cadernos BAD

7ª conferência luso-brasileira acessoaberto Instituto Politécnico de Viseu - Portugal

Vocabulários controlados na descrição de dados de investigação no Dendro

Yulia Karimova
INESC TEC – Faculdade de Engenharia da Universidade do Porto
ylaleo@gmail.com

João Aguiar Castro
INESC TEC – Faculdade de Engenharia da Universidade do Porto
joaoaguiarcastro@gmail.com

Resumo

Tendo em conta que a reutilização dos dados de investigação contribui para aumentar a visibilidade dos resultados científicos, a gestão de dados é cada vez mais valorizada. Neste contexto a descrição de dados, apesar de exigente, é uma tarefa muito importante. O Dendro, plataforma colaborativa de gestão de dados intermédia, está a ser desenvolvida na Universidade do Porto para auxiliar os investigadores na descrição de dados de investigação assim que começam a ser produzidos para posterior transferência para repositórios. A falta de tempo, ferramentas, experiência ou conhecimento na gestão dos dados, assim como as questões de criação de metadados de boa qualidade estão entre limitações identificadas na descrição e publicação de dados por parte de investigadores. Com este trabalho pretende-se introduzir vocabulários controlados para simplificar e normalizar a descrição de dados na plataforma Dendro. Estes são apresentados sob a forma de listas de escolhas e formalizados como ontologias. Os vocabulários controlados para descrição de dados de investigação foram testados em colaboração com investigadores do domínio da Produção de Hidrogénio. Os resultados de análise de qualidade de metadados sugerem que o uso de vocabulários controlados simplificou o processo de descrição, obtendo-se uma descrição mais completa e correta.

Palavras-chave: Gestão de dados de investigação, metadados, vocabulários controlados, qualidade de metadados, Dendro

Controlled vocabularies in the description of research data on Dendro

Abstract

Considering that the reuse of research data contributes to an increase in the visibility of scientific results, data management is increasingly valued. In this context, data description, despite being demanding, is an important task. Dendro, a collaborative data management platform, is being developed at the University of Porto to help researchers in the description of research data from their production to the subsequent transfer to repositories. The lack of time, tools, experience, knowledge of data management and supply of good quality metadata are among the limitations identified in the description and publication of data by researchers. This work proposes controlled vocabularies to simplify and normalize the data description in the Dendro platform. They are presented in the form of drop-down lists and formalized as ontologies. The controlled vocabularies for research data description were tested in collaboration with researchers from the "Hydrogen Production" domain. The results of the analysis of the metadata quality suggest that the use of controlled vocabularies has simplified the process of description, obtaining a more complete and accurate description.

Keywords: Research data management, metadata, controlled vocabularies, metadata quality, Dendro

Introdução

Atualmente, com o desenvolvimento de novas tecnologias de informação começam a surgir novas formas e métodos de cooperação científica. Consequência disso são o aparecimento de vários repositórios de dados de investigação que ajudam na promoção, divulgação e valorização dos mesmos. O acesso, partilha e reutilização de dados de investigação, ao diminuir a redundância na criação de novos dados, permite gerar conhecimento científico de forma mais expedita.

Devido à sua diversidade e complexidade, os dados de investigação, são por normas difíceis de interpretar (Smith, Seligman e Swarup, 2008), o que compromete a sua reutilização. Neste contexto os metadados têm um papel fundamental, uma vez que permitem o registo da informação necessária para que se possam compreender os dados, para além de auxiliarem na preservação, localização e recuperação dos objetos digitais. A publicação de dados, sem que as condições para a reutilização dos mesmos estejam assegurada, corresponde a um desperdício de recursos. Por este motivo, os investigadores devem assegurar que os dados são acompanhados por metadados suficientemente detalhados e rigorosos. No entanto, a descrição de dados é uma tarefa exigente e demorada,

que se não for devidamente suportada, pode desmotivar os investigadores a enveredar por atividades relacionadas com a gestão de dados.

A falta de tempo e ferramentas, falta de experiência e conhecimento técnico são alguns fatores que condicionam a partilha e publicação de dados de investigação em diversos domínios de investigação. Torna-se, portanto, importante fornecer aos investigadores ferramentas que facilitem na captura de metadados e potenciem a divulgação dos dados de investigação que estes recolheram.

Apesar de vária normas de metadados estarem disponíveis, estas são complexas, e por vezes inexistentes em certos domínios (Swan e Brown, 2008). Por outro lado, a usabilidade também influencia na utilização de um repositório de dados e pode afetar o envolvimento dos utilizadores (Zhang, Maron e Charles, 2013). Para além disso, muitos projetos não têm uma infraestrutura integrada de gestão dados, o que muitas vezes resulta no uso de ferramentas para gestão de dados não adequadas. Este problema surge especialmente em projetos pequenos, que têm exigências mínimas na utilização das ferramentas tecnológico, com recursos humanos e conhecimentos na curadoria digital limitados (Akmon *et al.*, 2011;Borgman, Wallis, Enyedy, 2007).

