A FALAR NOS ENTENDEMOS – A INTEROPERABILIDADE ENTRE REPOSITÓRIOS DIGITAIS

Ana Alice Baptista

INTRODUÇÃO

O Glossário da Dublin Core Metadata Initiative (DCMI) define interoperabilidade como "a capacidade de tipos diferentes de computadores, redes, sistemas operativos e aplicações trabalharem em conjunto com eficácia, sem comunicação prévia, de forma a trocarem informação de uma maneira útil e com significado" (Woodley, 2005 - tradução livre). Refere ainda o glossário que há três formas de interoperabilidade: semântica, estrutural e sintáctica.

Assim como nós, humanos, utilizamos vários idiomas para nos expressarmos, também as máquinas utilizam um variado conjunto de protocolos (linguagens de comunicação). Não nos reportando apenas à linguagem verbal, mas lembrando a linguagem gestual, e outras forma de expressão entre humanos (por exemplo, sinais de fumo ou o código morse), verificamos que as formas de expressão entre humanos se situam a vários níveis. O mesmo se passa com as máquinas: podem ser interoperáveis ao nível, por exemplo, dos protocolos de comunicação e não o ser, por exemplo, ao nível dos termos utilizados na mensagem que é transmitida. Ou seja, a roupagem é a mesma, mas os conteúdos veiculados são de natureza diferente. Seria algo semelhante a humanos tentarem comunicar verbalmente mas uns a falar italiano e outros a fa-

lar alemão. Com a excepção provável dos habitantes de alguns cantões suíços, poucos se entenderiam. Mesmo utilizando o mesmo idioma, poderá haver termos idênticos que têm significados diferentes ou, ao inverso, a referência ao mesmo conceito fazer-se utilizando termos diferentes. Por exemplo, falando todos em português, os portugueses utilizam o termo "logo" para significar "mais tarde", enquanto os brasileiros o utilizam para significar "já". Um exemplo relativo ao segundo caso é o facto de os portugueses designarem a primeira refeição da manhã como "pequeno-almoço" e os brasileiros a designarem como "café da manhã".

A inexistência, ou a falha, de interoperabilidade em qualquer um dos níveis de comunicação, compromete toda a tentativa de comunicação. O surgimento de iniciativas isoladas de interoperabilidade resulta na criação daquilo que apelido de "ilhas de interoperabilidade". Nestas ilhas existem várias máquinas que comunicam entre si e são interoperáveis, mas permanecem isoladas do resto do mundo. As máquinas, para se entenderem, necessitam de um idioma comum que possibilite a partilha não só da sintaxe e da estrutura, mas também, e isto é muito importante, do significado dos termos, ou seja, da sua semântica.

A interoperabilidade, afigurando-se uma questão meramente técnica, tem contudo grandes implicações em termos do acesso à informação disponível em repositórios, pois dela depende a capacidade de "comunicação" entre os mesmos. Se as plataformas de implementação e os dados presentes nos repositórios forem interoperáveis, as possibilidades de pesquisa simultânea entre repositórios é facilitada, permitindo maximixar o potencial dos recursos documentais arquivados individualmente em cada repositório, na medida em que se torna possível a pesquisa em simultâneo com significados partilhados nos vários repositórios, bem como a relação automática entre os resultados dessas pesquisas. A partir de uma pesquisa é possível manipular os seus resultados, agregando-os ou separando-os e expandir ou refinar pesquisas em termo semânticos, i.e. de significado.

O protocolo OAI-PMH, implementado na generalidade dos repositórios digitais, fornece uma base de interoperabilidade, mas deixa de fora as questões da interoperabilidade semântica (nem esse é o seu propósito).

Este texto aborda as questões da interoperabilidade, centrando-se na problemática, importantíssima, da interoperabilidade semântica. A segunda secção, que se segue a esta introdução, pretende clarificar alguns conceitos apresentando algumas definições que são relevantes no contexto deste texto. Na terceira secção aborda-se a temática "interoperabilidade e repositórios digitais", que informa sobre as razões e esforços de interoperabilidade a nível global (através de fronteiras de países, idiomas e tecnologias) e sobre as limitações actuais ao nível dos repositórios digitais. Na quarta secção analisam-se e comentam-se algumas das directrizes do projecto DRIVER 2.0, e termina-se com uma última secção onde se tecem considerações finais e se apresentam sugestões de trabalho futuro. Todo o texto se centra numa perspectiva fortemente influenciada pelo trabalho que tem vindo a ser desenvolvido no seio da DCMI.

