Modelação Semântica: O caso de modelação de poesia

Mariana Curado Malta

CEOS.PP, Politécnico do Porto, mariana@iscap.ipp.pt

LINHD, Universidad Nacional de Educación a Distancia, Madrid, Espanha, mariana.malta@linhd.uned.es

Helena Bermúdez Sabel

LINHD, Universidad Nacional de Educación a Distancia, Espanha, helena.bermudez@linhd.uned.es

Elena Gonzalez-Blanco

LINHD, Universidad Nacional de Educación a Distancia, Espanha, egonzalezblanco@flog.uned.es

Resumo

Este artigo tem como contexto o projecto POSTDATA cujo principal objectivo é o de publicar dados de poesia na Web de Dados (Linked Open Data - LOD). Estes dados são metadados dos recursos relacionadas com obras literárias, manuscritos, e tudo o que os caracteriza, como por exemplo: autor da obra, título da mesma, data de criação, idioma, relações de intertextualidade, e ainda questões métricas tais como rima, esquemas métricos ou número de estrofes, entre outro tipo de informação. De forma a promover interoperabilidade entre diferentes agentes que desejem publicar dados LOD relacionados com a poesia, o projecto está na presente data a desenvolver um perfil de aplicação de metadados (MAP) para a poesia. O ponto de partida do desenvolvimento do MAP é um conjunto de bases de dados de poesia e métrica da poesia existentes maioritariamente na Web de documentos. Este artigo apresenta o trabalho que está a ser realizado no sentido de desenvolver este MAP, focando esse desenvolvimento na actividade de definição do modelo de dados. O artigo apresenta ainda muito sucintamente o método utilizado para desenvolver o MAP e as técnicas aplicadas na definição do modelo de dados, e refere o processo de validação do modelo de dados junto da comunidade de prática. Uma vez definido o MAP, novas aplicações desenvolvidas com a tecnologia da Web Semântica poderão i) utilizar dados de diversas fontes (corpora) modelados tendo como base o MAP, e ii) ligar a outros dados LOD pertencentes a outras comunidades de prática. Tudo isto irá permitir que a comunidade de prática da poesia promova e desenvolva estudos mais complexos, cruzando corpora distintintos. O projecto POSTDATA prevê terminar o desenvolvimento do MAP em meados de 2018.

***Palavras-chave**: Dados Ligados, Web Semântica, Perfis de aplicação de metadados, poesia, métrica da poesia*

Abstract

This paper stems from the POSTDATA project. Its main goal is to publish European poetry data in the Web of Data (Linked Open Data – LOD). This data is metadata of different resources related to literary works, manuscripts, and any of their features, such as: author, title, creation date, language usage, intertextuality relationships, and even versification issues as metre and rhyme. In order to enhance interoperability among any agent willing to publish poetry data as LOD, this project is currently developing a metadata application profile (MAP) for poetry. The starting point is a set of poetry and metre databases –most of them available on the Web of Documents. This paper presents the work that is being done in order to develop a MAP, focusing this development in the data model definition activity. This paper also presents, shortly, the method used to develop the MAP and the techniques employed to define the data model. Finally, the paper refers the process of validation of the data model with the poetry community of practice. Once the MAP is defined,

new applications developed with Semantic Web technologies will be able to i) use data from different sources (corpora) modelled using the MAP for poetry, and ii) link to other LOD datasets belonging to other communities of practice. All this will allow the poetry community to promote and develop more complex studies across different corpora. POSTDATA plans to finish the MAP development around June 2018.

***Keywords**: Linked Data, Semantic Web, metadata application profiles, poetry, poetry metrics*

1. Introdução

A presente secção contextualiza o trabalho relatado neste artigo. A primeira sub-secção apresenta a comunidade de poesia Europeia através dos repertórios existentes ao longo do tempo; a segunda os desafios que a investigação de estudos literários enfrenta com as novas possibilidades que as tecnologias dos dados ligados na Web nos apresentam.

1.1. A comunidade de poesia na Europa

A literatura na Idade Média era um conjunto relativamente homogéneo no qual os temas, as fontes textuais e os motivos literários fluíam facilmente de um idioma para o outro. Este facto pode ver-se nas manifestações de formas poéticas comuns aos diferentes idiomas, assim como na existência de géneros literários semelhantes e na aparição de melodias musicais comuns. O estudo de fenómenos como os contrafacta musicais ou formais[1] ou a difusão de determinados temas associada a uma série concreta de estruturas estróficas deu muitos frutos. Eles revelam a existência de uma interacção constante de formas, idiomas e sistemas poéticos. Pensemos, por exemplo, nos trabalhos de Anton Touber (1999), Paolo Canettieri y Carlo Pulsoni (1995), Dominique Billy (1998) com estudos que analisam o fenómeno dos contrafacta. Este tipo de análise é muito complexo uma vez que as diferentes tradições poéticas, a organização distinta da investigação e a heterogeneidade bibliográfica tornam impossível a recompilação de todos os dados (González-Blanco y Seláf 2014). A preocupação pelo estudo e compilação de todas estas formas e fenómenos que permitam avançar na investigação, e compreender melhor a evolução das tradições literárias foi uma constante desde as primeiras investigações. Esta preocupação deu lugar à necessidade de agrupar os dados poéticos em forma de repertórios.

