Análise do Europeana Data Model no Contexto das Bibliotecas e de Conteúdos de Texto Integral

Nuno Freire[1], Valentine Charles[1,2], Sally Chambers[1]

[1]The European Library, [2]Europeana Foundation

Willem-Alexanderhof 5, 2509 LK The Hague, Netherlands

Tel: (+31)703140310

E-mail: nfreire@gmail.com, valentine.charles@kb.nl, sally.chambers@kb.nl

RESUMO

A Europeana oferece acesso a conteúdos digitais provenientes de bibliotecas, museus, arquivos e coleções áudio-visuais de toda a Europa. Uma das linhas de trabalho atuais da Europeana consiste em pôr em prática o Europeana Data Model (EDM), o qual consiste numa infraestrutura aberta para representação de dados, que se baseia na web semântica, e que permite representar as necessidades de informação dos vários domínios das organizações da herança cultural. O projeto Europeana Libraries, aborda especificamente a utilização do EDM para as necessidades de informação do domínio das bibliotecas. Este artigo apresenta alguns do resultados intermédios do projeto relativos ao alinhamento de formatos de dados frequentemente utilizados em bibliotecas, como UNIMARC, MARC21, MODS, METS e perfis de metadados baseados no Dublin Core. Um dos primeiros resultados do Europeana Libraries aborda algumas questões relevantes para o domínio das bibliotecas, bem como outros domínios interessados na criação de perfis de metadados para o EDM. Mais concretamente, este artigo aborda como o EDM pode ser especializado para atingir as necessidades da comunidade das bibliotecas. O artigo apresenta alguns detalhes das soluções encontradas para criar um modelo de dados, baseado no EDM, para monografias e publicações em série, e em particular realçará como as diferenças na estruturação dos dados destes materiais pode ser abordada. Questões relevantes que surgiram no decorrer deste trabalho e as soluções adotadas são descritas. São ainda mencionadas algumas questões que se levantaram durante a tentativa de especializar o EDM para o domínio das bibliotecas. O artigo foca também algumas questões relacionadas um tipo específico de conteúdos digitais: obras de texto integral. São descritos os resultados da análise de requisitos para a troca de dados num cenário de agregação de conteúdos com texto integral, baseada em EDM. Sendo este um passo essencial para o estabelecimento de normas para texto integral no contexto da Europeana.

PALAVRAS-CHAVE: Bibliotecas; Metadados; Europeana Data Model; Texto Integral.

INTRODUÇÃO

A Europeana (http://www.europeana.eu) oferece acesso a conteúdos digitais provenientes de bibliotecas, museus, arquivos e coleções áudio-visuais de toda a Europa. Uma das linhas de trabalho atuais da Europeana consiste em pôr em prática o Europeana Data Model, ou EDM (Europeana v1.0, 2012), o qual consiste numa infraestrutura aberta para representação de dados, que se baseia na web semântica, e que permite representar as necessidades de informação dos vários domínios das organizações da herança cultural. O EDM visa suportar as necessidades de informação do vários processos ligados à Europeana, como a ingestão de dados, a gestão de conteúdos digitais e da sua proveniência, a publicação para o utilizador final e também a publicação como recursos de dados abertos.

O EDM sucede ao Europena Semantic Elements, ou ESE (Europeana, 2012), que foi a solução original da Europena para atingir um nível mínimo de interoperabilidade para trocas de metadados sobre conteúdos digitais, baseado em Dublin Core. Comparativamente ao ESE, o EDM visa melhorar a preservação da riqueza dos dados (que muitas vezes existe nos formatos originais como os formatos MARC, EAD, etc.) a quando da troca com a Europeana, sem prejudicar a interoperabilidade. Para atingir este objetivo, o EDM faz uso das normas de representação de dados da web semântica.

O projeto Europeana Libraries (http://www.europeana-libraries.eu) inclui, entre outras atividades, a análise das necessidades de informação relativas aos conteúdos digitais oriundos de bibliotecas, no contexto da Europeana. Este projeto é coordenado pela The European Library (http://www.theeuropeanlibrary.org) e envolve a participação da Europeana e de mais duas importantes associações de bibliotecas: Association of European Research Libraries (LIBER - http://www.libereurope.eu) e Consortium of European Resarch Libraries (CERL - http://www.cerl.org). Até ao final de 2012, o projeto irá agregar, para a Europeana, metadados oriundos de 19 bibliotecas académicas e estabelecer uma infraestrutura que incorpore os recursos combinados das bibliotecas académicas europeias.

No projeto Europeana Libraries, uma das atividades em

curso aborda especificamente a utilização do EDM para as necessidades de informação do domínio das bibliotecas. Este artigo apresenta alguns do resultados intermédios do projeto relativos ao alinhamento de formatos de dados frequentemente utilizados em bibliotecas, como UNIMARC, MARC21, MODS, METS e perfis de metadados baseados no Dublin Core. Um dos primeiros resultados do Europeana Libraries aborda algumas questões relevantes para o domínio das bibliotecas, bem como outros domínios interessados na criação de perfis de metadados para o EDM.

Mais concretamente, este artigo aborda como o EDM pode ser especializado para atingir as necessidades da comunidade das bibliotecas. O artigo apresenta alguns detalhes das soluções encontradas para criar um modelo de dados, baseado no EDM, para monografias e publicações em série, e em particular realçará como as diferenças na estruturação dos dados destes materiais pode ser abordada. Questões relevantes que surgiram no decorrer deste trabalho e as soluções adotadas são descritas. São ainda mencionadas algumas questões que se levantaram durante a tentativa de especializar o EDM para o domínio das bibliotecas.