Tendo em conta a importância dos desafios reconhecidos tanto na gestão de dados de investigação como no processo de descrição de dados, a plataforma Dendro tem vindo a ser desenvolvida na Universidade do Porto. Através do Dendro pretende-se dar o apoio necessário aos investigadores na preparação dos dados de investigação, nomeadamente na captura atempada de metadados, para que os dados sejam transferidos, se possível, para repositórios externos, em condições de serem recuperados e reutilizados por terceiros.

Um dos objetivos deste trabalho consiste na agilização e facilitação do processo de descrição de dados na plataforma Dendro, de forma a motivar o interesse dos investigadores na organização e publicação dos dados que produzem. Neste contexto os vocabulários controlados apresentam-se como uma boa ferramenta, que pode facilitar o processo da descrição de dados e ao mesmo tempo contribuir para uma melhoria na qualidade dos metadados. Em muitos casos, os vocabulários controlados definem o conteúdo admissível para um determinado elemento descritivo e podem ser incorporados nos procedimentos de automatização de forma simples, contribuindo assim para o controlo de qualidade (Bermudez *et al.*, 2011). Com esta ideia em mente este estudo propõe a criação de vocabulários controlados para a descrição de dados em domínios científicos específicos no Dendro.

Dendro

O Dendro é uma plataforma colaborativa de gestão de dados intermédia, que permite a organização e documentação de dados desde o momento em que começam a ser

produzidos, sobretudo para investigadores na cauda longa da ciência, que se refere aos investigadores individuais e pequenos laboratórios, que não têm possibilidade de realizar gestão de grandes quantidades de dados. Esse tipo de dados é mais difícil de encontrar, reutilizar e preservar (Heidorn, 2008). O Dendro consiste numa interface web e tem a capacidade para exportar pacotes de dados e metadados para repositórios destinados à sua preservação e divulgação.

Esta plataforma facilita a criação de metadados com a utilização de vários esquemas de metadados existentes, como por exemplo *Dublin Core* e *Friend of a Friend*, assim como descritores específicos a domínios de investigação. A plataforma utiliza descritores, baseados em ontologias criadas em colaboração com investigadores dos mais diversos domínios e um curador de dados (Silva *et al.*, 2014). Os investigadores, como especialistas no seu domínio e produtores dos dados, fornecem conhecimento especializado sobre o contexto dos dados a descrever, enquanto o curador tem uma perspetiva mais abrangente sobre a gestão de dados de investigação.

De forma a adicionar novos descritores no Dendro, foram criadas ontologias para vários domínios científicos, tais como: Produção de Hidrogénio, Biodiversidade, Química Analítica, entre outros. Uma vez carregadas no Dendro, estas ontologias contribuem com novos descritores que podem ser combinados pelos investigadores para obter registos de metadados o mais abrangente quanto possível (Castro *et al.*, 2015; Silva *et al.*, 2014).

Decorrente da interação com os investigadores surgiu a necessidade de desenvolver vocabulários controlados, de modo a melhorar a qualidade das descrições e a tornar a tarefa de descrição mais apelativa para os investigadores.

Vocabulários controlados

Em muitos casos, a expressão vocabulário controlado define o conteúdo admissível para um determinado elemento de metadados e pode ser facilmente incorporado nos procedimentos de automatização, contribuindo para o controlo de qualidade, ao fornecer aos utilizadores uma lista de entradas permitidas para os elementos de metadados específicos (Bermudez et al., 2011).

O uso de vocabulários controlados permite ultrapassar as seguintes limitações (National Information Standards Organization, 2005):

- Diferença na interpretação do léxico (variações conceptuais);
- Diferença na utilização de expressões lexicais (variações sociais);
- Expansão da significação do léxico (polissemia);
- Desconhecimento do léxico.

Segundo Hedden (2010) há vários tipos de vocabulários controlados: a lista de conceitos, o ficheiro de autoridade, a taxonomia, os tesauros. Além disto, os vocabulários controlados podem ser abertos ("open-ended"), onde novos conceitos podem ser adicionados ao longo do tempo (Harpring, 2010), e fechados, onde não existe a possibilidade de inserir as novas sugestões.

