Mariana Curado Malta

Modelação de dados poéticos: uma perspectiva desde os dados abertos e ligados

The modelling of poetic data: A perspective from linked open data

Resumo: O presente capítulo apresenta o trabalho de um projecto de investigação (POSTDATA) que se focaliza no tratamento de dados relacionados com poesia europeia (PE). Apresenta em particular o trabalho realizado no âmbito da criação de um modelo de dados que represente as necessidades informacionais da comunidade científica da PE. Este modelo de dados é um marco no desenvolvimento de um perfil de aplicação de metadados (MAP). Um MAP é um constructo no contexto dos dados ligados e abertos (LOD) que fornece informações de referência sobre o contexto global dos dados relacionados e suas restrições. O MAP é um importante constructo porque potencia a interoperabilidade de dados de uma comunidade. Este capítulo apresenta a modelação RDF enquadrada no desenvolvimento particular do modelo de dados específico à comunidade referida. Como trabalho futuro perspectiva-se a continuação do desenvolvimento do MAP para fornecer a esta comunidade um instrumento que potencia a interoperabilidade dos dados.

Palavras-clave: modelação de dados; dados ligados e abertos; modelação RDF; perfil de aplicação de metadados; poesía.

Abstract: This chapter presents the work of the research project POSTDATA, which focuses on the processing of data related to European Poetry (EP). It describes work done in the creation of a data model that represents the information needs of the EP scholarly community. This data model is a milestone in the development of a metadata application profile (MAP). A MAP is a construct in the context of linked open data (LOD) that provides reference information about the global context of related data and its constraints. MAP is an important construct because it enhances the interoperability of data in a community. This chapter presents the

Mariana Curado Malta, CEOS.PP, Politécnico do Porto, Laboratorio de Innovación en Humanidades Digitales, UNED

https://doi.org/10.1515/9783110585421-004

RDF model applied to the particular development of the data model for the community of EP scholars. As future work, we intend to continue the development of the MAP to provide this community with an instrument that enhances the interoperability of data.

Keywords: Data Modelling; Linked Open Data; RDF Model; Metadata Application Profile; Poetry.

> Where is the Life we have lost in living?
> Where is the living we have lost in knowledge?
> Where is the knowledge we have lost in information?
> *T.S. Elliot in* "*Choruses from the Rock*"

1 Introdução

Este excerto de poema de 1934 da obra de teatro de T.S. Elliot separa claramente os conceitos de conhecimento (*knowledge*) e informação (*information*) (Eliot 2009).[1] Mais tarde em 1989, Ackkof, um teórico da gestão de conhecimento, apresenta a sua célebre pirâmide que tipifica os conteúdos da mente humana (Ackoff 1999: 170–172). Essa pirâmide tem na base os "dados" (*data*), e depois sucessivamente a "informação" (*information*), o "conhecimento" (*knowledge*), a "compreensão" (*understanding*) e, no topo da pirâmide, a "sabedoria" (*wisdom*). Ackoff explica que os "dados" são tudo o que obtemos em bruto, são dados não processados, factos como "Estão 14 graus centígrados". A "informação" é já um passo adiante depois do processamento dos dados, segundo uma determinada lógica que nos possa levar a responder a questões do tipo "quando", "onde", "o quê" e "quem". Como exemplo podemos dizer que uma base de dados, seja ela de que tipo for, possui informação a partir dos dados aí guardados. Essa informação são dados aos quais foi atribuído um determinado significado através de uma ligação relacional. O patamar seguinte da pirâmide é o do "conhecimento", que aparece como algo que o indivíduo guarda, ou que memoriza, como por exemplo o 2x3=6. O conhecimento já poderá responder a questões do tipo "como". Um indivíduo pode memorizar as tabelas de multiplicar na sua cabeça, no entanto há um limite para esse processo de memorização. Por exemplo será difícil a memorização de tabelas de multiplicação do tipo 1234x345. Para alcançar a resposta

1 Por opção da autora, a ortografia do capítulo ora apresentado não segue as regras do Acordo Ortográfico de 1990.

a essa questão o indivíduo já necessita de uma habilidade analítica e cognitiva, o que acontece no patamar seguinte, o da "compreensão". Nesse patamar já se responde a questões do tipo "porquê". A compreensão é a síntese de conhecimento em novo conhecimento, ou em nova informação. Para terminar, no topo da pirâmide está a "sabedoria", um patamar que pressupõe ir além da compreensão, e que congrega todos os patamares anteriores conjugados com outro tipo de vivências e construções (como código de ética, moral, etc), algo muito próprio da mente humana e que, segundo Bellinger/Castro/Mills (2004), os computadores nunca terão.