Duas questões relevantes desta analise relacionam-se com a forma de suportar a riqueza dos dados bibliográficos criados em bibliotecas no EDM. Por um lado, o detalhe dos dados em formatos MARC não é suportado em classes e atributos definidos pelo EDM, e por outro lado, a utilização de linguagem natural em alguns elementos de dados bibliográficos são difíceis de transpor para uma representação mais estruturada da web semântica e do EDM.

O artigo foca também algumas questões relacionadas um tipo específico de conteúdos digitais: obras de texto integral. A The European Library mantém um índice centralizado de texto integral contendo mais de 24 milhões de páginas de texto. Várias bibliotecas nacionais europeias disponibilizaram estes conteúdos durante o projeto TELplus[8]. Neste projeto, processos de reconhecimento ótico de caracteres foram aplicados a obras previamente digitalizadas, e o texto resultante foi recolhido e indexado por forma a melhorar a recuperação de informação no portal da The European Library.

Este resultado do projeto TELplus, resultou na primeira experiência prática de criação de um índice centralizado de texto integral, oriundo de várias bibliotecas nacionais. Partindo desta primeira experiência, no projeto Europena Libraries está-se a construir uma infraestrutura sustentável para a agregação de conteúdos com texto integral, e estabelecer normas para interoperabilidade destes conteúdos.

Neste artigo descrevemos os resultados da análise de requisitos para a troca de dados num cenário de agregação de conteúdos com texto integral. Sendo este um passo essencial para o estabelecimento de normas para texto integral no contexto da Europeana. Mais especificamente, o artigo aborda os requisitos para a troca de metadados sobre estes conteúdos entre as instituições detentoras, instituições agregadoras, e a Europeana.

A atual infraestrutura de interoperabilidade da Europeana baseia-se na troca de metadados sobre conteúdos digitais de vários tipos. As especificidades dos conteúdos de texto integral, bem como as várias formas de representação e disponibilização deste tipo de conteúdo, devem ser especificamente incorporados no EDM. Neste caso, um foco maior é dado às questões dos metadados estruturais, em particular na forma como o EDM pode ser utilizado de forma a permitir um serviço de pesquisa na Europeana tirando partido dos conteúdos digitais que têm disponível o texto integral em formatos digitais legíveis para computadores.

Este artigo começará por descrever a metodologia do grupo de trabalho, e introduzirá as principais classes do EDM. Serão depois apresentados os resultados da análise do EDM em monografias, e publicações periódicas. Serão depois apresentados alguns resultados para trabalho futuro no EDM incluindo a análise de objetos de texto integral.

METODOLOGIA

Um grupo de discussão foi formado no projeto Europeana Libraries com o objetivo de analisar a utilização do EDM para representação de metadados relativos aos tipos de objetos digitais mais comuns nos acervos das bibliotecas.

Este grupo foi constituído por um representante de cada uma das bibliotecas participantes no projeto Europeana Libraries, e também por representantes de bibliotecas nacionais que fazem parte do The European Library Metadata Working Group. Este grupo permitiu juntar um grupo de especialistas com experiência em normas de metadados, familiarizados com as questões relativas à agregação de metadados para a Europeana e com a diversidade de recursos existente em bibliotecas.

Este trabalho começou por abordar as normas de metadados mais comuns em bibliotecas, e avaliou como poderiam ser mapeados para EDM, baseando-se em casos concretos de metadados de objetos digitais das bibliotecas participantes. Foram analisados registos em UNIMARC, MARC21, MODS e outros formatos baseados em Dublin Core.

Um dos pontos identificados nesta primeira análise foi que a representação de todas as entidades do primeiro nível dos FRBR - Requisitos Funcionais para Registos Bibliográficos (Obra, Expressão, Manifestação e Item) não poderiam ser representadas na versão atual do EDM. Um registo EDM tipicamente representa uma manifestação ou um item, sendo difícil a separação entre os dados referentes às várias entidades FRBR.

Este trabalho avançou com a definição de como vários tipos de objetos das bibliotecas podem ser representados em EDM, e com identificação de diferenças entre os vários tipos de objetos. Este artigo descreve os resultados deste trabalho para monografias, obras de vários volumes, e publicações periódicas. Outra característica dos objetos das biliotecas, com implicações para a representação dos metadados em EDM, é se o objeto é nascido-digital ou se é uma digitalização de um objeto físico.

Como resultado, um perfil EDM para bibliotecas foi definido tendo em vista a infraestrutura de agregação de

metadados para a Europeana, em desenvolvimento no projeto Europeana Libraries. Este modelo cumpre as especificações do EDM e os requisitos de mapeamento de metadados em desenvolvimento para os vários formatos de metadaos em uso nas bibliotecas parceiras do projeto, tentando assegurar um mapeamento consistente entre formatos de metadados.

Foram definidos os seguintes requisitos para o perfil EDM para bibliotecas, no contexto do projeto Europeana Libraries:

- A The European Library é um agregador do sector das bibliotecas para a Europeana. O perfil EDM para bibliotecas deve suportar as especificidades das bibliotecas em tipos de materiais e formatos de dados.
- O projeto Europeana Libraries agrega metadados descrevendo objetos digitais para a Europeana. Assim, este perfil EDM irá utilizar classes e propriedades descritas nas especificações EDM v5.2.2.
- Os metadados serão utilizados num contexto de dados inter-ligados (*linked data*). Assim, é utilizado um modelo compatível com a Resource Description Framework (RDF) da web semântica.
- Os dados sobre os objetos digitais serão interligados entre subdomínios das bibliotecas, bem como outros domínios. Assim, o perfil EDM deve utilizar propriedades suportando a utilização de URIs (Uniform Resource Identifiers).
- O perfil deverá ser extensível e flexível para acomodar outros tipos de objetos digitais no futuro.

