U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

MESTRADO EM CIÊNCIA DA INFORMAÇÃO

VOCABULÁRIOS CONTROLADOS NA DESCRIÇÃO DE DADOS DE INVESTIGAÇÃO NO DENDRO

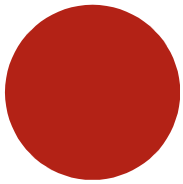
Yulia Karimova

M

2016

UNIDADES ORGÂNICAS ENVOLVIDAS

FACULDADE DE ENGENHARIA
FACULDADE DE LETRAS



YULIA KARIMOVA

VOCABULÁRIOS CONTROLADOS NA DESCRIÇÃO DE DADOS DE INVESTIGAÇÃO NO DENDRO

Dissertação realizada no âmbito do Mestrado em Ciência da Informação

Orientador: Prof. Dra. Maria Cristina de Carvalho Alves Ribeiro
Coorientador: João Aguiar Castro

Faculdade de Engenharia e Faculdade de Letras
Universidade do Porto

Junho de 2016

VOCABULÁRIOS CONTROLADOS NA DESCRIÇÃO DE DADOS DE INVESTIGAÇÃO NO DENDRO

YULIA KARIMOVA

Dissertação realizada no âmbito do Mestrado em Ciência da Informação

Orientador: Prof. Dra. Maria Cristina de Carvalho Alves Ribeiro

Coorientador: João Aguiar Castro

Membros do Júri

Presidente: Professor Doutor Gabriel de Sousa Torcato David
Faculdade de Engenharia – Universidade do Porto

Arguente: Professora Doutora Ana Alice Rodrigues Pereira Baptista
Departamento de Sistemas de Informação – Universidade do Minho

Orientador: Professora Doutora Maria Cristina de Carvalho de Alves Ribeiro
Faculdade de Engenharia – Universidade do Porto

Agradecimentos

Ao entrar para o Mestrado em Ciência de Informação sabia que me iria debater com inúmeras dificuldades. Tinha medo de não conseguir, mas encontrei no meu caminho diversas pessoas que se tornaram importantes e que me ajudaram a ultrapassar todas as dificuldades e medos que encontrei ao longo do meu percurso. Às seguintes expresso um sincero agradecimento…

À minha orientadora, Professora Cristina Ribeiro, que mais que uma professora, foi uma inspiração para mim. Ajudou-me a crescer, aprender e tornar-me uma pessoa melhor.

Aos investigadores do grupo InfoLab que se tornaram uma segunda família para mim. Agradeço em especial ao João Castro por estar sempre disponível para me ajudar e ao João Rocha, por todo o tempo e esforço que despendeu na implementação de vocabulários controlados no Dendro.

Aos investigadores do Grupo CEFT pela inestimável colaboração na criação de vocabulários controlados no domínio Produção de Hidrogénio, em particular ao Hélder Xavier pela paciência, disponibilidade e boa disposição, Josefina Figueira, Daniela Falcão, Professora Alexandra Pinto e à Vânia Oliveira pela valiosa participação.

Aos meus pais, Nina e Rinat, irmã, Valeria e família por me conhecerem e entenderem melhor do que ninguém. Por me apoiarem e me motivarem, não só em dias bons como nos maus e estarem sempre do meu lado.

Ao meu marido Fernando pela paciência, compreensão e apoio nos momentos bons e nos mais difíceis.

Aos meus amigos e amigas, Lia, Vanessa, Nuno, Luciano, entre outros, por me mostrarem o verdadeiro sentido da palavra Amizade - sempre presentes, sempre a apoiar, sempre a acreditar e a dar força.

A Deus por me ter dado a vida e estar sempre comigo.

A todos eles o meu muito obrigado pois tornaram possível a concretização deste mestrado e me mostraram que não existem impossíveis. Acreditar, sonhar, trabalhar, lutar, continuar aprender, crescer e nunca desistir é o melhor caminho para conseguir.

Resumo

Atualmente surgem cada vez mais repositórios digitais onde os investigadores podem preservar, descrever e partilhar os seus dados de investigação. O desenvolvimento de tecnologia neste ramo levanta muitos desafios e oportunidades. O processo de descrição de dados é um desafio particular. Os metadados desempenham um papel importante neste contexto, uma vez que ajudam a adicionar a informação valiosa que permite a interpretação de dados por terceiros, promovendo assim a reutilização de dados e preservação. No entanto, a descrição de dados pode ser uma tarefa muito exigente e demorada. De acordo com a literatura a criação de vocabulários controlados podem facilitar o processo da descrição de dados e ao mesmo tempo contribuir para uma melhoria na qualidade dos metadados. Quanto maior for a qualidade dos metadados maior será a qualidade dos serviços prestados aos utilizadores. Contudo, a qualidade dos metadados é uma tema que tem recebido menos atenção do que a qualidade dos dados.

Com esta ideia em mente este estudo propõe a criação de vocabulários controlados para a descrição de dados em domínios científicos específicos no Dendro. A plataforma Dendro tem vindo a ser desenvolvida na Universidade do Porto, de maneira a satisfazer as necessidades reconhecidas aos desafios existentes nos processos de descrição de dados de investigação.

O estudo implica a análise dos descritores existentes no Dendro, definindo os mais utilizados por cada grupo de investigadores de domínios específicos. Estes descritores tanto podem ter origem em esquemas de metadados genéricos, como ter sido criados em conjunto com os investigadores do domínio no contexto das experiências efetuadas na plataforma Dendro.

A abordagem metodológica do presente trabalho inclui os seguintes passos: revisão de literatura, a adaptação das práticas sugeridas na tese *Metadata Quality Issues in Learning Repositories* de Nikos Palavitsinis e outras publicações de outros autores para definição das métricas de qualidade de metadados e avaliação dos mesmos, criação de vocabulários controlados e sugestões de melhorias nos procedimentos existentes.

Os resultados obtidos demostram que a utilização dos vocabulários controlados diminui os erros na descrição, tornaram o registo de metadados mais completo e ao mesmo tempo diminuiu o tempo despendido na tarefa de descrição de dados.

Palavras-chave: *gestão de dados de investigação*; *metadados; vocabulários controlados; qualidade de metadados; dados de investigação; Dendro*

Abstract

Nowadays, there is a growing interest in research data platforms that can address research data management challenges. The development of tools in this field can be very demanding on the involved stakeholders and raise both challenges and opportunities. The data description process is a particular challenge. Metadata plays an important role in this regard as it helps to add valuable information that enables data interpretation by third-parties, thus promoting data reuse and preservation. Controlled vocabularies can ease the data description process and at the same time contribute to metadata quality. The higher the quality of metadata the higher will be the quality of the services provided to the users. However, metadata quality is a subjet that has less attention than the quality of the data.

This contribution can help improving research metadata, reducing description errors and facilitating metadata production in the Dendro platform, a staging area specifically designed with data description in mind. Dendro is currently being developed by researchers in the Infolab group at the University of Porto. For this purpose, we choose to create controlled vocabularies in order to improve the quality of description of the research data.

This study also involves analyzing the existing descriptors in Dendro and surveying the most used by each group of specific researcher's areas. These descriptors can either be extracted from both high-level and domain metadata schemas, or even created in collaboration with researchers in the context of the Dendro experiments.