Bellinger/Castro/Mills (2004) retiram da pirâmide o patamar "compreensão" e referem que essa dimensão é a necessária para passar de patamar em patamar. A pirâmide é para eles constituída por um patamar menos sendo então a base da pirâmide os "dados", depois a "informação" seguida do "conhecimento", estando a "sabedoria no topo:

- a passagem do patamar de "dados" para o de "informação" acontece devido à "compreensão de relações" (*understanding relations*)
- a passagem seguinte, isto é, entre "informação" e "conhecimento", acontece devido à "compreensão de padrões" (*understanding patterns*)
- a última passagem, entre "conhecimento" e "sabedoria", acontece devido à compreensão de princípios" (*understanding principles*)

O projecto POSTDATA enquadra-se na poesia europeia, mais especificamente num conjunto de projectos (cerca de vinte) de repositório digitais disponíveis na Web. Estes projectos estão alojados em servidores Web servidos por bases de dados locais, e que não comunicam uns com os outros (são chamados de "silos de informação"). O primeiro objectivo do projecto POSTDATA é o de fornecer meios para que estes sistemas possam tornar disponível no eco-sistema dos dados abertos e ligados (*Linked Open Data - LOD*), a informação que está confinada em cada uma das bases de dados. Esta informação dará à comunidade de investigadores de poesia novas possibilidades de relações de dados, e assim alcançar a dimensão de "sabedoria". À data da escrita deste capítulo os sistemas estão isolados, com dados estruturados modelados à necessidade local, e as relações entre os dados das diferentes bases de dados só poderão ser feitas manualmente ou através de tecnologias que utilizem APIs (*Application Programming Interface*), mas que estão aquém das possibilidades dos LOD. A partilha de dados através de API fornece um serviço bem menos aberto e menos inteligente que as possibilidades do paradigma LOD.

Para que estes dados possam estar disponíveis em LOD é essencial que sejam interoperáveis. A forma de atingir a máxima interoperabilidade de dados entre os diferentes sistemas é a definição de um perfil de aplicação de metadados, um

constructo da Web Semântica (ver: Coyle 2017; Nilsson/Baker/Johnston 2009). A equipa de trabalho do POSTDATA está a desenvolver um perfil de aplicação de metadados (*metadata application profile* - MAP), e esse trabalho inclui a definição de um modelo de dados comum à comunidade que deseja partilhar dados. Estamos conscientes de que almejar servir toda a comunidade de poesia europeia é um grande desafio, e de que não é possível apresentar uma solução perfeita à primeira, no entanto, integramos, como veremos, uma grande diversidade de tradições e de idiomas no nosso modelo.

O objectivo deste capítulo é apresentar os trabalhos que estão a ser desenvolvidos no momento da escrita deste capítulo no contexto LOD pelo projecto POSTDATA. Apresenta-se o que tem sido feito de forma a permitir no futuro a publicação de dados estruturados e interoperáveis de poesia no eco-sistema LOD.

Este capítulo divide-se em cinco secções. A segunda secção apresenta muito sucintamente o processo pelo qual tem de passar um trabalho de modelação de dados. A terceira secção apresenta em particular a modelação RDF. A quarta secção apresenta o conceito de interoperabilidade e o perfil de aplicação de metadados como o constructo que potencia a interoperabilidade semântica de dados, quando uma determinada comunidade de prática utiliza um para publicar dados LOD. Ainda nesta mesma secção se apresentam os trabalhos que estão a ser desenvolvidos para definir um perfil de aplicação de metadados para a poesia europeia. A quinta secção fecha este capítulo com considerações finais e trabalho futuro.

2 Modelação de dados

Qualquer sistema de recolha de dados deve estar pensado para guardar esses dados de uma forma estruturada. Para que os dados possam ser tratados é necessário que estejam modelados para servir um conjunto de requisitos funcionais. Os requisitos funcionais definem funções do sistema que se está a construir e que terá base informacional nos dados. Existem ocasiões em que na verdade os sistemas já existem, e portanto as base de dados já estão construídas mas por qualquer razão é necessário levantar o modelo de dados existente. Nestas situações realiza-se um processo de engenharia inversa (Müller *et al.* 2000) não havendo lugar ao levantamento dos requisitos funcionais. Este último caso é o do projecto POSTDATA que será apresentado na secção seguinte.

A modelação é um processo para organizar informação de uma realidade que queremos capturar que tem como objectivo a criação de uma representação

abstracta da realidade, para que seja possível criar bases de dados que sirvam sistemas de informação.

Na maior parte das vezes um processo de desenvolvimento de um modelo de dados é realizado para, como dissemos, responder a necessidades funcionais e por isso antes de realizar efectivamente a modelação é necessário levantar os requisitos funcionais. Para isso há técnicas na engenharia informática já bem definidas e documentadas, como por exemplo as apresentadas por Kotonya/ Sommerville (1998) ou Van Lamsweerde (2009).