AS ENTIDADES BÁSICAS

Esta secção apresenta as entidades básicas da primeira implementação do EDM, as quais são utilizadas nos modelos apresentados neste artigo.

edm:ProvidedCHO

Um edm:ProvidedCHO é definido no EDM como "o objeto de herança cultural sobre o qual a Europeana recolhe descrições. Esta classe tem um propósito funcional, não indicando nada sobre a natureza do objeto digital. Pode representar qualquer objeto que possa aparecer como um item individual numa lista de resultados de uma pesquisa no portal da Europeana.
Mesmo dentro do domínio das bibliotecas um ProvidedCHO pode representar uma edição de uma obra, ou um item especifico, como por exemplo um item de particular interesse por estar anotado.

edm:WebResource

O EDM define a classe edm:WebResource como "recursos de informação que têm pelo menos uma representação em linha e pelo menos um URI".

Um edm:WebResource representa uma versão digital de um objeto de herança cultural. Aplica-se tanto a obras nascidas digitais como a obras digitalizadas.

ore:Aggregation

Esta entidade é definida como "um conjunto de recursos relacionados, agrupados de forma a que possam ser tratados como um único recurso". Esta entidade encontra-se descrita no infraestrutura de interoperabilidade ORE por um Resource Map." (ver EDM v.5.2.2, página 6)

O EDM permite representar as relações fundamentais entre edm:WebResource e edm:ProvidedCHO, através da classe ore:Aggregation. Estas relações são:

edm:aggregatedCHO – relação estabelecida entre a ore:Aggregation e a edm:ProvidedCHO.

edm:hasView – relação estabelecida entre a ore:Aggregation e a edm:WebResource.

O MODELO PARA MONOGRAFIAS

O modelo para monografias encontra-se representado na Figura 1. Este modelo, embora simples, é bastante flexível e tem as seguintes vantagens:

- Pode ser utilizado tanto em obras nascidas digitais como obras digitalizadas.
- Pode ser facilmente expandido com outras entidades, como as classes não-informacionais edm:Event, edm:Agent, edm:Place, edm:TimeSpan e skos:Concept.
- Pode descrever estruturas hierárquicas complexas e obras de vários volumes, através da ligação destas entidades básicas por meio de um conjunto de propriedades criadas para exprimir relações estruturais (edm:isNextInSequence e dcterms:isPartOf) de acordo com a estrutura do objeto a representar.

As vantagens acima descritas podem ser visualizadas num exemplo de aplicação deste modelo básico numa obra de vários volumes, como se pode observar na Figura 2.

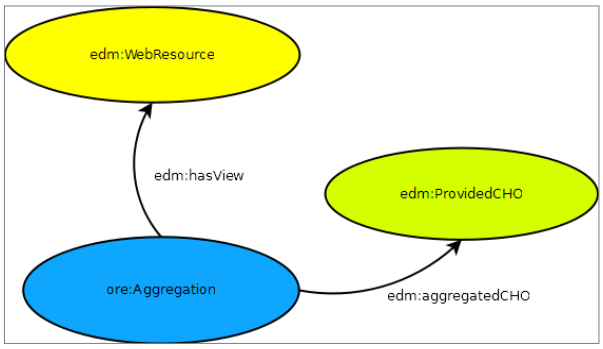

Figura 1 - Modelo definido no projeto Europeana Libraries para monografias (fonte: ANGJELI, 2011)

