

MARC VÀ XML

Ninh Xuân Phúc & Trần Thị Mộng Linh

Phòng Tài nguyên thông tin – Thư viện NH Khoa học Tự nhiên

Hiện nay, cùng với sự phát triển của công nghệ thông tin, người ta nói nhiều đến sự liên thông thư viện. “Chuẩn hóa – Hoãn nhập – Phát triển” hiện đang là hướng phát triển chủ đạo của các thư viện. Muốn hoãn nhập và phát triển theo hướng hiện đại thì các hệ thống phải “hiệu” nhau. Nếu nhất nhất nhiều nơi yêu cầu này thì chuẩn hóa, bao gồm các chuẩn thư mục và chuẩn kỹ thuật. Trong bài viết này, chúng tôi chia sẻ với đồng nghiệp công tác trong lĩnh vực thư viện một số kiến thức cơ bản về các tài liệu trong nước và nước ngoài về MARC và XML. Nội dung không nhằm tra cứu kiến thức tin học mà chỉ cung cấp những khái niệm cơ bản về ngôn ngữ cấu trúc dữ liệu ứng dụng trong kỹ thuật lưu trữ trình bày và trao đổi thông tin.

KHÁI NIỆM MARC

Chuẩn biên mục MARC (Machine Readable Cataloging) – Biên mục có thể đọc bằng máy, là hình thức cho phép máy tính lưu trữ và truy xuất thông tin mục lục bao gồm các biểu ghi theo dạng MARC. Nghĩa là người biên mục cần mã hóa hình thức thông tin trong biểu ghi.

Vào những năm 1960, nhân viên Thư viện Quốc hội Hoa Kỳ (Library of Congress) đã phát triển một hình thức tiêu chuẩn cho việc lưu trữ các thông tin biên mục trên máy tính. Cho đến nay một khối lượng thông tin biên mục khổng lồ đã được lưu trữ. Chẳng riêng tại Thư viện Quốc hội Hoa Kỳ số lượng biểu ghi dạng MARC đã lên tới 50 triệu.

1. Tính nhất quán

Tính nhất quán của biểu ghi thể hiện ở việc các giá trị tiêu đề nhân và các tiêu đề biên mục. Nếu nay cho phép người sử dụng mục lục có thể tìm thấy tất cả các tài liệu liên quan đối cùng một tiêu đề.

2. Biểu ghi MARC

Một biểu ghi MARC được cấu thành bởi 3 thành phần:

- Cấu trúc biểu ghi: được xuất phát từ Tiêu đề chuẩn quốc gia Hoa Kỳ về trao đổi thông tin.
- Chế độ nội dung: bao gồm các mã và các quy ước do MARC format nhận dạng. Chúng nhận dạng các yêu cầu dữ liệu trong một biểu ghi và cho phép máy tính xử lý dữ liệu này.
- Nội dung dữ liệu của biểu ghi được tạo ra qua quá trình sử dụng tiêu chuẩn biên mục mã ACR2 và biên mục hệ thống LCSH.

020 \$a0879612177 (alk. Paper)
 040 \$aLC\$beng\$cLC\$dLC
 043 \$an-us---\$ap-----
 050 00 \$aQH104.5.P32\$bB76 1996
 082 00 \$a574.979\$220
 100 1 \$aBrown, Vison,\$d1912-
 245 10 \$aExploring Pacific Coast tidepools /\$cby Vinson Brown, Ane Rovetta
 250 \$aRev. And expanded ed.
 260 \$aHappy Camp. Calif. :\$bNaturegraph Publishers,\$c1996
 300 \$av, 127 p. :\$bill. (some col.) ;\$c22 cm.
 504 \$aIncludes bibliographical references (p. 119) and index.
 650 0 \$aTide pool animals\$zPacific Coast (U.S.)\$xIdentification.
 650 0 \$aTide pool plants\$zPacific Coast (U.S.)\$xIdentification.
 650 0 \$aTide pool ecology\$zPacific Coast (U.S.)
 700 1 \$aRovetta, Ane.

