O Manual UNIMARC em linha

Hugo Manguinhas, José Borbinha e Nuno Freire

INESC-ID – Instituto de Engenharia de Sistemas e Computadores
Apartado 13069, 1000-029 Lisboa, Portugal
E-mail: hugo.manguinhas@ist.utl.pt, jlb@ist.utl.pt, nuno.freire@ist.utl.pt

RESUMO
O UNIMARC é uma família de formatos para representação de informação bibliográfica incluindo informação descritiva, classificação, autoridades e existências. Este artigo descreve uma iniciativa para a manutenção e publicação das descrições deste formato em diferentes línguas, versões, formatos e media. Este trabalho encontra-se integrado numa actividade pertencente a um projecto mais amplo para desenvolvimento de um "Metadata Registry" internacional para o formato UNIMARC, no âmbito do programa ICABS.

PALAVRAS-CHAVE: UNIMARC, formato, publicação, esquema, "Metadata Registry".

INTRODUÇÃO

O UNIMARC é uma família de formatos para representação de informação bibliográfica incluindo informação descritiva, classificação, autoridades e existências. O desenvolvimento dos formatos desenrola-se, no âmbito da IFLA[1], através do programa ICABS[2], herdeiro mais recente do programa tradicional UBCIM[3].

As descrições textuais dos formatos UNIMARC têm sido publicadas tradicionalmente em papel (como o caso do formato bibliográfico em língua Portuguesa), encontrando-se acessível normalmente no sítio da IFLA na Internet, em HTML, a versão de referência mais recente em língua inglesa.

A Biblioteca Nacional de Portugal, na sequência das suas responsabilidades próprias no âmbito da PORBASE e do ICABS, tem vindo a promover uma actividade para o desenvolvimento em suporte digital (com codificação XML) de uma representação formal da estrutura e das respectivas descrições dos formatos UNIMARC. Um dos objectivos desta actividade é permitir utilizar estas representações para suporte à manutenção da evolução dos formatos, permitindo formalizar a sua evolução e tornar mais flexível a publicação em diferentes línguas, formatos de publicação e media.

Estes desenvolvimentos encontram-se inseridos num objectivo mais vasto para desenvolvimento de um UNIMARC "Metadata Registry" compatível com a norma ISO 11179 [1], que se deverá tornar um ponto de referência para profissionais, organizações e sistemas de informação em seu redor.

Este artigo continua com uma genérica descrição do formato UNIMARC, seguido pela descrição dos esquemas para representação do formato desenvolvidos. O processo de publicação, que corresponde à principal parte deste trabalho é então descrito, seguido de uma apresentação relativa ao "Metadata Registry" que se encontra em fase de concepção. Por fim são apresentadas as conclusões e descrito o trabalho futuro.

UNIMARC

O principal propósito do UNIMARC consiste em facilitar a troca internacional de informação bibliográfica num formato legível por humanos entre sistemas de gestão de bibliotecas, especialmente entre agências bibliográficas nacionais. O UNIMARC pertence a uma família de outros formatos MARC, como o MARC21[4]. Estes são formato particularmente concebidos como formatos de transporte para troca de informação. Desta forma, não estipulam a forma, conteúdo, ou estrutura do registo de dados no interior de sistemas de informação individuais. Estes simplesmente providenciam recomendações na forma e conteúdo da informação quando esta é alvo de troca entre sistemas.

Como outros formatos MARC, UNIMARC é uma implementação específica do ISO2709[5], uma norma internacional que especifica a estrutura de registos contendo informação bibliográfica. Assim, cada registo neste formato deverá conter uma etiqueta de registo (um elemento textual de 24 caracteres), seguido de um número indefinido de campos. Um campo é identificado por uma etiqueta, podendo ser classificado num campo de controlo ou de dados. Campos de controlo contêm um conjunto bem definido de caracteres. Campos de dados poderão opcionalmente conter um ou mais indicadores, na forma de um carácter alfanumérico, adicionando informação sobre o conteúdo dos campos, relações entre campos, ou sobre informação necessária para a manipulação dos dados. Campos de dados, podem ser igualmente compostos por subcampos. Um subcampo é identificado por um carácter alfanumérico (código), podendo igualmente conter um conjunto de caracteres. A figura 1 mostra um modelo de dados para um formato MARC genérico.

O âmbito do UNIMARC é especificar as designações de conteúdo (etiquetas, indicadores, códigos e conteúdos) a serem utilizadas por registos bibliográficos e especificar

[1] IFLA – Internacional Federation of Library Associations and Institutions (http://www.ifla.org/)
[2] IFLA-CDNL Alliance for Bibliographic Standards (http://www.ifla.org/VI/7/icabs.htm)
[3] IFLA Universal Bibliographic Control and International MARC Core Activity (http://www.ifla.org/VI/3/ubcim.htm)
[4] http://www.loc.gov/marc/
[5] ISO 2709 – Information and documentation – Format for Information Exchange

o formato lógico e físico dos registos. Estes registos cobrem monografias, séries, materiais cartográficos, música, registos sonoros, gráficos, projecções visuais, registos de vídeo, livro antigo, recursos electrónicos, etc. O formato UNIMARC foi inicialmente publicado em 1977 sobre o título de "UNIMARC – Universal Marc Format". Em 1985 foi definitivamente adoptado pela IFLA como formato internacional para troca de registos entre agências bibliográficas nacionais e recomendado como modelo para futuros formatos nacionais baseados no MARC em países sem um formato oficial.