Podemos identificar três grandes etapas na criação de repertórios poéticos das literaturas medievais e renascentistas. A primeira diz respeito à época do positivismo (finais do século XIX), com obras como as de Gaston Raynaud (1884), Pillet y Carstens (1933) e Naetebus (1891). A segunda começa após a Segunda Guerra Mundial com o trabalho de referência de Frank (1966) sobre a poesia provençal dos trovadores, e prolonga-se até aos nossos dias com a edição de repertórios métricos em papel (o de Tavani (1967) sobre a poesia galego-portuguesa, Mölk e Wolfzettel (1972) sobre a lírica medieval francesa, os italianos: Solimena (1980) do *Stil Novo*, Antonelli (1984) sobre a escola siciliana, e recentemente Solimena (2000) combinando ambas as tradições, e Pagnotta (1995) sobre a *ballata* italiana; o catalão de Jordi Parramon i Blasco

1 Um contrafactum musical consiste na reutilização de uma música para fazer outro poema. Um contrafactum formal é quando somente o padrão métrico é igual.

(1992), e o castelhano de poesia de cancioneiro de Gómez Bravo (1999). Em línguas não romances destaca-se o alemão Brunner et al. (1986-2007). Os avanços tecnológicos possibilitaram a criação de uma terceira geração de repertórios, agora digitais, onde o trabalho de investigação é feito de uma forma mais eficiente. O primeiro repertório poético digitalizado em 1991 foi o "Répertoire de la Poésie hongroise ancienne jusqu'à 1600"[2] de Iván Horváth. Ao longo dos anos seguintes começaram a aparecer repertórios digitais baseados alguns nos já citados existentes em papel, e outros que surgiram de novo[3]. A maior parte deles estão representados no mapa https://goo.gl/9MCWrv[4], e contêm uma variedade linguística e cronológica muito ampla dentro do marco geográfico europeu. Estas três etapas identificam-se da seguinte forma: a primeira abarca a criação dos primeiros repertórios que foram desenvolvidos com escassos meios e infraestruturas, recompilando a informação de uma forma muito artesanal; a segunda traz uma nova geração de repertórios cujos conteúdos foram organizados graças ao apoio de um computador e umas bases de dados que permitiram a organização dos conteúdos; a terceira é a passagem dos repertórios já existentes para a consulta na Web de documentos.

Estes repertórios apresentam diferentes características em função do tipo de conteúdo das obras que recolhem, em alguns casos presta-se maior atenção à música, em outros à estrutura narrativa dos poemas ou ainda em outros à história crítica da peça em concreto.

1.2. As necessidades de investigação

A técnica da comparação é uma das operações académicas mais básicas, "a functional primitive of humanities research"[5] (Unsworth, 2000). De facto, existem disciplinas filológicas cujo método de trabalho recorre frequentemente à técnica da comparação, um exemplo é o caso da Literatura Comparada, que estuda o património cultural ultrapassando fronteiras linguísticas, nacionais e mesmo disciplinares. No entanto, as relações existentes entre as diferentes tradições literárias, especialmente no contexto europeu, é de tal importância, que é impossível estudar uma literatura nacional sem ter em consideração estas relações (Even-Zohar, 1978, p. 48) pelo que, mesmo fora do campo da Literatura Comparada, a rigorosidade da análise passa pelo conhecimento de literaturas afins.

Neste sentido, a possibilidade de fazer pesquisas que superem essas fronteiras linguísticas e nacionais abrem numerosas linhas de investigação dentro dos estudos poéticos. Apresentamos, por um lado, a oportunidade de realizar uma exploração de conteúdos maciça e transversal. Podemos assim pesquisar como e onde se desenvolvem certos temas

2 Ver http://rpha.elte.hu/rpha/ - acedido em 6 de Junho, 2017

3 É o caso, por exemplo, de um dos repertórios mais recente de nome ReMetCa – http://www.remetca.uned.es, acedido em 6 de Junho, 2017 -, repertório métrico castelhano, que só uma parte dos conteúdos da base de dados (os poemas de cancioneiro) têm um antecedente em papel (González-Blanco y Rodríguez 2015).

4 Acedido em 6 de Junho, 2017.

5 Tradução livre das autoras: "uma função primitiva da investigação humanística".

nas diferentes tradições poéticas. Através do estudo do desenvolvimento dum determinado tópico ao longo da História, tendo em consideração tanto as coordenadas geográficas como as cronológicas, pode mapear-se o interesse político-social e cultural que esse tema vai adquirindo. Como exemplo podemos desejar realizar um estudo sobre o fenómeno da peregrinação a Santiago de Compostela, e como parte deste estudo podemos desejar analisar a repercussão deste tema na literatura. Seria, portanto, necessário pesquisar todos os corpora literários existentes para poder saber em que contextos se explora mais (ou menos) esse tema, e assim entender os motivos pelos quais num determinado momento histórico esse tópico se torna relevante. Além do mais, uma pesquisa temática num corpus multi-linguístico e diacrónico proporciona materiais cuja análise traria à tona quais agentes estão a funcionar como propagadores de certos conteúdos em cada momento, informação que ultrapassaria o campo do meramente literário.