Algumas implementações de vocabulários controlados são implementadas através de listas de escolhas, que permitem mostrar todas as opções existentes e ajudam ao utilizador escolher um termo num conjunto de conceitos pré-definidos. Normalmente estas listas não incluem sinónimos e são mais fáceis de implementar, sendo adequadas para a melhoria de qualidade na descrição de dados e, portanto, optou-se por seguir esta abordagem no desenvolvimento deste trabalho.

Seleção do caso de estudo

O envolvimento dos investigadores no desenvolvimento das ferramentas que suportam as atividades de descrição de dados é tido como muito importante. Isto porque, os investigadores são as pessoas com maior aptidão para indicar a informação necessária para permitir interpretar os dados que produziram. O problema dos esquemas de metadados é a sua especialização, o que impede que os investigadores os usem, muitas vezes por não se sentirem familiarizados com a terminologia utilizada. Esta limitação leva a que os investigadores devam ser consultados tanto na seleção de descritores como na definição dos vocabulários controlados correspondentes.

O primeiro dos projetos selecionados, sem ferramentas adequadas para gestão de dados de investigação surge no contexto da Produção de Hidrogénio, num grupo de investigação da Universidade do Porto. Este grupo orienta a investigação à produção instantânea de hidrogénio através da hidrólise catalítica do borohidreto de sódio (*NaBH4*) para dispositivos portáteis (telemóvel, tablet, mp3, etc.). Os dados experimentais deste grupo são armazenados principalmente em folhas de Excel. Os sensores, conectados ao reator, estão ligados a um computador com software específico – *LabView*, que grava os dados de temperatura, pressão e outras medições relevantes obtidos durante as experiências no reator num ficheiro *Excel*. A partilha de dados é feita, sobretudo, através de *email*. Em outros casos, os dados são copiados do computador para discos externos.

O trabalho realizado juntamente com este grupo de investigadores incluiu várias entrevistas e experiências de descrição de dados, que permitiram identificar as necessidades existentes na gestão de dados deste grupo. Durante as entrevistas os investigadores afirmaram que a descrição dos dados seria uma grande vantagem, nomeadamente para incentivar a sua partilha, publicação e reutilização dos dados. Esta colaboração levou à definição de descritores como *Additive*, *Catalyst*, *Reactor Type*, entre outros, entretanto

utilizados em experiências de descrição de dados no Dendro, em que estes investigadores foram participantes.

Os resultados desta experiência foram então avaliados para aferir a qualidade dos metadados criados por este grupo. A Tabela 1 mostra que, por exemplo, para o mesmo registo, diferentes investigadores usam diferentes conceitos. Isto é o caso da descrição do tipo de reator usado na experiência (Reactor Type), em que *ovoid* é o mesmo que *Egg Reactor,* o que levanta problemas de consistência nas descrições.

Descritor	Valor de descritor do utilizador 1	Valor de descritor do utilizador 2	Valor de descritor do utilizador 3
Temperature	25ºC	24	28ºC
Hydration Factor	16	16	15
Reactor Type	RG/RM	ovoid	Egg Reactor / Conical Small Reactor
Reagent	NaBH4	Sodium Borohydride	NaBH4
Catalyst	Ni-Ru	Nickel-ruthenium	NiRu
Gravimetric Capacity	<5wt%	2,3	1.9wt%
Hydrolysis	Hidrolise clássica	alkali	Classic hydrolysis

Tabela 1: Exemplos da diferença na descrição

A análise dos registos de metadados recolhidos reforçou a necessidade da criação de uma ferramenta que além de simplificação de processo de descrição no geral, ajude a melhorar a qualidade dos metadados no Dendro. Por isso mesmo, partiu-se para a elaboração de vocabulários controlados.

Elaboração de vocabulários controlados

A formalização de vocabulários controlados tirou partido da ontologia desenvolvida para o domínio Produção de Hidrogénio e da avaliação das descrições feitas por parte dos investigadores em Produção de Hidrogénio. Com os investigadores foram definidos os conceitos a incluir no vocabulário controlado para determinado descritor (Tabela 2).

Descritor	Conceitos definidos para vocabulários controlados
Additive	CMC SDS
Catalyst	Co-B Co-B/Ni Co-Mn-B Ni-Ru Pt/C
Hydrolysis	Acid hydrolysis Alkali-free hydrolysis Classic hydrolysis
Reactor Type	EggR – ovoid mini reactor LR – large reactor MRc – conical medium reactor

	MRf – flat medium reactor SRc – conical small reactor SRf – flat small reactor
Reagent	KBH4 LiAlH4 LiBH4 NH3BH3 NaBH4

Tabela 2: Descritores e conceitos definidos para vocabulários controlados

De acordo com literatura, existem várias maneiras de modelar vocabulários controlados numa ontologia.