The applied methodology includes several steps to ensure a better fitness to existing practices in Research Data Management: literature revision, an adaptation of the practices presented in the *Metadata Quality Issues in Learning Repositories* thesis, by Nikos Palavitsinis, and other authors to define the metadata quality metrics and correspondent evaluation, creating controlled vocabularies and suggesting improvements in existing procedures.

The results show that the use of controlled vocabularies reduce the data description errors, make the metadata records more comprehensive and has decreased the time spent in data description activities.

Keywords: *research data management; metadata; controlled vocabularies; quality metadata; research data; Dendro*

Tabela de Conteúdo

Lista de figuras

Lista de tabelas

Lista de anexos

Lista de Abreviaturas e Siglas

CEFT	Centro de Estudos de Fenómenos de Transporte
CODATA	Comittee on Data for Science and Technology
DCC	Digital Curation Centre
FEUP	Faculdade de Engenharia da Universidade do Porto
INFOLAB	Information Systems Research Group
ISO	International Organization for Standardization
MANTRA	Research Data Management Training
MQACP	The Metadata Quality Assurance Certification Process
NISO	National Information Standards Organization
OECD	Organization for Economic Cooperation and Development
SKOS	Simple Knowledge Organization System
TGN	Thesaurus of Geographic Names
VRE	Virtual Research Environment

Introdução

A comunicação científica engloba todos os processos de representação, transmissão e receção de informação, sendo o principal mecanismo de funcionamento e desenvolvimento da ciência, bem como uma condição necessária para a formação e desenvolvimento dos investigadores. A partilha científica entre investigadores estimula o surgimento de conhecimentos teóricos e práticos, assegurando a difusão do conhecimento científico.

Atualmente, com a introdução de novas tecnologias de informação começam a surgir novas formas e métodos de cooperação científica que têm vindo a mudar a estrutura da comunicação científica. Consequências disso são o aparecimento de infraestruturas para computação científica, a criação de repositórios de dados de investigação e o aumento do espaço *on-line* de comunicação científica. A expressão curadoria digital é cada vez mais utilizada nas ações ligadas à adição de valor aos dados e preservação dos mesmos a longo prazo. A representação digital de informação abre grandes possibilidades de armazenamento e processamento, mas pode introduzir problemas relacionados com a gestão de dados de investigação, como a preservação, descrição, partilha e acesso.

Os metadados podem ajudar a resolver este problema, permitindo a identificação, criação de pontos de acesso, interpretação e avaliação dos dados. Para qualquer investigador o processo de partilha de dados recolhidos durante a investigação é muito importante, mas estes devem estar associados a metadados com qualidade, de forma a serem interpretados e reutilizados sem dificuldade.

Os dados publicados devem estar acessíveis e ter uma descrição completa. O formato da descrição, sintaxe e semântica devem ser normalizados, para facilitar a interpretação e partilha dos dados. Existem várias normas e esquemas para a criação de metadados que ajudam na descrição de dados, sobretudo de caráter genérico, mas uma vez que os investigadores irão em geral descrever os dados em áreas mais específicas, coloca-se um problema: que descritores se devem usar tendo em conta as grandezas envolvidas na investigação, condições da recolha de dados e metodologias utilizadas, por exemplo.

Muitos projetos não têm uma infraestrutura integrada de dados, que muitas vezes resulta em uso de ferramentas para gestão de dados não normalizadas. Falta de descritores em domínios específicos pode provocar a descrição incompleta, com metadados de mal qualidade

e desta forma os investigadores não conseguem assegurar que os mesmos possam ser interpretados e reutilizados por outros. A plataforma Dendro tem vindo a ser desenvolvida na Universidade do Porto, de maneira a satisfazer as necessidades reconhecidas aos desafios existentes nos processos de descrição de dados de investigação.

Esta plataforma facilita a criação de metadados com a utilização de vários esquemas existentes, como por exemplo *Dublin Core* e *Friend of a Friend*, assim como descritores específicos criados neste contexto para fornecer aos investigadores de diferentes domínios descritores que lhes permitam obter registos de metadados abrangentes e precisos. O Dendro utiliza os conjuntos de descritores baseados em ontologias criadas através de colaboração entre os investigadores dos domínios e um curador de dados. Os investigadores dos domínios fornecem conhecimento especializado sobre o contexto de produção de dados, enquanto o curador tem uma perspetiva mais abrangente sobre gestão de dados de investigação.

De forma a se poder adicionar novos descritores específicos para cada domínio no Dendro, foram criadas ontologias para vários domínios científicos, tais como: Produção de Hidrogénio, Biodiversidade, Química Analítica (Castro, Silva, e Ribeiro 2014; Castro et al. 2015). Os metadados obtidos após as experiências realizadas com os grupos de investigadores dos respetivos domínios fazem parte da análise de qualidade dos metadados e ajudam na escolha de caso de estudo desta dissertação. Para análise de qualidade irão ser definidas as métricas de avaliação e indicados os descritores mais utilizados por cada grupo de investigadores.

Para além das ontologias e da sua escolha ou desenho, outra questão que se coloca é a necessidade de vocabulários controlados para registo de metadados com qualidade. De maneira a se poder garantir a qualidade em áreas e domínios específicos, é fundamental ter-se uma política de descrição e uso de vocabulários controlados, de forma a se facilitar a descrição de dados, diminuir os erros na descrição e consequentemente melhorar a qualidade da mesma.

Uma parte deste trabalho é a colaboração com a equipa do projeto Dendro e com investigadores de diversos domínios científicos, a fim de conseguir integrar os resultados na plataforma Dendro e efetuar os testes necessários. A plataforma Dendro encontra-se em desenvolvimento e em fase de testes de utilização na descrição de dados por investigadores de várias áreas da Universidade do Porto. Este trabalho poderá ajudar na continuação do desenvolvimento do Dendro, nomeadamente ao facilitar a experiência dos investigadores e ao melhorar a qualidade na descrição de dados que produzem.

Objetivos específicos e resultados esperados

O objetivo principal deste trabalho é contribuir para melhorar a qualidade dos metadados criados no Dendro.

Para atingir esse objetivo são propostas as seguintes tarefas:

1. *Definir as métricas de qualidade da descrição de dados de investigação*. Para realizar esta tarefa pretende-se fazer uma revisão da literatura, definir domínios de investigação para análise e efetuar o levantamento de dados de descrições existentes no Dendro.

2. *Criar os vocabulários controlados*. A revisão de literatura sobre os vocabulários controlados, normas existentes e esquemas genéricos ajuda na execução desta tarefa.

3. *Analisar a qualidade da descrição de dados de investigação no Dendro*. Para realização desta tarefa é necessário recolher os metadados existentes no Dendro, indicar os descritores mais utilizados por cada grupo de investigadores e aplicar as métricas de qualidade definidas.

As tarefas estão interligadas e ajudam a responder as questões relacionadas com a produção de registos de metadados no Dendro com qualidade por parte dos investigadores. Acredita-se que com recurso a vocabulários controlados os processos de descrição são facilitados, o que pode significar um aumento na qualidade dos metadados.