Entretanto, foi desenvolvido um esquema XML, intitulado MARCXML [2], para substituir o ISO2709 como solução de transporte para registos MARC. Este esquema foi inicialmente desenvolvido para o MARC21, mas tem sido utilizado naturalmente para o UNIMARC. Finalmente, o desafio de definir um esquema XML para que qualquer formato existente baseado no ISO2709 pudesse ser representado foi satisfeito pelo esquema MarcXchange [3]. A utilidade da estrutura base do MARCXML foi reconhecido, tendo sido copiado para o MarcXchange, tornado os dois formatos facilmente compatíveis. As figuras 2 e 3 mostram a representação de um registo UNIMARC utilizando respectivamente, o esquema MarcXchange e o formato ISO2709.

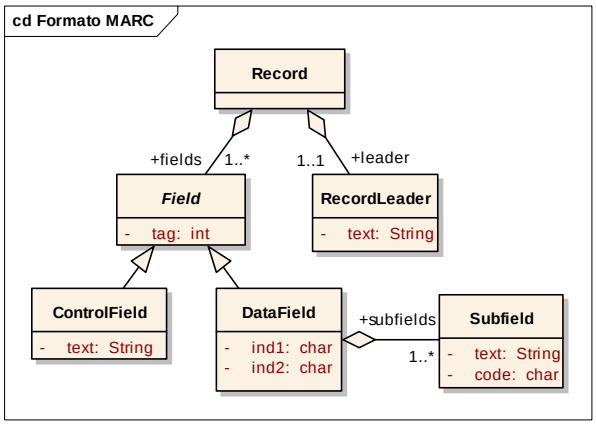

Figura 1: Modelo genérico para o formato MARC

```
<?xml version="1.0" encoding="UTF-8" ?>
- <collection xmlns="http://www.bs.dk/standards/MarcXchange"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.bs.dk/standards/MarcXchange
  http://www.bs.dk/standards/MarcXchange.xsd">
 - <record format="Unimarc" type="bibliographic">
    <leader>00604cam 02200205 04500</leader>
    <controlfield tag="001">3</controlfield>
    <controlfield tag="005">20020701015800.0</controlfield>
   - <datafield ind1="" ind2="" tag="095">
      <subfield code="a">PTBN00000003</subfield>
    </datafield>
   - <datafield ind1="" ind2="" tag="100">
      <subfield code="a">19790701d1978 m y0pora0103 ba</subfield>
    </datafield>
   - <datafield ind1="0" ind2="" tag="101">
      <subfield code="a">por</subfield>
    </datafield>
   - <datafield ind1="" ind2="" tag="102">
      <subfield code="a">PT</subfield>
    </datafield>
   - <datafield ind1="1" ind2="" tag="200">
      <subfield code="a"><O >Brasil</subfield>
      <subfield code="e">evolução no Império Português</subfield>
      <subfield code="f">Orlando Ribeiro</subfield>
    </datafield>
   - <datafield ind1="" ind2="" tag="210">
      <subfield code="a">Coimbra</subfield>
      <subfield code="c">Inst. de História Económica e Social</subfield>
      <subfield code="d">1978</subfield>
    </datafield>
```

Figura 2: Registo UNIMARC representado em MarcXchange

```
00833c   02200253
0450000100050000000500170000509500170002210000410003910100080008010200070008820001220009521000300021721500290024722500300027667500350030667500550034170000310039670200350042780100160046296600260047896600290050496600280053399800180056l-5355-19981110210300.0-  aPTBN00005492-  -a19890123d1979    m  y0pora0103    ba-1 apor-aPT-1 a<A >tradição de resistência em Moçambiquee<o >vale do Zambeze, 1850-1921fAllen F. Isaacmangcolab. Barbara Isaacman-  aPorto-cAfrontamentof1979-  a353, [3] p.cmap.d20 cm-2 a<As >armas e os varöesv9-  a325vBNzporvmed-zpor3288004-  a94(469)(045)"1806/1815"vBNzpor-vmedzpor3288146- 1aIsaacmanbAllen F.3132954- 1-aIsaacmanbBarbara40703207566- 0aPTbBNgRPC-  -lBNmMGsS.C. 30372 P.-  lDARMADmDARMADs9/B/417 -c5lULLEmULLEsHP 996 P-  aCR111 - 00444-
```

Figura 3: Registo UNIMARC codificado em formato ISO2709

REPRESENTAÇÃO DO UNIMARC

Para tornar possível a gestão dos processos de manutenção e publicação dos formatos UNIMARC foram desenvolvidos dois esquemas de dados. Um esquema para representação da informação estrutural do formato ajustada ao domínio particular do UNIMARC e não genérico do MARC. Outro esquema para representação da informação descritiva de cada parte constituinte do formato.