Por outro lado, a exploração dos elementos formais duma maneira contrastiva e abrangente permite criar uma imagem mais verídica do nosso objecto de estudo. Apesar da prosódia e das figuras estilísticas estarem intimamente vinculadas com a língua, existe um espaço comum que permite estabelecer comparações entre as diferentes tradições europeias. Este facto fornece novo conhecimento no que respeita às diferentes estratégias que as tradições poéticas exprimem com o fim de criar uma linguagem poética própria. A História da cultura europeia está marcada pela existência de focos de cultura que criam modelos que são imitados e readaptados até que um novo foco vê a luz. Isto significa que se geram vinculações entre obras em todos os níveis da composição poética. Por exemplo, no século XI o Midi francês surgiu como foco de cultura, e o trovadorismo occitano converteu-se no modelo literário lírico do Ocidente. Isto significa que as canções provençais eram conhecidas, imitadas e traduzidas noutras regiões. De facto, a partir do esquema métrico-rimático das cantigas occitanas podemos descobrir que composições noutras línguas utilizaram a mesma música adaptando as letras à tradição concreta.

Estas complexas redes de relacionamentos, característica partilhada por qualquer produto cultural, dificultam as análises literárias, daí a importância de uma ferramenta que facilite o estudo dum contexto literário abrangente.

Para que a técnica de comparação seja realizada de uma forma completa e global a nível europeu, é necessário encontrar um paradigma tecnológico que permita implementar a interoperabilidade destes dados que estão fechados nas diferentes bases de dados por toda a Europa. O paradigma que o projecto POSTDATA[6] escolheu foi a Web Semântica, e o constructo que permite essa interoperabilidade de dados é o perfil de aplicação de metadados.

Este artigo tem como objectivo apresentar uma tarefa específica das várias etapas de desenvolvimento deste Perfil de Aplicação de Metadados para a poesia europeia. A secção seguinte

apresenta em primeiro lugar uma visão geral do paradigma em que o projecto POSTDATA se situa, a Web Semântica, e a razão pela qual esse paradigma foi o escolhido para resolver a questão de partilha de dados na comunidade Europeia

6 Ver http://postdata.linhd.es – acedido em 6 de Junho, 2017.

de poesia, e em segundo lugar a forma particular como o projecto está a definir um modelo de dados semântico para a comunidade de prática referida. As conclusões do artigo são apresentadas na última secção, juntamente com os trabalhos que prevemos realizar no futuro.

2. Web Semântica e o caso da Poesia

2.1 A Web Semântica

Os vários repertórios referidos anteriormente possuem dados guardados em servidores que são acedidos através de uma interface Web. Cada interface Web apresenta vistas dos seus dados, onde investigadores podem retirar informação de uma forma não sistemática. Para obter dados dos vários repertórios, eles terão de ser retirados das diferentes interfaces Web, comparando posteriormente estes dados através de ferramentas arcaicas como papel ou outras mais sofisticadas como programas de computador (onde terão de passar sempre por um processo manual de importação de dados, estando expostos a possíveis erros de manipulação), para utilizar as ferramentas existentes nos programas. Estes dados estão assim fechados em silos de informação não sendo possível reutilizá-los de uma forma automática.

O projecto POSTDATA nasce então desta necessidade de encontrar uma maneira de libertar estes dados presentes nestes silos de informação, de forma a torná-los comparáveis e reutilizáveis. A esta possibilidade de reutilizar os dados acrescenta-se a interoperabilidade dos dados, isto é, a "possibilidade de diferentes sistemas com diferentes programas, tecnologia, estruturas de dados e interfaces poderem trocar dados sem comunicar previamente e sem perder um mínimo de conteúdo e funcionalidade" (DCMI, 2011)

Em 2001 foi apresentado um novo paradigma num artigo seminal na revista *Scientific American* onde um utilizador poderia num futuro próximo organizar a sua vida sem ter de se preocupar em transportar dados entre sistemas diferentes. Estes sistemas eram capazes de comunicar e retirar conclusões, dando respostas em tempo real às necessidades de um utilizador (Berners-Lee *et al.*, 2001). Mais do que isso, máquinas inteligentes poderiam utilizar dados provenientes de várias fontes de informação, inferir sobre os dados, e apresentar respostas e conclusões a pedidos mais ou menos complexos efectuados pelo utilizador final. A este novo paradigma chamou-se a Web Semântica. Mais tarde aos dados estruturados publicados na Web Semântica deu-se o nome de dados ligados (em Inglês *Linked Data*) ou dados ligados e abertos (LOD) (em Inglês *Linked Open Data*) caso os dados sejam abertos, isto é, livres de serem utilizados gratuitamente.

Esta visão outrora futurista, é hoje realidade. Os dados publicados na Web Semântica formam um grande conjunto de dados disponíveis para serem usados utilizando tecnologias LOD. A este conjunto de dados LOD dá-se o nome de nuvem LOD[7]. Hoje já existe uma panóplia de programas de computador disponíveis, gratuitos ou para serem comprados, e existem bastantes empresas de suporte especializadas no assunto já com um historial de sucesso no mercado. Exemplos

7 Ver LOD cloud – http://lod-cloud.net (acedido em 23 de Maio, 2017)

de projectos pioneiros e com grande êxito são os da Dbpedia[8], da BBC[9] ou da Europeana[10].