Bieù ghi mã MARC trên ñöôc hieù nhö sau:

000 00952nam##2200265#a#4500

Ñäù bieù (000) gôm 24 kyütöi, yinghóa cái vò trí:

vò trí 1-5 là ñöäi logic của bieù ghi (00952) do mã y táo ra;

vò trí 6: tình trạng bieù ghi (n: môi);

vò trí 7: Loại bieù ghi (a: Tài liệu bằng ngôn ngữ văn bản, gồm cả microform, microfilm, microfiche);

vò trí 8: cấp thö tìch (m: chuyên khảo);

vò trí 9: Kiểu kiểm tra (#: kiểu kiểm tra không cui the);

vò trí 10: không xác ñònh;

vò trí 11: số chæ thö luôn là 2;

vò trí 12: mã ñöông con, luôn là 2;

vò trí 13-17: Mã chæ DSDL (00265);

vò trí 18: Mã ñö mã ñö (#: mã ñö này ñu);

vò trí 19: Hình thöc bién mức mã ñö (a: Qui tắc mã ñö Anh-My);

vò trí 20: Yêu cầu bieù ghi liên quan (#: không yêu cầu);

vò trí 21-24: Số ñö thö tìch ñnh xai, luôn là 4500.

001 00012188151

Số kiểm soát (001) ñöôc gán cho tài chöc táo lập, söi dùng hoac phân phối bieù ghi, ñöôc thể hién trong ñöông 003.

003 CaOOAMICUS

Nhãn ñang số kiểm soát (003)

005 9970910000000.0

082 00 \$a574.979\$d220

Kyùhieäu phaïn loaï Dewey, coi2 chaè thò

Chaè thò 1: Daïng aïn baïn (0: aïn baïn ñaày ñu)

Chaè thò 2: nguoaïn kyùhieäu Dewey (0: ñoïc gaïn boï LC)

\$a: Phaïn loaï (574.979)

\$2: Soátai lieäu (20)

100 1 \$aBrown, Vison,\$d1912-

Daïn muïc chính – taïc giaicaiïnhaïn (100), chaè thò 1 dung cho teïn rieïng

\$a: Hoï teïn caiïnhaïn (Brown, Vison)

\$d: Naïm lieïn quan vooï teïn caiïnhaïn (1912-)

245 10 \$aExploring Pacific Coast tidepools /\$cby Vinson Brown, Ane Rovetta

Thoïng tin veïnhaïn ñeà(245), coi2 chaè thò

Chaè thò 1: Boäsung daïn muïc (1: Coïboäsung)

Chaè thò 2: Caïc kyütöï baät ñaäu khoïng saïp xep (0: soákyütöï baät ñaäu khoïng xöï lyüsap xep, nghóa laïsaïp xep töïkyütöï ñaäu tieïn)

\$a: Nhan ñeà

\$c: Thoïng tin traïch nhieïm

250 \$aRev. And expanded ed.

Laiïn xuaät baïn.

260 \$aHappy Camp. Calif. :\$bNaturegraph Publishers,\$c1996

Thoïng tin xuaät baïn: \$a – Nôi xuaät baïn, \$b – Nhaøxuaät baïn , \$c – Naïm xuaät baïn

300 \$av, 127 p. :\$bill. (some col.) ;\$c22 cm.

Moätai vaät lyü \$a – Khoá lööng tai lieäu (v, 127 p. : 5 trang ñaïn soáLa maïvaø127 trang ñaïn soáAïlrap), \$b – Caïc chi tiet khai (ill. : minh hoä, some col. : moät soáminh hoä mau), \$c – Khoá(22 cm.)

504 \$aIncludes bibliographical references (p. 119) and index.

Tham khaiö thò töch vaøchaè muïc (504)

650 0 \$aTide pool animals\$zPacific Coast (U.S.)\$xIdentification.