CLARIFICAÇÃO DE CONCEITOS

O quadro 1 clarifica alguns conceitos utilizados no contexto deste texto. Estes são retirados ou baseados em documentos emanados da DCMI ou do World Wide Web Consortium (W3C) ou de artigos relevantes na área. Incluo a definição de interoperabilidade que apresentei no início da Introdução.

Termo / acrónimo	**Definição**
DCMES	Acrónimo de Dublin Core Metadata Element Set . Também apelidado apenas de Dublin Core, DC simples ou apenas DC. É um conjunto nuclear de propriedades (elementos de metadados) desenvolvido, mantido e recomendado pela DCMI. Além de ser um conjunto de propriedades estável desde 1996, é uma recomendação DCMI desde 1998, com a sua versão 1.0. A especificação do DCMES vai, neste momento na sua segunda versão (versão 1.1) e é endossada formalmente pelas seguintes normas (Dublin Core Metadata Initiative, 2008): * ISO Standard 15836-2003 de Maio de 2003; * ANSI/NISO Standard Z39.85-2007 de Maio de 2007; * IETF RFC 5013 de Agosto de 2007.

DC Terms	DCMI Metadata Terms. Conjunto de todos os termos de metadados mantidos pela DCMI. Inclui as 15 propriedades do DCMES (Dublin Core Metadata Initiative, 2008a).
Elemento de metadados	Um elemento de metadados é a propriedade de um recurso. É um termo definido formalmente que é utilizado "para descrever atributos ou propriedades de um recurso" (Woodley, 2005 – tradução livre). Ver também "Propriedade".
Esquemas de codificação	"Um esquema de codificação fornece informação contextual ou regras de parsing (descodificação) que ajudam na interpretação de um valor de um termo. Tal informação contextual pode tomar a forma de vocabulários controlados, notações formais ou regras de parsing" (Woodley, 2005 – tradução livre). Existem dois tipos de esquemas de codificação: esquemas de sintaxe e esquemas de vocabulários.
Esquemas de sintaxe	Os esquemas de sintaxe indicam que uma string está formatada de acordo com uma notação formal (Woodley, 2005). Por exemplo, permitem não só identificar a string "2002-05-07" como uma data, como identificar a forma como essa data deve ser processada. De acordo com o esquema de sintaxe em causa, é atribuída semântica a cada um daqueles conjuntos de dígitos.
Esquemas de vocabulários	Os esquemas de vocabulário permitem identificar um valor no contexto de um vocabulário controlado (Woodley, 2005). Por exemplo, o valor "K.4.1 Public Policy Issues " do ACM Computing Classification System disponível a partir de http://www.acm.org/about/class/1998.
Interoperabilidade	"A capacidade de tipos diferentes de computadores, redes, sistemas operativos e aplicações trabalharem em conjunto com eficácia, sem comunicação prévia, de forma a trocarem informação de uma maneira útil e com significado" (Woodley, 2005 - tradução livre).
Metadados	Metadados são dados sobre os dados (Woodley, 2005) ou informação sobre recursos. «É o termo da era da Internet para a informação que tradicionalmente os bibliotecários põem nos seus catálogos e, a maior parte das vezes refere-se a informação descritiva sobre recursos Web» (Hillmann, 2001 – tradução livre).

Propriedade	"Um aspecto específico, característica, atributo ou relação utilizadas para descrever um recurso. Os elementos de metadados Dublin Core são propriedades" (Woodley, 2005 – tradução livre).Por exemplo, title. Ver também "Elemento de Metadados". Neste texto será utilizado preferencialmente o termo "propriedade" em detrimento de "elemento de metadados".
Recurso	"Um recurso é qualquer coisa que tenha uma identidade. Exemplos comuns incluem um documento electrónico, uma imagem, um serviço (...), e uma colecção de outros recursos. Nem todos os recursos são recuperáveis pela Internet; por exemplo, seres humanos, empresas e livros numa biblioteca física podem também ser considerados recursos" (Woodley, 2005 – tradução livre).
Registo de metadados	"Uma representação sintacticamente correcta da informação descritiva (metadados) para um recurso de informação" (Woodley, 2005 – tradução livre).
Valor	Valor associado a uma propriedade aplicada à descrição de um determinado recurso. Por exemplo, o valor associado à propriedade title para a descrição deste texto seria "A Falar nos Entendemos - a Interoperabilidade entre Repositórios Digitais".[1]
Vocabulário Controlado	"Um conjunto prescrito de termos cuidadosamente definidos e utilizados de forma consistente" (Woodley, 2005 – tradução livre).