Uma das grandes vantagens desta abordagem LOD (porque existem outras abordagens para partilhar os dados entre as várias organizações que os possuem) é que, sendo a Web Semântica um paradigma onde se publicam dados ligados (e abertos), as possibilidades de ligações são infinitas. De facto, o termo "ligados" é o que abre este novo paradigma a imensas possibilidades, imaginadas e não imaginadas. Quando estamos confinados à nossa organização, ou grupo de organizações não há mais dados além dos que esse grupo de organizações possui. Assim como nós humanos acedemos a uma página Web (em HTML) e clicamos em ligações e navegamos por outras páginas de outros sítios Web, que já nada têm a ver com a primeira página acedida, isto é, as novas páginas estão num outro servidor, pertencendo a outra organização e provavelmente estarão num outro país, o mesmo acontece no paradigma da Web Semântica. As máquinas podem percorrer ligações entre os dados (dados descritos numa sintaxe que fornencem dados estruturados. Exemplo de sintaxes são o Turtle[11] ou RDF/XML[12], entre outros) e navegar pela Web de Dados[13] através dos diferentes *datasets* LOD, ou seja, navegar na nuvem LOD. No caso da Web Semântica teremos sempre dados de outras comunidades disponíveis para enriquecer os nossos dados. O potencial da "ligação" é infinito.

De forma a que estes dados de todos os repertórios digitais de poesia se tornem interoperáveis, é necessário desenvolver um perfil de aplicação de metadados (MAP – *Metadata Application Profile*, em Inglês) (Nilsson, Baker e Johnston, 2008) Um MAP não é mais do que um modelo de dados semântico que associa a cada propriedade do modelo um termo de um vocabulário RDF[14], e que define restrições específicas a cada propriedade. Para saber mais sobre perfis de aplicação de metadados ler: Heery e Patel (2000), Baker e Coyle (2009), e Coyle (2017). O projecto POSTDATA está neste momento de escrita deste artigo a desenvolver um MAP para a comunidade europeia de poesia.

2.1 Modelação semântica e o caso da poesia

Para que seja possível publicar dados de poesia interoperáveis na nuvem LOD é necessário definir um perfil de aplicação

8 Ver http://wiki.dbpedia.org/ (acedido em 23 de Maio, 2017)

9 Ver http://www.bbc.co.uk/things/ (acedido em 23 de Maio, 2017)

10 Ver http://www.europeana.eu/portal/en (acedido em 23 de Maio, 2017)

11 Ver https://www.w3.org/TR/rdf-primer/ - acedido em 23 de Maio, 2017

12 Ver https://www.w3.org/TR/rdf-syntax-grammar/ - acedido em 23 de Maio, 2017

13 Uma outra forma de denominar o contexto dos dados ligados e abertos (LOD).

14 Também chamado de esquema de metadados (*metadata scheme,* em Inglês). Um vocabulário RDF é um conjunto de termos de metadados com as respectivas definições. Por exemplo o Dublin Core Metadata Terms (ver http://dublincore.org/documents/dcmi-terms/ - acedido em 24 de Maio, 2017) é um vocabulário multi-domínio (*cross-domain*, em Inglês) composto por 55 propriedades e 22 classes (por exemplo: dc:title, dc:subject, dc:creator, etc) (Nilsson et al., 2006). O *Dublin Core Metadatat Terms* é um importante standard da Web Semântica. Existem muitos vocabulário RDF, por exemplo, o projecto *Linked Open Vocabularies* (LOV) (http://lov.okfn.org/dataset/lov/ - acedido em 24 de Maio, 2017) disponibiliza uma ferramenta que nos permite procurar vocabulários RDF por domínio de aplicação e obter informação específica sobre eles.

de metadados (MAP) para a comunidade descrita. O desenvolvimento de um MAP implica acima de tudo a estruturação de dados. Esta estruturação de dados tem a ver com a modelação dos dados que é necessária efectuar de forma a organizá-los em conceitos semelhantes. O problema que se nos apresenta de modelação de dados poéticos e métricos é mais complexo do que se poderia pensar à partida já que a comunidade de prática com quem estamos a trabalhar - organizações Europeias que estudam poesia e a métrica da poesia – é muito heterogénea. Estas organizações vêm de diferentes países com diferentes culturas de abordagem sobre a forma de estudar poesia, e ainda, utilizam diferentes idiomas. Esta heterogeneidade traz problemas de padronização à cabeça, isto é, só é possível realizar a modelação de dados se toda a comunidade se puser de acordo em relação aos princípios de estudo e de classificação dos conceitos. Não existe um repertório de termos métricos comummente admitido e utilizado em cada um dos repertórios. As ambiguidades são constantes. Um exemplo simples de entender é o caso do nome dado a uma "linha" num poema. Em Portugal chamamos "verso" a uma linha assim como em espanhol, no entanto um "verso" na tradição anglo-saxónica (*verse*) pode referir-se tanto a um só verso como a um agrupamento de versos, sendo esta segunda acepção equivalente à nossa "estrofe". Temos então duas questões a resolver, a primeira é a padronização em termos dos conceitos filológicos, a segunda a modelação desses conceitos.