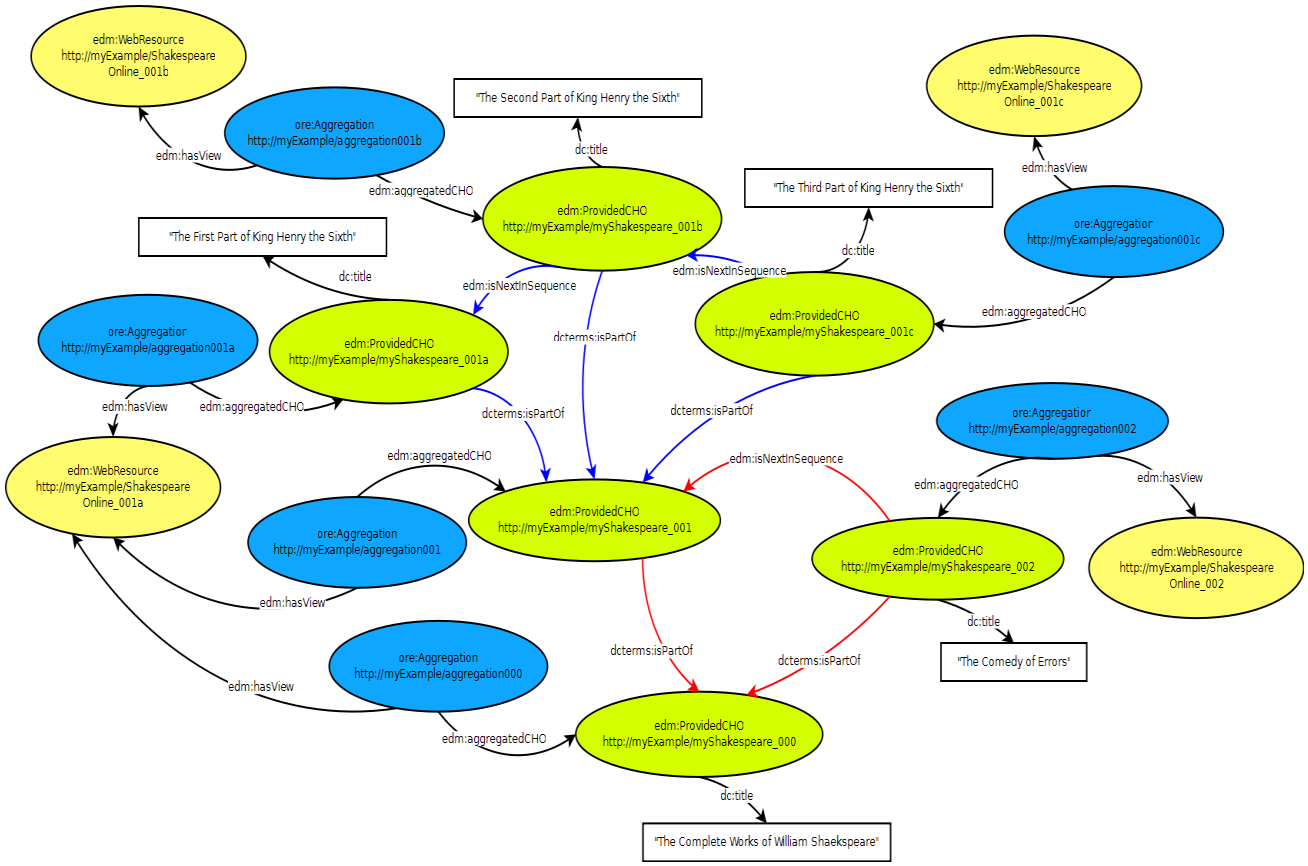

Figura 2 - Exemplo de uma obra multi volume (fonte: ANGJELI, 2011)

O MODELO PARA PUBLICAÇÕES PERIÓDICAS

Uma publicação periódica é um recurso estruturado que pode seguir uma estrutura hierárquica, de sequência ou ambas.

As práticas em uso nas bibliotecas europeias para a catalogação e digitalização de publicações periódicas são muito variadas. Por esta razão, não foi possível definir um modelo único para a representação de publicações periódicas em EDM.

Foram identificados diferentes níveis de descrição que podem dar origem a entidades ProvidedCHO a enviar para a Europeana, ou que têm relações entre ProvidedCHOs, sendo estas relaçoes relevantes para a Europeana ou para o utilizador final da Europeana:

- Artigo
- Número
- Volume
- Título

Cada um destes níveis pode ser um potencial objeto de herança cultural, e podem ser representados através das três classes principais definidas pelo EDM, ou seja um ProvidedCHO, uma ou mais WebResources, e uma Aggregation.

Na Figura 3 pode ser visualizada um diagrama da uma representação EDM para um artigo.

Para além dos vários níveis de descrição possíveis, as publicações periódicas apresentam também uma maior complexidade de relações entre os objetos digitais.

Uma vez que cada um dos níveis pode ser considerado um ProvidedCHO, é possível conceber um modelo complexo considerando a representação de relações hierárquicas.

No diagrama da Figura 4, os diferentes níveis de uma publicação periódica são ligados utilizando propriedade dcterms:isPartOf. Esta propriedade permite representar as relações existentes entre os vários objetos culturais.

As publicações periódicas podem também estar estruturadas de forma sequencial. Este tipo de estruturas pode ser expressa em EDM através da propriedade edm:isNextInSequence. Através do uso desta propriedade, o portal da Europeana poderá, por exemplo, oferecer uma funcionalidade de navegação entre os vários ProvidedCHO pertencentes a uma mesma estrutura.

Na Figura 5 pode ser visualizado um diagrama da uma representação EDM para um objeto com uma estrutura sequencial.

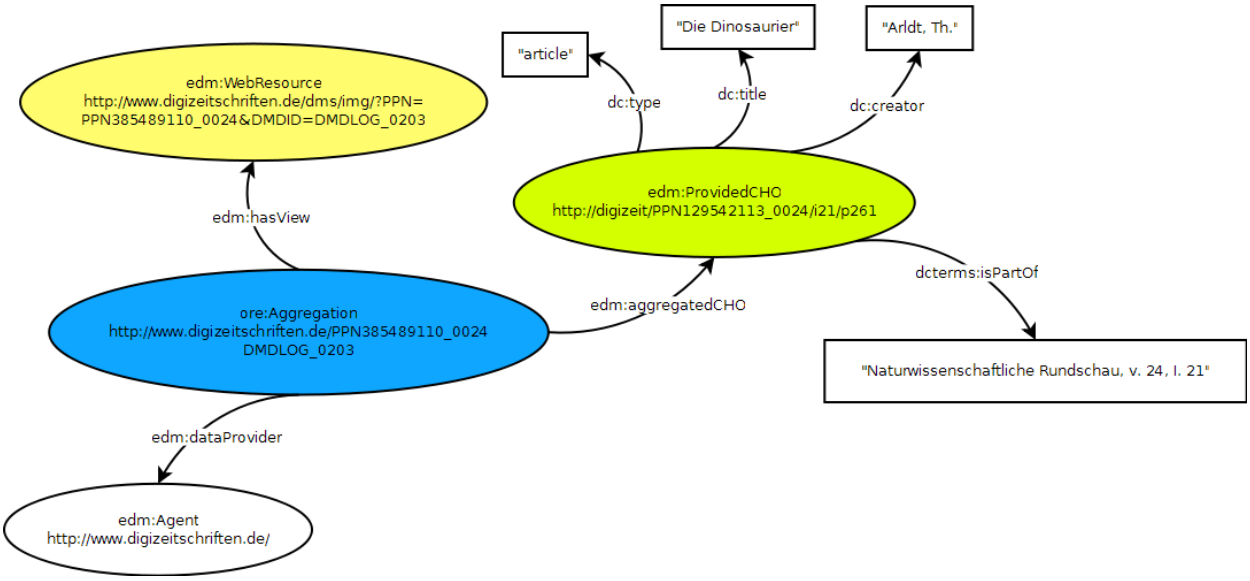

Figura 3 – Exemplo de um artigo (fonte: ANGJELI, 2011)