A análise da qualidade é verificada através do processo *The Metadata Quality Assurance Certification Process* (Palavitsinis 2013) e *Automatic evaluation of metadata quality in digital repositories* (Ochoa e Duval 2009). O processo MQACP tem ferramentas para a recolha de dados, tais como questionários relacionados com elementos de metadados, que estão a ser utilizados, bem como formas de avaliação, que incluem métricas de qualidade, para avaliar a qualidade de descrição de dados. O estudo de Ochoa e Duval inclui propostas de métricas e fórmulas de cálculo de qualidade de metadados em repositórios digitais.

Os vocabulários controlados vão ser criados de tal forma que vão disponibilizar todos os valores possíveis e aceitáveis para determinado descritor.

Por fim a avaliação da qualidade é feita de acordo com métricas definidas durante o trabalho. Para conseguir obter os resultados concretos e ver se conseguimos melhorar a qualidade de descrição de dados de investigação no Dendro, serão comparados os resultados das descrições antes e após a implementação de vocabulários controlados.

Estrutura da dissertação

Este trabalho encontra-se estruturado em cinco capítulos abaixo detalhados.

O Capítulo 1, intitulado "Revisão de literatura" serve para fornecer uma visão dos temas principais que serão abordados neste trabalho, e obter as direções para a concretização do objetivo principal. São apresentados conceitos, definições e estudos sobre dados de

investigação, metadados na gestão de dados, qualidade de metadados, critérios e métricas para avaliação da mesma e usabilidade de repositórios digitais, vocabulários controlados e expressões regulares. A seguir fala-se dos trabalhos publicados pelo grupo referentes ao Dendro, fornecendo o contexto do projeto.

O Capítulo 2, intitulado "Seleção de caso de estudo e definição de métricas de qualidade de metadados" descreve a fase preparatória, que inclui a escolha de domínio de investigação para estudo, recolha de dados existentes no Dendro no domínio escolhido, definição dos descritores mais utilizados pelo grupo de investigadores deste domínio e as métricas de qualidade de descrição de dados de investigação mais adequadas ao nosso caso. Em seguida, é descrita uma experiência efetuada por um investigador de domínio escolhido, com o objetivo de completar os dados para análise de qualidade de descrição.

O Capítulo 3, intitulado "Elaboração de vocabulários controlados para domínio Produção de Hidrogénio explica a metodologia de criação e implementação de vocabulários controlados, concretamente modelação da ontologia em *Protégé*, que se baseia na análise de estudos semelhantes, tal como o caso de *FMA (Foundational Model of Anatomy)*. A seguir, é descrito o processo das experiências, realizadas por um grupo de investigadores de domínio escolhido, após de implementação de vocabulários controlados e recolha dos dados para respetiva análise.

O Capítulo 4, "Análise de qualidade de descrição de dados de investigação" se dedica a documentar o processo realizado de análise de qualidade de descrição de dados de investigação antes e após a criação e implementação de vocabulários controlados. Mais especificamente este capítulo inclui os cálculos realizados para avaliação de qualidade de metadados após as experiências realizadas no Dendro por investigadores de Produção de Hidrogénio, sem e com uso de vocabulários controlados. Por fim, são demostrados os resultados da comparação de valores obtidos.

Esta dissertação termina com uma reflexão global sobre o trabalho realizado, num capítulo dedicado às conclusões e perspetivas futuras. Em anexo pode ser consultada a documentação produzida para apoiar este trabalho.

Capítulo 1. Revisão de literatura

A revisão de literatura inclui as seguintes secções:

- A primeira secção apresenta a definição dos dados de investigação; como estes podem ser classificados e geridos; que tipos de modelos de ciclo de vida dos dados existem e a importância da curadoria digital; o papel de investigadores no contexto de gestão de dados de investigação.
- A segunda secção mostra as definições existentes de metadados, que foram identificados durante a revisão de literatura; quais os esquemas e normas que existem para a descrição de dados; a necessidade e a importância de qualidade de metadados e quais as métricas que podem ajudar na análise de qualidade dos metadados.
- A terceira secção define os vocabulários controlados; indica a necessidade da criação dos mesmos e demonstra os tipos e exemplos de vocabulários controlados existentes.
- A quarta secção contém a informação sobre mais uma ferramenta – expressões regulares, que pode ser utilizada na simplificação do processo de descrição de dados de investigação.
- Por fim, na quinta secção fala-se sobre o desenvolvimento de plataforma Dendro e ações realizadas até agora.

No geral, este capítulo reflete o estado atual dos problemas existentes, ligadas aos metadados e qualidade de metadados para recursos e repositórios digitais, e identifica as metodologias, critérios e métricas utilizados na análise da qualidade de metadados. Além disto, mostra-nos vantagens de utilização de vocabulários controlados e expressões regulares.

1.1 Dados de investigação

Os dados de investigação, necessários à validação dos resultados de investigação e base de novos projetos, tendem a ser cada vez mais valorizados pela comunidade científica. Os dados de investigação podem ser classificados de várias maneiras: pela natureza, método de recolha, entre outros, o que representa um desafio a gestão dos mesmos. Os dados de investigação não são fáceis de organizar, descrever e disponibilizar para que sejam compreensíveis e reutilizáveis. Neste contexto, a curadoria digital poderá ajudar na gestão de dados ao propor os modelos que irão dar suporte à gestão de dados de investigação em vários contextos científicos.

1.1.1 Definição de dados de investigação

Atualmente verifica-se um avanço tecnológico que coincide com o desenvolvimento científico. Por exemplo, surgem com maior frequência repositórios, que criam um novo espaço de informação científica. Assim, ocorrem novas formas de colaboração, por exemplo, um ambiente virtual de investigação – *Virtual Research Environment*. O *VRE* representa ambientes de trabalho inovadores, que visam o reforço da cooperação e colaboração entre os investigadores em todos os cenários da investigação moderna. Através do VRE os investigadores obtêm apoio na elaboração e partilha dos resultados de investigação, e no desenvolvimento de novas abordagens de investigação (Candela 2011).

De acordo com a literatura os dados de investigação podem ser definidos como:

"recorded factual material commonly accepted in the scientific community as necessary to validate research findings…"[1]*; "on which an argument, theory or test is based"*[2]*; e podem ser "numerical, descriptive, visual or tactile. It may be raw, cleaned or processed, and may be held in any format or media…"*[3].

Assim sendo, podemos afirmar que os dados de investigação resultam das experiências ou simulações, estatísticas, observações, entrevistas, anotações, entre outros, que estão ligados ao processo de investigação científica; e, de acordo com, Corti et al. (2011), os dados de investigação são uns recursos valiosos, que geralmente necessitam de muito tempo e dinheiro para serem produzidos (Van de Eynden et al. 2013). Na opinião destes autores a partilha de dados de investigação é muito importante por vários aspetos:

- Incentiva a investigação científica e debates;

- Promove a inovação e potenciais novos usos de dados;

- Leva a novas colaborações entre utilizadores de dados e criadores de dados;

- Maximiza a transparência e a responsabilização dos investigadores;

- Permite o controlo dos resultados da investigação;

---

[1] Federal Register Notice re OMB Circular A-110, https://www.whitehouse.gov/omb/fedreg_a110-finalnotice

[2] Management of Research Data and Records Policy, The Univesrity of Melbourne, https://policy.unimelb.edu.au/MPF1242#section-3.1

[3] Queensland University of Technology, Management of research data. http://www.mopp.qut.edu.au/D/D_02_08.jsp

- Incentiva a melhoria e validação de métodos de investigação;

- Reduz o custo de duplicação das coleções de dados;

- Aumenta o impacto e visibilidade da investigação;

- Promove o investigador, que criou, partilhou os dados e recebeu os resultados;

- Fornece recursos importantes para educação e formação.