Neste artigo falamos da modelação dos conceitos e não das questões de padronização de conceitos filológicos (para saber mais sobre a padronização dos conceitos filológicos ver Bermúdez-Sabel, Curado Malta e Gonzalez-Blanco (2017)).

O desenvolvimento do MAP segue um método específico para o desenvolvimento de MAPs, o Me4MAP[15] – ver Curado Malta (2017), Curado Malta e Baptista (2013). Este método é o único, tanto quanto é do nosso conhecimento, verdadeiramente formalizado como um método para o desenvolvimento de MAPs. O Enquadramento de Singapura (ver Nilsson, Baker e Johnston (n.d)) é uma declaração que define o que é o MAP da Dublin Core Metadata Initiative (também chamado de DCAP[16]). Ele define o conjunto de documentos que formalizam o perfil de aplicação, mas não diz como se desenvolvem os artefactos que esses documentos enquadram. O Me4MAP define as várias actividades a realizar durante todo o processo de desenvolvimento de um MAP, e os artefactos que elas produzem. Além disso este método define a forma como as actividades interagem entre si e aconselha por vezes as técnicas a utilizar para desenvolver os artefactos. Não é o objectivo deste artigo expor detalhadamente o Me4MAP, no entanto de forma a entender melhor o enquadramento do processo de modelação semântica, iremos apresentá-lo muito brevemente. O Me4MAP define cinco actividades principais, S1 a S5 – "S" de Singapura, porque estas actividades têm como resultado (*Delivarable)* os artefactos de Singapura – (que depois se desdobram em outras actividades mais simples), as primeiras três obrigatórias, as duas últimas facultativas:

15 Method for the development of Metadata Application Profiles

16 Dublin Core Application Profile

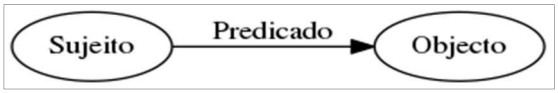
- S1 - Desenvolvendo os Requisitos Funcionais: um Requisito funcional define uma função de um sistema de software ou seu componente. Uma função é descrita como um conjunto de entradas, seu comportamento e as saídas. Nesta actividade identificam-se os requisitos funcionais;
- S2 - Desenvolvendo o modelo de domínio: Um modelo de domínio, também apelidado de modelo de dados, apresenta os conceitos do contexto que queremos capturar e as relações entre esses conceitos. Nesta actividade desenvolve-se o modelo de domínio;
- S3 - Desenvolvendo o *Description Set*: O *Description Set Profile* (DSP) define os termos dos vocabulários RDF que descrevem cada um dos termos do modelo de domínio. Define ainda as restrições sobre os termos. Nesta actividade define-se o DSP;
- S4 - Desenvolvendo os guias de sintaxe;
- S5 - Desenvolvendo o guia do utilizador.

Este artigo apresenta brevemente a forma como o modelo de domínio está a ser desenvolvido (actividade S2 do Me4MAP).

O processo de definição do modelo de domínio parte de um conjunto de bases de dados que servem os repertórios digitais de poesia referidos na secção 1. Essas bases de dados têm todas elas um modelo de dados associado que está a ser estudado pela equipa de modeladores do projecto de forma a compreender os conceitos que representam, num processo de re-engenharia inversa. Para além dessa fonte de informação, os modeladores recorrem à análise das interfaces Web dos repositórios digitais para retirar informação sobre os dados que suportam as funcionalidades existentes (ver Curado Malta, Centenera e Gonzalez-Blanco (2017) para saber mais sobre as técnicas aplicadas). Finalmente um inquérito foi posto em marcha para conhecer as necessidades informacionais dos utilizadores finais de futuros recursos que irão utilizar os dados poéticos LOD[17].

Todo o processo de desenvolvimento do modelo de domínio a partir desta fonte de informação é altamente interativo.

O modelo de domínio é um modelo de dados que pode ser representado com um diagrama UML relacional de classes. No entanto a modelação em LOD é uma modelação semântica, e não relacional, que está baseada no modelo de dados Resource Description Framework (RDF) (W3C, 2014). Este modelo baseia-se em triplos onde um triplo é composto por sujeito, predicado e objecto (ver Figura 1).



17 Ver http://postdata.linhd.es/limesurvey/index.php/113575?lang=en – acedido em 24 de Maio, 2017.

Figura 1. Triplo RDF

Um exemplo de triplo no contexto da modelação semântica seria o apresentado na Figura 2. Este triplo refere que uma determinada obra (um poema que aqui é identificado através de um URI[18]) tem como título "Señores que me mandáis". Na Figura 3 apresentamos já um conjunto de dados mais complexo, um conjunto de triplos com mais informação sobre essa mesma obra: o título, os idiomas em que se pode ler, o livro que o refere. Este livro é identificado por um URI que remete para um servidor LOD que está no Arquivo Nacional da Torre do Tombo[19]. Aqui podemos ver o potencial dos dados ligados, ao definir-se que um determinado livro (http://antt.dglab.gov.pt/12876) refere a obra em questão recorre-se a dados que estão presentes num outro servidor de dados ligados (no caso seria o servidor LOD da Torre do Tombo).

