Pesquisa sobre ferramentas de conversão de registros catalográficos padrão MARC para a linguagem XML

A survey on tools the for the conversion of MARC cataloging records into XML language

Mauricio Barcellos Almeida[1]
Beatriz Valadares Cendon[2]

Resumo

Com a automatização das bibliotecas a partir da década de 70, o MARC (*Machine Readable Cataloging Record)* foi adotado como padrão para representar documentos em bases de dados catalográficas. Novas necessidades apareceram a partir da emergência da Internet e da explosão da disseminação da informação nos anos 90. Nesse contexto informacional, a linguagem XML (*Extensible Markup Language*) passou a ser uma alternativa às limitações do padrão MARC na representação de registros catalográficos. O propósito desse artigo é apresentar ferramentas de código aberto que realizam a conversão de registros do padrão MARC para a linguagem XML. São estudadas características básicas do padrão MARC, da linguagem XML e limitações do padrão MARC nos ambientes atuais. Apresentam-se nove ferramentas de conversão disponíveis na Internet, comentários sobre testes em cinco dessas ferramentas e exemplos de fragmentos XML gerados no processo. Conclui-se analisando a forma com que as ferramentas efetuam a conversão e apontando direções para trabalhos futuros.

Palavras-chave: MARC, XML, registros catalográficos, registros bibliográficos

Abstract

With the automation of libraries in the seventies, the MARC (Machine Readable Cataloging Standart) was adopted as a standard to cataloging data in eletronic databases. New needs arose with the emergence of the Internet and with the information boom phenomenon in the nineties. In that context, XML(Extended Markup Language) became an alternative to the limitations of the MARC standard for the cataloging records representation. The purpose of this paper is to present open-source tools for converting MARC standard into the XML language. It presents the basic features of both, MARC standard and the XML language and limitations of the MARC standard in current enviromments. The paper presents nine available-on-the-net conversion tools and comments the tests performed in five of them. It also presents XML fragments generated by conversion tools. The paper concludes analysing the conversion process used by each tool and pointing directions to future works.

Keywords: MARC, XML, cataloging records, bibliographic records

1 Introdução

Os catálogos têm sido a forma com que as pessoas buscam o acesso aos documentos e, finalmente, à informação. Na década de 70, catálogos foram convertidos para o formato eletrônico e organizados em registros compostos por campos padronizados, para que pudessem ser manipulados em computadores. O padrão MARC (*Machine Readable Cataloging Record*) se

[1] Mestre em Ciência da Informação pela UFMG. Professor assistente da PUCMINAS. mba@pucminas.br.
[2] Doutora em Filosofia e professora adjunta da ECI-UFMG. cendon@eci.ufmg.br.

tornou a principal forma de representar registros catalográficos e seu formato de comunicação (que possibilita a transferência dos registros entre diferentes mídias) foi intensamente utilizado para o intercâmbio entre sistemas de bibliotecas.

O tamanho das bibliotecas e o acesso a elas cresceu a partir da explosão da disseminação da informação via Internet nos anos 90, sugerindo a necessidade de um novo paradigma para a busca em catálogos (WELTY e JENKINS, 2000). Em ambientes abertos e distribuídos como a Internet, onde um catálogo pode ser acessado a partir de diversos locais, as linguagens de marcação são a principal forma de representar e transportar dados. Nesse contexto, destaca-se a linguagem XML (*Extended Markup Language),* a qual é uma linguagem de marcação publicada em 1997 pelo *W3Consortium,* que tem aplicação na descrição de conteúdo e transporte de dados.

Tornar disponíveis os registros catalográficos diretamente em ambientes abertos e distribuídos não é possível em função de características do padrão MARC (MILER, 2000; FIANDER, 2001; TENNANT, 2002). Os registros do padrão MARC não podem ser publicados na Internet, pois seu formato complexo não pode ser interpretado pelos navegadores. Entretanto, o legado composto por registros do padrão MARC não pode ser simplesmente descartado. A linguagem XML têm sido vista como uma alternativa para a representação desses registros, possibilitando que eles estejam disponíveis nos ambientes atuais.

Este artigo apresenta uma pesquisa sobre ferramentas de *software* especializadas na conversão de registros MARC para a linguagem XML e está organizado da seguinte forma: na Seção 2 apresentam-se características básicas do padrão MARC e da linguagem XML; na Seção 3, discutem-se as limitações do padrão MARC e o uso da linguagem XML como alternativa na representação dos registros; as iniciativas pioneiras para conversão do padrão MARC para a linguagem XML, os *softwares* de conversão pesquisados e comentários sobre os testes executados são apresentados na Seção 4; a Seção 5 identifica características dos arquivos XML resultantes da conversão, apresenta fragmentos XML obtidos com as ferramentas, conclui sobre as possíveis vantagens para as bibliotecas na conversão do padrão MARC para a linguagem

XML e apresenta direções para estudos futuros.

2 Características básicas do padrão MARC e da linguagem XML

Nessa seção apresentam-se algumas características básicas do padrão MARC e da linguagem XML. Um conhecimento mínimo sobre os duas formas de representação é necessário para se tratar da conversão entre eles. O leitor iniciado nos dois assuntos pode dispensar a leitura dessa seção pois o seu conteúdo é introdutório.

2.1 O padrão MARC

O padrão MARC foi criado nos anos 70 pela *Library of Congress*, com a finalidade de possibilitar que registros bibliográficos pudessem ser manipulados em computadores. O MARC recebeu modificações e passou a ser denominado USMARC nos anos 80 e MARC 21 no final dos anos 90. É utilizado na organização de catálogos de bibliotecas em todo o mundo.