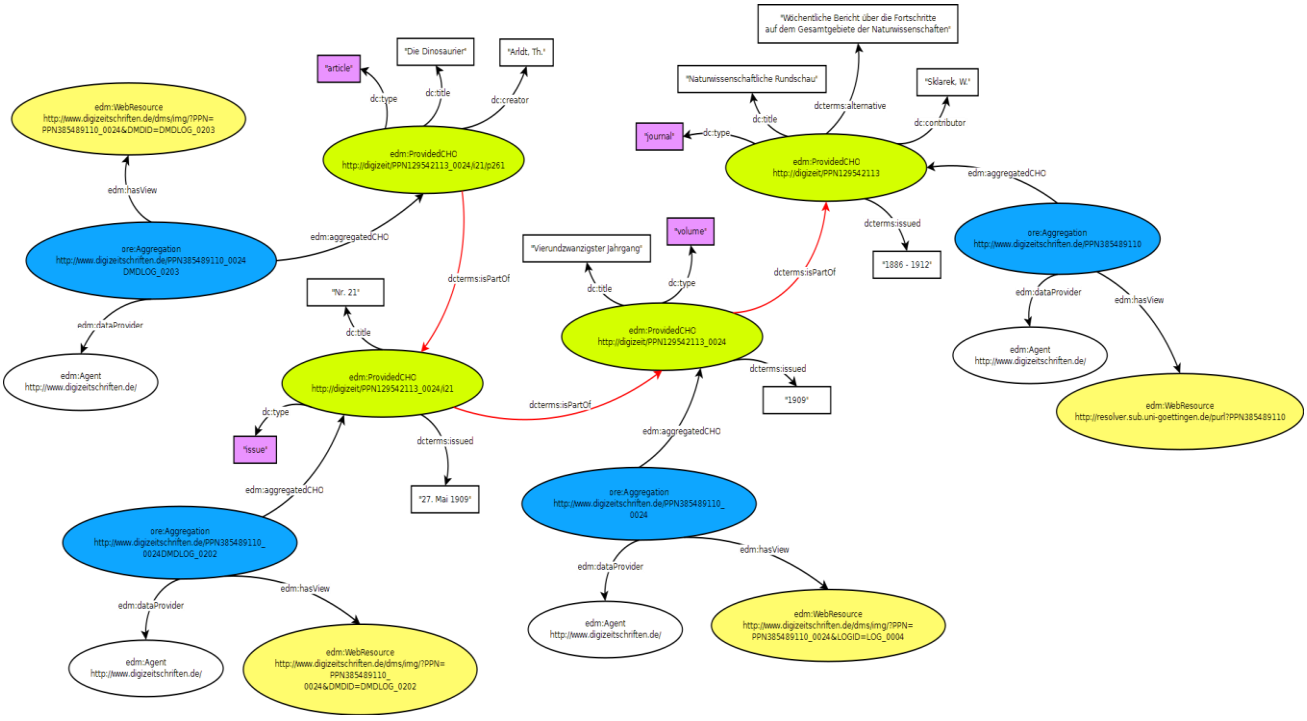

Figura 4 - Exemplo de uma publicação periódica demonstrando a utilização de relações hieráquicas (fonte: ANGJELI, 2011)