Os dados de investigação podem ser classificados em várias maneiras: pela sua natureza (imagens, vídeos, etc.), pela disciplina de origem, por método de recolha, de acordo com o tipo de investigação (observacional, computacional ou experimental) (Willis, Greenberg, e White 2012). Através da classificação dos dados podemos compreender as semelhanças e diferenças entre as diversas tipologias, de forma a potenciar o valor dos mesmos.

Os dados observacionais são dados capturados em tempo real, geralmente únicos e insubstituíveis, não podem voltar a ser recolhidos, por isso geralmente são arquivados indefinidamente, como é o caso do registo da temperatura do mar numa data específica (Simberloff et al. 2005; MANTRA 2014).

Por outro lado, os dados computacionais são resultado da execução de algum modelo ou simulação no computador, incluindo a descrição completa do hardware, software utilizado e dados de entrada, nos quais os metadados podem ter mais importância do que os dados de saída do modelo, como são disso exemplo os modelos climáticos. A preservação num repositório de longo prazo poderá não ser necessária, pois os dados podem ser reproduzidos. Contudo poderá ser essencial arquivar o próprio modelo e um conjunto de metadados (Simberloff et al. 2005; MANTRA 2014).

Já os dados experimentais são obtidos, sobretudo, em ambientes controlados e geralmente são reproduzíveis, e portanto podem não precisar de ser armazenados. Os dados experimentais, podem contudo ser caros de se obter, tornando complicado a reprodução de todas as condições experimentais. Por essa razão, nestes casos, a preservação a longo prazo de dados é necessário. (Simberloff et al. 2005; MANTRA 2014).

Compreender as tipologias e variação dos dados de investigação, torna-se importante para explorar os esquemas de metadados disponíveis para a sua descrição (Willis, Greenberg, e White 2012). Os dados de investigação variam muito, assim como o seu uso pretendido e potencial ao longo do tempo. Esta diferença reflete-se nos esquemas de metadados, suportando o uso e manipulação de dados. O estudo de normas e esquema de metadados é uma das tarefas deste trabalho, que será vista com mais detalhe num dos próximos parágrafos.

1.1.2 Curadoria digital

Tendo em conta os desafios levantados pelo aumento constante do volume de dados a comunidade científica lançou uma série de iniciativas destinadas à gestão dos dados de investigação. Uma destas iniciativas é o *Committee on Data for Science and Technology*[4] (CODATA) que concentra-se na melhoria da qualidade, confiabilidade, gestão e acessibilidade de dados em todas as áreas da ciência e tecnologia, facilita a cooperação entre investigadores e promove a importância da recolha, organização e utilização de dados de investigação.

Uma boa gestão de dados é fundamental para a qualidade dos dados de investigação. Se os dados são bem organizados, documentados, preservados e acessíveis, como resultado temos uma maior eficiência na gestão dos dados o que contribui para que estes sejam de melhor qualidade (Van de Eynden et al. 2013).

No geral, a curadoria digital refere-se às ações necessárias para manter os dados da investigação e outros objetos digitais em todo o seu ciclo de vida e ao longo do tempo para as gerações atuais e futuras utilizações.

Existem vários modelos de ciclo de vida que demostram o processo em geral com algumas diferenças, começando com a criação, seguindo com análise e finalizando com preservação e reutilização.

A curadoria digital envolve a manutenção, preservação e agregação de valor aos dados digitais de investigação em todo o ciclo da vida, reduzindo a ameaça de desvalorização, ou mesmo a perda dos dados. Além disso, reduz a duplicação do esforço na criação de dados de investigação, deixando os investigadores livres para outras atividades. A preservação digital requer inteligibilidade a longo prazo, resistindo às mudanças de tecnologias. A curadoria digital responde às necessidades de longo prazo, evoluindo em resposta às preocupações sobre a viabilidade, autenticidade e sustentabilidade dos conteúdos digitais (Walters e Skinner 2011; Higgins 2008).

Segundo o Digital Curation Centre (DCC), curadoria digital e preservação de dados são processos em curso, que exigem reflexão considerável, investimento de tempo e recursos adequados. O DCC apresenta um modelo do ciclo de vida da curadoria digital (Figura 1), que inclui várias etapas para o sucesso do processo de curadoria e de preservação de dados de investigação. O modelo está a ajudar na modelação de atividades de curadoria digital em vários ambientes como repositórios institucionais, arquivos digitais, entre outros. Possui o elemento central é dado (objeto digital ou base de dados), que mostra a importância do que está sendo

[4] CODATA International Council for Science: Committee on Data for Science and Technology. http://www.codata.org/about-codata/our-mission

curado e três rodelas de conjuntos de atividades: atividades completas do ciclo de vida, sequenciais e ocasionais. A primeira rodela de etapas trata da descrição e representação da informação, planeamento da preservação, observação e participação da comunidade, curadoria e preservação. A seguinte rodela cobre a conceitualização, criação e/ou receção dos dados, avaliação e seleção, arquivamento, acesso, utilização e reutilização e transformação. O terceiro conjunto inclui eliminação, reavaliação e migração. Este modelo é apenas uma sugestão e não obriga as organizações cumprir o ciclo deste a primeira etapa, pode ser alterada depende das necessidades de cada organização (Higgins 2008).