Sintaxe URN

Ambos os esquemas descritos nesta secção utilizam identificadores URN [4] (Uniform Resource Name) para identificação e referência. A sua fonte consiste num "Metadata Registry", que será uma importante parte do nosso trabalho (assunto aprofundado no final deste artigo). A figura 4, mostra em notação BNF, a sintaxe utilizada para construção de identificadores URN para as partes constituintes do formato.

```
<uri>       ::= "urn:" <body> ":/" <path>

<body>      ::= <format> ":" <subformat> ":"
                <version>

<format>    ::= "unimarc"

<subformat> ::= "bibliographic" | "authority" |
                "item" | "classification"

<version>   ::= <edition> | <edition> "." <revision>

<edition>   ::= <integer>

<revision>  ::= <integer>

<path>      ::= <leader> | <block> | <field> | <unit>

<leader>    ::= "leader" | "leader" "/" <element>

<block>     ::= <digit> ( <digit> "-" | "-" "-")

<field>     ::= 3*<digit> ( | "." <subfield> | "/"
                <indicator> | "/" <element>)

<indicator> ::= "i" ("1" | "2")

<subfield>  ::= <character> | <character> "/"
                <element>

<element>   ::= <integer> | <integer> "." <element>

<unit>      ::= <string> | <string> "." <character>
```

Figura 4: Sintaxe dos identificadores URN para o formato UNIMARC (codificada em notação BNF)

Esquema do UNIMARC

Este esquema tem como objectivo a representação da informação estrutural específica do formato UNIMARC, bem como, a representação de restrições ao seu conteúdo.

Foram analisadas algumas linguagens para representação de esquemas tradicionais, mas constatou-se que estas iriam tornar o seu desenvolvimento extremamente complexo e extenso. Por outro lado, os profissionais responsáveis pelas tarefas de gestão do formato não se encontram familiarizados com os conceitos intrínsecos a estas linguagens. Desta forma, optou-se por desenvolver uma linguagem capaz de uma forma simples e clara, representar os conceitos partilhados por estes profissionais (etiqueta de registo, blocos de campos, campos, subcampos, indicadores e códigos). Esta linguagem abrange todo o domínio do UNIMARC, permitindo ainda definir restrições que correlacionam diversos elementos no formato. A figura 5 mostra o esquema desenvolvido para representação dos formatos UNIMARC.

A principal motivação para o desenvolvimento deste esquema consiste em facilitar o processo de validação de registos codificados em formatos compatíveis com MARC (ISO2709, MARCXML ou MarcXchange) de acordo com a norma específica UNIMARC. Contudo, este esquema veio igualmente facilitar a tarefa de publicação. Ao estar orientado aos conceitos específicos do UNIMARC, permitiu que fosse utilizado como fonte de informação para a construção do esqueleto base do documento de publicação, oferecendo a forma e estrutura ao mesmo.