Quadro 1 - Clarificação de conceitos

INTEROPERABILIDADE E REPOSITÓRIOS DIGITAIS

Conforme facilmente se compreende, o maior ou menor grau de interoperabilidade está dependente do grau de obediência a normas. Estas podem ser definidas localmente, por tipo de aplicação, negócio, ou tendo como base qualquer outro segmento. As recomendações da DCMI e do W3C pretendem ser universais e, por isso, facilitadoras da interoperabilidade a nível global. Relativamente à DCMI, reporto-me a todas as recomendações, com especial ênfase no DCMES (DCMI, 2008), no DCTerms (DCMI, 2008a), no modelo abstracto da DCMI (Powell et al, 2007), no

enquadramento de Singapura (Nilsson, Baker e Johnston, 2008), na definição dos níveis de interoperabilidade (Nilsson, Baker e Johnston, 2009) e na definição de perfis de aplicação Dublin Core (Coyle e Baker, 2009). No que toca ao W3C, reporto-me aos trabalhos em desenvolvimento no seio da actividade da Web Semântica[1], em particular às normas relacionadas com o RDF[2], o OWL[3] (Web Ontology Language) e o SKOS[4] (Simple Knowledge Organization System). Estas recomendações baseiam muito do trabalho que tem vindo a ser desenvolvido por todo o mundo e que visa a interoperabilidade, incluindo a interoperabilidade entre repositórios digitais.

A maioria dos repositórios digitais relacionados com o movimento de Acesso Livre (AL) implementam o Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH – Lagoze, Van de Sompel, Nelson e Warner, 2002). A última versão da especificação, datada de 14 de Junho de 2002, obriga à utilização do DCMES e deixa ao critério dos desenvolvedores a utilização de outros conjuntos de propriedades. A especificação é, em geral, agnóstica relativamente a esquemas de codificação. Algumas excepções têm a ver com a execução do próprio protocolo. Por exemplo, a obrigatoriedade de utilização da norma ISO8601 expressa em UTC (Universal Coordinate Time – uma actualização do GMT – Greenwich Mean Time) para codificar informação temporal.

A necessidade da especificação semântica é conhecida há muito tempo, mas apenas agora se começa a tornar evidente para muitos dos gestores de repositórios digitais. A pergunta que se segue é pertinente: como retirar informação com significado a partir de conjuntos de dados tão diversos, que utilizam a mesma roupagem (o protocolo OAI-PMH e as propriedades Dublin Core utilizadas no seu âmbito), mas com distintos e, por vezes incompatíveis, conteúdos?

Quando, por exemplo, eliminando para já as questões relacionadas com a utilização de diferentes idiomas, associado o elemento type

---

1 Ver http://www.w3.org/2001/sw/.

2 Ver http://www.w3.org/RDF/.

3 Ver http://www.w3.org/2007/OWL/wiki/OWL_Working_Group.

4 Ver http://www.w3.org/2004/02/skos/.

do DC na base de dados de um repositório se colocado o valor "artigo científico", na de outro se coloca "artigo", na de outro se coloca "texto" e na de outro se coloca "artigo de revista", como se podem interpretar e relacionar os dados agregados provenientes destes repositórios? Outro exemplo: quando, relativamente ao elemento date, num repositório se coloca o valor "10-12-06", o que significa? 10 de Dezembro de 2006, 12 de Outubro de 2006?, 6 de Dezembro de 2010? Como interpretar este valor e relacioná-lo com o valor "12-Out-06" de outro repositório qualquer?

Trata-se de interoperabilidade semântica e a menos que sejam utilizadas regras precisas para resolver esta questão, os dados agregados dos repositórios digitais serão praticamente inúteis. Depois essas regras têm, também elas, de estar definidas e descritas não apenas em papel/PDF/Word/html (inteligível apenas por humanos), mas codificadas de uma forma ela própria interoperável (inteligível por máquinas). Nessas circunstâncias a aplicação cumprirá algumas das condições para ser compatível com o nível 4 de interoperabilidade definido pela DCMI5: interoperabilidade do perfil de conjuntos de descrições.