Segundo a *W3C Recommendation – OWL Web Ontology Language Reference* os vocabulários controlados podem ser criados na ontologia como *Annotation Property*. O mesmo acontece no caso da ontologia sobre anatomia humana, da *Foundational Model of Anatomy* (Golbreich, Zhang e Bodenreider, 2006), que tem como base o *NCI-Thesaurus de National Cancer Institute* (Coronado, de e Fragoso, 2004). Já o *OWL FULL* não coloca quaisquer restrições sobre as anotações numa ontologia e o *OWL DL* permite anotações em classes, propriedades, indivíduos e cabeçalhos de ontologias. Além disso, afirma-se que é possível especificar o tipo de valor de um literal numa indicação da *Annotation Property*. Existem cinco *Annotation Property* pré-definidas, que podem ser utilizadas para anotação de DataProperties (Bechhofer *et al.*, 2004), estas são: *VersionInfo*; *label*; *comment*; *seeAlso*; *isDefineBy*.

Com base nestes estudos decidiu-se modelar os vocabulários controlados da ontologia de Produção de Hidrogénio através de *Annotation Property – has Alternative*, que é uma das alternativas possíveis para o valor de um descritor como conceito de vocabulário controlado, que por sua vez descreve determinada *Data Property*. A modelação da ontologia com os vocabulários controlados foi realizada utilizando o software Protégé (http://protege.stanford.edu/), onde os conceitos definidos foram associados aos respetivos descritores (Figura 1).

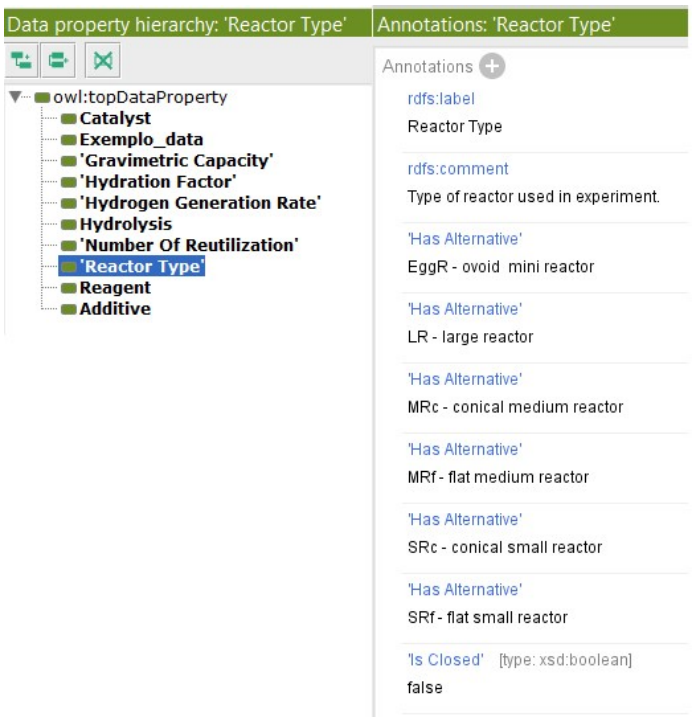

Figura 1: Conceitos de vocabulários controlados de descritor *Reactor Type* no *Protégé*

De acordo com esta modelação os conceitos de vocabulários controlados estão na ontologia e não é no Dendro, o que os permite expandir a utilização da plataforma para qualquer outro domínio de investigação.

Após a implementação da ontologia, os descritores com vocabulários controlados surgem no Dendro, tal como ilustrado na Figura 2.

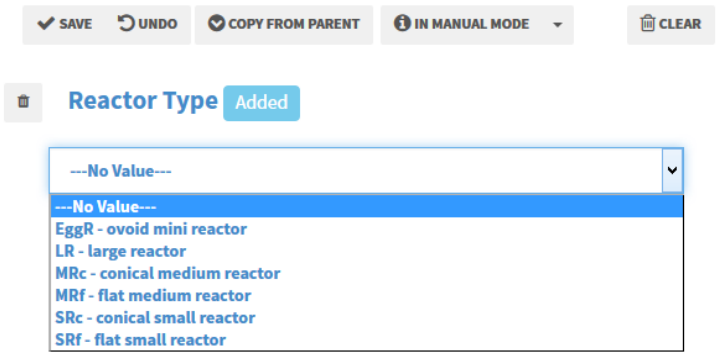

Figura 2: Vocabulários controlados de descritor *Reactor Type* no *Dendro*

Definição de métricas de qualidade de metadados

De modo a verificar se a implementação de vocabulários controlados facilita o processo de descrição no geral e ajuda a aumentar a qualidade dos metadados, foram definidas métricas de avaliação. Estas métricas servem para determinar a qualidade dos

metadados através da comparação entre descrições feitas sem o uso de vocabulários controlados e descrições com o uso dos mesmos.