A sigla MARC significa *Machine Readable Cataloging Record*, ou seja, registro catalográfico legível por computador. Registro catalográfico significa um registro bibliográfico[3], tradicionalmente apresentado em uma ficha catalográfica que inclui uma descrição (título, responsabilidade, edição, dados sobre o material, descrição física, etc.), a entrada principal e as entradas secundárias ("pontos de acesso" que permitem recuperar itens em um catálogo), cabeçalhos de assunto (descritores retirados de listas padronizadas de termos que descrevem o conteúdo do item) e os números de chamada (código de classificação, em geral alfanumérico, que reúne itens de mesmo assunto em um mesmo local físico) (FURRIE, 1994).

O padrão MARC é composto por diversos campos padronizados, que contém representação de dados e metadados bibliográficos. Cada campo é identificado por uma seqüência de três dígitos (etiqueta), por exemplo: 100 para o campo autor, 130 para o campo título, 300 para o campo descrição física, etc. Os campos podem conter subcampos como

[3] Daqui em diante, adotar-se-á o termo registro bibliográfico.

explicado adiante.

O registro MARC contém sinalizadores que marcam o registro armazenado e auxiliam na leitura e interpretação desse registro. Os sinalizadores indicam o início e o término dos campos e subcampos. Por exemplo, ao invés de palavras, usam-se os códigos 260 $a $b $c, para marcar o campo que contém os subcampos "área de publicação", "local de publicação", "nome da editora" e "data de publicação" em cada registro. Os "sinalizadores" MARC auxiliam os computadores na leitura e interpretação do registro, marcando o registro bibliográfico para armazenamento em meio magnético.

A Figura 1 apresenta um fragmento de um registro, mostrando os sinalizadores de texto e seus correspondentes no padrão MARC:

Dados	Sinalizadores de texto	**Sinalizadores MARC**		
		Campos	Indicadores	Subcampos
Arnosky, Jim.	Entrada principal, nome pessoal com sobrenome simples	100	1#	$a
Raccoons and ripe corn / Jim Arnosky.	Área de título Menção de responsabilidade	245	10	$a $c
New York : Lothrop, Lee & Shepard Books, c1987.	Local de publicação Nome da editora. Data de publicação	260	##	$a $b $c
25 p. : col. ill. ; 26 cm.	Paginação Ilustrações Dimensão	300	##	$a $b $c

Figura 1 –Simbologia do padrão MARC (Fonte: FURIE, 1994)

A Figura 1, mostra os sinalizadores para campos, indicadores e subcampos. O numero "100" corresponde à etiqueta que representa o campo onde está o nome do autor.

Os indicadores correspondem a duas posições de caracteres localizados após cada etiqueta. Na primeira linha da Figura 1, os indicadores para o campo "100" são os caracteres "1" e "#" (o símbolo # significa que o indicador não é usado). Um indicador de valor "1" no campo título, correspondente ao "100" e significa que deverá haver uma entrada de título no catálogo.

Cada tipo de dado em um campo é chamado subcampo e é precedido pelo código do subcampo, representado por letras minúsculas. Na Figura 1, o campo "300" tem o subcampo "a" que representa o número de páginas. O código do subcampo é precedido por um delimitador. Delimitadores são caracteres usados para separar subcampos e podem ser representados por

diferentes símbolos (@, ( ), $, _, etc). Na Figura 1, os delimitador é o sinal "$".

Existem diferentes formas pelas quais um registro bibliográfico pode ser representado: uma ficha catalográfica tradicional (cartão), as telas dos sistemas informatizados de bibliotecas OPAC (*Online Public Access Catalog*) e as telas de edição de dados de *softwares* que trabalham com o padrão MARC. Além desses, o padrão MARC possui um formato de comunicação, que segue a norma ISO 2709 e é utilizado quando o objetivo é o intercâmbio de registros bibliográficos. O formato de armazenamento interno é convertido para o formato de comunicação para que os registros possam ser transferidos entre sistemas (MARCONDES; SAYÃO, 1992).

```
01041cam  2200265 a 4500001002000000003000400020005001700024008004100041010002400082020002500106020004400131040001800175050002400193082001800217100003200235245008700267246003600354250001200390260003700402300002900439500004200468520022000510650003300730650001200763^###89048230#/AC/r91^DLC^19911106082810.9^891101s1990####maua###j######000#0#eng##^##$a###89048230#/AC/r91^##$a0316107514 :$c$12.95^##$a0316107506 (pbk.) :$c$5.95 ($6.95 Can.)^##$aDLC$cDLC$dDLC^00$aGV943.25$b.B74 1990^00$a796.334/2$220^10$aBrenner, Richard J.,$d1941-^10$aMake the team.$pSoccer :$ba heads up guide to super soccer! /$cRichard J. Brenner.^30$aHeads up guide to super soccer.^##$a1st ed.^##$aBoston :$bLittle, Brown,$cc1990.^##$a127 p. :$bill. ;$c19 cm.^##$a"A Sports illustrated for kids book."^##$aInstructions for improving soccer skills. Discusses dribbling, heading, playmaking, defense, conditioning, mental attitude, how to handle problems with coaches, parents, and other players, and the history of soccer.^#0$aSoccer$vJuvenile literature.^#1$aSoccer.^\
```

Figura 2 – Exemplo de um registro MARC, ainda como uma string no formato de comunicação MARC
Fonte: FURIE, 1994.

No formato de comunicação, precedendo a parte do registro bibliográfico que contém os dados, existem duas seqüências de caracteres chamadas "líder" e "diretório". O líder corresponde aos vinte e quatro primeiros caracteres de cada registro e é usado pelo computador. O diretório informa quais etiquetas existem no registro e sua localização no formato de comunicação.