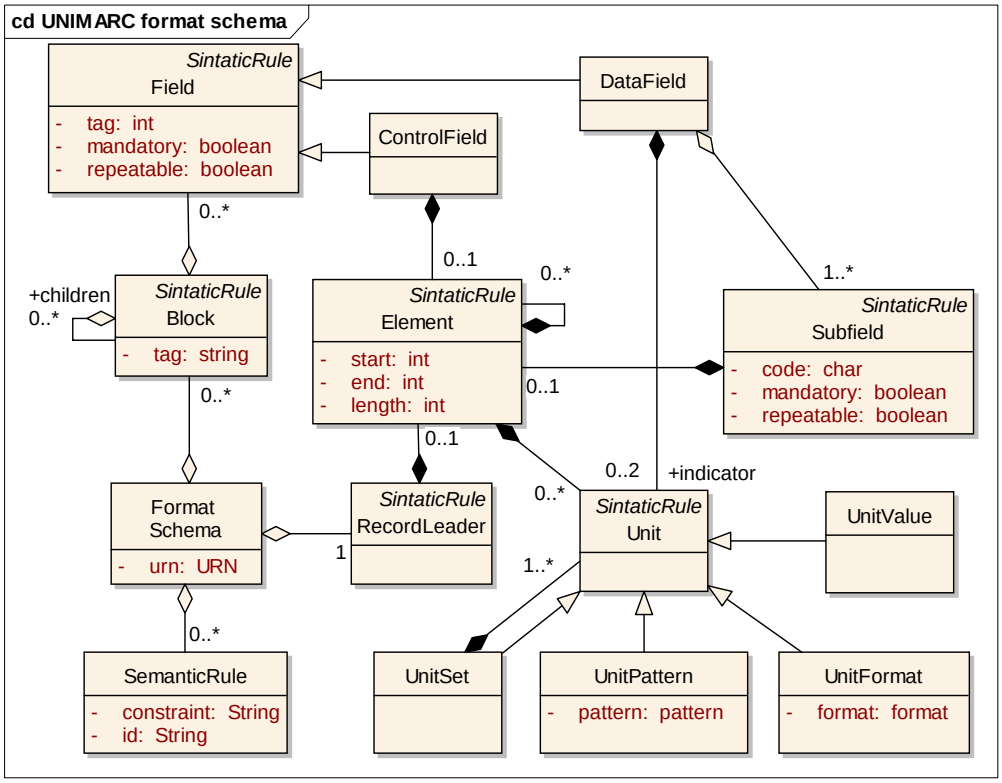

Figura 5: Modelo do esquema para o formato UNIMARC

Descrição do UNIMARC

Cada esquema UNIMARC contém um esquema complementar com as descrições dos termos (ver figura 6). Este esquema contém os mapeamentos, recorrendo aos identificadores URN, entre a informação estrutural do formato e as descrições textuais correspondentes. Cada entidade estrutural é descrita por: uma etiqueta, um conjunto limitado de caracteres que definem o nome dado ao elemento; uma definição, um conjunto ilimitado de caracteres que descreve a entidade; exemplos, um conjunto de dados que providenciam informação sobre possíveis utilizações da entidade; e relações, informação estruturada para representar ligações de proximidade com outras entidades presentes no formato.

Esta informação, combinada com a estrutura fornecida pelo esquema permitem a construção do documento base da publicação.

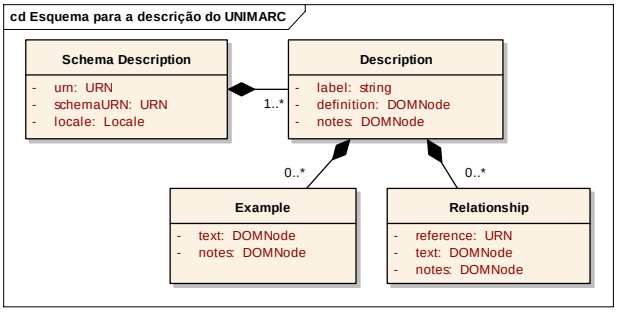

Figura 6: Modelo do esquema para o formato UNIMARC

PUBLICAÇÃO

O processo de publicação (ver figura 7) compreende um conjunto de tarefas responsáveis pela produção de um documento legível por humanos ou maquinas contendo a informação necessária para a compreensão do formato.

Para a execução destas tarefas foi desenvolvida uma infra-estrutura específica, composta por um módulo por tarefa. Todos estes módulos foram desenvolvidos recorrendo à tecnologia XSLT [5]. Esta tecnologia permite uma grande flexibilidade e facilita o acesso e utilização por outras aplicações. Por motivos de desempenho foi utilizada a tecnologia JAVA para implementação das etapas de compilação e fusão.

Esta infra-estrutura utiliza, para representação do documento de publicação, o esquema DocBook[6]. DocBook é uma norma largamente utilizada para representação de livros técnicos em XML. Esta solução facilita a execução de forma independente de cada uma das tarefas permitindo ajustar o processo de acordo com as necessidades de publicação, bem como, adicionar ou remover etapas ao processo. A existência de uma vasta colecção de ferramentas para publicação técnica de documentos DocBook facilitou ainda, a criação de diversos formatos de publicação para o UNIMARC.

Processo

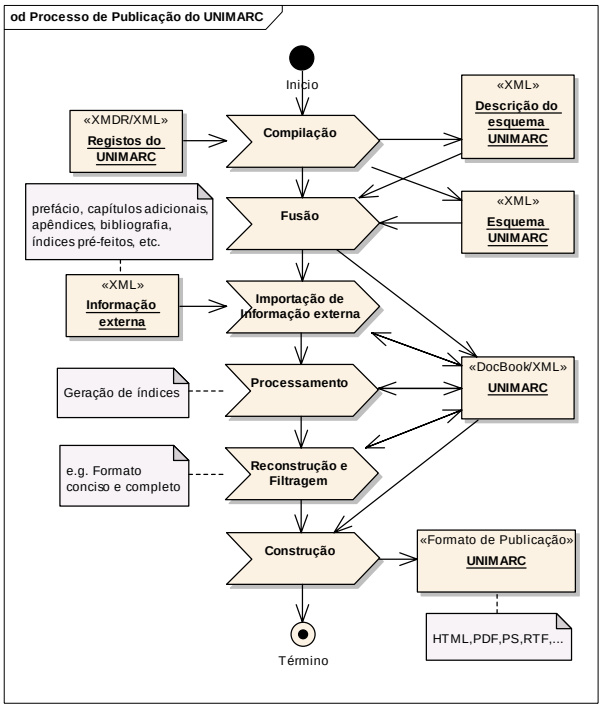

Figura 7: Processo de publicação do formato UNIMARC

O processo de publicação estabelecido é composto por 5 etapas chave. A primeira etapa corresponde à **compilação** dos 2 formatos anteriormente descritos (esquema e suas descrições). Esta etapa necessita para tal de recolher informação para a sua construção proveniente de um "Metadata Registry". Esta primeira etapa poderá ser ultrapassada se ambos os formatos já tiverem sido produzidos (caso se trate de uma publicação noutra língua do mesmo formato, ou o esquema já tenha sido produzido para fins de validação).