De acordo com vários estudos (Alkhattabi, Neagu e Cullen, 2010; Bruce e Hillmann, 2004; Moreira *et al.*, 2009; Ochoa e Duval, 2006; Palavitsinis, 2013; Stvilia *et al.*, 2007) foram identificadas as métricas que melhor se adequam ao nosso caso:

1.***Correctness*** – grau em que a linguagem utilizada nos metadados é sintaticamente e gramaticalmente correta;

2.***Completeness*** – número de descritores preenchidos em comparação com o número total de descritores;

3.***Conformance to expectations*** – grau em que o registo de metadados preenche os requisitos de uma determinada comunidade de utilizadores;

4.***Overall Rating*** – pontuação geral do registo de metadados, tendo em conta as métricas anteriores.

E as métricas de avaliação da usabilidade de plataforma Dendro:

5.***Satisfaction*** – grau de satisfação de utilizador após da experiência;

6.***Task time*** – grau de rapidez da descrição de dados.

Para proceder à análise da qualidade dos metadados no domínio de Produção de Hidrogénio, de acordo com as métricas escolhidas, realizou-se uma série de experiências de descrição de dados na plataforma Dendro.

A primeira análise teve em conta metadados obtidos manualmente em três experiências de descrição de dados, anteriores à implementação dos vocabulários controlados, efetuadas por utilizadores do domínio indicado. Para determinar a qualidade dos metadados a utilização dos conceitos presentes nos vocabulários controlados é tida como uma descrição com cem por cento de qualidade. O registo de metadados, por exemplo, aplicando a métrica *Correctnes* é calculado através da comparação entre o número total de palavras na inserção de um conceito em linguagem natural, com o numero total de palavras do mesmo conceito definido no vocabulário controlado para determinado descritor. O cálculo de *Correctness* não tem em conta a existência de pequenas discrepâncias na escrita, tais como a diferença em letras maiúsculas e minúsculas, vírgulas, acentos, entre outros.

Para além disso foram consideradas as respostas dos investigadores a um inquérito realizado neste contexto. A análise realizada mostrou que a qualidade de descrição sem utilização de vocabulários controlados estava aquém das expectativas, quer seja pela diferença na descrição de dados de investigação, lapsos e registos incompletos. Apesar disto, os utilizadores ficaram satisfeitos com a utilização da plataforma Dendro e não necessitaram de muito tempo para descrever os dados de investigação que produziram.

A segunda ronda de experiências contou com a participação dos mesmos utilizadores, desta vez com recurso a vocabulários controlados. Para possibilitar a comparação de resultados entre as experiências, a análise após a implementação de vocabulários controlados baseou-se nas mesmas métricas e regras definidas na análise anterior. Contudo, a utilização dos conceitos presentes nos vocabulários controlados é tida como uma descrição com cem por cento de qualidade.

Os resultados da segunda análise mostram que a qualidade de descrição de dados de investigação com utilização de vocabulários controlados aumentou. Aliás, todos os valores ficaram muito próximos dos cem por cento. Isto indica que a utilização de vocabulários controlados diminui as diferenças na descrição, erros, e ajudou a obter uma descrição mais completa.

As experiências da descrição dos dados de investigação realizadas no Dendro com utilização de vocabulários controlados deixaram os utilizadores satisfeitos com a usabilidade da plataforma. Para além disso, o tempo que cada investigador precisou para completar a tarefa diminuiu.

Avaliação do uso de vocabulários controlados

A comparação dos resultados de análises de descrição de dados de investigação mostra que a maioria dos valores de qualidade dos metadados aplicando as métricas definidas subiu. Por exemplo, o valor de qualidade de metadados de descritor *Reactor Type*, aplicando métrica *Correctness* passou de 28% para 100% e aplicando a métrica *Conformance to expectations* aumentou de 36% para 100%. Tal como se compreende, através da Tabela 3, a qualidade da descrição após a implementação dos vocabulários controlados melhorou.

Durante a realização das experiências da descrição os investigadores utilizaram praticamente a mesma quantidade de descritores e não tiveram grandes dificuldades na utilização da plataforma. O nível da satisfação teve um aumento pouco significativo. Contudo, afirmaram que a utilização de vocabulários controlados facilita o processo da descrição. Analisando as respostas de inquéritos e comentários de utilizadores conclui-se que os investigadores preferem criar os metadados com auxílio de vocabulários controlados, pois a comparação dos valores na Tabela 3 obtidos aplicando a métrica *Task Time*, mostram que o tempo médio necessário para realização de uma tarefa diminuiu.