Conforme mencionado, o padrão MARC apresenta limitações para a representação de dados bibliográficos em um contexto como o atual, caracterizado por grande quantidade de fontes de dados distribuídas e pela utilização intensiva de redes e da Internet na disseminação da informação. Uma alternativa para a representação dos registros no padrão MARC é a utilização

da linguagem XML. Conceitos básicos sobre a linguagem XML são apresentados na seção seguinte.

2.2 A linguagem XML

Historicamente, a palavra "marcação" descreve anotações ou marcas que informavam a um desenhista ou datilógrafo sobre a maneira como parte de um texto deveria ser representada. A marcação em textos impressos (sinais de pontuação, letras maiúsculas e minúsculas, disposição do texto na página, espaço entre as palavras, etc) ajuda as pessoas a determinar onde uma palavra termina ou onde outra começa, a identificar características estruturais (por exemplo, cabeçalhos) ou simples unidades sintáticas (por exemplo parágrafos e sentenças).

Com a automatização da formatação e impressão de textos, o termo marcação passou a ser usado em textos eletrônicos. Codificar ou "marcar" um texto para processamento em computadores é o processo de tornar explícito como o conteúdo do texto deve ser interpretado. Uma "linguagem de marcação" é um conjunto de convenções utilizadas para a codificação de textos.

Atualmente, a aplicação mais popular para linguagens de marcação são os arquivos HTML (*Hypertext Markup Language)*, os quais são lidos por *softwares* denominados navegadores. A linguagem XML é uma linguagem de marcação publicada em 1997 pelo *W3Consortium*. A linguagem XML e a linguagem HTML são simplificações da SGML (*Standard Generalized Markup Language).*

A estrutura da linguagem XML é caracterizada por uma marcação inicial inserida em sinais "<" e ">" e uma marcação final inserida nos sinais "<" e "\>". Um exemplo de dados textuais e o correspondente em linguagem XML é apresentado na Figura 3:

Dados textuais	Correspondente em XML
Catálogo de endereços João Silva Rua Carijós, 135 Belo Horizonte, MG 30.000 Brasil 31 3335-5556 (preferido) 31 3549-4446 joaosilva@net.com.br José Almeida jalmeida@net.com.br	<?xml version=”1.0”?> <catálogo de endereços> <entrada> <nome> João Silva </nome> <endereço> <rua> Carijós, 135</rua> <estado> MG </estado> < cep> 30.000 </cep> <país> Brasil </pais> <endereço> <telefone preferido=”true”>31 3335-4456</telefone> <telefone> 31 3594-4446 </telefone> <email> joaosilva@net.com.br </email> </entrada> <entrada> <nome><prim>José</prim> <sobren>Almeida</sobren> <email> jalmeida@net.com.br </email> </entrada> </catalogo de endereço>

Figura 3: dados textuais e o correspondente em linguagem XML

A diferença da linguagem XML para a linguagem HTML, atual padrão em uso na Internet, é que as marcações da linguagem XML não são fixas, ou seja, podem ser criadas de acordo com a necessidade do autor. A linguagem HTML foi desenhada para descrever apresentação e a linguagem XML para descrever conteúdo. A linguagem XML permite maior facilidade para interpretação dos dados por computadores, maior facilidade para a criação de aplicativos e é um formato livre de relações com fabricantes de *software* e *hardware.* A Figura 4 apresenta um HTML e seu correspondente em linguagem XML.

Fragmento HTML	Correspondente em XML

Pessoas que estudam na UFMG
  João , 30 anos, joao@ufmg.br
  Jose , 27 anos, jose@ufmg.br
	<mestrado>
	<descrição> Pessoas que estudam na UFMG </descrição>
	<turma>
	<pessoa>
	<nome> João </nome>
	<idade> 30 </idade>
	<email> joao@ufmg.br </email>
	</pessoa>
	<pessoa>
	<nome> Jose </nome>
	<idade> 25 </idade>
	<email> jose@ufmg.br </email>
	</pessoa>
	<turma>
	<mestrado>

Figura 4 – Fragmento HTML e seu correspondente em linguagem XML

Como a linguagem XML permite a criação das marcações pelo autor do documento, para que seja possível a comunicação entre diferentes instituições, é necessária a utilização de um conjunto de regras chamado de DTD (*Data Type Definition*). Diz-se que um documento XML é bem formado, quando ele tem um único elemento raiz, os elementos e entidades são aninhados adequadamente, os atributos estão entre aspas e as entidades são declaradas. Um documento XML válido deve ser bem formado, deve ter uma DTD e seguir as regras dessa DTD.

Além de outras aplicações, a linguagem XML tem sido estudada como uma alternativa ao padrão MARC na representação de registros bibliográficos, possibilitando, dentre outras vantagens, que o legado em padrão MARC se torne disponível para atuais aplicações e ambientes informacionais. A seção seguinte apresenta motivos que tornam interessante a idéia da conversão.

3 Problemas do padrão MARC nos ambientes informacionais atuais

A criação de índices e catálogos objetiva auxiliar os usuários a encontrar e localizar documentos, além de possibilitar acesso organizado às coleções. Entretanto, as necessidades dos usuários e as ferramentas de busca e recuperação se modificaram a partir da emergência da Internet e do avanço crescente nas técnicas computacionais.

Atualmente, os usuários buscam relações entre os dados que vão muito além do que pode oferecer um catálogo tradicional. O catalogo não é mais apenas uma ferramenta limitada aos visitantes da biblioteca, mas um nó em uma rede, o qual os usuários podem visitar de qualquer local do mundo via Internet (WELTY; JENKINS, 2000). Nesse contexto, verificaram-se as primeiras limitações do padrão MARC, até então dominante no registro bibliográfico via computadores.