Existem diversas iniciativas que pretendem estabelecer algumas regras a este nível. Uma é (foi) o SWAP que, como já foi referido antes e segundo o documento "Interoperability Levels for Dublin Core Metadata", define um perfil de aplicação para os trabalhos científicos6 e atinge o nível 4 de interoperabilidade[7]. Outras iniciativas de relevo são as desenvolvidas no âmbito do projecto DRIVER e as em desenvolvi-

---

[5] A DCMI define quatro níveis distintos de interoperabilidade, conforme especificado na recomendação de Maio de 2009 intitulada "Interoperability Levels for Dublin Core Metadata" (Nilsson, Baker e Johnston, 2009). O nível mais baixo, definições partilhadas de termos, refere-se á utilização "informal" dos quinze elementos (propriedades) do DCMES. O mais alto, interoperabilidade do perfil de conjuntos de descrições, implica a conformidade com o modelo de informação e a expressão XML de restrições estruturais num conjunto de descrições definido na norma "Description Set Profiles: A constraint language for Dublin Core Application Profiles". Um exemplo de uma aplicação neste nível é a Scholarly Works (Eprints) Application Profile.

[6] Tradução de "scholarly works".

[7] Neste momento, segundo email pessoal enviado em 10 de Junho de 2010 por Talat Chaudri, investigador do UKOLN, estão a desenvolver-se esforços para tornar a documentação do SWAP menos técnica e, por isso, mais legível.

mento no âmbito do projecto OpenAIRE. Em particular, as directrizes enunciadas no âmbito do DRIVER são de extrema relevância para o que aqui se pretende advogar: a necessidade de atribuir significados comuns às propriedades utilizadas e a necessidade de definir regras e esquemas de codificação a relacionar com as propriedades no âmbito dos repositórios digitais.

Na secção seguinte apresentam-se e analisam-se algumas das directrizes do projecto DRIVER 2.0 relacionadas com a questão da interoperabilidade semântica.

DIRECTRIZES DO PROJECTO DRIVER 2.0

O projecto DRIVER (Digital Repository Infrastructure Vision for European Research), "é um projecto dinamizado por um consórcio financiado pela União Europeia (UE) que visa a constituição de uma estrutura organizacional e tecnológica para implementar uma camada de dados pan-europeia que permita o uso avançado de recursos de conteúdos na área da investigação no ensino superior" (Vanderfeesten, Summann e Slabbertje, 2008).

No âmbito da segunda iteração deste projecto (chamada de DRIVER 2.0), foram estabelecidas algumas directrizes para fornecedores de conteúdos, com o foco sobre os recursos textuais – estas estão patentes no documento "Directrizes DRIVER 2.0: Directrizes para fornecedores de conteúdos - Exposição de recursos textuais com o protocolo OAI-PMH" (op. cit), que a partir de agora será referido apenas por "directrizes". As directrizes emanadas do projecto DRIVER são aquelas a que os repositórios digitais europeus estão a obedecer. Foram pensadas cuidadosamente, fornecem bases de interoperabilidade e estão (salvo as excepções que aponto) alinhadas com o DCMES e o DCTerms. Adicionalmente, permitem vislumbrar um futuro alinhamento com outras recomendações da DCMI e do W3C.

As directrizes estão focadas em cinco aspectos: colecções, metadados, implementação do protocolo OAI-PMH, práticas recomendadas e vocabulários e semântica. Embora estes cinco aspectos sejam da maior

importância, realçam-se dois: 1) metadados (correspondente ao capítulo "Uso de metadados OAI_DC") e 2) vocabulários e semântica ("Uso de vocabulários e semântica"). O primeiro porque descreve as propriedades a serem utilizadas, a sua semântica e o contexto da sua utilização. O segundo porque descreve esquemas de codificação ou cuidados a ter em conta na definição de alguns contra-domínios das propriedades, i.e. da gama de valores que será possível associar às propriedades.

Uso de metadados OAI_DC

Relativamente ao uso de metadados (propriedades), é considerado requisito mínimo obrigatório a utilização do DCMES com a semântica definida nas directrizes. É obrigatória[8] a utilização das seguintes propriedades: title, creator, date, type, identifier e, quando aplicável, subject e description.

A propriedade description tem a semântica pré-definida da sup-propriedade abstract, i.e. no campo description deve ser colocado o resumo do documento. A propriedade date tem a semântica pré-definida da sup-propriedade published, i.e. no campo date deve ser colocada a data de publicação.

São recomendadas as propriedades publisher, format, language e rights. São opcionais as propriedades contributor, source, relation, coverage e audience. A propriedade coverage tem a semântica pré-definida da sup-propriedade period, i.e. no campo coverage deve ser colocada informação sobre as características temporais do recurso. A propriedade audience tem a semântica pré-definida da sub-propriedade educationLevel, i.e. no campo audience deve ser colocado informação sobre o nível de educação dos utilizadores do recurso.