Após a fase de compilação, estes formatos irão ser sujeitos a uma **fusão**, na qual irá ser utilizada a informação contida no esquema para servir de esqueleto do documento, bem como, a descrição para acréscimo do conteúdo. Esta fusão é conseguida recorrendo ao sistema de URNs para procura das descrições a partir de cada elemento estrutural. A necessidade de uma prévia compilação e de uma posterior fusão deve-se à partilha dos esquemas produzidos por outras aplicações, nomeadamente aplicações utilizadas para fins de validação. A execução destas tarefas permite ainda reduzir a complexidade para o processo de fusão.

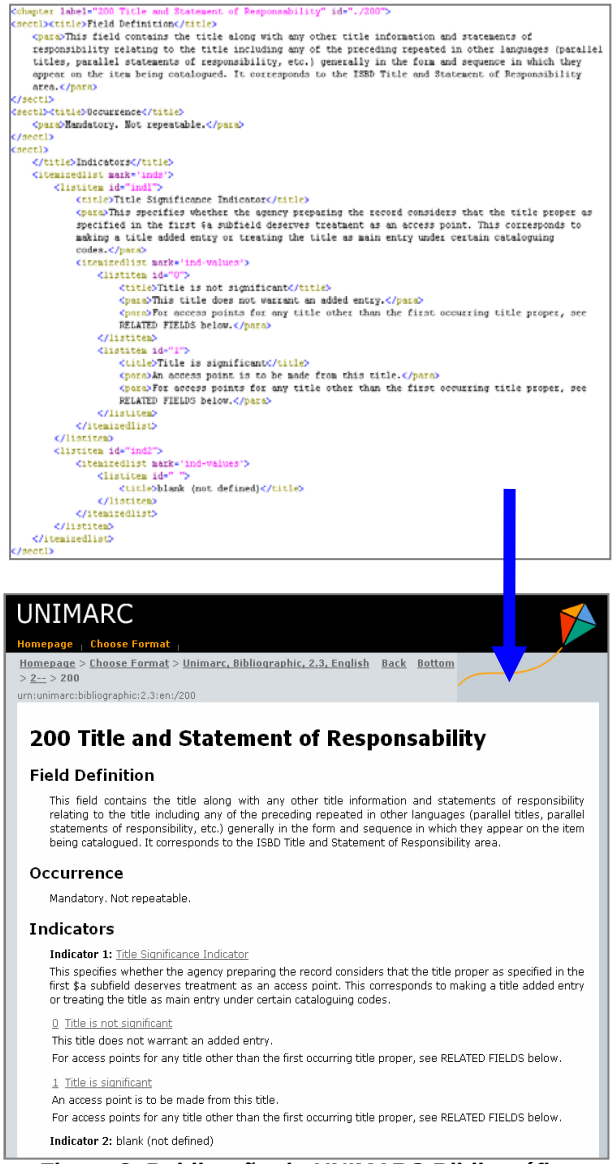

Figura 8: Publicação do UNIMARC Bibliográfico (Conversão entre os formatos de visualização DocBook/XML e HTML)

Após a execução deste processo obtêm-se o primeiro artefacto de publicação, uma primeira versão do documento em DocBook/XML que será alvo de tratamento complementar.

[6] http://www.docbook.org

A etapa de **importação de informação externa** consiste em colocar nos locais destinados, informação complementar ao documento como, informação bibliográfica (título, autoria, etc.), índices, prefácio, capítulos adicionais, apêndices, bibliografia, entre outros. Após a importação, o documento será ainda alvo de **processamento** para construção dos mais diversos índices a partir da informação adicionada no anterior processo.

Após as duas últimas etapas obtém-se o documento de publicação completo. Este será então alvo de **reconstrução e filtragem**, de forma a construir a publicação desejada. Existem até ao momento duas opções de publicação definidas, a publicação **concisa**, contendo apenas as etiquetas para os elementos do formato, removendo todas as informações complementares, como descrições, notas, exemplos, relações, apêndices, etc. (este é um formato bastante útil para consulta rápida do formato) e, **completo**, contendo toda a informação detalhada relativa ao formato.

Após todas as tarefas de construção e estruturação do documento dá-se a **construção** do documento final em formato de visualização. Este é efectuado recorrendo a transformações XSLT e XSLT-FO (XSLT Formating Objects) específicas para o formato de visualização pretendido. A figura 8 mostra um exemplo da publicação em formato de visualização HTML, a partir de DocBook/XML.