Depois de identificar esses requisitos, realiza-se a modelação. Inicialmente realiza-se uma modelação conceptual que é independente da implementação tecnológica. A modelação conceptual é o processo de identificar as coisas do domínio que desejamos capturar. Por "coisa" referimo-nos à figura que reúne um conjunto de indivíduos que têm atributos em comum, ou dito de outra forma, que partilham as mesmas características. Tomando como exemplo a realidade de uma biblioteca, as "coisas" cuja informação queremos capturar são os livros e sua localização em determinada prateleira da biblioteca, os seus autores e editores, e as editoras. Cada uma destas coisas tem de facto atributos ou características em comum, por exemplo podemos identificar o livro como uma coisa (cujo conjunto de indivíduos seriam todos os livros da biblioteca) com os seguintes atributos (ou propriedades): título, sub-título, número total de páginas, data de edição, entre muitos outros. Um autor ou editor pode em termos abstractos representar uma mesma "coisa" porque ambos representam uma pessoa. Uma pessoa tem como atributos: apelido, nome, data de nascimento, local de nascimento, nacionalidade, e muitas outras coisas que dependem dos requisitos funcionais. Por exemplo caso seja importante enviar cartas convite para o lançamento de novos livros, será necessário ter o endereço postal, caso esse convite seja feito digitalmente possivelmente somente será necessário incluir o endereço eletrónico. No caso de outros domínios, uma pessoa pode ainda ter outras características como peso, altura, cor dos olhos, mais uma vez isso tudo depende dos requisitos funcionais.

As coisas estão associadas, e essa associação é também parte do modelo, no exemplo apresentado, um livro tem autor(es) e editor(es), e é lançado por uma editora, ou por várias ao longo do tempo (mas nunca por várias ao mesmo tempo). Estas associações têm cardinalidade, essa cardinalidade também é registada no modelo. Por exemplo, um livro tem um ou mais autores e um autor pode escrever um ou mais livros. No caso das editoras, se o registo for o de uma edição de um livro em particular, o livro tem somente uma e só uma editora, mas uma editora pode editar um ou mais livros.

O pioneiro na modelação de dados foi Peter Chen no seu artigo de 1976, onde apresenta o modelo Entidade-Relação (modelo ER) que ainda hoje é utilizado.

Nesse artigo seminal Chen propõe um modelo de dados que "incorpora informação semântica do mundo real" (Chen 1976: 9) e chama às coisas, "entidades", às características das coisas, "atributos", e "relacionamentos" à forma como as entidades se associam. Muito sucintamente, Chen apresenta uma técnica que inclui uma linguagem gráfica (ou notação) que define a entidade como um rectângulo, as associações como losangos ligados às entidades por linhas. Dentro do losango deve dar-se nome à relação, e no fim de cada linha deve escrever-se a cardinalidade da relação (Figura 3.1). Mais tarde surgiram outras notações para o modelo ER de Chen, nomeadamente como exemplo, a *Integration DEFinition for information modeling* (IDEF1X) (ver Bruce 1992), o pé-de-galinha (*crow's foot*) (ver Everest 1976), e a Unified Modeling Language (UML) (ver OMG 2009) com os seus diagramas de classe. Esta última é hoje muito utilizada em todas as suas vertentes na engenharia informática e é definida como "uma linguagem gráfica para visualização, especificação, construção, e documentação de artefactos de sistemas de objectos distribuídos" (OMG 2009). Um exemplo de diagrama de classes é apresentado na Figura 3.2 Este diagrama representa um excerto do modelo de dados desenvolvido pelo Projecto POSTDATA. No modelo podemos ler, por exemplo, "Uma L*ine hasFirstSyllable* uma (ou nenhuma) *Syllable*" e "Uma *Syllable isFirstSyllable* de uma (ou nenhuma) *Line*", significa isto que podemos definir as linhas de um poema (com as propriedades aí descritas: *lineNumber*, *nextPageNumber*, *nextColumnLabel*, *content*, etc.) e associar a primeira sílaba da linha (com todas as características da sílaba: *content*, *nucleus*, *onset*, *coda*, *weight*, *positionInWord*, etc). As sílabas seguintes da linha são definidas através da ligação "nextSyllable", isto é "Uma *Syllable nextSyllable* uma (ou nenhuma, no caso de ser a última sílaba da linha em questão) *Syllable*".

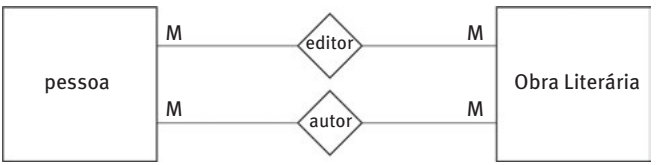

Figura 3.1: Um exemplo do modelo Entidade-Relação de Chen (1976).