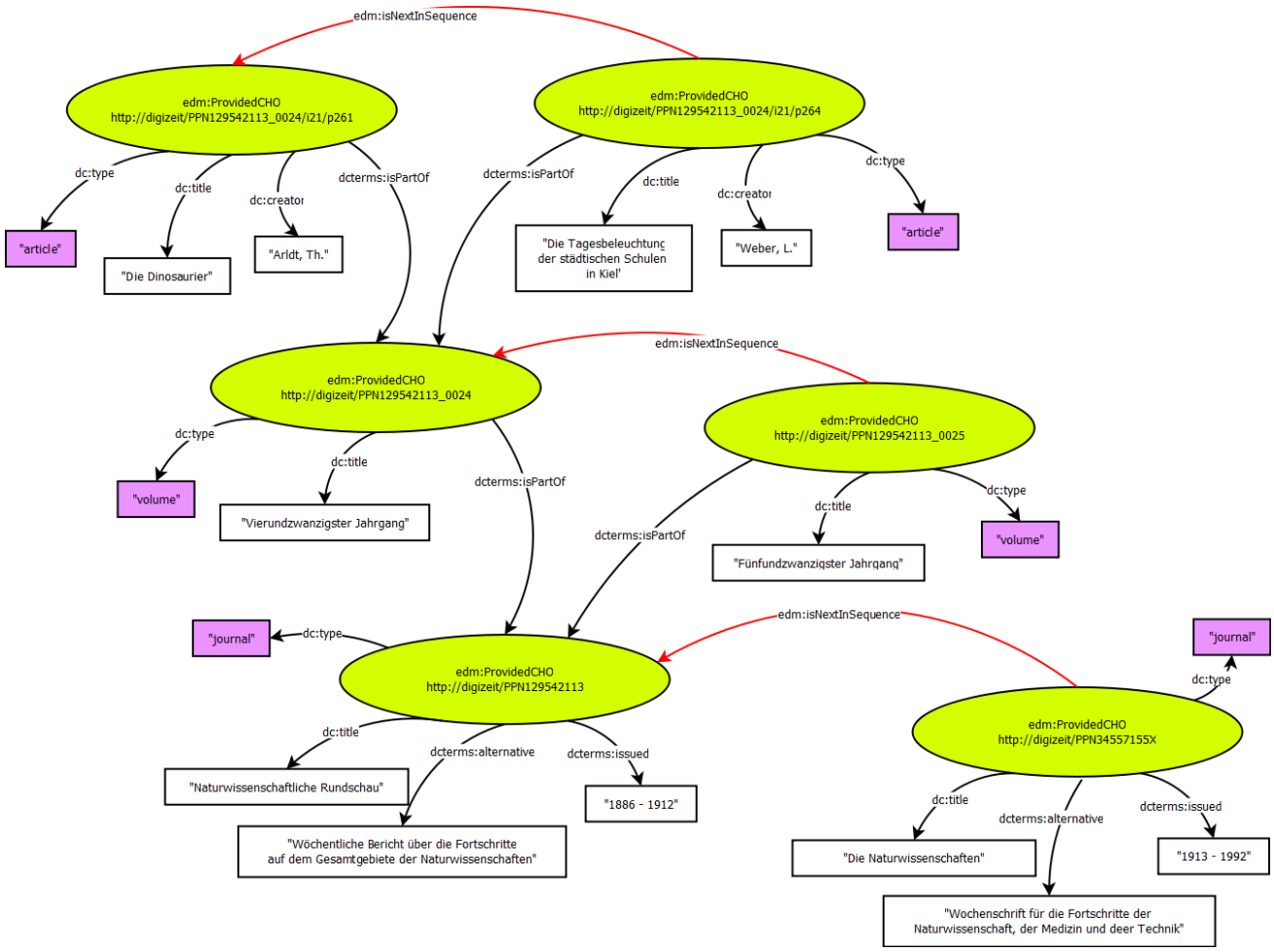

Figura 5 - Exemplo de uma publicação periódica demosntrando a utilização de relações de sequência (fonte: ANGJELI, 2011)

TRABALHO FUTURO

No acervos das bibliotecas podemos encontrar mais tipos de objetos de herança cultural que não foram ainda abordados pelo grupo de trabalho, e que poderão ser alvo de futura investigação. Alguns exemplos são as teses, os manucritos e os mapas. Os recursos áudio e vídeo fazem também parte do acervos de muitas bibliotecas, e poderão também ser analisados em cooperações com outros projetos de outros sectores da cultura, ou projetos que cobrem os variados sectores de interesse para a Europeana.

Alguns aspetos relevantes para o perfil das bibliotecas foram identificados nestes trabalho, mas não foi possível colocá-los nos modelos devido a limitações da versão atual do EDM. Alguns destes aspetos estão já previstos na versão mais recente do EDM, enquanto outros serão ainda objeto de análise futura.

As entidades FRBR são muito relevantes no contexto das bibliotecas. Normalmente, um ProvidedCHO representa uma entidade que faz parte de uma estrutura complexa de entidades FRBR e as suas relações. Uma proposta para a integração de entidades FRBR no EDM, usando classes FRBRoo, pode ser visualizada no diagrama da Figura 6.

Uma característica distinta dos metadados das bibliotecas, é a forma como o evento da publicação é descrito. As bibliotecas detalham e estruturam esta informação ao nível das entidades envolvidas, os locais e as datas. A versão mais recente do EDM já permite a representação da publicação como um evento (através da classe edm:Event), permitindo que a estrutura dos metadados produzidos nas bibliotecas não se perca ao ser representada em EDM. Um exemplo encontra-se na Figura 7.

Para além da representação de estruturas hierárquicas e sequenciais entre objetos digitais, o EDM permite também agrupar diferentes recursos (possivelmente oriundos de diferentes organizações) que contêm conteúdos semelhantes. Um exemplo da utilização de EDM para este fim encontra-se representado na Figura 8.

Outra linha de trabalho em curso no projeto Europeana Libraries visa analisar o caso concreto de obras de texto integral, em que se pretende que o texto, em forma legível por computador, seja partilhado com a Europeana por forma a permitir a criação de um serviço de pesquisa em texto integral na Europeana.

A Figura 9 apresenta as classes e relações do EDM que permitem representar este tipo de recursos. Foi identificada a necessidade de mais uma classe (Full-text Resource). Esta é uma subclasse de Information Resource, e permite a representação dos conteúdos de texto integral de forma independente das versões de acesso para o utilizador final, as quais são modelada em EDM na classe Web Resource.

As versões de texto integral são incluidos nas representações dos ProvidedCHO por intermédio de relações "hasFormat" das Web Resources para Full-text Resources. Desta forma, é possível a Europeana indexar o texto integral, e providenciar aos utilizadores o acesso às versões de consulta.

Os resultados desta análise levaram-nos a concluir que as normas em que o EDM se baseia, podem ser também utilizadas para cumprir os requisitos de interoperabilidade para texto integral. Muitos dos requisitos são já cumpridos pelo EDM, sendo apenas necessária a a extensão do EDM em mais uma classe e uma propriedade.