O padrão MARC foi desenvolvido em um período em que as principais necessidades, na recuperação e manipulação de dados, eram restritas a acomodar dados de tamanho variável, a representar semântica para identificação de elementos de dados, a acomodar os modelos de

dados bibliográficos existentes e possibilitar o desenvolvimento de novos modelos no futuro (MCCALUM, 2000). Além disso, também era necessário racionalizar recursos de armazenamento, memória e processamento, em função de seu alto custo. Os recursos tecnológicos são hoje bem mais acessíveis do ponto de vista financeiro.

Em geral, localizados no que se convencionou chamar de "*Web* oculta" (PRICE, 2002) e utilizando sistemas ditos "legados"[4], o padrão MARC não possui uma linguagem de fácil aplicação e que possa ser interpretada pelos navegadores da Internet. Dessa forma não é possível tornar disponível um catálogo com registros no padrão MARC para consulta via rede. Grande parte dos acervos de bibliotecas em todo o mundo estão em formatos proprietários de sistemas integrados de bibliotecas ou no padrão MARC, ambos restritos aos serviços e sistemas de bibliotecas e inacessíveis para pesquisa direta de seu conteúdo via Internet.

A linguagem XML tem sido utilizada como um formato genérico para representação e transporte de dados. Sua utilidade tem-se comprovado em diversas áreas de estudo tais como: *Web* Semântica (BERNERS-LEE, 2000), manutenção de *web sites*, troca de informação entre organizações, comércio eletrônico, aplicações científicas, dispositivos de comunicação (MARCHAL, 2000), na representação de bancos de dados (ABITEBOUL; BUNEMAN; SUCIU, 2000) na representação de registros bibliográficos contendo diacríticos[5], caracteres especiais e dados em formato não-romano (LAM, 2001).

Nos novos ambientes, pode-se usar a linguagem XML e a XSL *(Extensible Stylesheet Language),* a qual permite apresentar a XML em um navegador comum, para criar registros uma única uma vez e depois manipulá-los em diferentes sistemas. Além disso, os registros apresentados diretamente em navegadores da Web, podem ser submetidos aos mecanismos de busca e a diferentes sistemas de bibliotecas, sem esforço de programação e sem perda de informações.

---

4 Sistemas construídos no passado em linguagens e padrões já em desuso, para os quais não se encontram mais profissionais qualificados ou os próprios autores para manutenção.

5 Sinais gráficos com que se marcam os caracteres alfabéticos para lhe dar um valor especial.

Visto a complexidade dos dados bibliográficos, o processamento por computadores será naturalmente beneficiado com a adoção da linguagem XML. A importância da linguagem XML parece inquestionável simplesmente pelo fato de que trata-se de uma linguagem capaz de representar estruturas complexas de uma forma não-proprietária e auto-explicativa (CARVALHO; CORDEIRO, 2002).

A utilização intensiva do padrão MARC nos últimos anos e o conseqüente volume de registros acumulados nesse formato, sugere que converter registros do padrão MARC para a linguagem XML pode tornar disponível o acervo das bibliotecas à Internet, sem grandes investimentos.

Algumas iniciativas para conversão de registros do padrão MARC em linguagem XML geraram ferramentas de *software* de código aberto. A disponibilidade dessas ferramentas contribui para que a conversão possa ser feita com facilidade pelas bibliotecas. A seção seguinte apresenta algumas iniciativas pioneiras de conversão do padrão MARC para a linguagem XML, ferramentas para conversão e algumas de suas características.

4 A conversão MARC/XML: iniciativas, s*oftwares* e exemplos

O desenvolvimento de ferramentas para tradução de registros do padrão MARC para a linguagem XML facilitou a conversão de registros em diversas instituições. Apresentam-se nas seções seguintes as principais iniciativas de conversão, os *softwares* disponíveis e comentários sobre os testes realizados com estes *softwares*.

4.1 Iniciativas pioneiras

Algumas iniciativas para utilização da linguagem XML para representar registros originalmente representados em MARC são apresentados abaixo:

A NLM (National Library of Medicine) utiliza a linguagem XML como formato para disseminação de dados de citações bibliográficas MEDLINE (United States National Library of

Medicine - Bibliographic Services Division);

- O *ADS (Astrophysics Data System)* da *NASA-National Spacial Agency* utilizou a linguagem XML para reformatar seus registros bibliográficos sobre observações astronômicas (ACCOMAZZI et.al., 2001);
- A *DialogWeb*, empresa que comercializa informações financeiras, industriais, governamentais, sobre patentes, sobre ciência e tecnologia, dentre outros, utiliza a linguagem XML em seu banco de dados de registros bibliográficos;
- No *WIPO (World Intellectual Property Organization)*, organização internacional de proteção a propriedade intelectual, a linguagem XML se tornou o formato padrão para submissão de documentos;
- A *Lane Medical Library* da *Stanford University Medical Center* criou o *Medlane Project,* para conversão de registros de seus catálogos para a linguagem XML, de forma a possibilitar a integração com outros recursos;
- O governo francês criou o projeto *BibloML,* para intercâmbio de dados de registros UNIMARC entre aplicações;
- A *Library of Congress* produziu um mapeamento *MARC x XML* e criou um módulo escrito em *PERL* (*Practical Extraction and Report Language)*, que efetua a conversão entre os formatos;
- A *Logos Research Systems* desenvolveu um conversor *MARC - XML - MARC* que converte registros do padrão MARC em XML bem formado. Pode também converter o documento XML em um registro do padrão MARC válido;
- A *Portia Systems* e o *DBC (Danish Bibliographic Center)*, implementaram a ferramenta *VisualCat.* O *DBC* produziu aplicativos XSL para validar os registros dinamarqueses *danMARC2/XML*. A conversão dos registros do padrão MARC para a linguagem XML é feita a partir das DTDs produzidas pela *Library of Congress* e pelo pacote JAVA XMLMARC desenvolvido pela *Stanford University* (o mesmo utilizado no projeto