	Qualidade	
Descritor	**Qualidade antes de implementação de vocabulários controlados (*Correctness*)**	**Qualidade após de implementação de vocabulários controlados (*Correctness*)**
Reactor Type	28%	100%
Hydrolysis	51%	100%
Catalyst	50%	100%
Reagent	50%	100%
Additive		100%

	Qualidade antes de implementação de vocabulários controlados (*Conformance to expectations*)	Qualidade após de implementação de vocabulários controlados (*Conformance to expectations*)
Reactor Type	36%	100%
Hydrolysis	69%	100%
Catalyst	75%	100%
Reagent	75%	100%
Additive		50%
	Qualidade antes de implementação de vocabulários controlados (*Completeness*)	**Qualidade após de implementação de vocabulários controlados (*Completeness*)**
Média	75%	82%
	Qualidade antes de implementação de vocabulários controlados (*Overall Rating*)	**Qualidade após de implementação de vocabulários controlados (*Overall Rating*)**
Reactor Type	46%	94%
Hydrolysis	65%	94%
Catalyst	67%	94%
Reagent	67%	94%
Additive		77%
	Usabilidade sem utilização de vocabulários controlados	**Usabilidade com utilização de vocabulários controlados**
Satisfaction	4	4.5
	***Task Time* de tarefa (média) sem utilização de vocabulários controlados**	***Task Time* de tarefa (média) com utilização de vocabulários controlados**
	17 min	10 min
	***Task Time* por descritor (média) sem utilização de vocabulários controlados**	***Task Time* por descritor (média) com utilização de vocabulários controlados**
	62 seg	38 seg

Tabela 3: Comparação de resultados de avaliação de qualidade de dados antes e após a implementação de vocabulários controlados

Sendo assim, a comparação dos resultados demostra que o uso de vocabulários controlados na criação dos metadados facilita o processo de descrição em geral, obtendo-se uma descrição mais completa e correta.

Conclusões e trabalhos futuros

O estudo realizado sobre gestão de dados de investigação, criação de metadados e importância de qualidade dos mesmos ilustrou vários problemas existentes nesta área. O processo de descrição de dados exige competências, esforço, tempo e ferramentas adequadas, pois só os metadados de qualidade garantem a precisão e acesso completo aos recursos digitais e permitem aos utilizadores finais encontrar e recuperar os recursos que precisam. O interesse e a motivação de investigadores tanto na gestão de dados como na descrição de dados e na escolha de um sistema para criação dos metadados depende de vários fatores, tais como a usabilidade, utilidade, simplicidade, facilidade, entre outros.

O trabalho desenvolvido enquadrou o desafio da descrição de dados no processo de desenvolvimento da plataforma de gestão de dados de investigação Dendro da Universidade do Porto. Pretendeu-se encontrar uma solução para agilizar e facilitar todo o processo de descrição de dados de investigação e, assim, contribuir para a melhoria da qualidade dos metadados criados no Dendro.

Em suma, os objetivos deste trabalho foram segmentados em três vertentes. Em primeiro lugar procedeu-se à escolha do domínio Produção de Hidrogénio como caso de estudo, a realização de uma série de experiências de descrição de dados no Dendro e recolhidos os registos de metadados. Em segundo lugar foram elaborados vocabulários controlados para o domínio escolhido e após a sua implementação realizou-se a segunda ronda de experiências de descrição de dados de investigação. Em terceiro lugar definiram-se as métricas para a avaliação de qualidade dos metadados e procedeu-se à análise da qualidade dos metadados criados sem e com utilização de vocabulários controlados, com a finalidade de demostrar que a implementação de vocabulários controlados facilita o processo de descrição e melhora a qualidade de descrição de dados de investigação no Dendro.

Mediante os resultados obtidos pode-se afirmar que os objetivos foram alcançados. Em particular, a descrição efetuada com uso de vocabulários controlados simplificou todo o processo de criação de metadados, permitiu obter descrições completas e mais rigorosas, com a vantagem de o permitir sem o aumento do tempo necessário para o efeito.

Uma das perspetivas de trabalho futuro é a elaboração de vocabulários controlados para outros domínios de investigação existentes no Dendro e para descritores genéricos, tal como *Language*, *Format*, entre outros. Outro dos objetivos é implementar vocabulários controlados abertos e fechados.

A elaboração de expressões regulares é mais uma estratégia que pode ser utilizada na simplificação do processo de descrição de dados no Dendro. (Grimalovskii, 2013; Skoglund, 2011; Standen, 2010). De acordo com a literatura, as expressões regulares podem reduzir o esforço manual na introdução da informação e ajudar na qualidade de dados. Elas são complexas aquando da sua boa configuração e funcionamento trazem sempre benefícios da sua utilização (Friedl, 2006).