Como requisitos mínimos as directrizes informam que nos valores associados às propriedades deve ser utilizado o Unicode e não devem

[8] Para mais informação sobre o significado do termos "obrigatório", "recomendado" e "opcional", aconselho a consulta de (Vanderfeesten, Summann e Slabbertje, 2008).

ser utilizadas linguagem de marcação HTML ou XML. Algumas recomendações adicionais que destaco são:

- Utilizar sub-propriedades (elementos de refinamento);
- Utilizar o Inglês como linguagem de descrição – neste caso parece-me que prescindir do Português não será boa opção, dado que muitos utilizadores fazem pesquisas em Português. Recomendo, antes, a utilização do atributo xml:lang para definição do idioma do valor associado à propriedade e, assim, utilizar os dois idiomas, português e inglês, repetindo as propriedades ora com valores em Português, ora com valores em inglês; Por exemplo:

```
<dc:titlexml:lang="eng">Interoperability</
dc:title>
<dc:titlexml:lang="por">Interoperabilidade</
dc:title>
```

Resta-me referir que os valores associados ao atributo xml:lang devem ser codificados utilizando a mesma norma que as directrizes recomendam para a codificação de valores associados à propriedade Language: a ISO 639-3.

- Utilizar apenas um registo de metadados para efectuar a descrição de várias manifestações de um determinado objecto (e.g., .DOC e .PDF). As directrizes, no entanto, não sugerem que elemento de metadados utilizar. Na maioria dos casos não será surpreendente que se utilize a propriedade hasFormat (refinamento de relation) para identificar outras manifestações do recurso, utilizando a propriedade identifier para identificar a versão preferencial e a propriedade format para identificar o formato da versão preferencial. Por exemplo,

```
<dc:identifier>http://meu.exemplo.pt/
meuTexto.html</dc:identifier>
```

```
<dc:format>html</dc:format>
<dcterms:hasFormat>http://meu.exemplo.pt/
meuTexto.rtf</dcterms:hasFormat>
```

É necessário ter em conta que o *namespace* dc terms deve ser convenientemente identificado.

1) Para diferentes versões de um mesmo documento, recomenda-se a utilização de registos separados e o recurso ao elemento relation. Uma alternativa, não sugerida nas directrizes mas que vai de encontro à recomendação de utilizar propriedades de granularidade superior, será a utilização das sup-propriedades de relationreplaces, isReplacedBy, hasVersion e isVersionOf. Por exemplo,

```
<dc:identifier>http://meu.exemplo.pt/
meuTexto2.html</dc:identifier>
<dcterms:replaces>http://meu.exemplo.pt/
meuTexto1.html</dcterms:replaces>
```

Como no caso anterior, é necessário ter em conta que o *namespace* dc terms deve ser convenientemente identificado.

A obediência a boas práticas na utilização de propriedades é fundamental. Em particular, deve ser respeitada a semântica que lhes está associada. No caso das propriedades recomendadas pela DCMI deve ser consultada a recomendação "DCMI Metadata Terms" (DCMI, 2008a) e verificado o significado de cada um dos termos, bem como o seu contexto de utilização. Como já referi anteriormente, as directrizes estão, em geral, alinhadas com as recomendações da DCMI. As directrizes devem ser obedecidas também nos casos em que refinam os significados das propriedades (em description, date, coverage e audience).

Para além das questões relativas à semântica das propriedades, são também bastante relevantes as questões relativas à semântica dos conteúdos, i.e. dos valores a associar às propriedades. Na subsecção seguinte apresento uma análise relativa aos esquemas de codificação mencionados nas directrizes. Uma parte dessa análise reporta, no entanto,

à questão da semântica das propriedades, ou seja, às questões que trato na secção actual. Coloquei-as ali por uma questão de melhor compreensão e de fluidez do texto.

Uso de vocabulários e semântica

A identificação (através dos URI) e consequente processamento dos esquemas de codificação (esquemas de sintaxe e esquemas de vocabulários) é a forma mais correcta e eficaz para garantir a interoperabilidade semântica de repositórios digitais. Seguindo este princípio, as directrizes recomendam a utilização de URI para identificar os *namespaces* dos esquemas de vocabulários controlados. A obtenção de melhores resultados de interoperabilidade só será possível através da utilização de esquemas de codificação comuns. Mas, mesmo neste caso, não se poderá prescindir da identificação dos esquemas de codificação.

O *namespace* criado no âmbito do DRIVER 2.0 está registado em http://info-uri.info/ e identifica-se como info:eu-repo.