3 Modelo de dados RDF

A modelação em RDF (ver W3C 2004) baseia-se na figura do triplo: um triplo é composto por um sujeito, um objecto e um predicado (Figura 3.3).

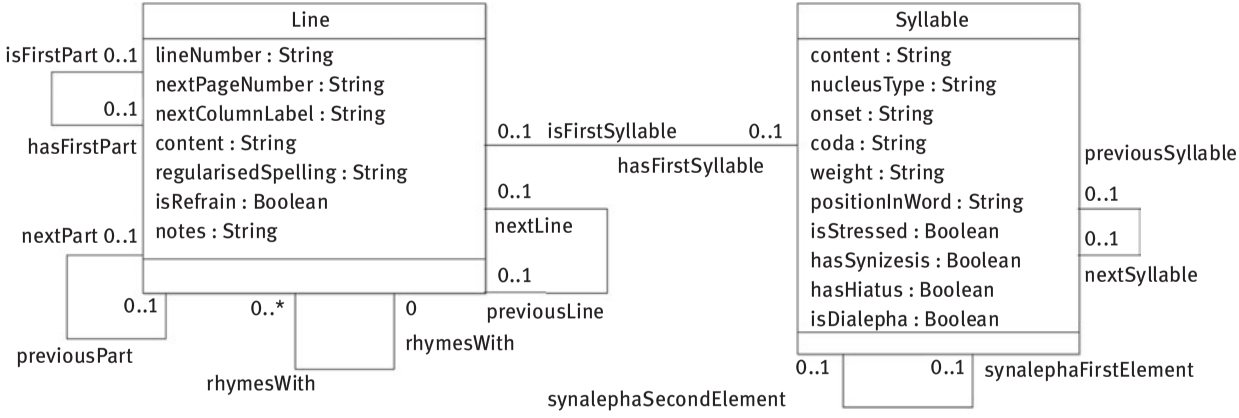

Figura 3.2: Exemplo de um diagrama de classes UML.

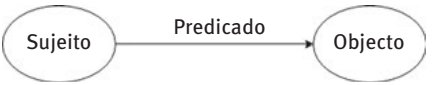

Figura 3.3: O triplo do modelo de dados RDF.

A Figura 3.4 apresenta um exemplo de um triplo representando dados, aí definimos o título de uma obra literária como sendo "Ondas do mar de Vigo". *Opus* é um conceito definido pelo modelo de dados POSTADATA como uma criação artística (ex: poema) na sua concepção abstracta.

Figura 3.4: Triplo que define o título de uma obra literária.

Um modelo de dados utilizando o paradigma RDF é composto por um conjunto de triplos, a esse conjunto chama-se *grafo*. A Figura 3.5 apresenta um grafo definindo uma obra literária a partir do modelo de dados POSTDATA.[2]

Num triplo o "Sujeito" e o "Predicado" são sempre representados por um URI (identificador uniforme de recurso), o "Objecto" pode ser representado tanto por uma cadeia de caracteres como por um URI. Um URI é uma forma de identificar um recurso no eco-sistema LOD.[3] Podemos ver que o título da obra literária é representado por um rectângulo na Figura 3.5, isto significa que o "Objecto" é, neste caso, uma cadeia de caracteres. Na Figura 3.5 temos outros "Objectos" que são URIs, por exemplo os valores do vocabulário controlado RFC4646[4] para definir o idioma, ou do vocabulário controlado que define os tipos de "Redaction",[5] ou ainda uma entidade existente num outro *dataset*[6] (por exemplo *datasets* de domínio público como o projecto http://dbpedia.org (acedido em 10/01/2018), ou uma qualquer entrada num *dataset* disponível no eco-sistema LOD). Num triplo os URIs são representados por um círculo. Os "Predicados" são termos de

2 Para simplificar a apresentação e porque o modelo POSTDATA ainda não está concluído, toda a informação aqui apresentada é fictícia, tanto a nível dos dados da obra como dos vocabulários RDF.
3 Equivale à chave primária de uma tabela numa base de dados relacional, que identifica univocamente uma entrada na tabela.
4 Um vocabulário controlado define um conjunto de possíveis valores para uma propriedade ou atributo. Cada valor é definido por um URI. Para o RFC4646 ver http://www.ietf.org/rfc/rfc4646.txt, acedido em 10/012/2018.
5 Uma "Redaction" é uma manifestação física de uma "Opus". O vocabulário controlado que define os possíveis valores para a propriedade "typeOfRedaction" está a ser desenvolvido desenvolvido no momento de escrita deste texto pelo que ainda não foi implementado informaticamente. O nome do vocabulário e o URI indicados no grafo são por essa razão fictícios.
6 Um *dataset* é uma colecção de dados.