Para terminar, pode-se afirmar que os trabalhos futuros definidos e a continuação de interação e elaboração com os investigadores de diferentes domínios científicos vão continuar a trazer benefícios, tais como o aumento da usabilidade de plataforma Dendro, de forma a incluir os investigadores progressivamente em atividades de gestão de dados de investigação.

Referências bibliográficas:

AKMON, D. *et al.* (2011) – The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs. *Archival Science* [Em linha]. Vol.11, Nº3, p.329-348. [Consult. 23 jul. 2016]. Disponível na Internet: <URL: http://link.springer.com/article/10.1007/s10502-011-9151-4>.

ALKHATTABI, M.; NEAGU, D.; CULLEN, A. (2010) – Information quality framework for e-learning systems. *Knowledge Management & E-Learning: An International Journal* [Em linha]. Vol.2, Nº4, p.340-362. [Consult. 20 jul. 2016]. Disponível na Internet: <URL: http://www.kmel-journal.org/ojs/index.php/online-publication/article/view/21/62>.

BERMUDEZ, L. *et al.* (2011) – The importance of controlled vocabularies. *The MMI Guides: Navigating the World of Marine Metadata* [Em linha].[Consult. 19 abr 2016]. Disponível na Internet: < https://marinemetadata.org/guides/vocabs/vocimportance >.

BORGMAN, C.L.; WALLIS, J.C.; Enyedy, N. (2007) – Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital librarioes. *International Journal on Digital Libraries* [Em linha]. Vol.7, Nº1, p.17-30. [Consult. 20 abr. 2016]. Disponível na Internet: <URL:

https://pages.gseis.ucla.edu/faculty/enyedy/assets/Projects/Teaching%20and%20Learning%20Science/borgman%20enyedy.pdf>.

BRUCE, Thomas R.; HILLMANN, Diane I. (2004) – The Continuum of Metadata Quality: Defining, Expressing, Exploiting. *Metadata in Practice* [Em linha]. [Consult. 24 jul. 2016]. Disponível na Internet: <URL: https://ecommons.cornell.edu/handle/1813/7895 >.

CASTRO, João Aguiar *et al.* (2015) – Ontologies for research data description: a design process applied to vehicle simulation. *Proceedings of the 9th Metadata and Semantics Research Conference (MTSR 2015)*, CCIS 544, pp.348-354

CORONADO, Sherri.; FRAGOSO, Gilberto. (2004) – Enterprise Vocabulary Development in Protege/OWL: Workflow and Concept History RequirementsNCI Center for Bioinformatics. [Em linha]. [Consult. 09 abr. 2016]. Disponível na Internet: <URL: http://protege.stanford.edu/conference/2004/abstracts/DeCoronado.pdf>.

DATAMARTIST.COM, Using regular expressions to check data quality. Part 2. James Standen. [Em linha]. [Consult. 21 jun. 2016]. Disponível na Internet: <URL: http://www.datamartist.com/how-to-use-regular-expressions-to-check-data-quality-part-2 >.

FRIEDL, Jeffrey E. F. (2006) – *Mastering Regular Expressions*. 3rd Edition. O'Reilly Media, Inc., 544p. ISBN 978-0-596-52812-6

GOLBREICH, Christine; ZHANG, Songmao; BODENREIDER, Olivier (2006) – The Foundational Model of Anatomy in OWL: Experience and Perspectives. *Web Semantics: Science, Services and Agents on the World Wide Web* [Em linha] Vol.4, Nº3, p.181-195. [Consult. 06 abr. 2016]. Disponivel na Internet: <URL:

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.9032&rep=rep1&type=pdf>.

GRIMALOVSKII, Alexandr (2013) – Expressões regulares [Em linha].[Consult. 30 abr. 2016]. Disponível na Internet: <URL: http://www.codenet.ru/webmast/php/regexps.php >.

HARPING, P. (2010) – *Introduction to controlled Vocabularies : Terminology for Art, Architecture, and Other Cultural Works.* [Em linha]. 1st ed. Los Angeles: Getty Research Institute. [Consult. 14 mar 2016]. Disponível na Internet: <URL: http://d2aohiyo3d3idm.cloudfront.net/publications/virtuallibrary/160606018X.pdf>. ISBN 978-1-60606-018-6.

HEDDEN, H. (2010) – Taxonomies and Controlled Vocabularies Best Practices for Metadata. *Journal of Digital Asset Management* [Em linha] Vol.6, N°5, p.279-284. [Consult. 30 abr. 2016]. Disponível na Internet: < URL: http://link.springer.com/article/10.1057%2Fdam.2010.29>.