650 0 \$aTide pool plants\$zPacific Coast (U.S.)\$xIdentification.

650 0 \$aTide pool ecology\$zPacific Coast (U.S.)

Tieäu ñeàñeàmuïc (650) \$a: Ñeàmuïc chính, \$b: Tieäu phaïn muïc, \$x: Phui ñeàchung, \$z: Phui ñeàñoä lyü

700 1 \$aRovetta, Ane.

Daïn muïc boäsung (700), soáchaè thò 1- xaïc ñoïn teïn rieïng, \$a: teïn caiïnhaïn (Rovetta, Ane).

6. Moät vai toaïn tai cuï MARC

XU HƯỚNG MỚI - XML

Biểu ghi MARC có thể chứa hơn 800 trường mô tả dữ liệu chi tiết nên hình dạng thế hệ những phần lớn các trường này là không cần thiết và hầu như không được sử dụng trong các thư viện hiện tại. Trong thư viện hiện tại quá trình tách rời các kho dữ liệu luôn diễn ra trên các hệ quản trị CSDL với chức năng giúp người sử dụng nhanh chóng tìm kiếm và truy cập thông tin trực tiếp thông qua mạng máy tính, không dùng nữa các thẻ in sẵn. Hạn chế của biểu ghi MARC về phương pháp trao đổi dữ liệu là nó rất khó khăn và viết bởi con người, mô tả cũng như một số trường với mã đại số, hạn chế trong việc hỗ trợ các ngôn ngữ theo chuẩn UNICODE khi cần sử dụng đến 04 bytes cho một ký tự.

Xu hướng hiện nay trên thế giới, các biệt lập các quốc gia cho hình thành các kho dữ liệu thông tin – thư viện theo dạng MARC, là chuyển qua sử dụng **chuẩn biểu thức MARC với ngôn ngữ mô tả dữ liệu XML**. Dữ liệu được trực tiếp trao đổi nên liên thông trên mạng Internet từ các website, không cần sử dụng nữa giao thức liên thư viện quốc tế Z39.50.

1. XML là gì?

XML (eXtensible Markup Language) ngôn ngữ hình dấu môi trường có nguồn gốc từ ngôn ngữ hình dạng siêu văn bản HTML (Hyper Text Markup Language), cái hai ngôn ngữ này đều bắt nguồn từ chuẩn ngôn ngữ hình dạng văn bản tổng quát có cấu trúc SGML (Structured General Markup Language).

XML là ngôn ngữ được hình thành bởi tổ chức mạng toàn cầu (World Wide Web Consortium), thông tin viết tắt theo cách chơi chữ là W3C. Đây là tổ chức quốc tế hình ra các chuẩn của Web và Internet.

Một văn bản XML hình thành từ các thẻ (tag) với tên gọi phần tử (element). Khác với ngôn ngữ HTML, số lượng và tên gọi các phần tử trong XML là không hạn chế XML là ngôn ngữ tổng quát dùng hình thức dữ liệu thông qua các thẻ. Trong HTML các thẻ được hình thành và qui định trước. Trong khi với XML ta có thể tùy ý hình thành mọi thẻ. Nhờ vậy có thể coi XML như tập cha của ngôn ngữ HTML. Do vậy một số qui tắc, XML tồn tại và phát triển thì thành các ngôn ngữ hình thức khác.

Nhiệm vụ quan trọng nhất là XML cho phép dễ dàng xử lý chuyển tải và trao đổi dữ liệu giữa rất nhiều ứng dụng và tài liệu người dùng với các hình dạng khác nhau. Nếu đã quen với máy tính, bạn sẽ biết rằng có rất nhiều hình dạng file khác nhau. Việc chuyển đổi dữ liệu giữa chúng quả là nan giải mà để được không ít trình ứng dụng hỗ trợ. Ví dụ như file .DOC (Văn bản Word), .XLS (Bảng tính Excel), .DBF (Lập trình FoxPro), .MDB (Lập trình Access), .TXT (File văn bản), .RTF (Rich Text Format) và mỗi đây là .HTML. Chẳng riêng các file văn bản thôi mà nếu gặp khó khăn, nếu bạn nhận được một file Word 2000 mà máy tính của bạn còn dùng Word 7.0, cố gắng làm cũng chẳng được phần văn bản còn các nội dung khác thông bị biến dạng.