Vocabulários controlados relativos ao assunto

A utilização de vocabulários comuns é uma forma escorreita de colocar todos os repositórios "a falar a mesma língua". Os vocabulários controlados identificados nas directrizes como sendo os mais utilizados pela comunidade dos repositórios digitais são:

- Library of Congress Classification (LoC)[9];
- Dewey Decimal Classification (DDC)[10];
- Universal Decimal Classification (UDC)[11];
- Library of Congress Subject Headings (LCSH)[12];

[9] Ver http://www.loc.gov/catdir/cpso/lcc.html.

[10] Ver http://www.oclc.org/dewey/.

[11] Ver http://www.udcc.org/.

[12] Ver http://www.loc.gov/cds/lcsh.html.

- Schlagwortnormdatei (SWD)[13];
- os disciplinares Mathematics Subject Classification (MSC)[14] e Medical Subject Headings (MeSH)[15];
- e outros locais.

À margem das directrizes, devo informar que os vocabulários controlados locais podem ser utilizados sem comprometer a interoperabilidade se:

- Estiverem definidos num esquema de *namespace* utilizando uma linguagem normalizada como o RDF Schema[16]OWL ou o SKOS;
- Esse esquema de *namespace* estiver devidamente identificado por um URI;
- Houver um mapeamento entre os vocabulários controlados locais e outros de âmbito global como, por exemplo, qualquer um dos identificados na lista anterior;
- Esse mapeamento estiver definido no esquema de *namespace* do vocabulário controlado local; e, por último,
- O esquema de *namespace* local estiver identificado no registo de metadados que o repositório envia ao fornecedor de serviços.

Recomendo, ainda, a inscrição do esquema de *namespace* local num registo apropriado como o utilizado pelo DRIVER 2.0.

Vocabulário controlado relativo ao tipo de documento

O DRIVER 2.0 criou, com base no trabalho previamente desenvolvido noutros projectos, um vocabulário controlado específico para a propriedade type. Os seus termos são: article, bachelorThesis, masterThe-

[13] Ver http://www.d-nb.de/standardisierung/normdateien/swd.htm.

[14] Ver http://www.ams.org/mathscinet/msc/msc2010.html.

[15] Ver http://www.nlm.nih.gov/mesh/.

[16] Ver http://www.w3.org/TR/rdf-schema/.

sis, doctoralThesis, book, bookPart, review, conferenceObject, lecture, workingPaper, preprint, report, annotation, contributionToPeriodical, patent e other.

O URI http://purl.org/info:eu-repo/semantics resulta num redireccionamento para um ficheiro OWL/RDF[17] que se encontra associado ao URI http://wiki.surffoundation.nl/download/attachments/852421/info-eu-repo.rdf. Aqui estão definidos, numa linguagem interpretável por máquinas, todos os termos deste vocabulário controlado. Repare-se na relação estabelecida de todos os termos com o termo document do esquema FOAF[18]. Ao serem estabelecidos mapeamentos deste género, está a potenciar-se o aumento de interoperabilidade.

Vocabulário controlado relativo à versão

O vocabulário controlado relativo à versão é mais curto que o relativo ao tipo e surgiu para que a informação sobre o tipo do documento não estivesse misturada com a informação sobre a versão. Os termos deste vocabulário controlado são: draft, submittedVersion, acceptedVersion, publishedVersion e updatedVersion.

As directrizes apontam para uma associação deste vocabulário controlado com o elemento type do DC, tal como o vocabulário relativo ao tipo. Assim, as directrizes recomendam que, sempre que aplicável, se proceda a uma repetição do elemento type indicando a versão do documento, como no exemplo seguinte (valor para o tipo como conteúdo do primeiro elemento dc:type e valor para a versão como conteúdo do segundo elemento dc:type):

```
<dc:type>info:eu.repo/semantics/
doctoralThesis</dc:type>
<dc:type>info:eu.repo/semantics/
acceptedVersion</dc:type>
```

[17] Ver http://www.w3.org/TR/owl2-overview/#Syntaxes.

[18] Ver http://www.foaf-project.org/.

Apesar de reconhecer que esta é uma solução engenhosa para resolver a questão das versões, parece-me melhor solução a criação de uma nova propriedade específica. Advogo esta solução mais uma vez por questão de semântica. O elemento type do DC está definido como "A natureza ou género do recurso" (tradução livre – DCMI, 2008a). Ora, uma versão não diz respeito nem à natureza, nem ao género do recurso. Não duvido que esta solução funcione no âmbito restrito dos repositórios digitais mas, quando estes pretenderem ser interoperáveis com os restantes serviços e aplicações da Web, haverá aqui um desvio semântico desnecessário.