HEIDORN, P.Bryan (2008) – Shedding light on the dark data in the long tail of science.*Library Trends* [Em linha] Vol. 57, N°2, p.280-299. [Consult. 20 set. 2016]. Disponível na Internet: < URL: https://www.researchgate.net/publication/49175975_Shedding_Light_on_the_Dark_Data_in_the_Long_Tail_of_Science >.

KARIMOVA, Yulia (2016) – *Vocabulários controlados na descrição de dados de investigação no Dendro.* Porto: Faculdade de Engenharia da Universidade do Porto. 120p. Dissertação do mestrado. Disponível na Internet: < URL: https://repositorio-aberto.up.pt/bitstream/10216/85221/2/140750.pdf >.

LYNDA.COM. Using regular expressions. Kevin Skoglund. [Em linha]. [Consult. 02 abr. 2016]. Disponível na Internet: <URL: http://www.lynda.com/Regular-Expressions-tutorials/Using-Regular-Expressions/85870-2.html >.

MARINE METADATA INTEROPERABILITY. The Importance of Controlled Vocabularies. [Em linha]. [Consult. 03 jun. 2016]. Disponível na Internet: <URL: https://marinemetadata.org/guides/vocabs/vocimportance >.

MOREIRA, Bárbara L. *et al.* (2009) – Automatic evaluation of digital libraries with 5SQual. *Journal of Informetrics*. [Em linha] Vol.3:, N°2, p.102-123. [Consult. 06 abr. 2016]. Disponivel na Internet: <URL: http://www.sciencedirect.com/science/article/pii/S1751157708000734>.

NATIONAL INFORMATION STANDARDS ORGANIZATION (2005) – *ANSI/NISO Z39.19-2005: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies* [Em linha]. NISO: Maryland, USA [Consult. 08 jun 2016]. Disponível na Internet: < URL: http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf>. ISBN: 978-1-937522-22-3.

OCHOA, Xavier; DUVAL, Erik (2006) – Quality Metrics for Learning Object Metadata. *World Conference on Educational Multimedia, Hypermedia and Telecommunications.* [Em linha]. [Consult. 06 abr. 2016]. Disponivel na Internet: <URL: https://www.researchgate.net/publication/254358241_Quality_Metrics_for_Learning_Object_Metadata>.

PALAVITSINIS, Nikos (2013) – *Metadata Quality Issues in Learning Repositories.* Universidad

de Alcalá: Departamento de Ciencias de la Computación. 295p. Tese de doutoramento. Disponível na Internet: <URL:

http://dspace.uah.es/dspace/bitstream/handle/10017/20664/Thesis%20Palavitsinis.pdf?sequence=1&isAllowed=y>.

SILVA, João Rocha Da *et al.* (2014) - Dendro: Collaborative Research Data Management Built on Linked Open Data. *The Semantic Web: ESWC 2014 Satellite Events*. p.483-487

SMITH, K.; SELIGMAN, L.; SWARUP, V. (2008) - Everybody Share: The challenge of data-sharing systems. *ComputerComputer*. [Em linha] Vol.41, p.54-61. [Consult. 06 abr. 2016]. Disponivel na Internet: <URL:

http://ieeexplore.ieee.org/document/4623223/?reload=true&arnumber=4623223>.

STVILIA, Besiki *et al.* (2007) - A framework for Information Quality Assessment. *Journal of the American Society for Information Science and Technology.* [Em linha] Vol.58, Nº12, p.1720-1733. [Consult. 06 abr. 2016]. Disponivel na Internet: <URL: http://myweb.fsu.edu/bstvilia/papers/stvilia_IQFramework_p.pdf>.

SWAN, Alma; BROWN, Sheridan (2008) - To share or not to share: Publication and quality assurance of research data outputs. *Report commissioned by the Research Information Network*. [Em linha] June, Vol.56. [Consult. 06 abr. 2016]. Disponivel na Internet: <URL: http://eprints.soton.ac.uk/266742/>.

W3 RECOMMENDATION. Annotations, ontology header, imports and version information. [Em linha]. [Consult. 07 jun. 2016]. Disponível na Internet: <URL: https://www.w3.org/TR/owl-ref/#Header >.

ZHANG, Tao; MARON, Deborah J.; CHARLES, Christopher C. (2013) - Usability Evaluation of a Research Repository and Collaboration Web Site. *Journal of Web Librarianship*. [Em linha]. Vol.7, Nº1, p.58-82 [Consult. 06 abr. 2016]. Disponivel na Internet: <URL: http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1061&context=lib_fsdocs>.