Trong XML, dữ liệu vẫn ở dạng mã nguồn mở và vẫn có thể dễ dàng cấu hình cũng như thay đổi bằng các trình soạn thảo thông thường nếu không có trong tay trình soạn thảo XML chuyên nghiệp. Dữ liệu vẫn các thể trong XML không mã hóa, không nội hội bản quyền.

Tháng 12/1997, phiên bản đầu tiên XML 1.0 (Extensible Markup Language – Ngôn ngữ đánh dấu mở rộng) ra đời và là chuẩn nền tảng của SGML. Từ đó, nhiều công ty phần mềm đã chạy đua ứng dụng XML vào nhiều lĩnh vực. Hàng trăm ngôn ngữ đánh dấu chuyên dụng dựa trên XML đã ra đời. Dưới đây là một số tiêu chuẩn ngôn ngữ đánh dấu dựa trên XML cho thấy sức mạnh của XML:

- BITS – Banking Industry Technology Secretariat : Ngôn ngữ và phòng về kỹ thuật nghiệp vụ ngân hàng
- IFX – Financial Exchange : Trao đổi dữ liệu tài chính
- BIPS – Banking Internet Payment System : Hệ thống thanh toán qua Internet của nghiệp vụ ngân hàng
- TIM – Telecommunication Interchange Markup : Mã đánh trao đổi viễn thông
- CBL – Common Business Library : Thư viện kinh doanh phổ thông
- ebXML – XML kinh doanh điện tử
- PDML – Product Data Markup Language : Ngôn ngữ đánh dấu dữ liệu sản phẩm
- FIX – Financial Information eXchange Protocol : Giao thức trao đổi thông tin tài chính
- CML – Chemical Markup Language : Ngôn ngữ đánh dấu trong lĩnh vực hóa học, cho phép biểu diễn các công thức hóa học, hóa trị phân tử đang nghiên cứu,
- v.v.

Dữ liệu thông tin – thư viện mở rộng cũng là một lĩnh vực tiềm năng. Dung chuẩn XML:

- có thể tạo biểu ghi thư tịch một lần và xuất bản chung theo các dạng khác nhau;
- hiện thì biểu ghi thư tịch trực tiếp trên trình duyệt Web, search engines (công cụ tìm kiếm) và các hệ thống thư viện tiềm năng khác mà không cần chuyển đổi;
- biểu ghi thư tịch có thể dễ dàng chuyển đổi qua lại giữa XML và MARC mà không bỏ sót thất.
- nhiều vấn đề tồn tại trong dạng MARC phức tạp, ví dụ như việc kiểm soát tiêu chuẩn.

Năm 1995, TVQH Mỹ bắt đầu nghiên cứu tính khả thi của việc dùng SGML (Standard Generalized Markup Language – Chuẩn ngôn ngữ đánh dấu và bản tổng quát của cấu trúc) để mã hóa dạng MARC 21. Sau đó, phiên bản MARC DTDs (Document Type Definitions) định nghĩa loại văn bản MARC để định nghĩa dữ liệu MARC 21 trong dạng thức SGML được phát hành năm 1998. Cùng năm này TVQH Mỹ công bố phần mềm chuyển đổi giữa MARC 21 và SGML.