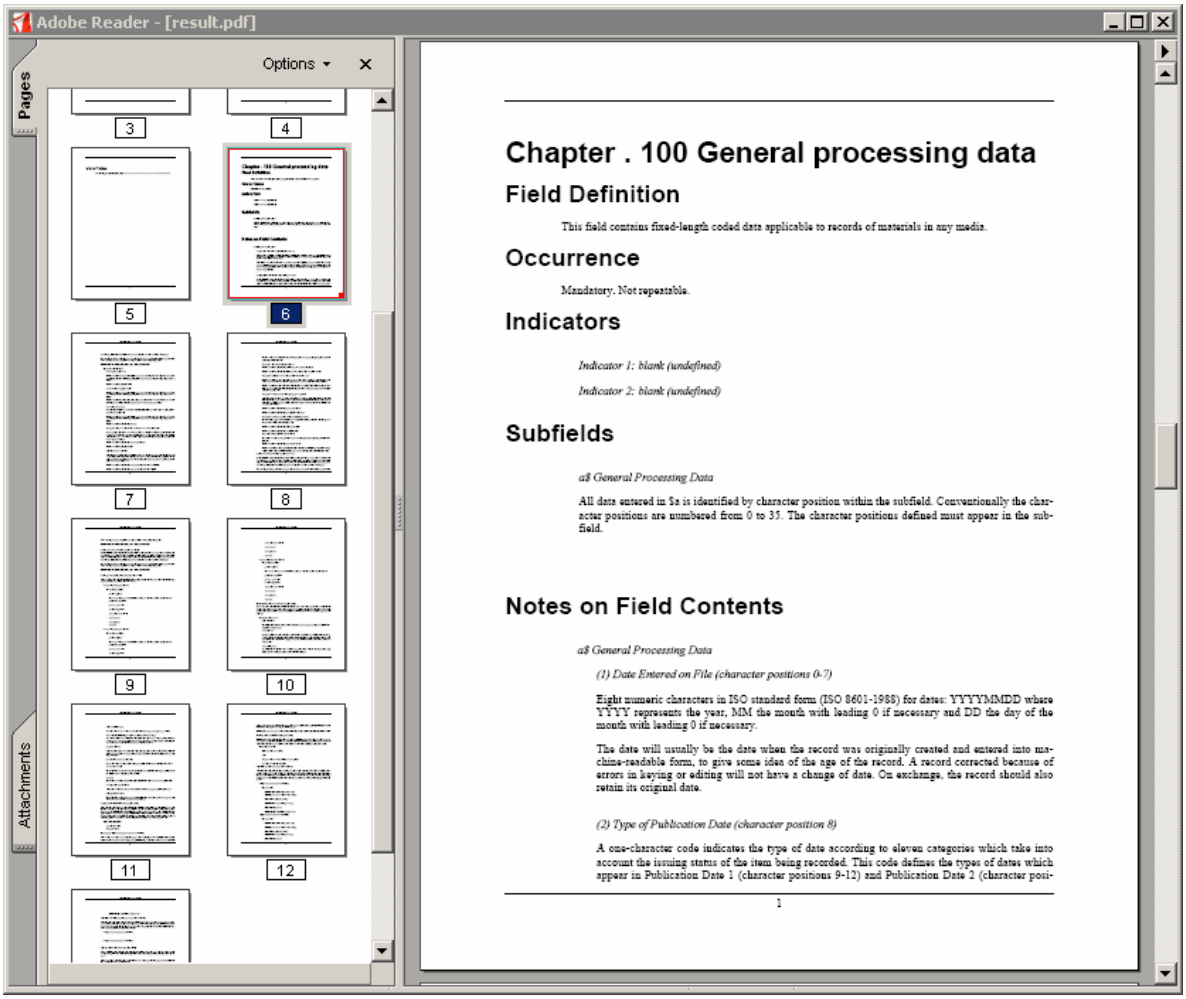

Figura 9: Publicação do UNIMARC em PDF

Formatos de Visualização

O processo de publicação permite a produção de um conjunto diversificado de formatos de visualização (ver figura 10). Este aspecto é conseguido devido à construção do formato DocBook/XML e de um conjunto de ferramentas disponíveis para processamento deste formado, permitindo a publicação em diversos formatos de visualização. Actualmente existem disponíveis ferramentas para conversão para PDF, PS e texto simples. A figura 9 mostra a publicação em PDF do UNIMARC, formato bibliográfico, 2ª edição, terceira revisão em Inglês.

No entanto, outros formatos legíveis por humanos poderão ser construídos a partir dos formatos existentes, recorrendo a outras ferramentas de conversão. Entre estes destacam-se os formatos RTF e DOC. Para formatos legíveis por maquinas apenas o DocBook/XML é produzido, no entanto, outros formatos como RDF encontram-se em fase de avaliação.

Todas as publicações estão acessíveis on-line[7]. Este sítio está reservado ao acesso às publicações do UNIMARC. Presentemente estão publicadas as versões textuais em

[7] www.unimarc.info

Português e Inglês dos formatos bibliográfico, autoridades e de existências. Existe a hipótese de vir a estender o leque disponível de publicações, a línguas provenientes de países onde o UNIMARC foi adoptado, no entanto, este aspecto requer uma avaliação mais aprofundada do seu impacto.