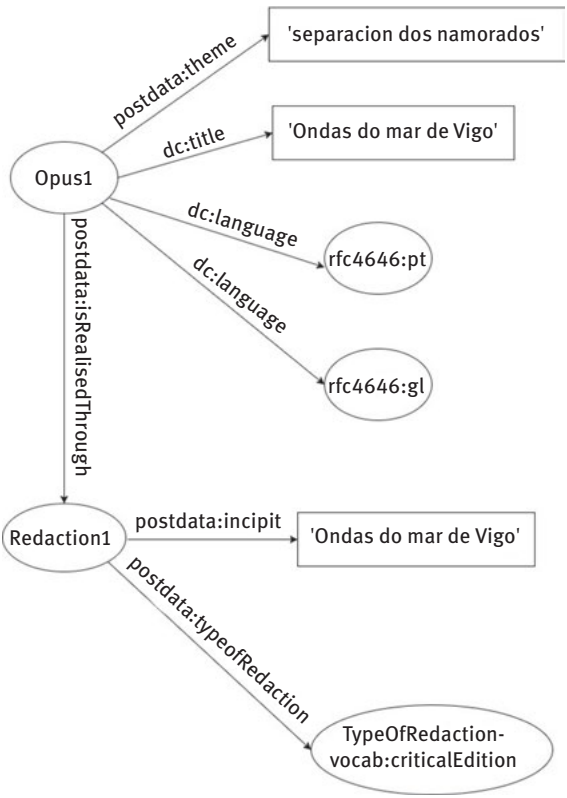

Figura 3.5: Um grafo de triplos definindo a obra literária Opus1 (modelo de dados POSTDATA).

vocabulários RDF[7] que são identificados também por um URI. Um vocabulário RDF é um conjunto de elementos de metadatos, com um conjunto de regras para a sua utilização. O primeiro vocabulário RDF padrão a surgir, e muito utilizado hoje em dia pela sua importância e generalidade de contexto de aplicação, é o Dublin Core Metadata Terms[8] mais comummente conhecido por "dcterms" ou "dc". Existem outros vocabulários RDF tais como *foaf*, *schema.org*, *darwin core*, *good relations*,[9] entre muitos outros.

A vantagem dos URIs sobre as cadeias de caracteres é a de que os URIs permitem a ligação dos dados, conectando diferentes *datasets*, que por sua vez

7 Também chamados de "Esquema de metadados".

8 Ver http://dublincore.org/documents/dcmi-terms/, acedido aem 12/01/2018.

9 Ver http://xmlns.com/foaf/spec/, http://schema.org, http://rs.tdwg.org/dwc/ e http://www.heppnetz.de/projects/goodrelations/, respectivamente acedidos em 12/01/2018.

também estão ligados a outros *datasets*. Esta possibilidade, que não tem limites, torna o eco-sistema LOD uma grande base de dados mundial, onde é possível colocar máquinas inteligentes a inferir sobre os dados de uma forma muita mais eficaz, já que têm acesso a muitos dados estruturados e interoperáveis. Passamos de um mundo fechado em servidores Web, na Web de Documentos, para um mundo aberto e ligado. A Web de Documentos é a Web que utilizamos todos os dias através de *browsers* que lêem documentos em HTML. A Web de dados ou eco-sistema LOD é um espaço onde máquinas comunicam com máquinas e onde, utilizando tecnologias LOD, inferem sobre os dados apresentando resultados ao utilizador final na Web de Documentos. Estas duas Webs não se substituem, como vemos, mas co-habitam de facto (Heath/Bizer 2011). A modelação RDF aliada às tecnologias LOD permite então esta abertura dos dados, possibilitando novas descobertas às comunidades de prática, potenciando a possibilidade de ascender ao patamar "sabedoria" apresentado por Ackoff (1999) na sua pirâmide.

Na próxima secção explicaremos a importância da interoperabilidade dos dados e de que forma ela pode ser potenciada.

4 Interoperabilidade e perfil de aplicação de metadados

A IEEE define no seu glossário[10] "interoperabilidade" como a "capacidade de um sistema, ou de um produto, trabalhar com outros sistemas ou produtos sem esforço especial por parte do cliente. A interoperabilidade é possível graças à implementação de padrões". No caso do eco-sistema LOD lidamos com interoperabilidade semântica. A interoperabilidade semântica é "a capacidade de diferentes tipos de computadores, redes, sistemas operativos e aplicações, trabalhar em conjunto de forma eficaz, sem comunicação prévia, para trocar informações de forma útil e significativa".[11]

No paradigma LOD existe um constructo que potencia a interoperabilidade semântica: trata-se do perfil de aplicação de metadados (MAP) (ver Coyle 2017). Um MAP fornece informações de referência sobre o contexto global dos dados

10 Ver https://www.standardsuniversity.org/article/standards-glossary/, acedido em /10/01/2018.
11 Definição retirada do glossário da DCMI: http://www.dublincore.org/documents/usageguide/glossary/, acedido em 12/01/2018.

relacionados e suas restrições. Na verdade um MAP não é mais do que um modelo de dados que associa a cada uma das entidades, propriedades e relações desse modelo, os termos de vocabulários RDF, e que define ainda restrições sobre cada um destes termos.