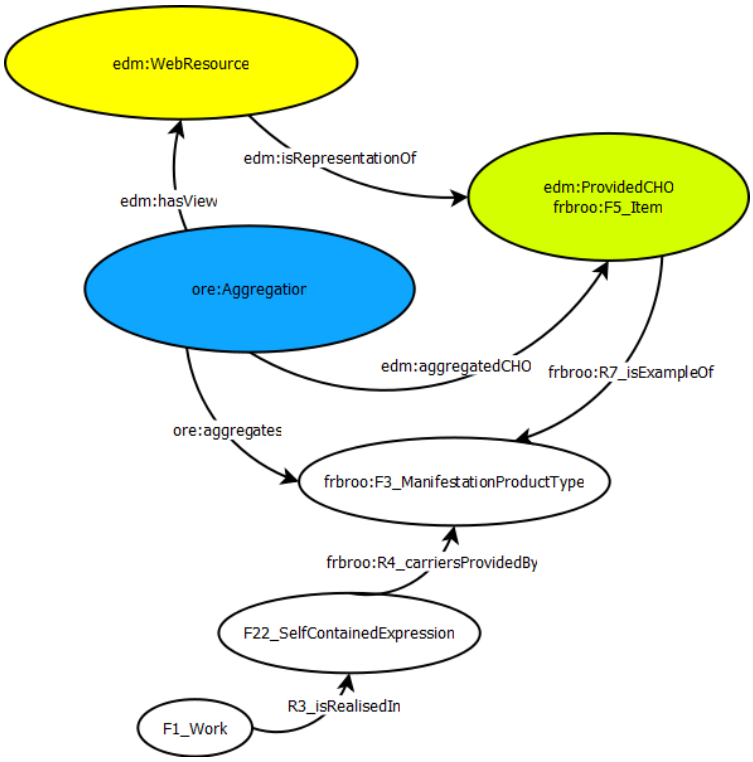

Figura 6 – Possível utilização de entidades FRBRoo em EDM (fonte: ANGJELI, 2011)

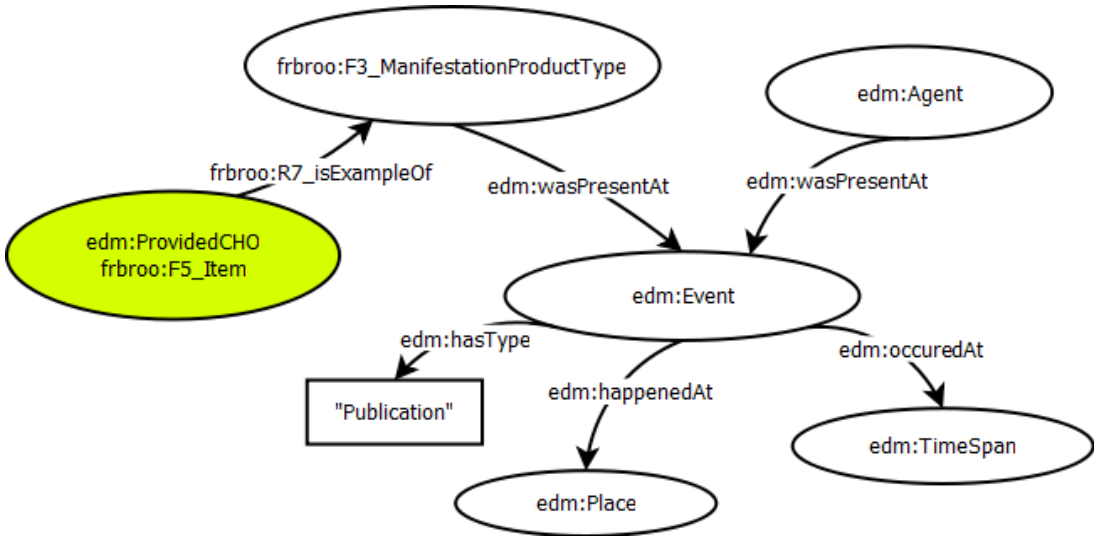

Figura 7 – Proposta de representação baseada em eventos para os dados sobre a publicação (fonte: ANGJELI, 2011)