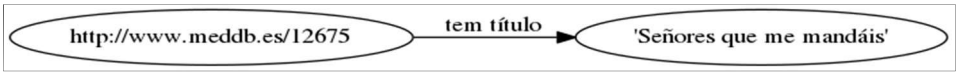


Figura 2. Um exemplo de triplo com os dados sobre o título de uma determinada obra poética

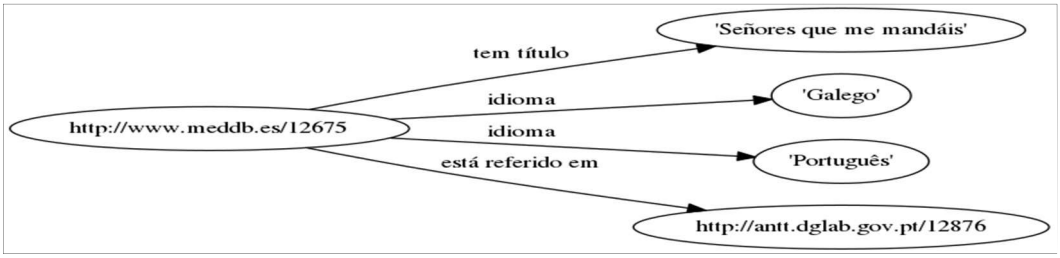


Figura 3. Um exemplo de várias triplos com dados de uma determinada obra poética

Das vinte e cinco bases de dados, a maior parte delas são relacionais, outras são XML e uma está organizada numa folha de cálculo. Uma vez que a maior parte das bases de dados são relacionais, a equipe de trabalho decidiu construir modelos conceptuais de todos os modelos de dados para haver uma forma de os comprar a todos, e a partir de aí construir um modelo conceptual comum. Como podemos verificar, a modelação semântica nada tem a ver com a modelação relacional, nem mesmo com a modelação hierárquica XML. O processo de passagem de um modelo relacional ou de um modelo XML para um modelo conceptual, para depois ser passado ao semântico, é uma técnica simples, mas que retira a obrigação de aplicar determinadas regras obrigatórias na modelação relacional, por exemplo. Um MAP é um modelo orientado à propriedade, ao contrário do modelo relacional que está baseado na definição das entidades e suas relações. No processo de transformação, o importante é definir cada uma das propriedades, e definir o seu domínio e contra-

18 De referir que os URIs são exemplos não reais, só servem para questões de demonstração. "URI» significa *Uniform Resource Identifier*, em Português Identificador Uniforme de Recurso. Um URI identifica univocamente um recurso na Web de Dados. São equivalentes às cheves primárias dos modelos relacionais de dados.

19 Exemplo fictício.

domínio (recorrendo ao exemplo da Figura 3: i) a propriedade "está referido em" terá um domínio uma "obra poética" e um contra-domínio um "recurso bibliográfico" (livro); ii) a propriedade "idioma" tem como domínio uma "obra poética" e um contradomínio uma cadeia de caracteres). As questões de normalização (formas normais) não se colocam precisamente porque o modelo não é orientado à entidade, mas sim à propriedade. Nos modelos que eram relacionais, o processo de re-engenharia inversa retirou essa necessidade de estarem na 3ª forma normal.

A obtenção do modelo comum (modelo de dados ou modelo de domínio do MAP) foi realizado a partir de todos os modelos conceptuais, inicialmente este modelo tinha todas as entidades e propriedades, e relações existentes. E depois, num processo iterativo e de crescente abstracção, o modelo foi ficando mais pequeno e mais eficiente – ver Bermúdez-Sabel, Curado Malta e Gonzalez-Blanco (2017) para mais detalhes sobre este processo iterativo.

O processo de desenvolvimento do modelo de domínio definirá várias versões do modelo antes de chegarmos à versão final. De facto, estão planeadas pelo menos quatro versões distintas; isto acontece porque a equipe de modelação deseja o seu modelo validado pela comunidade de poesia Europeia antes de avançar para a actividade S3. O processo de validação tem a seguinte sequência:

- A primeira versão do modelo de domínio (V0.1) foi apresentada numa workshop em Março de 2017 onde vários delegados dos repertórios digitais de poesia Europeia estiveram presentes. Durante um dia e meio envolveram-se em actividades de validação e discussão dando retorno à equipe de modelação do projecto da forma como o modelo de domínio respondia às suas necessidades informacionais;
- A versão 0.2 nasce da validação à versão 0.1;
- A versão 0.3 está a ser desenvolvida no momento de escrita deste artigo, e recebe informação de mais bases de dados, mas também a informação proveniente dos resultados do inquérito já referido. O mapa https://goo.gl/O0mqhl[20] apresenta o planeamento de análise de bases de dados de repertórios digitais segundo as versões do modelo de domínio. Este mapa é um *work-in-progress*, isto é, pode mudar ao longo do tempo (até à versão final) porque há ainda bases de dados que não responderam ao pedido de informação técnica e é possível que novas bases de dados sejam adicionadas ainda até ao fim do processo. A versão 0.3 será validada por outras bases de dados que não foram usadas no processo de desenvolvimento (ver *Validation 2nd Phase* no mapa – aqui serão adicionadas novas base de dados), para que a validação não esteja viciada;
- A versão 0.4 nasce da validação à versão 0.3. Espera-se que a versão 0.4 seja a final, estável, para seguir com o processo de desenvolvimento do MAP.