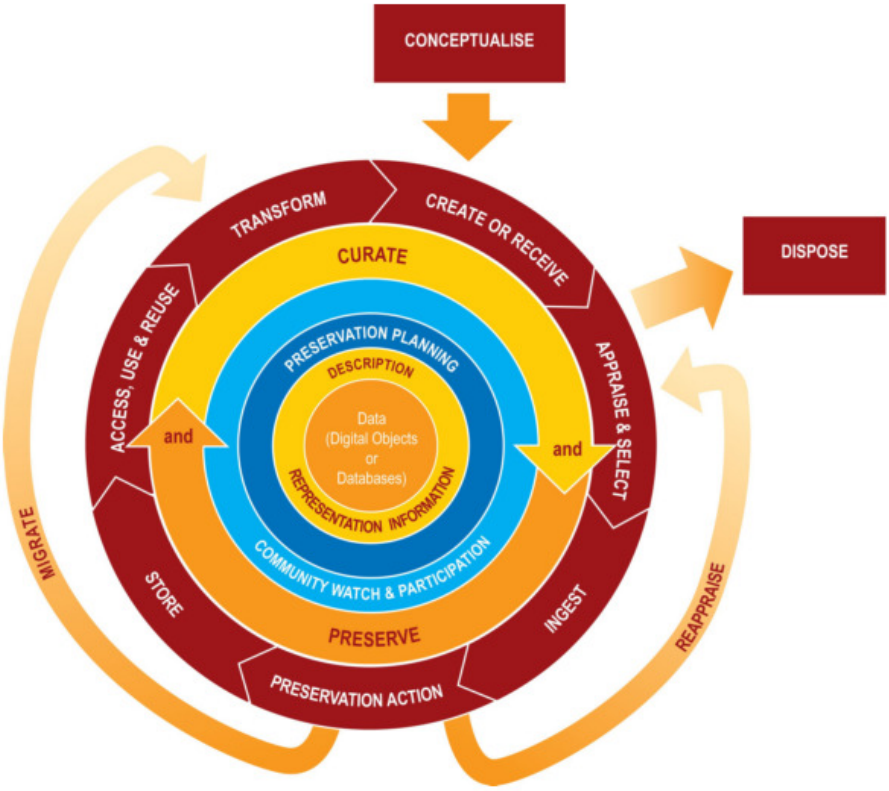


Figura 1 - Ciclo da vida da curadoria digital (Higgins 2008).

Além disto, no artigo "Digital Curation: A Life-Cycle Approach to Managing and Preserving Usable Digital Information" (Pennock 2007) a curadoria digital define-se como a gestão ativa e avaliação da informação digital em todo o seu ciclo de vida. O processo é importante pois os materiais digitais são frágeis e suscetíveis a mudanças durante o seu ciclo de vida. A curadoria digital é reconhecida como a forma de ajudar nestes processos e manter os recursos digitais autênticos e reutilizáveis.

Do ponto de vista de um investigador o ciclo da vida de dados possui 8 componentes[5], que podem ser vistos na Figura 2 (Strasser et al. 2012):

1. Planear: planear o ciclo da vida dos dados (como estes vão ser geridos e preservados).
2. Recolher: é importante recolher os dados de tal forma que garanta a sua facilidade de utilização posterior.
3. Assegurar: a qualidade dos dados é assegurada através de controlos e inspeções.
4. Descrever: os dados são descritos com precisão e com utilização de esquemas padronizados de metadados apropriados.
5. Preservar: os dados são submetidos a um arquivo adequado a longo prazo.
6. Descobrir: os dados potencialmente úteis estão localizados junto das informações relevantes sobre os dados (metadados).
7. Integrar: os dados de fontes diferentes são combinados para formar um conjunto homogéneo de dados que podem ser analisados.
8. Analisar: os dados são analisados.

Figura 2 - O ciclo de vida dos dados a partir da perspetiva de um investigador (Strasser et al. 2012).

Para motivar um ambiente de investigação colaborativo, a comunidade científica exige cada vez mais o planeamento da gestão de dados com políticas definidas para o tratamento de dados[6],[7].Tendo em conta que os metadados são fundamentais para a gestão de dados, os dados devem ser descritos com o maior detalhe possível. Sem uma descrição minuciosa é pouco provável que os dados possam ser facilmente compreendidos e efetivamente reutilizados. Em

[5] DataOne. Data Observation Network for Earth. https://www.dataone.org/data-life-cycle

[6] NSF: National Science Foundation. http://www.nsf.gov/bfa/dias/policy/dmp.jsp

[7] H2020 Model Grant Agreement for SME Instrument Phase 1 – Multi. 2015. http://ec.europa.eu/research/participants/data/ref/h2020/mga/sme/h2020-mga-sme-1-multi_en.pdf

consequência, as instituições de investigação começam dar mais atenção à qualidade de metadados, da mesma maneira que à qualidade dos próprios dados de investigação. Com a implementação de boas práticas de gestão de dados integradas desde o início do ciclo de vida dos dados, os investigadores podem dar resposta as exigências da comunidade científica.

1.1.3 Investigadores e gestão de dados de investigação

Atualmente a ciência tem a escala mundial, que não tem fronteiras, mas tem o constante crescimento de conhecimento. O acesso, partilha e reutilização de dados de investigação permitem diminuir a redundância e o tempo gasto na criação de novos dados, e gerar conhecimento de forma mais expedita. Por norma os dados de investigação são complexos, o que pode dificultar a sua interpretação por outros investigadores (Smith, Seligman, e Swarup 2008). Para reutilizar os dados é necessário ter a informação sobre as condições de produção, quais as condições associadas a essa reutilização, entre outros detalhes (Juan 2012). Por isso mesmo a gestão dos dados engloba um conjunto de atividades desafiantes para os investigadores.

Os metadados têm um papel fundamental neste contexto, uma vez que fornecem a informação necessária para garantir a preservação, localização e recuperação dos objetos digitais. Por este motivo, os investigadores devem assegurar que aos dados estão associados metadados com qualidade para desta forma assegurar que os dados possam ser interpretados por outros, ou mesmo por quem os produziu a longo prazo (Vickers 2006).

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 eles precisam (Sayogo e Pardo 2013; Barton, Currier, e Hey 2003). Por isso o formato da descrição, sintaxe e semântica devem ser normalizado para promover a partilha dos dados.

Devido ao desenvolvimento das tecnologias de informação, apareceram novos tipos de recolha, partilha e preservação dos dados. A comunidade científica está a criar novas maneiras de partilhar os dados e lidar com questões de qualidade de dados, normas e proteção, uso ético e responsabilidade dos dados partilhados. As respostas para estas questões podem influenciar a motivação dos investigadores para publicação dos dados de investigação. O interesse dos investigadores pode depender de gestão de dados. A falta de ferramentas adequados para gestão, falta de tempo ou experiência na gestão de dados, restrições legais, medo de exploração ou uso inadequado dos dados, limitação de conhecimento das normas podem afetar e diminuir a vontade de investigador para partilhar os dados (Sayogo e Pardo 2013; Swan e Brown 2008). Mas apesar de existirem lacunas na motivação dos investigadores, a crescente procura por transparência dos dados de investigação, juntamente com a pressão de uma grande

variedade de interessados combinam-se para alimentar a partilha e citação de dados (Mayernik 2012).

De acordo com Arzberger et al. (2004), a falta de publicação e partilha dos dados de investigação tem razões específicas, desde aspetos tecnológicos, organizacionais, incluindo elementos financeiros e orçamentais, aspetos legais, políticos e comportamentais. Analisando essas razões, os autores definem o controlo de qualidade dos dados e a descrição dos mesmos como uma das características de aspeto tecnológico, a satisfação e a usabilidade como características de aspeto organizacional (Arzberger et al. 2004).

O acesso aberto e a partilha dos dados reforçam a investigação científica aberta, incentivam a diversidade de análise e opinião, promovem novas investigações, possibilitam a verificação de hipóteses e métodos de análise, apoiam estudos sobre métodos de recolha de dados e medição dos mesmos, facilitam a formação de novos investigadores e permitem a criação de novos conjuntos de dados. Em geral, a partilha dos dados e o acesso aberto não só ajudam a maximizar o potencial das novas tecnologias, mas fornecem maior retorno do investimento público em investigação. A limitação de acesso aos dados de investigação pode ser uma barreira à inovação e cooperação interdisciplinar e internacional, e pode bloquear o processo de verificação da exatidão dos dados. Como resultado, a qualidade dos dados pode ser mais baixa do que seria no caso se estivessem livremente disponíveis, reduzindo mais uma vez o retorno do investimento em produção de dados (Fienberg, Martin, e Straf 1985; Arrison et al. 2009).