2. Chuyển biểu ghi MARC sang XML

Những cuốn tranh luận về việc mã hóa biểu thị thô tích theo chuẩn MARC về các biểu ghi máy móc môi trường hơn và có thể ảnh hưởng tới trong môi trường Internet. Tác giả K. T. Lam, Nhà học Khoa học và Kỹ thuật Hồng Kông đã nghiên cứu ứng dụng XML về các tiến mô liên kết giữa tiêu đề chính và các tiêu đề không nội thất lập và trình bày kết quả như sau:

Một trong những vấn đề lớn trong mục lục MARC là cách trình bày và kết nối tiêu đề các tài liệu có nhiều dạng tên khác nhau, nội dung ghi bằng nhiều ngôn ngữ. Có những tài liệu có nhiều hình thức tên gọi và tên những tác giả nổi tiếng thông qua phiên âm bằng nhiều ngôn ngữ khác nhau.

Trong biểu ghi MARC cho phép chọn một hình thức tên trong CSDL thô tích. Nhu cầu hiện tại là phải thể hiện nội thất các hình thức tên gọi của cùng một tác giả. Sau đây là một biểu ghi tiêu đề các tài liệu của một nhà báo, nhà văn Hồng Kông Kim Dung tới TVQH Mỹ

```
....
100 1 $a Jin, Yong,$d1924-
400 1 $aChin, Yung,$d1924-
400 1 $aZha, Liangyong,$d1924-
400 1 $a Cha, Louis,$d1924-
400 1 $a Cha, Liang-yung,$d1924-
400 0 $a Kim-Dung,$d1924-
400 1 $a Kim, Dung,$d1924-
400 0 $aJinyong,$d1924-
400 1 $a Yong, Jin,$d1924-
400 1 $a Kin, Yo , $d1924-
....
```

Biểu ghi này gặp phải những vấn đề sau:

- Chưa bao gồm những hình thức tên theo chữ cái La tinh, còn tên tiếng Hoa không ghi nội thất.
- Trong ví dụ này, "Jin, Yong" nội thất lập danh mục chính (100), còn các hình thức tên khác của tác giả Kim Dung nội thất nữa vào thông tin tung thô (400) . Tuy nhiên, những quốc gia khác lại muốn chọn hình thức tên khác làm tiêu đề chẳng hạn Hồng Kông sẽ chọn hình thức tên tiếng Hoa làm tiêu đề

Nếu nữa tên tiếng Hoa của tác giả Kim Dung vào biểu ghi MARC, tác giả K. T. Lam đã sử dụng hai cách, nội thất gọi Kiểu A và Kiểu B trong môi trường UCS/Unicode (UCS – Universal Character Set – Bộ ký tự toàn cầu) như sau:

Kiểu A

Dung Kieu A hay Kieu B neu conhong thuan loi va bat loi rieng. Tuy nhien, neu dung XML neinh dau sieu doi lieu tieu neichuan, nhong rat roi nay hoan toan noch khac phuc. Vi doi, chung ta coi moitaisau:

```
<?xml version="1.0" encoding="UTF-8"?>
<marc name="authority" cdate="19980625" udate="19980625" rcn="ABrG">
....
<fd id="0" name="001" ind1="" ind2="" label="Control Number">
  <sf name="">oca00560270</sf>
</fd>
<fd id="1.1" script="cjk.chinese" name="100" ind1="1" ind2="b" label="Author">
  <sf name="a"> ,</sf><sf name="d">1924-</sf>
</fd>
<fd id="1.2" script="latin.pinyin" name="100" ind1="1" ind2="b" label="Author">
  <sf name="a">Jin, Yong,</sf><sf name="d">1924-</sf>
</fd>
<fd id="1.3" script="latin.wadegiles" name="100" ind1="1" ind2="b" label="Author">
  <sf name="a">Chin, Yung,</sf><sf name="d">1924-</sf>
</fd>
<fd id="2.1" script="cjk.chinese" name="400" ind1="1" ind2="b" label="See From Author">
  <sf name="a"> ,</sf><sf name="d">1924-</sf>
</fd>
<fd id="2.2" script="latin.pinyin" name="400" ind1="1" ind2="b" label="See From Author">
  <sf name="a">Zha, Liangyong,</sf><sf name="d">1924-</sf>
</fd>
<fd id="2.3" script="latin.wadegiles" name="400" ind1="1" ind2="b" label="See From Author">
  <sf name="a">Cha, Liang-yung,</sf><sf name="d">1924-</sf>
</fd>
<fd id="3" script="latin.english" name="400" ind1="1" ind2="b" label="See From Author">
  <sf name="a">Cha, Louis,</sf><sf name="d">1924-</sf>
</fd>
....
</marc>
```