Segundo Coyle/Baker (2009) para formalizar um MAP é necessário (obrigatório) definir os seus requisitos funcionais, o modelo de dados, e a descrição do perfil (chamado de *description set profile*). Para além disso é importante (mas não obrigatório) fornecer guias de utilização e de sintaxe para que os agentes que irão aplicar o MAP aos dados que querem publicar possam fazer essa aplicação o mais correctamente possível.

4.1 O perfil de aplicação de metadados do POSTDATA

No momento de escrita deste capítulo existem na Web de Documentos muitos repositórios com informação sobre *corpora* de diferentes tradições, com dados que não são interoperáveis. Na verdade, cada repositório foi definido à sua necessidade, à medida de um paradigma egoísta (Jannidis/Flanders 2013). Isso aconteceu porque os desenvolvedores das bases de dados que servem os repositórios não tinham em mente a partilha interoperável dos dados, preocupando-se somente com os seus próprios requisitos funcionais. Uma vez que é desejável que os dados da comunidade de prática referente à investigação em poesia possam ser partilhados por toda a comunidade, será importante passar desse paradigma egoísta para o altruísta (Jannidis/Flanders 2013). Um paradigma altruísta tem em conta os requisitos de toda a comunidade.

Na prática não é possível responder às necessidades funcionais de toda a comunidade de poesia uma vez que é impossível conhecer quem são todos os agentes que irão partilhar dados de poesia na Web de Dados. Este problema tem a ver com a forma aberta e universal da Web de Dados, ao contrário de uma modelação para um sistema dentro de uma organização, onde as fronteiras estão bem definidas e os requisitos funcionais são finitos e definidos pelas pessoas da organização em questão. Para ultrapassar da melhor forma esta questão a equipa POSTDATA tratou de integrar no seu estudo um conjunto de repositórios que apresentassem *corpora* de diferentes tradições e culturas e de diferentes períodos no tempo. Os *stakeholders* do projecto podem ser localizados geograficamente no mapa https://goo.gl/9MCWrv (acedido em 12/01/2018). Estes projecto digitais trabalharam com a equipa POSTDATA partilhando informação e as estruturas das base de dados dos seus projectos.

A equipa POSTDATA está a desenvolver um MAP e a seguir o método Me4MAP (ver Curado Malta/Baptista 2013) nesse trabalho.

Na verdade a obrigatoriedade da definição dos requisitos funcionais não nos parece adequada num caso em que os dados já existam. Isto porque o levantamento dos requisitos funcionais já foi realizado por quem desenvolveu as base de dados, tendo realizado a modelação de dados baseado nesses requisitos. O trabalho pode então partir mais à frente, analisando os modelos já existentes no caso de estes estarem bem documentados e executados, ou realizando um processo de engenharia inversa no caso de não haver acesso às estruturas de dados. No caso de POSTDATA os requisitos funcionais não foram efectivamente definidos. Realizou-se um conjunto de trabalhos de recolha de informação para definição do modelo de dados POSTDATA: o processo de re-engenharia já referido que constou no estudo de algumas interfaces Web de repositórios de *stakeholders* (para mais detalhes ver Curado Malta/Centenera/González-Blanco 2017; Bermúdez-Sabel/Curado Malta/González-Blanco 2017), a análise de um inquérito a utilizadores finais de dados poéticos para investigação em poesia, e a elaboração de um conjunto de casos de uso com recursos reais retirados das bases de dados dos *stakeholders*. Este modelo de dados POSTDATA representa todas as necessidades informacionais de todos os *stakeholders*. Representando estes *stakeholders* uma importante parte das diferentes tradições, culturas e idiomas da poesia europeia, pensamos que o modelo que iremos obter irá servir uma grande parte da comunidade de investigadores em poesia europeia. O modelo de dados POSTDATA está no momento da escrita deste artigo em fase final de desenvolvimento.