De acordo com o conjunto de diretrizes desenvolvidas pelo *Organisation for Economic Co-operation and Development (OECD)* para simplificação de acesso e partilha de dados de investigação, a qualidade, interoperabilidade e eficiência são áreas importantes na gestão de dados. Além disto, os metadados com qualidade podem aumentar a confiança para reutilização dos dados. Os dados devem ser acompanhados por metadados suficientes para fácil localização e reutilização dos mesmos; têm de ser armazenados em repositórios, utilizando atualizadas. Para que dados sejam utilizados no futuro, os investigadores têm de estabelecer um plano de gestão de dados e fornecer a maior parte dos metadados, porque só eles conhecem os seus dados bem e conseguem avaliar o que deve ser preservado e descrito para mantem o seu valor a longo prazo (Arrison et al. 2009).

Geralmente, os dados são mal documentados porque os investigadores não os descrevem no momento oportuno, ou seja no momento da recolha, podendo perder-se informação importante (Fienberg, Martin, e Straf 1985; Akmon et al. 2011). Além disto, muitos projetos de investigação não tem uma infraestrutura integrada de dados, o que muitas vezes resulta em uso de ferramentas para gestão de dados não normalizados. Este problema surge especialmente em projetos pequenos, que têm exigências mínimas na utilização das ferramentas tecnológicos, recursos humanas e conhecimentos na curadoria digital (Akmon et al. 2011).

A falta de tempo e ferramentas, falta de experiência e conhecimento na gestão dos dados, dificuldade na descrição de dados, questões de fornecimento de metadados de boa qualidade são alguns fatores que desmotivam a partilha e publicação de dados de investigação. Metadados de má qualidade ou inadequados são obstáculos para localização e acesso aos conjuntos de dados de investigação por outros investigadores e organizações. Os centros de dados aplicam os procedimentos rigorosos para garantir que os conjuntos de dados de investigação possuem padrões de qualidade em relação à estrutura e formato dos dados, tal como os metadados associados. Mas muitos investigadores não têm as competências para atender a essas normas, sem ajuda de especialistas (Swan e Brown 2008).

Assim sendo, podemos afirmar que, em geral, os problemas de qualidade, dificuldade do processo de descrição de dados de investigação, falta de ferramentas adequadas para criação de metadados, são problemas comuns em diversos domínios de investigação. O objetivo principal deste trabalho é melhorar a qualidade na descrição de dados e facilitar o processo da descrição na plataforma Dendro, aumentado a usabilidade e simplicidade da mesma e ajudando a motivar os investigadores para partilha e publicação dos dados.

1.2 Metadados na gestão de dados

Os metadados são informação descritiva ou contextual que está associada aos objetos digitais, que garante a preservação, localização, recuperação e reutilização dos mesmos. Em geral os dados de investigação são de difícil compreensão e reutilização, por isso a descrição de dados de investigação é essencial. Ao longo dos anos foram desenvolvidos diferentes esquemas e normas de metadados com o objetivo de normalizar a descrição. O nível de detalhe de descrição de dados de investigação muitas vezes é escassa, desta forma afirma-se que os metadados de qualidade podem contribuir para garantir o acesso, interpretação e consequente reutilização dos mesmos. Desta forma em particular a atenção deverá ser dada à qualidade de metadados.

1.2.1 Definição de metadados

Em geral, os dados de investigação são produzidos durante experiências ou observações para posterior interpretação. Para conseguir esta finalidade eles devem ser acompanhados com a máxima informação possível para auxiliar a sua compreensão. Esta informação adicional pode

ser fornecida através dos metadados, a fim de ser devidamente interpretada por humanos ou *software*. Os elementos de metadados devem ser descritos com precisão[8].

Inicialmente o conceito "*metadata*" foi definido pelo Jack E. Myers em 1969 para descrever conjuntos de dados e produtos. Seguidamente foi adotado por diversas comunidades, entre outras, as da ciência da informação, da biblioteconomia e da ciência de computação (Caplan 1995; Greenberg 2005). Os metadados são classificados como "*the backbone of digital curation*" (Higgins 2007), porque permitem identificar, localizar, compreender, preservar e utilizar os objetos digitais (Ercegovac 1999; Harvey 2005). Geralmente, os metadados são definidos como dados sobre outros dados (Caplan 1995), dados estruturados sobre os dados (Woodley, Clement, e Winn 2003; Duval et al. 2002), ou informação sobre a informação (Guenther e Radebaugh 2004).

Contudo, alguns autores afirmam que a definição não é muito precisa, principalmente no que diz respeito à diferença entre dados e metadados, sendo necessário invocar um contexto ou seja, um ponto de referência para identificar o que queremos dizer com metadados em uma dada situação. (Bargmeyer e Gillman 2000).

Numa perspetiva global os metadados são informação descritiva ou contextual que está associada a um objeto ou recurso. Segundo Harvey (2005), Guenther e Radebaugh (2004) os metadados podem ser classificados de acordo com a finalidade da descrição, quer seja uma descrição mais técnica que inclui o tamanho e formato de ficheiros, ou administrativa que descreva processos aplicados ao longo do tempo. Por outro lado metadados descritivos, estruturais e para efeitos de preservação vão permitir a localização, identificação e a relação com outros objetos digitais (Qin, Ball, e Greenberg 2012).

Todavia existem os autores que classificam os metadados por 4 tipos principais, excluindo o tipo "*preservation metadata*" ou por 6 tipos, adicionando "*provenance e rights metadata*". Mas no geral, não importa em qual tipo os metadados estão, desde que estes sejam fornecidos, constantemente criados e acessíveis através do sistema (Treloar e Wilkinson 2008; Corrado e Moulaison 2014).

De acordo com (Willis, Greenberg, e White 2012) os metadados para a representação e descrição de dados de investigação são uma componente essencial da comunicação científica contemporânea. Ao longo das últimas décadas diferentes comunidades desenvolveram esquemas de metadados para facilitar a documentação, partilha, arquivo e reutilização de metadados de dados de investigação. Muitos desses desenvolvimentos estão associados a repositórios de dados específicos de cada disciplina. Abertura, partilha e interoperabilidade de

[8] Davenhall C., Scientific Metadata.National e-Science Centre. DCC. http://www.dcc.ac.uk/resources/curation-reference-manual/chapters-production/scientific-metadata

sistemas de metadados podem ajudar efetivamente a preservar e fornecer acesso a dados científicos através das disciplinas.