MEDLANE);

- A OAI (*Open Archives Iniciative),* instituição dedicada à integração de bibliotecas digitais e a *Virginia Tech's DLRL (Digital Library Research Laboratory)* desenvolveram o projeto *MARC - XML - DTD onde foram* criadas classes *JAVA* para lidar com traduções entre os formatos de comunicação do padrão MARC e o *OAI-XML;*
- O projeto da União Européia *ONE-2* estuda o uso de *XSL* para conversões MARC XML;
- A *ICCU (Biblioteca Nacional Italiana)* e a *Bristish Library* participam da iniciativa *Shared Cataloguing Trial* que utiliza ferramentas do projeto ONE-2 e o *VisualCat;*

Na seção seguinte são apresentados nove *softwares* para conversão MARC XML disponíveis na Internet e comentários sobre os testes realizados em cinco deles. A escolha dos *softwares* que seriam testados baseou-se no critério simplicidade de instalação e os registros MARC utilizados nos testes foram exemplos obtidos na *Virginia Tech's DLRL.*

4.2 *Softwares* de conversão

Os *softwares* de conversão do padrão MARC para a linguagem XML estudados são ferramentas de código aberto disponíveis na Internet. Em alguns casos, são o resultado de iniciativas de instituições que desejavam converter seus registros para a linguagem XML.

Algumas ferramentas para conversão do padrão MARC para a linguagem XML foram pesquisadas, sendo apresentadas a seguir suas características básicas, procedimentos de instalação e uso, problemas e impressões. Não se pretende aqui apresentar uma lista completa de ferramentas. Testes experimentais foram feitos em cinco dessas ferramentas. Os *softwares* testados são JAMES (Seção 4.2.2), TIGRIS (Seção 4.2.3), MARCXML *Converter* (Seção 4.2.5), MARC *Tools* (Seção 4.2.6) e *MARC - XML - DTD* da *Virginia Tech's DLRL* (Seção 4.2.7).

O processo de teste consiste em utilizar, como entrada, um arquivo no formato de comunicação MARC e obter uma saída em código XML. Não se procurou aqui obter dados

quantitativos dos testes e não são apresentados resultados para cada processo de conversão. Pretende-se continuar a pesquisa e apresentar esse detalhes em outras publicações posteriores. Na Seção 5, são apresentados exemplos de fragmentos XML obtidos com o uso de ferramentas de conversão.

4.2.1 OAI (Open Archives Iniciative) / LOC (Library of Congress)

A ferramenta é um programa PERL que gera um documento XML o qual reflete as marcações e os subcampos que ocorrem no arquivo de entrada do padrão MARC. O arquivo de saída é um documento XML bem formado, mas não necessariamente válido. A DTD reflete apenas as marcações presentes no MARC 21. O *software* requer o ambiente PERL V5.003 ou superior[6] e o *parser* nsgmls versão 1.2 ou superior[7].

4.2.2 James (Java Marc Events)

A JAMES é uma API (*Application Program Interface)* útil na conversão do padrão MARC para a linguagem XML. É inspirado no SAX, uma API simples para a XML. Ao usar o JAMES, pode-se escrever programas que envolvem registros do padrão MARC sem saber detalhes sobre sua estrutura.

A JAMES fornece um modelo de acesso seqüencial a uma coleção de registros do padrão MARC no formato de comunicação. O objetivo é proporcionar uma interface genérica para uma aplicação que suporta o formato de comunicação ISO-2709. O software foi testado e executou a conversão. O uso do programa requer conhecimento básico do ambiente JAVA[8].

4.2.3 Tigris MARC4J

O MARC4J é uma biblioteca (de códigos) para trabalhar com registros do padrão MARC em JAVA. A biblioteca consiste de um *parser* MARC e de um modelo de objetos para edição de objetos do registro do padrão MARC e SAX2. O MARC4J não necessita de bibliotecas adicionais. É necessário o ambiente JAVA e que ele possa localizar a biblioteca. O software foi

---

6 Disponível em http://www.perl.com
7 Disponível em http://www.jclark.com/sp
8 Disponível em http://java.sun.com/

testado e executou a conversão. O uso do programa requer conhecimento básico do ambiente JAVA para seu uso. A instalação é similar a do JAMES.

4.2.4 MARC.pm

O MARC.pm é um módulo *PERL* que possibilita a leitura, manipulação e conversão de registros do padrão MARC. O MARC::XML é subclasse do MARC.pm que fornece métodos para conversão entre MARC e XML. O arquivo XML gerado não está associado a um DTD, o que significa que os documento precisam ser bem formados, mas não serão validados (conforme definições da Seção 2.2);

4.2.5 MARC XML Converter

Trata-se de um conversor MARC para XML de instalação simples. Um arquivo executável (*install.exe*) para Windows, MAC, Linux ou UNIX é fornecido pela Internet[9]. É necessário o ambiente JAVA 1.1.8 ou superior. O *software* foi testado e executou a conversão. Não é necessário nenhum conhecimento prévio, pois o *software* foi feito para usuários finais.

4.2.6 MARC Tools

Trata-se de um conversor do padrão MARC para XML de instalação simples. Um arquivo executável (*setup.exe*) para *Windows* é obtido pela Internet[10]. O programa têm uma interface simples, onde deve-se escolher a opção *MarcMaker*, indicar os arquivos de entrada, saída e que a saída deve ser em XML. O *software* foi testado e executou a conversão. Não é necessário nenhum conhecimento prévio, pois o *software* foi feito para usuários finais.