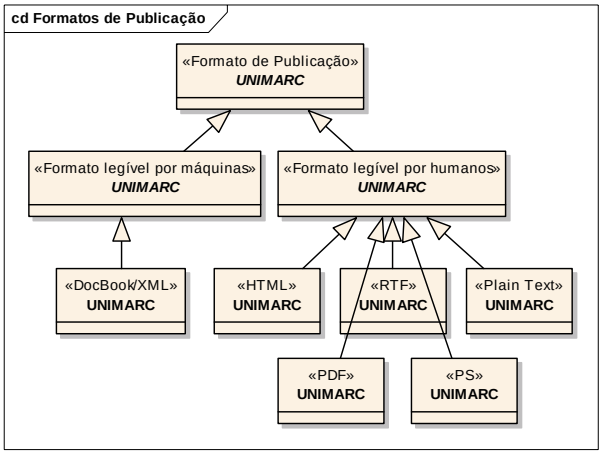

Figura 10: Resumo dos formatos de publicação contemplados para o UNIMARC

Ligação com outros serviços

O sítio de publicação referido na secção anterior utiliza uma sintaxe simples para a construção de URLs. Estes URLs são persistentes (PURL), ou seja, não alteram com actualizações e novas versões do sistema. A sintaxe utilizada provém do modelo conceptual definido para o esquema com o acréscimo da informação relativa à língua da publicação. Desta forma, esta sintaxe é em grande parte partilhada pela sintaxe utilizada para a construção dos URNs no esquema do formato. Este aspecto permite aos serviços externos, um acesso imediato e permanente às publicações disponíveis on-line. A figura 11 mostra uma ligação do sistema IRIS (sistema para integração de bases de dados cooperantes na PORBASE) para a página de publicação on-line do formato.

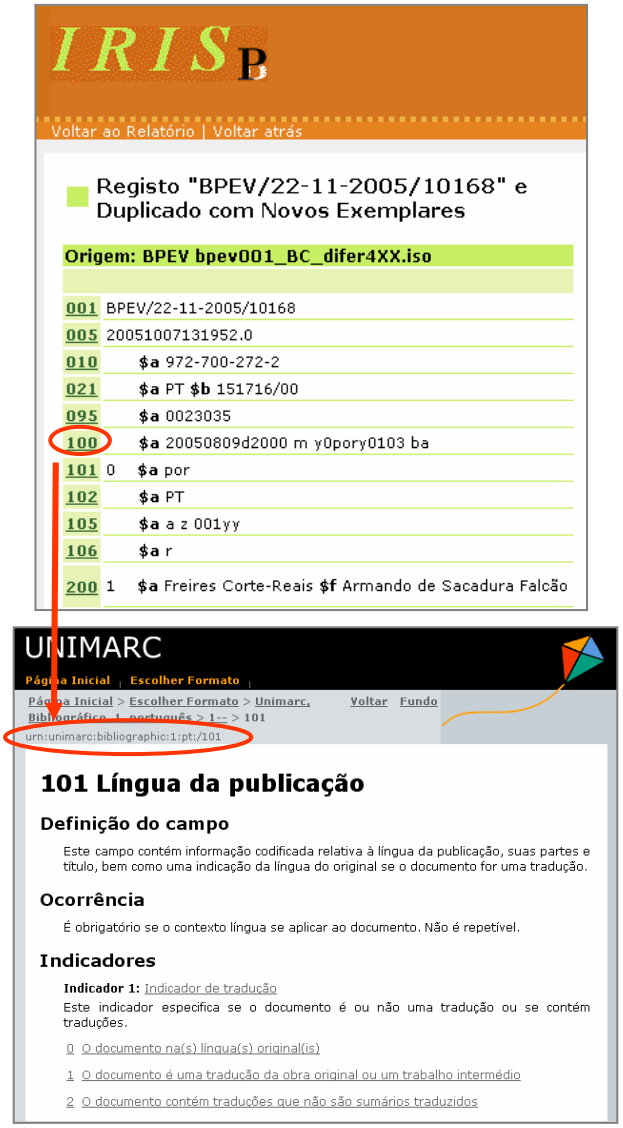

Figura 11: Ligação do Sistema IRIS à página de publicação do formato por ligação URL persistente.

UNIMARC METADATA REGISTRY

Para a gestão dos processos de registo, manutenção, evolução, acesso, descoberta e interoperabilidade do formato, encontra-se em desenvolvimento um "Metadata Registry" (MDR) de acordo com a norma ISO/IEC 11179. Parte deste sistema já se encontra construído, no entanto, ainda necessita de uma avaliação e concepção mais estruturada. Contudo, o trabalho produzido até ao momento foi essencial para, por um lado produzir os resultados anteriormente mencionados, e por outro, deu-nos a experiência necessária e elementos de análise para desempenhar esta nova tarefa de uma forma mais produtiva.

ISO/IEC 11179 é uma norma que define os conceitos por detrás dos "Metadata Registries", como a semântica da informação, a representação da informação e do registo das descrições para a informação. Este encontra-se dividido em 6 partes distintas: Infra-estrutura; Classificação; Metamodelo de um "registry" e atributos básicos; Formulação das definições da informação; Princípios de nomeação e identificação; e Registo.