1.2.2 Esquemas e normas genéricas

No contexto dos metadados, o termo *schema* define os elementos admissíveis, os seus tipos de dados, os valores admissíveis, o formato que um valor deve tomar e em alguns casos o comprimento máximo e mínimo. Um esquema de metadados destina-se a ser compartilhado com outras pessoas. Como resultado, vários registos de esquema de metadados podem servir em repositórios públicos para facilitar a partilha na comunidade em geral (Mitchell 2015). O principal objetivo dos esquemas é facilitar a troca e processamento automático de informação de diferentes fontes. Semântica, regras de conteúdo e sintaxe são três aspetos de metadados que podem ser especificados em esquemas de metadados (Caplan 2003).

Ao longo dos anos foram desenvolvidos diferentes esquemas e normas de metadados, de que são exemplo: MARC (Machine Readable Cataloging) para itens bibliográficos, EAD (Encoding for Archival Data) para coleções de arquivos intactas com uma proveniência comum, CDWA (Categories for the Description of Works of Art) para os objetos de arte, VRA Core (Visual Resources Association Core) para substitutos visuais de obras de arte e arquitetura, Dublin Core para recursos da Web e outros (Baca 2003). Alguns deles são projetados para catalogar documentos de bibliotecas, outros foram criados para localizar os recursos eletrónicos num ambiente web.

Ao mesmo tempo as comunidades científicas também estão a propor esquemas de metadados. Por exemplo: EML (Ecological Metadata Language) e DwC (Darwin Core) para descrição de coleções biológicas e ecológicas; CIF (Crystallographic Information File) e mmCIF (Macromolecular Crystallographic Information File) para descrição de estruturas cristalográficas físicas e biológicas; DDI (Data Documentation Initiative) para descrição os dados das ciências sociais; ThermoML para descrição de propriedades termofísicas e termoquímicas e NeXMLpara descrição das árvores filogenéticas. Estes esquemas são usados para descrever os dados experimentais, observacionais e conjuntos de dados estatísticos, incluindo coleções físicas e digitais. (Willis, Greenberg, e White 2012).

Uma melhor compreensão dos objetivos de metadados articulados por comunidades individuais ajuda a definir uma abordagem universal para a descrição de dados de investigação. No seu estudo Willis, Greenberg e White definem 11 requisitos fundamentais para os esquemas de metadados para documentação de dados de investigação, dos quais integralidade e simplicidade, por exemplo, são aplicáveis para esquemas de dados de investigação.

A escolha do esquema de metadados mais adequado ou normas para descrever os objetos e materiais específicos é apenas o primeiro passo na construção de um recurso de informação

eficaz e utilizável (Baca 2003). Infelizmente os repositórios que contêm diferentes domínios científicos confrontam-se com vários problemas na descrição de dados de investigação de cada domínio. Por isso a combinação dos elementos através de perfis de aplicação, dá mais possibilidade para descrever os dados de investigação com mais precisão e adequação a necessidade de descrição particulares (Heery e Patel 2000). No caso de não haver elementos para descrição específica, existe a possibilidade de criar os metadados para esta finalidade (Gattelli 2015).

Tendo em conta a importância e desafios reconhecidos ao processo de descrição de dados, a plataforma Dendro tem vindo a ser desenvolvida na Universidade do Porto. Através do Dendro pretende-se apoiar os investigadores na preparação dos dados de investigação, nomeadamente na captura atempada de metadados para que os dados sejam transferidos para repositórios externos em condições de serem preservados. O Dendro utiliza os conjuntos de descritores dependentes do domínio e representa-os na forma de ontologias que permitem descrever os dados de investigação de domínios específicos, por exemplo Produção de Hidrogénio ou Oceanografia Biológica.

1.2.3 Qualidade de metadados

Atualmente milhares de recursos digitais são publicados na Web todos os dias e garantir a qualidade dos metadados beneficia os potenciais utilizadores. Uma vez que as propriedades de recursos digitais são refletidas nos metadados a qualidade de metadados também deve ser de primordial importância.

Qualidade de metadados é um assunto que tem recebido menos atenção do que qualidade de conteúdo e dados. Os metadados têm qualidade quando servem o seu propósito – permitir ao utilizador encontrar e compreender os dados que são descritos (Bargmeyer e Gillman 2000).

De acordo com a revisão da literatura existem vários argumentos, que apoiam a necessidade de garantir a qualidade dos metadados, por exemplo:

- Metadados de má qualidade podem significar que um recurso é essencialmente invisível dentro de repositório e se torna inutilizável (Barton, Currier, e Hey 2003);
- Metadados de alta qualidade garantem a precisão e acesso completo aos recursos digitais e permitem aos utilizadores finais encontrar e recuperar os recursos que eles precisam com maior facilidade (Shankaranarayanan e Even 2006);
- O nível de descrição muitas vezes é deficiente: a maioria dos elementos de metadados nunca ou raramente são utilizados para descrição (Sanz-Rodriguez, Dodero, e Sánchez-Alonso 2010);
- A facilidade de utilização e eficácia de qualquer biblioteca digital é afetada pela qualidade de metadados. Metadados de baixa qualidade podem torná-la quase

inutilizável, enquanto a alta qualidade de metadados podem levar a uma maior satisfação e maior utilização dessas bibliotecas (Stvilia et al. 2004).

Erros de metadados em bibliotecas digitais ou em repositórios podem ocorrer de várias formas. Existindo, estes podem facilmente bloquear o acesso ao material disponível. Na opinião de Jeffrey Beall (2006) esses erros são mais graves, porque através de metadados acontece a pesquisa dos objetos. As bases de dados de imagens são mais vulneráveis neste caso, porque praticamente todo o acesso às imagens ocorre através de metadados (Beall 2006).

Contudo os autores Shreves, Riley e Milewicz (2006) caracterizam os metadados de qualidade como sendo metadados compartilháveis, qualificados para troca com outros sistemas distribuídos. Metadados compartilháveis devem ter os seguintes aspetos: o conteúdo de metadados deveria ser otimizado para partilha, contendo uma descrição suficiente; o registo de metadados deve ser consistente tanto na sua presença como ausência. Por exemplo, se um campo não está preenchido em todos os registos, um agregador é capaz de ignorar esse campo no ecrã e pesquisa; os registos compartilháveis são coerentes. Ou seja, os utilizadores devem ser capazes de interpretá-los à primeira-vista. Os metadados compartilháveis devem ter contexto e devem ser entendidos do contexto do domínio ou local da criação; os registos têm de confiar no estabelecimento de comunicação entre os prestadores de serviços e fornecedores de dados; finalmente, devem estar em conformidade com normas reconhecidas (Shreeves, Riley, e Milewicz 2006).

Alguns metadados são criados automaticamente, outros com uso de esquemas, mas os erros podem ocorrer em todos os tipos de descrição de dados, por isso é importante dar mais importância à criação de metadados com qualidade e tentar controlar este processo com mais precisão.

De acordo com a revisão de literatura, existem várias dimensões, critérios, métricas e indicadores de qualidade que tem papel importantíssimo na avaliação de qualidade de metadados e de usabilidade de repositórios digitais, tal como precisão, conformidade com as expectativas, exaustividade, satisfação e tempo de tarefa. Para conseguir escolher as métricas para avaliação de registos de metadados no Dendro, no próximo parágrafo iremos analisá-los com mais detalhe.