4.2.7 Virginia Tech DLRL

O *Virginia Tech DLRL* fornece um grupo de classes JAVA para traduções entre o formato de comunicação do padrão MARC e o XML-OAI. Uma DTD também é fornecida. O uso do programa requer um conhecimento básico de JAVA.

O *software* foi testado e apresentou problemas, com mensagens de erro na tentativa de

---

9 Dispomível em http://www.caspr.com/MarcXml_Install.html

10 Disponivel em http://ucs.orst.edu/~reeset/marcedit/software/setup.exe

conversão. Examinando-se o código, verificou-se que durante a compilação uma classe chamada *oaihandler* não estava presente. Fez-se contato com a *Virginia Tech DLRL*, que informou que o código utilizado originalmente para teste nesse artigo estava obsoleto[11]. Foi fornecido novo endereço[12] para se obterem novas classes. O *software* passou a funcionar normalmente, mas apenas para registros do padrão MARC que continham o campo OCLC do padrão MARC (um campo de identificação da *Library of Congress,* nem sempre presente em registros MARC).

4.2.8 BiblioML

A ferramenta BiblioML, já citada na Seção 4.1, é uma aplicação XML para registros bibliográficos, baseada no formato UNIMARC. Estão disponíveis uma ferramenta de conversão, DTDs e folhas de estilo XSL para conversão de registros UNIMARC codificados em linguagem XML para o formato BiblioML. O BiblioML é um formato baseado na XML, para intercâmbio entre registros bibliográficos UNIMARC.

4.2.9 *Medlane* XMLMARC

O projeto *MedLane*, já citado na Seção 4.1, utiliza a ferramenta XMLMARC desenvolvida pela *Stanford University*. O esquema da linguagem XML utilizado para modelar dados do padrão MARC é conhecido por XOBIS.

4.2.10 Quadro sinótico

O quadro abaixo sumariza as informações apresentadas sobre os *softwares*:

Software	Tipo de conversão	Ambiente de necessário	Testado nesse artigo	Apresentou problemas
OAI/LOC	Direta	PERL	N	-
JAMES	Direta	JAVA	S	N
TIGRIS	Direta	JAVA	S	N
MARC.pm	Direta	PERL	N	-
MARC XML Converter	Direta	Sistema operacional	S	N
MARC Tools	Direta	Sistema operacional	S	N
Virginia Tech DLRL	Direta	JAVA	S	S
Biblio-ML	Semântica	JAVA	N	-
Medlane e *Danish B.C*	Semântica	JAVA	N	-

Figura 5 – Quadro sinótico dos softwares pesquisados

[11] Disponível no endereço www.dlib.vt.edu

[12] Disponível em http://csgrad.cs.vt.edu/~rkelapur/MARC.zip

5 Conclusões e trabalhos futuros

A partir da avaliação dos códigos XML gerados pelos *softwares* apresentados (a partir dos testes ou de exemplos fornecidos pelos autores), observam-se duas formas principais em que a conversão é feita. A diferença ocorre em função de como o registro do padrão MARC é mapeado para a linguagem XML.

Dessa forma, classificam-se as ferramentas avaliadas em duas categorias principais:

- *Ferramentas de conversão direta:* aquelas que constroem as marcações do documento XML como uma equivalência direta dos elementos do padrão MARC;
- *Ferramentas de conversão semântica:* aquelas em que a estrutura do documento XML reflete o significado dos campos do padrão MARC;

Alguns exemplos de ferramentas de *conversão direta* existentes pesquisadas nesse artigo são:

- OAI (*Open Archives Iniciative)* / LOC (*Library of Congress*);
- *James(Java Marc Events)*;
- *Tigris;*
- *MARC.pm*;
- *MARC XML Converter*;
- *MARC Tools*;
- *Virginia Tech*.

Na Figura 6, apresenta-se um fragmento de código XML, com as marcações XML representando o mapeamento direto dos campos do padrão MARC, gerado pela ferramenta OAI-LOC:

```
<marc:record>
    <marc:leader>00925njm 22002777a 4500</marc:leader>
    <marc:controlfield tag="001">5637241</marc:controlfield>
    <marc:controlfield tag="003">DLC</marc:controlfield>
    <marc:controlfield tag="005">19920826084036.0</marc:controlfield>
    <marc:controlfield tag="007">sdubumennmplu</marc:controlfield>
    <marc:controlfield tag="008">910926s1957 nyuuun eng</marc:controlfield>
    <marc:datafield tag="010" ind1="" ind2="">
        <marc:subfield code="a">91758335</marc:subfield>
```

```
        </marc:datafield>
        <marc:datafield tag="028" ind1="0" ind2="0">
            <marc:subfield code="a">1259</marc:subfield>
            <marc:subfield code="b">Atlantic</marc:subfield>
        </marc:datafield>
        <marc:datafield tag="040" ind1="" ind2="">
            <marc:subfield code="a">DLC</marc:subfield>
            <marc:subfield code="c">DLC</marc:subfield>
        </marc:datafield>
        <marc:datafield tag="050" ind1="0" ind2="0">
            <marc:subfield code="a">Atlantic 1259</marc:subfield>
        </marc:datafield>
        <marc:datafield tag="245" ind1="0" ind2="4">
             <marc:subfield code="a">The Great Ray Charles</marc:subfield>
            <marc:subfield code="h">[sound recording].</marc:subfield>
        </marc:datafield>
        <marc:datafield tag="260" ind1="" ind2="">
            <marc:subfield code="a">New York, N.Y. </marc:subfield>
            <marc:subfield code="b">Atlantic,</marc:subfield>
            <marc:subfield code="c">[1957?]</marc:subfield>
        </marc:datafield>
...
    </marc:record>
</marc:collection>
```

Figura 6 – Fragmento XML gerado por ferramenta de conversão direta

Algumas ferramentas de conversão semântica existentes aqui pesquisadas são:

- *Danish Bibliographic Center and PortiaSystems* (citada na Seção 4.1);
- BiblioML;
- *Medlane* XML MARC.