A primeira parte da norma diz respeito a conceitos de desenho, implementação e gestão da infra-estrutura em geral, que se encontra fora do âmbito deste artigo. A segunda parte trata de aspectos de classificação da informação contida neste sistema e não será igualmente alvo de reflexão.

Relativamente à terceira parte, esta diz respeito ao metamodelo definido para representação de modelos específicos. O modelo a ser utilizado para o nosso caso em particular será o UNIMARC e rege-se pelos mesmos conceitos definidos para o esquema do UNIMARC definido no início deste artigo. É importante realçar que cada entidade de informação contida nesse esquema corresponde a um registo e será, desta forma, alvo de identificação, descrição, classificação, gestão e manutenção.

A quarta parte diz respeito à determinação das descrições da informação gerida, estabelecendo um conjunto de atributos utilizados para esse fim. Grande parte desses atributos é partilhada com o esquema anteriormente definido para a descrição, de forma a facilitar a sua conversão.

Para além do modelo definido pela quinta parte da norma para identificação dos registos foi adicionada uma identificação sobre a forma de um URN. Estes URNs irão ser posteriormente representados no esquema do UNIMARC. Baseados nestes identificadores, o sistema irá providenciar PURLs para acesso imediato ao registo correspondente à informação mais actual pertencente a cada elemento do esquema, tal como para as anteriores versões.

A última e sexta parte, o registo, é efectuada de acordo com determinados níveis de administração. Profissionais responsáveis pela edição e revisão do formato podem aceder ao sistema on-line e editar as alterações nas descrições do formato. Alterações às descrições são efectuadas de imediato, no mesmo registo de informação e sobrepostas à informação anteriormente registada. Usualmente existe um profissional responsável por cada língua e formato, em todas as suas versões.

Por outro lado, profissionais responsáveis pela gestão da evolução do formato podem igualmente editar on-line o esquema do formato. Todas as alterações à informação estrutural dão origem a novas versões do mesmo registo de forma a manter a preservar a evolução. Usualmente, não existe um grupo de profissionais responsáveis por todas as versões do formato, visto evolver um grande nível de interligação de responsabilidades e tarefas, diferente dos editores e revisores.

Um "Metadata Registry" serve ainda como a principal fonte para a descoberta do formato e providencia acesso a toda a informação e serviços relativos ao mesmo.

Publicação

Relativamente à publicação, este "registry" actua como a principal fonte para a construção dos esquemas necessários para o processo de publicação (e validação, não descrito neste artigo), e em segundo lugar, como fonte de descoberta, providenciando elos hiper-textuais para as publicações disponíveis on-line.

Alterações efectuadas ao formato podem ser imediatamente disponibilizadas no sítio de publicação on-line em todos os formatos de visualização disponíveis.

Este sistema partilha ainda uma outra importante particularidade, permitindo que se registe a evolução das entidades geridas e suas descrições, permitindo a comparação entre diferentes versões do formato. Toda esta informação pode igualmente ser publicada juntamente com a publicação dos formatos.

CONCLUSÕES E TRABALHO FUTURO

Publicar os formatos UNIMARC em todas as línguas cujos países tenham adoptado o adoptado como formato nacional. Este trabalho envolve a recolha de todas as descrições textuais, usualmente utilizadas para impressão, e representá-las de acordo com o modelo definido. Esta actividade encontra-se em curso, e será seguida da produção dos esquemas do formato e das descrições.

Por outra vertente, será necessário analisar a viabilidade da utilização de uma linguagem de ontologias (e.g. RDF) para representação da informação descritiva.

Encontra-se igualmente em curso o desenvolvimento de um maior número de formatos de visualização, utilizando para tal alguns dos formatos já produzidos.

De uma perspectiva mais alargada, os requisitos funcionais para o desenvolvimento de um "Metadata Registry" para o UNIMARC encontra-se em curso. A infra-estrutura utilizada para produzir os resultados reportados neste artigo foi desenhada para ser compatível com estes requisitos, mas necessita de ser revista à luz da norma ISO/IEC 11179.

NOTAS

1. ISO/IEC – ISO/IEC 11179, Information Technology – Metadata Registries (MDR). Março 2006. Disponível em WWW: http://metadata-standards.org/11179/.

2. LoC – MARC Standards, MARC in XML. Setembro 2004. Disponível em WWW: http://www.loc.gov/marc/marcxml.htmlISO.

3. ISO/IEC – ISO/DIS 25577: Information and documentation – MarcXchange. Novembro 2005. Disponível em WWW: http://www.niso.org/international/SC4/n577.pdf.

4. Moats, R. – URN Syntax. RFC 2141. Maio 1997. Disponível em WWW: http://tools.ietf.org/html/rfc2141.

5. Clark, James. XSL Transformations (XSLT) Version 1.0. Novembro 1999. Disponível em WWW: http://www.w3.org/TR/xslt.