Adicionalmente, considero o termo "estado" (*state*, em Inglês) mais apropriado do que o termo "versão" para designar este conceito, uma vez que a versão do documento pode ser a mesma se este não tiver sofrido alterações, apesar de, por exemplo, poder ter mudado do estado "submetido" para o estado "aceite". Contudo, e apesar das minhas reservas, por uma questão de coerência com o texto das directrizes e por questões de legibilidade, utilizei e utilizarei neste texto o termo "versão".

Controlo de Autoridade

Um dos problemas das bibliotecas e repositórios digitais é a identificação unívoca de cada autor quer a nível local, quer a nível global. As directrizes recomendam a criação de listas dinâmicas de publicações por autor através da criação de DAI[19] (Digital Author Identifier), utilizando o formato ISNI[20] (International Standard Name Identifier). Os DAI devem ser persistentes de modo a não criar incongruências nos dados agregados. As directrizes são claras: "é da exclusiva responsabilidade de cada RI [Repositório Institucional] garantir que um autor pode ser identificado através de um DAI e que cada DAI atribuído é único num repositório institucional".

[19] Ver http://www.surffoundation.nl/en/themas/openonderzoek/infrastructuur/Pages/digitalauthoridentifierdai.aspx.

[20] Ver http://www.isni.org/.

Sintaxe e esquemas de sintaxe

Conforme acontece com os vocabulários controlados, é essencial o esquema de sintaxe estar adequadamente identificado. É o caso da data que mencionei anteriormente: como interpretar o valor "2002-06-11"? Tem de existir alguma especificação que forneça informação sobre como fazer esta interpretação. A sua identificação faz-se através de um URI.

Os esquemas de sintaxe recomendados são os apresentados no quadro 2.

Propriedade	**Sintaxe ou Esquema de sintaxe**
Title	Título:Subtítulo (sem espaços)
Creator, contributor	Sintaxe: apelido, iniciais (primeiro nome); opcionalmente utilizar DAI. APA Style[2].
Date	Sintaxe: AAAA-MM-DD ISO 8601 W3C DTF[3]
Format	MIME Types[4]
Identifier, Relation	URI[5], URN, handle[6], DOI[7]
Source	Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata[8]
Language	ISO 639-3[9]
Coverage	DCMI Period[10], Getty Thesaurus of Geographic Names[11], ISO 3166[12], DCMI Box[13].
Rights	URI (para identificar licenças Creative Commons[14]), DAI ou ISNI (para identificar pessoas ou organizações a relacionar com os direitos).

Quadro 2 - Relação entre propriedades, sintaxe e esquemas de sintaxe recomendados.

Como aconteceu relativamente ao relacionamento do elemento type com a versão do documento, coloco também reservas ao relacionamento do elemento source com a citação bibliográfica por questões de interoperabilidade futura com outros tipos de serviços/aplicações. A definição do elemento source é "um recurso relacionado a partir do qual

o recurso descrito derivou" (tradução livre – DCMI, 2008a). Ora, existe uma sub-propriedade de identifier (e não de source), denominada de bibliographicCitation cuja definição é "uma referencia bibliográfica do recurso" (DCMI, 2008a). Não sendo utilizadas as sub-propriedades do DC, devem ser utilizadas as propriedades que lhes estão associadas. E a propriedade associada à sub-propriedade bibliographicCitation é identifier e não source.

Seja qual for a opção dos implementadores relativamente a esta questão e à da associação da propriedade type com o estado do documento, é muito importante que as restantes directrizes sejam obedecidas fielmente. Conforme informei antes, elas estão perfeitamente alinhadas com recomendações da DCMI e permitem vislumbrar algum alinhamento futuro com outras recomendações da DCMI e do W3C. Estas directrizes restringem a liberdade na utilização das propriedades e dos valores a associar-lhes mas, por outro lado, potenciam a interoperabilidade semântica entre repositórios digitais que, como já referi, é de vital importância para o tratamento de dados agregados.

CONSIDERAÇÕES FINAIS E TRABALHO FUTURO

Este texto trata das questões da interoperabilidade, em particular da interoperabilidade semântica entre repositórios digitais, ou seja, do significado das propriedades (elementos de metadados) e dos valores com elas relacionados. A interoperabilidade entre repositórios digitais não se resume à utilização do protocolo OAI-PMH: este fornece apenas um nível de interoperabilidade de base.

Para que se obtenha interoperabilidade semântica, é necessário estabelecer e obedecer a regras apropriadas. Tanto a DCMI, como o W3C têm vindo a desenvolver trabalho que visa o estabelecimento de algumas destas regras. O projecto DRIVER 2.0, baseado nestas iniciativas, entre outras, estabeleceu um conjunto de directrizes para fornecedores de conteúdos. Este texto analisa e discute algumas dessas directrizes que estão directamente relacionadas com as questões de interoperabili-

dade semântica, tanto a nível das propriedades, como a nível dos valores com elas relacionados.

