).

Implications for practitioners

This summary has been provided to allow a rapid appreciation of the significance of the content of this article. Browsers may then choose to read the article *in toto*, to derive full benefit from the authors' work.

It is often said that communications and technology have combined to make the world a smaller place.

Such ideas, and with them the concept of the global village, are attractive. However, they are not much consolation to library professionals faced with issues of global name access control.

The Library of Congress used the Machine Readable Cataloging (MARC) 21 standard to establish one of the world's largest name authority files. It was designed for Anglo-American libraries but its use has been far wider.

So, it is when some worldwide users adopt the Library of Congress records and standards for non-Latin names that problems occur. Chinese, Japanese and Korean are just three examples of scripts which create difficulties for MARC21 in Congress authority records, because of the limitations imposed by Latin transliterations.

Yong Jin, the Chinese novelist and journalist, serves as a good example. That appellation would be the entry taking the Pinyin romanization of his pen name for martial arts fiction. One could make eight more entries, including the Chinese script version of that name to Japanese and Vietnamese romanizations to his English name.

There are many weaknesses in the major fields of the authority record found in the OCLC database, and MARC21 can at the moment handle only some of them. The writer provides two different models which find ways of addressing some of the issues. Among them is the Established Linking Entry created by one model. This stores the Chinese script and makes it an alternate established form. The other allows better linking of the vernacular script with its transliterations.

However, bibliographic utilities such as OCLC, to take just one example, would not want to support it for authority control. Neither model is sufficient for use in a global setting; something more radical is needed.

Extensible Markup Language (XML) and related technologies can help. XML has been

rapidly adopted and implemented by many industries during the last four years. XML and other Web technologies offer some answers, including conversion between XML and MARC formats without data loss and with resolution of many problems inherent in MARC.

What we need now is a globally agreed-upon use of these opportunities. A MARC-XML feasibility was conducted last year at the Hong Kong University of Science and Technology Library. It suggests that, through globalization information encoded in the MARC-XML name metadata, a virtual international authority file can become a reality.

Other areas for further investigation include Simple Object Access Protocol (SOAP), an emerging World Wide Web Consortium standard. SOAP carries the dual advantage of being easy to follow while introducing no new technology. It could serve MARC metadata in XML format, and its use could pave the way for XML-based name access control metadata to be published as Web services.

The ultimate aim has to be creation of a true global name access control environment. That requires a global effort to work towards the standards described here.

That shrinking world or global village, which nevertheless is witness to an information explosion, demands nothing less.

(Précis supplied to MCB UP Limited by consultants.)