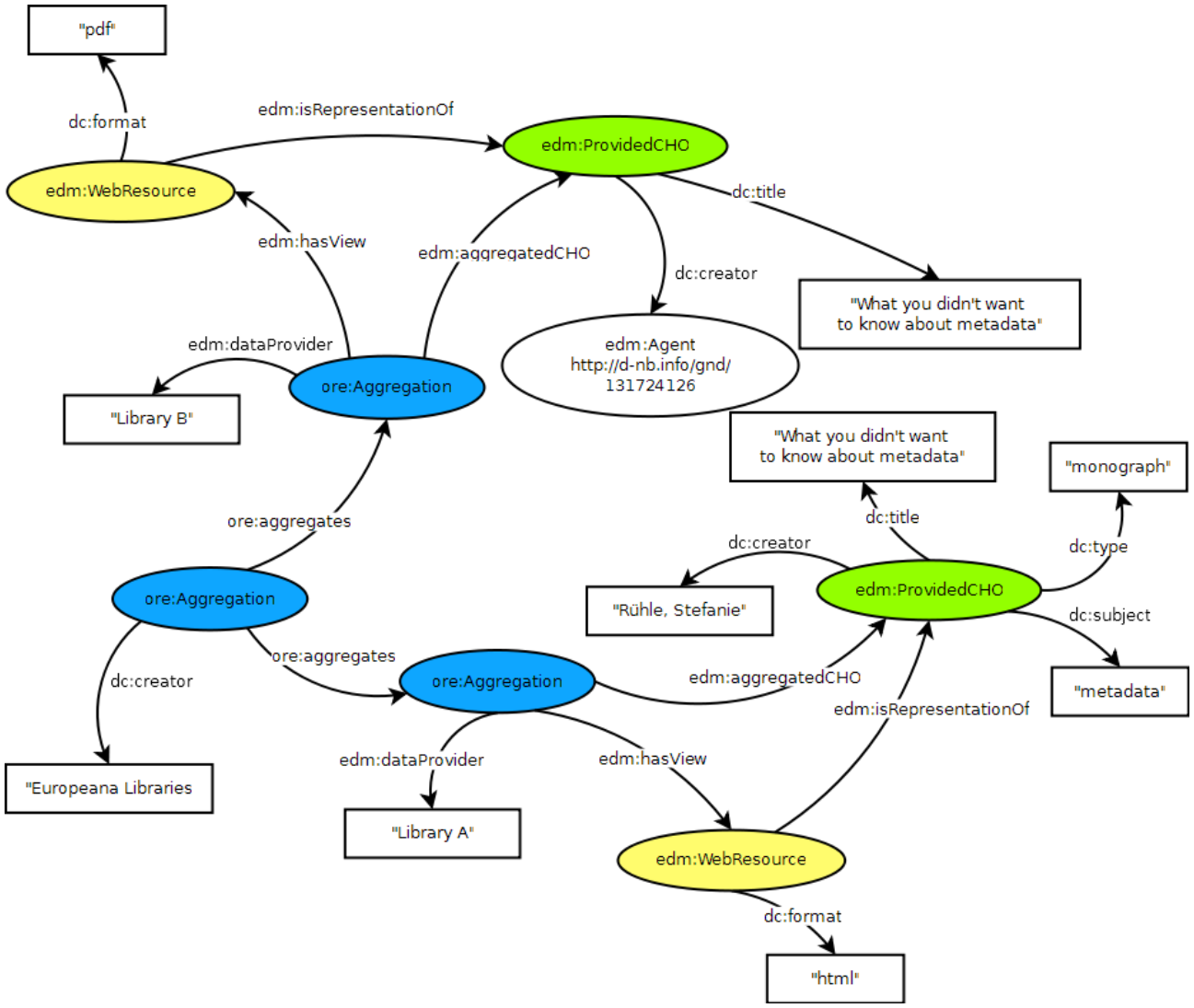

Figura 8 – Agrupamento de objectos culturais semelhantes (fonte: ANGJELI, 2011)

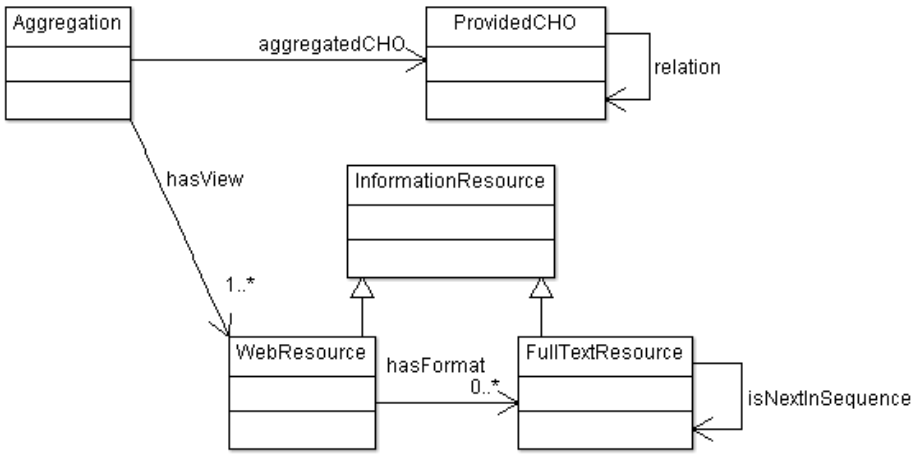

Figura 9 - Modelo de classes do subconjuto do EDM para texto integral (fonte: CHARLES, 2011)

CONCLUSÃO

O trabalho aqui apresentado deve ser considerado como mais um passo intermédio na continua análise do EDM no contexto das bibliotecas. Outros tipos de objetos de herança cultural podem ser encontrados nos acervos das bibliotecas e estes deverão ser alvos de análise futura.

O trabalho aqui apresentado encontra-se agora em fase de validação, estando a ser aplicado em coleções de metadados de bibliotecas, a partir dos seus formatos originais (UNIMARC, MARC21, MODS, etc.). Onze bibliotecas participam nesta validação, incluindo a Biblioteca Nacional de Portugal, e irá ser recolhida a opinião destas bibliotecas sobre a representação EDM dos metadados das suas coleções digitais.

AGRADECIMENTOS

Este trabalho foi executado no contexto do projeto Europeana Libraries. Este projeto foi co-financiado no programa *ICT Policy Support Programme* da Comissão Europeia (Grant Agreement 270933).
O trabalho apresentado neste artigo contou com a contribuição de vários participantes no projeto Europeana Libraries e do The European Library Metadata Working Group: Anila Angjeli (Bibliothèque nationale de France), Martin Baumgartner (Bayerische Staatsbibliothek), Robina Clayphan (The European Library / Europeana), Corine Deliot (The British Library), Jörgen Eriksson (Lunds Universitet), Alexander Huber (University of Oxford), Alexander Jahnke (Consortium of European Research Libraries), Gilberto Pedrosa (Instituto Superior Ténico), Vicky Phillips (National Library of Wales), Natalie Pollecutt (Wellcome Library), Glen Robson (National Library of Wales), Wolfram Seidler (Universität Wien), Stefanie Rühle (Consortium of European Research Libraries), e Andreas Juffinger (The European Library).

REFERÊNCIAS

ANGJELI, Anila et al. - Report on the alignment of library metadata with the European Data Model (EDM). Projeto Europeana Libraries D5.1, 2011.

CHARLES, Valentine, Robina Clayphan, Sally Chambers, Nuno Freire, Andreas Juffinger, Gilberto Pedrosa - Report on how the full-text content will be made available to Europeana. Projeto Europeana Libraries D4.3. 2011.

EUROPEANA v1.0 - Definition of the Europeana Data Model elements: Version 5.2.3 [Em linha]. 2012. Disponível em: http://pro.europeana.eu/documents/900548/bb6b51df-ad11-4a78-8d8a-44cc41810f22

EUROPEANA - The ESE Specifications: Version 3.4.1. [Em linha]. 2012. Disponível em: http://pro.europeana.eu/documents/900548/dc80802e-6efb-4127-a98e-c27c95396d57