Estas directrizes são muito importantes porque, ao serem implementadas, proporcionam um nível de interoperabilidade que permite aos fornecedores de serviços trabalhar com eficácia os dados agregados provenientes dos diversos repositórios. Isto significa, basicamente, que a qualidade dos serviços baseados em repositórios digitais prestados à comunidade científica tem potencialidades para melhorar bastante.

Contudo, existem alguns pontos de melhoria. Começo por lembrar o potencial comprometimento de interoperabilidade a nível global (fora do ambiente dos repositórios digitais) por causa da utilização das propriedades `type` e `source` para fins diferentes dos advogados pela DCMI. Outra questão que me levanta dúvidas é a recomendação para utilização do Inglês como idioma das descrições, já que tal opção compromete as pesquisas feitas em Português. Recomendo, por isso, a criação de registos bilingues através da utilização do atributo `xml:lang` e da duplicação das propriedades. Sugiro ainda a utilização de algumas propriedades nos casos em que as directrizes são omissas.

Como trabalho futuro, proponho o estabelecimento de directrizes semelhantes para as sub-propriedades e valores com elas relacionados. Outros trabalhos interessantes são a criação de mapeamentos directos para descrições em RDF (é possível porque o identificador do recurso é obrigatório), a criação de perfis de aplicação locais, o mapeamento entre vocabulários controlados locais e outros globais, entre outros. Na verdade, muito ainda há a fazer nesta área. Contudo, seja o que for que se faça, é necessário ter em atenção que é muito fácil perder a interoperabilidade. Para que esta se mantenha é necessário obedecer às normas apropriadas e relacionar sempre os novos desenvolvimentos com os mais antigos da comunidade. Enfim, o que é necessário resume-se a uma expressão bem conhecida: pensar globalmente, agir localmente.

REFERÊNCIAS

Coyle, K. & Baker, T. (2009, Maio 18). *Guidelines for Dublin Core Application Profiles*. Obtido em Junho 9, 2010, de http://dublincore.org/documents/2009/05/18/profile-guidelines/

Dublin Core Metadata Initiative. (2008, Janeiro 14). *Dublin Core Metadata Element Set, Version 1.1*. Recomendação da DCMI. Obtido em Junho 9, 2010, de http://dublincore.org/documents/dces/.

Dublin Core Metadata Initiative. (2008a, Janeiro 14). *DCMI Metadata Terms*. Recomendação da DCMI. Obtido em Abril 13, 2009, dehttp://dublincore.org/documents/dcmi-terms/.

Hillmann, D. (2001, Abril 12). *Using Dublin Core*. Obtido em Junho 9, 2010, de http://www.dublincore.org/documents/2001/04/12/usageguide/.

Lagoze, C., Van de Sompel, H., Nelson, M. & Warner, S. (2002, Junho 14). *The Open Archives Initiative Protocol for Metadata Harvesting* (Specification No. 2.0). Obtido de http://www.openarchives.org/OAI/openarchivesprotocol.html#MetadataNamespaces

Nilsson, M., Baker, T. & Johnston, P. (2008, Janeiro 14). *The Singapore Framework for Dublin Core Application Profiles*. Recomendação da DCMI. Obtido em Junho 9, 2010, de http://dublincore.org/documents/singapore-framework/

Nilsson, M., Baker, T. & Johnston, P. (2009, Maio 1). *Interoperability Levels for Dublin Core Metadata*. Recomendação da DCMI. Obtido em Junho 9, 2010, de http://dublincore.org/documents/interoperability-levels/

Powell, A., Nilsson, M., Naeve, A., Johnston, P. & Baker, T. (2007, Junho 4). *DCMI Abstract Model*. DCMI. Recomendação da DCMI. Obtido em

Junho 9, 2010, http://dublincore.org/documents/2007/06/04/abstract-model/

Vanderfeesten, M., Summann, F. & Slabbertje. M. (2008). *Directrizes DRIVER 2.0: Directrizes para fornecedores de conteúdos - Exposição de recursos textuais com o protocolo OAI-PMH* (p. 144). Obtido em Junho 9, 2010, de http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_PT.pdf

Woodley, M. S. (2005, Novembro 7). *DCMI Glossary*. Obtido em Junho 9, 2010, de http://dublincore.org/documents/usageguide/glossary.shtml.