A terceira etapa de desenvolvimento do MAP é a definição da descrição de perfil que declara os termos dos vocabulários RDF que mais se adequam à entidade/propriedade/relação do modelo de dados. Boas práticas indicam que se deve sempre procurar primeiro nos vocabulários RDF padrão e mais utilizados. Isto porque a utilização de vocabulários RDF padrão é essencial para a implementação de interoperabilidade. No caso de alguma das entidades/propriedades/relações do modelo de dados não poder ser descrita por nenhum termo de nenhum vocabulário RDF existente, então será necessário criar um vocabulário RDF novo. Esta solução deve ser seguida somente como último recurso, porque ela faz necessariamente diminuir a interoperabilidade dos dados com comunidades fora da comunidade de prática que esteja a desenvolver os trabalhos. A ferramenta *Linked Open Vocabularies* (LOV)[12] é uma iniciativa que reúne à data da escrita deste capítulo 627 vocabulários RDF, e que pode servir de ponto de partida para a procura dos termos de vocabulários RDF necessários a descrever o modelo de dados. Há no entanto outros recursos para a busca que o desenvolvedor de MAPs não deve descurar, como por exemplo o *Open Metadata*

12 Ver http://lov.okfn.org/dataset/lov/, acedido em 10/01/2018.

Registry, ou o *Basel Register of Thesauri* ou *Ontologies & Classifications*.[13] A equipe POSTDATA está, à data da escrita deste capítulo, a iniciar os trabalhos de definição da descrição do perfil.

O modelo de dados está ainda em desenvolvimento mas há duas entidades que são o âmago do modelo e que podem ser apresentadas aqui nesta secção. Tratam-se da "Opus" e da "Redaction". A primeira representa uma criação artística no seu sentido abstracto que tem obrigatoriamente de ser em verso (ex: poema, peça de teatro, cantiga). A segunda representa já a manifestação física dessa criação artística. Estas duas entidades estão associadas através de uma relação de nome "isRealisedThrough".

O modelo de dados POSTDATA está a ser terminado e a sua descrição estará brevemente disponível no site do projecto[14] pelo que convidamos o leitor a consultá-lo.

5 Considerações finais e trabalho futuro

O projecto POSTDATA nasceu de uma bolsa *Starting Grant* do *European Research Council* atribuída à investigadora Elena González-Blanco. Este projecto tem como um dos seus objectivos a criação de meios para que os investigadores em poesia possam partilhar os seus dados no eco-sistema *Linked Open Data* (LOD), ou dados abertos e ligados. A vantagem deste eco-sistema sobre todos os outros existentes é a sua abertura e globalidade. Os dados abertos e ligados são uma grande base de dados mundial disponível para consulta e utilização dos dados, e que permite a agentes inteligentes (computadores com tecnologia LOD) inferir sobre dados que porventura nunca antes tinham estado juntos numa mesma base de dados. Ao disponibilizarmos dados de diferentes fontes num mesmo espaço e tendo como suporte a mesma tecnologia estamos a abrir novos caminhos para a investigação.

O projecto POSTDATA está neste momento a desenvolver um perfil de aplicação de metadados (MAP), um constructo que potencia a interoperabilidade semântica em LOD. O objectivo é que todos os *stakeholders* do projecto em particular, e todos os consumidores de dados de poesia europeia em geral possam partilhar dados no eco-sistema LOD. Um MAP fornece informações de referência sobre o contexto global dos dados relacionados e suas restrições.

13 Ver http://metadataregistry.org/ e http://www.bartoc.org/, respectivamente. Acedidos em 15/01/2018.

14 Ver http://postdata.linhd.es, acedido em 15/01/2018.

Este trabalho de desenvolvimento de MAP segue um método existente de nome Me4MAP (ver Curado Malta/Baptista 2013). Um dos marcos do Me4MAP é a definição de um modelo de dados.

A criação de um modelo de dados é um trabalho já muito documentado na comunidade de desenvolvimento de *software*, no entanto estes trabalhos estão enquadrados em contextos de desenvolvimento realizados em organizações fechadas, com necessidades e utilizadores bem definidos. O eco-sistema LOD é um contexto muito aberto onde é impossível definir todo o tipo de utilizadores e suas necessidades. Esta questão coloca um grande desafio aos desenvolvedores de MAPs.

Para desenvolver um MAP que sirva a comunidade de poesia europeia a equipa do POSTDATA analisou as estruturas de dados das bases de dados que servem os repositórios de poesia de alguns *stakeholders*, e ainda utilizou técnicas de engenharia inversa de forma a obter as necessidades informacionais de interfaces Web de *stakeholders* que não disponibilizaram as estruturas de bases de dados à equipa POSTDATA. Além disso ainda foi criado um conjunto de casos de estudo a partir de recursos reais retirados de outros repositórios digitais de poesia, e ainda se analisou um inquérito a utilizadores finais de dados poéticos. O modelo de dados final será em breve publicado no site do projecto POSTDATA.