3. Conclusões

A publicação de dados sobre poesia como dados abertos enlaçados apresenta inúmeras vantagens no que respeita ao

20 Acedido em 24 de Maio, 2017

questionamento dos dados literários: não só permite a extracção de informação duma maneira mais eficiente, mas abre as portas para a formulação de novas perguntas.

A materialização desse objectivo passa pelo desenvolvimento dum Perfil de Aplicação de Metadados (MAP) da poesia europeia. Este artigo relata a experiência de desenvolvimento deste MAP, mais especificamente na etapa de desenvolvimento do modelo de dados. O artigo contextualiza este trabalho na Web Semântica, apresentando muito brevemente este paradigma assim como o método utilizado para desenvolver o MAP. Segue depois com a descrição da forma de modelação dos dados e do processo de validação do modelo.

Neste momento o projecto está a terminar a terceira versão do modelo de dados. Uma vez determinado o modelo de dados final, o desenvolvimento do MAP segue com os trabalhos de descrição do *description set*, onde irá cumprir um conjunto de micro-etapas, onde as principais são a definição do mapeamento do ambiente (*environmental scan*), a definição do alinhamento de vocabulários RDF e a elaboração da matriz de restrições. Todo o processo é iterativo e os artefactos desenvolvidos são sempre validados junto da comunidade de prática de poesia. Este MAP irá ser integrado em todas as actividades de desenvolvimento de software do projecto POSTDATA. O desenvolvimento do MAP é, portanto, uma etapa muito importante do projecto POSTDATA uma vez que influencia tudo o que se virá realizar a seguir.

Por fim, gostaríamos de referir a importância deste trabalho como um projecto de investigação que supera o âmbito dos Estudos Literários, na medida em que proporciona materiais de interesse para a História das Mentalidades e a História Cultural, além do conhecimento empírico obtido dentro do campo da modelação semântica.

4. Agradecimentos

O desenvolvimento deste trabalho foi possível graças ao projecto *Poetry Standardization and Linked Open Data: POSTDATA* (ERC-2015-STG-679528), financiado com uma Starting Grant do Conselho Europeu de Investigação (ERC) sob o programa de investigação e inovação Horizon2020 da União Europeia (http://postdata.linhd.es).

Mariana Curado Malta agradece ao Politécnico do Porto pela atribuição da licença sem vencimento equiparada a bolseiro que lhe permite trabalhar no projecto POSTDATA, uma excelente experiência de investigação em Madrid.

As autoras agradecem aos responsáveis das diferentes bases de dados por lhes darem acesso a informação técnica das mesmas e por terem contribuído activamente para a compreensão dos respectivos modelos de dados.

5. Fontes de Informação

ANTONELLI, R. Repertorio metrico della scuola poetica siciliana. Palermo: Centro di Studi Filologici e Linguistici Siciliani, 1984.

BAKER, T.; COYLE, K. - Guidelines for Dublin Core Application Profiles [em linha] [Consult. 12 abr. 2016]. Disponível em WWW:URL:http://dublincore.org/documents/profile-guidelines/.

BERMÚDEZ-SABEL, H.; CURADO MALTA, M.; GONZALEZ-BLANCO, E. - Towards Interoperability in the European Poetry Community: the Standardisation of Philological Concepts. Em Lecture Notes in Computer Science [em linha]. [S.l.]: Springer International Publishing, (2017), p. 156–165. [Consult. 30 mai. 2017]. Disponível em WWW:URL:https://doi.org/10.1007%2F978-3-319-59888-8_14.

BERNERS-LEE, T. *et al.* - The semantic web. Scientific American. *284*:5 (2001), 28–37.

BILLY, D. Contrafactures des modèles troubadouresques dans la poésie catalane. Le rayonnement des troubadours. ed. by Anton Touber. Amsterdam: Rodopi, 1998 (pp.51-74).

BRUNNER, H.; Wachinger, B.; Klesatschke, E., Repertorium der Sangsprüche und Meisterlieder des 12. bis 18. Jahrhunderts. Tubingen: Niemeyer, 1986-2007.

CANETTIERI, P; PULSONI C. Contrafacta gallego-portoghesi. Medioevo y Literatura. Actas del V Congreso de la Asociación Hispánica de Literatura Medieval. ed. by Juan Paredes. Granada, 1995. Vol. I, (pp. 479-498).

COYLE, K. - Application Profiles: An Overview. Em CURADO MALTA, M.; BAPTISTA, A. A.; WALK, P. (Eds.) - Developing Metadata Application Profiles. 1st. ed. Hersey, PA: IGI Global. (2017), 1–15.

CURADO MALTA, M. - Me4MAP: a method for the development of Dublin Core Application ProfilesDublin Core Metadata Initiative. [Em Linha]. DCMI. [Consult. 1 jun. 2017]. Disponível em WWW:URL: https://www.youtube.com/watch?v=Ff73npqlx7A.