Na Figura 7, apresenta-se um fragmento de código XML, com as marcações XML representando o mapeamento semântico dos campos do padrão MARC, gerado pela ferramenta Biblio-ML:

```
<?xml version="1.0" ?>
<!DOCTYPE BOOKLIST (View Source for full doctype...)>
<BOOKLIST>
    <BOOKS>
        <ITEM CAT="S">
            <TITLE>Number, the Language of Science</TITLE>
            <AUTHOR>Danzig</AUTHOR>
            <PRICE>5.95</PRICE>
            <QUANTITY>3</QUANTITY>
         </ITEM>
        <ITEM CAT="F">
            <TITLE>Tales of Grandpa Cat</TITLE>
            <PUBLISHER>Associated Press</PUBLISHER>
            <AUTHOR>Wardlaw, Lee</AUTHOR>
            <PRICE>6.58</PRICE>
            <QUANTITY>5</QUANTITY>
         </ITEM>
        <ITEM CAT="S">
            <TITLE>Language & the Science of Number</TITLE>
            <AUTHOR>Danzig</AUTHOR>
            <PRICE>8.95</PRICE>
            <QUANTITY>5</QUANTITY>
         </ITEM>
...
```

```
</BOOKLIST>
```

Figura 7 – Fragmento XML gerado por ferramenta de correspondência semântica

A linguagem XML foi concebida para representar conteúdo e assim, as ferramentas de conversão direta não parecem adequadas para aproveitar essa funcionalidade. Já as ferramentas de conversão semântica, parecem mais adequadas pois vão atender a principal audiência da linguagem da XML, ou seja, os próprios computadores, a partir do que se pressupõe na *Web* Semântica (BERNERS-LEE, 2000).

Apesar da profusão de iniciativas e ferramentas para a conversão, ainda não parece bem estabelecido, em termos práticos, como utilizar a linguagem XML nos sistemas de informação das bibliotecas. Não parece claro se a conversão direta produz resultados adequados ou se é essencial que a conversão leve em conta a semântica contida nos registros do padrão MARC. Acredita-se que o processo que considera a semântica na conversão deva ser o mais adequado (ARMS, 2000; CAPLAN, 2001; DILLON, 2001; KIM; CHOI, 2000; QIN, 2000).

Os registros do padrão MARC representam uma parte da "*Web* oculta" formada por registros que representam dados acadêmicos e de pesquisa, entre outros. Possibilitar que esses recursos estejam disponíveis para pesquisa na Internet pode auxiliar na busca por informações cientificas e parece uma iniciativa louvável. As bibliotecas poderão dessa forma tornar seu acervo mais facilmente disponível. Existem ainda *softwares* de código aberto que possibilitam às bibliotecas publicar na Internet os arquivos XML resultantes[13].

Em trabalho futuro espera-se apresentar resultados quantitativos de testes nos nove *softwares* de conversão pesquisados e caso práticos de registros no padrão MARC de bibliotecas reais. Propostas para a conversão de registros do padrão MARC para linguagem XML, aliadas a criação de bases de conhecimento[14], organizadas a partir de mapeamentos dos campos do padrão MARC para elementos de ontologias (WEINSTEIN, 1998), parecem um bom direcionamento para a continuidade das pesquisas na área.

---

12 Um exemplo de *software* de biblioteca digital de código aberto é o Greenstone, desenvolvido pela Universidade da Nova Zelândia e disponível em www.greenstone.org .

14 Dados organizados e representados através de alguma linguagem formal.

6 Referências bibliográficas

ABITEBOUL, S.; BUNEMAN, P.; SUCIU, D. *Data On the Web – from relations to semistructured data and XML.* San Francisco: Morgan Kaufman, 2000. 257 p.

ACCOMAZZI, A. et.all. *The NASA Astrophysics Data System: Architecture* - Harvard-Smithsonian Center for Astrophysics, Cambridge, MA 02138. Disponível em: <http://wwwxxx.lanl.gov/abs/astro-ph/0002105>. Acesso em: 27 nov. 2001.

ARMS, C. Some observations on metadata and digital libraries. In: CONFERENCE ON BIBLIOGRAPHIC CONTROL IN THE NEW MILLENNIUM (LIBRARY OF CONGRESS), 2000.

BERNERS-LEE, T. (2000). *Rules and Facts*: Inference engines vs Web. Personal Note. Disponível em: <http://www.w3.org/DesignIssues/Rules.html>. Acesso em: 21 set. 2001.

CAPLAN, P. 2001. *International metadata Iniciatives*: lessons in bibliographic control. Disponível em: <http://www.lcweb.loc.gov/catdir/bibcontrol/dillon_paper.html>. Acesso em: 20 fev. 2003.

CARVALHO, J. CORDEIRO, M. I. *XML and bibliographic data*: the TVS (Transport, Validation and Services). In: 68TH IFLA COUNCIL AND GENERAL CONFERENCE, Glascow, 2002.

CROSSNET Systems Limited. *Using XLST for XML MARC Conversin Records*. Disponível em: <http://www.crxnet.com/one2/xslt_marc_report.pdf>. Acesso em: 03 dez. 2001.