O modelo de dados é um marco no desenvolvimento do MAP. Depois de terminado, segue a definição da descrição do perfil. A descrição do perfil define os termos dos vocabulários RDF que mais se adequam à entidade/propriedade/relação do modelo de dados e adiciona as restrições necessárias a cada termo. A equipa POSTDATA está neste momento a iniciar este trabalho. As restrições irão implicar possivelmente a definição de vocabulários controlados para determinados termos, permitindo um controlo maior sobre os dados introduzidos nos *datasets*, potenciando assim ainda mais a interoperabilidade dos dados. Além disso, será possivelmente necessário desenvolver um vocabulário RDF para descrever as propriedades do modelo de dados para as quais não sejam encontrados termos de vocabulários RDF adequados.

Referências bibliográficas

Ackoff, R. L. (1999): *Ackoff's Best: His Classic Writings on Management* (1 edition). New York: Wiley.

Bellinger, Gene/Castro, Durval/Mills, Anthony (2004): *Data, Information, Knowledge, & Wisdom*. [Em linha: http://www.systems-thinking.org/dikw/dikw.htm, 09/01/2018]

Bermúdez-Sabel, Helena/Curado Malta, Mariana/González-Blanco, Elena (2017): "Towards Interoperability in the European Poetry Community: The Standardization of Philological Concepts". Em: Gracia, Jorge *et alii* (eds.): *Language, Data and Knowledge. First*

International Conference. LDK 2017, Galway, Ireland, June 19-20, 2017, Proceedings. Cham: Springer International Publishing, pp. 156–165. https://doi.org/10.1007/978-3-319-59888-8_14.

Bruce, Thomas A. (1992): *Designing quality databases with IDEF1X information models*. New York: Dorset House Publishing.

Chen, Peter P.-S. (1976): "The Entity-relationship Model – Toward a Unified View of Data". *ACM Transactions on Database Systems*, 1.1, pp. 9–36. [https://doi.org/10.1145/320434.320440]

Coyle, K. (2017): "Application Profiles: an overview". Em: Curado Malta, Mariana/Baptista, Ana Alice/Walk, Paul (eds.): *Developing Metadata Application Profiles*, IGI Global, pp. 1–15.

Coyle, Karen/Baker, Thomas (2009): *DCMI: Guidelines for Dublin Core Application Profiles (Working Draft)*. [Obtido 15/01/2018, de http://dublincore.org/documents/profile-guidelines/]

Curado Malta, Mariana/Baptista, Ana Alice (2013): "A Method for the Development of Dublin Core Application Profiles (Me4DCAP V0. 2): A Description". Em: *International Conference on Dublin Core and Metadata Applications*, Lisbon: DCMI, pp. 90–103.

Curado Malta, Mariana/Centenera, Paloma/González-Blanco, Elena (2017): "Using reverse engineering to define a domain model". Em: Curado Malta, Mariana/Baptista, Ana Alice/Walk, Paul (eds.): *Developing Metadata Application Profiles*, IGI Global, pp. 146–180). [Obtido de http://recipp.ipp.pt/handle/10400.22/9541]

Eliot, Thomas S. (2009): *Collected Poems 1909–1962*, London: Faber & Faber.

Everest, Gordon (1976): "Basic data structure models explained with a common example". Em: *Proceedings Fifth Texas Conference on Computing Systems, Austin, TX, 1976 October 18–19*. Long Beach, CA: IEEE Computer Society Publications Office, pp. 39–46.

Heath, Tom/Bizer, Christian (2011): *Linked Data: Evolving the Web into a Global Data Space* (1st ed., Vol. 1). Morgan & Claypool Publishers.

Jannidis, Fotis/Flanders, Julia (2013): "A concept of data modeling for the humanities". Em: *Digital Humanities 2013: Conference Abstracts*. Lincoln, USA: Center for Digital Research in the Humanities, pp. 237–239.

Kotonya, Gerald/Sommerville, Ian (1998): *Requirements Engineering: Processes and Techniques* (1st ed.). Wiley Publishing.

Lamsweerde, Axel Van (2009): *Requirements Engineering: From System Goals to UML Models to Software Specifications* (1 st ed.). Chichester, England/Hoboken, NJ: Wiley.

Müller, Hausi A./Jahnke, Jens H./Smith, Dennis B./Storey, Margaret-Anne/Tilley, Scott R./Wong, Kenny (2000): "Reverse Engineering: A Roadmap". Em: *Proceedings of the Conference on The Future of Software Engineering*. New York: ACM, pp. 47–60. [https://doi.org/10.1145/336512.336526]

Nilsson, Mikael/Baker, Thomas/Johnston, Peter (2009): "DCMI: Interoperability Levels for Dublin Core Metadata". [Obtido 09/01/2018, de http://dublincore.org/documents/interoperability-levels/]

OMG (2009): "About the Unified Modeling Language Specification Version 2.2." [Obtido 10/01/2018,de http://www.omg.org/spec/UML/2.2/]

W3C (2004): *RDF Primer*. [Em linha, https://www.w3.org/TR/rdf-primer/, consultado 10/01/2018]