Dung bang ninh kieu XSL (eXtension Style Sheet – Bang ninh kieu moiroing) nho tren se tao ra kieu doi lieu ma moit he toi nhong tho vien coitheachuyen thanh cai Kieu A va Kieu B ma khong gap phai khoikhan na. Moi noan <fd id ... > nen </fd> la thong tin ve moit hinh thoi tieu nei trong noi bao gom cai kieu choi nhan trong va trong con MARC, v.v., con ho the ta ghi lai noch tat cai cai hinh thoi ten bang moi moi hinh thoi ngoi ngoi cua moit tac gia va coitheatiep cai bieu ghi to bat ky hinh thoi tieu nei ten canhan na.

Minh hoĩa nhoĩtreĩn cho thay khainang khac phuc nhöng bat cap ve ket noi hoac chuyen giao döo lieu con toĩn tai trong nöhnh dang MARC.

Thiet ngho, söi phat trien nhanh choa töng coĩc cuĩa công nghệ thöng tin nãovãñang chi phoi hoat nöhng thöng tin – thö vieĩn. Trong quaitrinh phat trien, MARC nãonöhng vai tro quan troĩng, giai quyet nöhöc nhieu van ñeĩ öĩnöhng bööc ñau tin hoc hoĩ thö vieĩn, tao ra hieu qua xoi hoĩ khöng nöhĩnãp öĩng nhanh choĩng nhu cau löu troĩvã tìm kiem thöng tin. Khoa hoc thöng tin khöng döng lai ma tiep tuc phat trien, nhöng nöhöc ñiem cuĩa MARC format se nöhöc khac phuc. Công nghệ XML ñang dan thay theĩ MARC. Nöhieu nay khöng coĩng hĩa MARC bö loai böi hoĩn toĩn ma nöhöc phat trien len möc ñoĩ cao hön, hieĩn ñãĩ hön, ñoĩ la söi döĩng **chuan bien muc MARC voi ngoĩ ngoĩ moĩta döo lieu XML.**

TAI LIEU THAM KHAO

HOANG LE MINH. *Giai phap kyö thuaĩ hoĩtroĩ lieĩn thöng tin – Soĩtay quan ly thöng tin – thö vieĩn* / Nguyẽn Minh Hiẽp chui bieĩn. – TP. HCM : Nãĩ hoc Quoc gia, 2002.

K. T. Lam. *Moving from MARC to XML* <http://ihome.ust.hk/~lblkt/xml/marc2xml.html>

MILLER, DICK R.. *XML and MARC : A Choice or Replacement?* <http://elane.stanford.edu/laneauth/ALChicago2000.html>

MORTIMER, MARY. *Kieĩn thöic coĩ baĩn veĩ MARC 21.* – H. : Cty Nam Hoang, 2001.

NGUYỄN PHÖÔNG LAN. *XML : Neĩn taĩng vã öĩng döĩng* / Nguyẽn Phöông Lan, Hoang Nöhic Hai. – TP. HCM : Giaĩp Düic, 2001.

“Coĩ theĩ hình dung MARC nhö möĩ ngoĩ nhaĩ tieĩn nghi, röĩng rãĩ, nhöng cuĩ kyö vã nöhöc söĩ chöĩ nhieu laĩ. Nöhöng troĩ öĩc nhu cau hieĩn ñãĩ hoĩ, noĩ gãĩ cho ngoĩ ta caĩm giai lööng löĩ giöĩ trung tu hay xay möĩ”

Dick R. Miller