CURADO MALTA, M.; BAPTISTA, A. A. - A method for the development of Dublin Core Application Profiles (Me4DCAP V0.2): detailed description. Em FOULONNEAU, M.; ECKERT, K. (Eds.) - Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013 [Em linha]. Lisbon: Dublin Core Metadata Initiative, (2013), p. 90–103. [Consult. 30 mai. 2017]. Disponível em WWW:URL:http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/178/81.

CURADO MALTA, M.; CENTENERA, P.; GONZALEZ-BLANCO, E. - Using Reverse Engineering to to Define a Domain Model: The case of Development of a Metadata Application Profile for the European Poetry. Em CURADO MALTA, M.; BAPTISTA, A. A.; WALK, P. (Eds.) - Developing metadata application profiles. 1st. ed. Hershey PA: IGI Global. (2017), 146–180.

DCMI - DCMI Glossary [em linha]. Dublin Core Metadata Initiative. [Consult. 22 mai. 2017]. Disponível em WWW:URL:http://wiki.dublincore.org/index.php/Glossary.

EVEN-ZOHAR, Itamar - Papers in Historical Poetics. [S.l.]: Tel Aviv: Porter Institute for Poetics and Semiotics, 1978.

GÓMEZ BRAVO, A. M. Repertorio métrico de la poesía cancioneril del siglo XV. Alcalá de Henares: Universidad de Alcalá de Henares, 1999.

GONZALEZ-BLANCO, E; del RIO, G.; MARTÍNEZ CANTÓN, C.; Linked open data to represent multilingual poetry collections. A proposal to solve interoperability issues between poetic repertoires. Proceedings of the LREC (Tenth International Conference on Language Resources and Evaluation) 2016 Workshop "LDL 2016 – 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources". 24 May 2016 – Portorož, Slovenia Edited by John P. McCrae, Christian Chiarcos, Elena Montiel Ponsoda, Thierry Declerck, Petya Osenova, Sebastian Hellmann. http://www.lrec-conf.org/proceedings/lrec2016/index.html.

GONZALEZ-BLANCO, E.; SELÁF, L. Megarep: A comprehensive research tool in medieval and renaissance poetic and metrical repertoires. Humanitats a la xarxa: món medieval / Humanities on the web: the medieval world, eds. L. Soriano - M. Coderch - H. Rovira - G. Sabaté - X. Espluga. Oxford, Bern, Berlin, Bruxelles, Frankfurt am Main, New York, Wien. Peter Lang, 2014 (pp. 321-332).

HEERY, R.; PATEL, M. - Application profiles: mixing and matching metadata schemas. Ariadne. (2000) *25*, 27–31.

MÖLK, U.; WOLFZETTEL, F. Répertoire métrique de la poésie lyrique françaisedes origines à 1350. Munchen: W. Fink Verlag, 1972.

NAETEBUS, G. Die nicht lyrischen Strophenformen des Altfranzösischen. Leipzig: S. Hirzel, 1891.

NILSSON, M.; BAKER, T.; JOHNSTON, P. - The Singapore Framework for Dublin Core Application Profiles [em linha]. Dublin Core Metadata Initiative. [Consult. 15 abr. 2017]. Disponível em WWW:URL:http://dublincore.org/documents/singapore-framework/.

NILSSON, M.; JOHNSTON, P.; POWELL, A. - Towards an Interoperability Framework for Metadata Standards. Em International Conference on Dublin Core and Metadata Applications [em linha]. [Consult. 30 Mai. 2017]. Disponível em WWW:URL:http://dcpapers.dublincore.org/pubs/article/view/835.

PAGNOTTA, L. Repertorio metrico della ballata italiana. Milano-Napoli: Ricciardi, 1995.

PARRAMON i BLASCO, J. Repertori mètric de la poesia catalana medieval. Barcelona: Curial, Abadia de Montserrat, 1992.

PILLET A.; CARSTENS, H., Bibliographie der Troubadours. Halle: M. Niemeyer, 1933.

RAYNAUD, G. Bibliographie des chansonniers français des XIII. [treizième] et XIV. [quatorzième] siècles: comprenant la description de tous les manuscrits, la table des chansons classées par ordre alphabétique de rimes et la liste des trouvères. Paris: Vieweg, 1884.

SOLIMENA, A. Repertorio metrico dello Stil novo. Roma: Presso la Societa, 1980.

TAVANI, G. Repertorio metrico della lingua galego-portoghese. Roma: Edizioni dell'Ateneo, 1967.

TOUBER, A. Les formes métriques dans la poésie médiévale en France et en Allemagne. Métriques du moyen âge et de la Renaissance, ed. by Dominique Billy. Paris: L'Harmattan, 1999 (pp. 288-301).

SOLIMENA, A. Repertorio metrico dei poeti siculo-toscani. Palermo: Centro di studi filologici e linguistici siciliani in Palermo, 2000.

UNSWORTH, John - Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? [em linha]. Kings's College, London, 2000. [Consult. 24 mar. 2017]. Disponível em WWW:URL:http://www.people.virginia.edu/~jmu2m/Kings.5-00/primitives.html.

W3C - RDF 1.1 Semantics - W3C Recommendation 25 February 2014 [em linha] [Consult. 22 maio. 2017]. Disponível em WWW:URL:http://www.w3.org/TR/rdf11-mt/.