DANISH Bibliographic Center. Disponível em: <http://www.dbd.dk>. Acesso em> 03 dez. 2001.

DILLON, M. 2001. *Metadata for web resources*: how metadata works on the web. Disponível em: <http://www.lcweb.loc.gov/catdir/bibcontrol/dillon_paper.html>. Acesso em: 20 fev. 2003.

EU-PROJECT in DG XIII. Disponível em: <http://www.one-2.org/>. Acesso em: 02 dez. 2001.

FIANDER, D. J. (2001). Applying XML to the bibliographic description. *Cataloguing & Classification Quartely*, v. 33, n. 2, p.17-28, 2001.

FURRIE, B. (1994). Título original. *Understanding MARC bibligraphics*. O MARC bibliográfico: um guia introdutório; catalogação legível por computador. Betty Furrie: Tradução de Beatriz Valadares Cendón, Sonia Burnier, Maria Helena Santos e Natália Guiné de Mello Carvalho. Brasília: Thesaurus, 2000.

ISTITUTO Centrale per il Catalogo Unico delle Biblioteche Italiane e per le Informazioni Bibliografiche - ICCU. *Gateway Z39.50*. Disponível em <http://www.opac.sbn.it/>. Acesso em: 05 jan. 2002.

KIM, H.; CHOI, C. XML: how it will be applied to digital library systems. *The Eletronic Librart*, v. 18, n. 3, p. 188-189, 2000.

LAM, K. T. (2001). *Hong Kong University of Science and Technology*. Disponível em: <http://www.ihome.ust.hk/~lblkt/xml/marc2xml.html>. Acesso em: 02 dez. 2001.

LIBRARY of Congress. *Marc Standards*. Disponível em: <http://www.ftp.loc.gov/pub/xmldtd/marcconv_xml.zip>. Acesso em: 28 nov. 2001.

LIBRARY of Congress. *Marc Standards*. Disponível em: <http://www.loc.gov/marc/marc.html>. Acesso em: 20 nov. 2001.

LOGOS Research Systems. *MARC Record Resources*. Disponível em: <http://www.

logos.com/marc>. Acesso em: 28 nov. 2001.

MARCHAL, B. *XML by example*. Indianapolis: QUE, 2000. 504 p.

MARCONDES, C. H.; SAYÃO, L. F. Situação Brasileira dos Formatos de Intercâmbio e dos Softwares de Suporte. In: SEMINÁRIO NACIONAL DE BIBLIOTECAS UNIVERSITÁRIAS, 7, 1991, Rio de Janeiro. *Anais*... Rio de Janeiro: SIBI/UFRJ, 1992. V. 1, p. 241-255.

MCCALLUM, S. (2000). Extending Marc for bibliographic control in the web enviromment: challenges and alternatives. Disponível em: <http://lcweb.loc.gov /catdir/ bibcontrol/mccallum_paper.html>. Acesso em: 02 dez. 2002.

MILLER, D.R. (2000). *XML and MARC*: a choice or replacement? Disponível em: <http://elane.stanford.edu/laneauth/ALAChicago2000.html>. Acesso em: 04 mar. 2003.

MINISTERE de la culture et de la communication, France. Mission de la recherche et de la technologie. *BiblioML*. Disponível em: <http://www.culture.fr/BiblioML>. Acesso em: 28 nov. 2001.

OPEN Archives Initiative. Disponível em: <http://www.openarchives.org/index.html>. Acesso em: 03 dez. 2001.

PORTIA System. Disponível em: <http://ww.portia.dk>. Acesso em: 03 dez. 2001.

PRICE, G. (2002). The invisible Web. Disponível em: <http://www.freeprint.com/ gary/ili2002.htm>. Acesso em: 22 mar. 2003.

QIN, J. Representation and organization of information the web space: from marc to xml. *Special Issue on Information Science Research*, v. 3, n. 2, 2000.

SOURCE Forge.net. Disponível em: <http://marcpm.sourceforge.net/>. Acesso em: 29 nov. 2001.

STANFORD University Medical Center: *Medlane Projetc*. Disponível em: <http://xmlmarc.stanford.edu/>. Acesso em: 28 nov. 2001.

TENNANT, R. (2002). *MARC mustdie*. Disponível em: <http://libraryjournal.reviews news.com/index.asp?layout=article&articleId=CA250046&display=searchResults&stt=001&text =miller>. Acesso em: 06 mar. 2003.

THE Dialog Corporation. Disponível em: <http://www.dialogweb.com>. Acesso em: 27 nov. 2001.

UNITED States National Library of Medicine. *Bibliographic Services Division*. Disponível em: <http://www.nlm.nih.gov/bsd/licensee.html>. Acesso em: 17 dez. 2001.

UNIVERSITY of Virginia Tech Digital Library Research Laboratory. Disponível em: <http://www.dlib.vt.edu/>. Acesso em: 08 jan. 2002.

W3CONSORTIUM, *Extensible Markup Language (XML)*. Disponível em: <http://www.w3.org/XML/>. Acesso em: 15 nov. 2001.

WEINSTEIN, P. C. (1998). Ontology-based metadata: transforming the MARC legacy. In: PROCEEDINGS OF THE 3RD ACM INTERNATIONAL CONFERENCE ON DIGITAL LIBRARIES. Pittsburgh, PA, USA. ACM, June 23-26, 1998.

WELTY, C.; JENKINS, J. (2000). *Untangle*: a new ontology for card catalog systems. Disponível em: <http://untangle.cs.vassar.edu/>. Acesso em: 16 dez. 2002.

WORLD Intellectual Property Organization. Disponível em: <http://www.wipo.org.>. Acesso em 27 nov. 2001.