1.2.4 Dimensões, métricas e indicadores de qualidade de metadados

O crescimento de repositórios digitais levou a um aumento no número de estudos ligados ao problema de qualidade de metadados e mecanismos de avaliação dos mesmos (Stvilia et al. 2007). Para conseguir avaliar a qualidade de metadados é preciso entender como medir esta qualidade e para isto temos de definir as dimensões, métricas e indicadores de qualidade.

A dimensão de qualidade é uma série de aspetos relacionados com a qualidade de dados que queremos analisar. São usadas para definir, medir e gerir a qualidade dos dados e informação. As métricas de qualidade são medidas específicas que expressam cada dimensão de qualidade e descrevem o que está a ser medido e como vai ser medido. Os indicadores são medidas estatísticos que expressam o grau (nível) de métricas de qualidade. No geral, os indicadores são uma maneira quantitativa de expressar as métricas de qualidade (McGilvary 2008; Palavitsinis 2013; Heldman 2005).

No seu trabalho, Lee et al (2002) e Stvilia et al (2007) indicam as seguintes dimensões de qualidade de informação:

- *Intrinsic Metadata Quality,* a dimensão que reconhece que os metadados podem ter exatidão (*correctness)*, independentemente do contexto em que estão a ser usados;

- *Contextual Metadata Quality,* que reconhece que a qualidade pode variar de acordo com a tarefa específica. Mede relações entre a informação e alguns aspetos do seu contexto de utilização;

- *Representational Metadata Quality,* que permite entender se os metadados são fáceis de entender e apresentados de uma forma clara;

- *Accessibility Metadata Quality*, que trata a facilidade com que os metadados são obtidos (Lee et al. 2002; Stvilia et al. 2007).

Além disto, existem outras dimensões de qualidade, por exemplo: *policy, users, technology, contents, standards* e *rules* (1997), Moen, Stewart, and McClure (1997, 1998)), e mais métricas para avaliar a qualidade de metadados. Há métricas que servem para avaliação de objetos digitais (*accessibility, significance, similarity, timeliness*), outros para avaliação de especificações de metadados (*completeness, conformance*) e para avaliação de serviços (*efficiency, confidence*) (Moreira et al. 2009).

A Tabela 1 mostra quais as métricas são mais comuns e populares na visão de autores diferentes (Bruce e Hillmann 2004; Stvilia et al. 2007; Moreira et al. 2009; Palavitsinis 2013; Ochoa e Duval 2006; Alkhattabi, Neagu, e Cullen 2010):

Tabela 1 – Métricas para avaliação de qualidade de metadados

Métricas de qualide de metadados	Bruce & Hillmann (2004)	Ochoa et al. (2006)	Stvilia & Gasser (2007)	Moreira & Gonçalves (2009)	Alkhattabi et al. (2010)	Palavitsinis (2013)
Accessibility	+	+	+	+	+	
Accuracy	+	+	+			+
Appropriateness					+	+
Cohesiveness			+			
Completeness	+	+	+	+	+	+
Complexity			+			
Confidentiality				+		
Conformance to expectations	+	+		+		
Consistency	+	+	+		+	
Correctness						+
Currentness (atualidade)			+			
Efficiency (eficiência)				+		
Informativeness			+			
Naturalness			+			
Objectiveness					+	+
Overall Rating						+
Precision			+			+
Provenance	+	+				
Significance				+		
Similarity				+		
Timeliness	+	+		+	+	

Com esta análise podemos ver que as métricas *accessibility, accuracy, completeness, conformance to expectations* e *consistency* são mais populares. Basicamente, as métricas mostram onde podem ocorrer problemas com metadados.

Além da qualidade de metadados a usabilidade também tem grande importância na avaliação de sistemas de *software*, tal como sistemas de informação e web sites. No geral a usabilidade de um repositório científico pode afetar a motivação e envolvimento de utilizadores, bem como a utilização do repositório (T. Zhang, Maron, e Charles 2013).

De acordo com a literatura, existem várias métricas de avaliação de usabilidade, tais como: *effectiveness*, *efficiency, satisfaction, task time, productivity.* (Seffah et al. 2006). Essas métricas podem ajudar a avaliar o desempenho de um repositório digital, analisando o conteúdo, funcionalidade e interface do ponto de vista de utilizador, tal como podem ajudar a identificar onde existem falhas e onde as melhorias precisam ser feitas.

A análise de usabilidade da plataforma Dendro, pode ajudar a avaliar se a criação e implementação de vocabulários controlados facilitam o processo de descrição de dados de investigação no Dendro. Seguidamente, verificar se há um incremento na satisfação dos

investigadores e diminuição do tempo gasto na criação dos metadados. Por isso, junto com as métricas de qualidade de descrição de dados, vamos definir as métricas para avaliação de usabilidade da plataforma Dendro.

1.3 Vocabulários controlados

Os vocabulários controlados são um conjunto de conceitos padronizados e agrupados de acordo com o seu significado, que ajudam a evitar ambiguidades na interpretação léxica e visam facilitar a entrada da informação. O processo de descrição de dados de investigação que exige muito esforço e tempo, pode diminuir o interesse e motivação de investigadores para descrição e publicação dos seus dados. Neste contexto, os vocabulários controlados apresentam-se como uma ferramenta que pode ajudar simplificar o processo de descrição de dados na plataforma Dendro e, provavelmente, melhora a qualidade dos metadados.

1.3.1 Definição de vocabulários controlados

Os vocabulários controlados são importantes, porque podem facilitar o processo de descrição de dados e diminuir os erros (Fidel 1992). 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).

Um vocabulário controlado é um instrumento de informação que contém palavras padronizadas e frases usadas para se referir a ideias, características físicas, pessoas, lugares, eventos, assuntos, e muitos outros conceitos (Harpring 2010); é uma lista restrita de palavras ou conceitos, normalmente utilizados para catalogação descritiva, etiquetagem ou indexação (Hedden 2010); é um instrumento para indicar as atividades ou funções, criando a confiança no sistema (Smit e Kobashi 2003). Assim sendo, os vocabulários controlados permitem a categorização, indexação e recuperação de informação.

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 (*pick-list*) é frequentemente utilizada para elementos de metadados administrativos e estruturais. Muitas vezes são exibidos como *drop-down lists* ou como caixa de seleção de itens com botão, que permite mostrar todas as opções existentes. O ficheiro de autoridade (*authority file*) é um vocabulário controlado que inclui sinónimos ou variantes para cada conceito. A taxonomia é um vocabulário controlado, em que todos os termos pertencem a uma única estrutura hierárquica e tem relação do tipo pai-filho com outros termos. Os tesauros é uma espécie de dicionário que contém sinónimos ou expressões alternativas para cada termo e possivelmente até antónimos (Hedden 2010). 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.

Da mesma maneira Amy J. Warner (2002) afirma que as taxonomias, tesauros e sistemas de classificação são vocabulários controlados e são listas organizadas de palavras e frases, usadas para etiquetar o conteúdo e em seguida conseguir encontra-lo através de pesquisa. Literalmente, existem centenas e milhares de vocabulários controlados criados em formatos eletrónicos, mas a maioria foi criada dentro das empresas para satisfazer objetivos próprios. Com esta ideia em mente é sugerido que provavelmente o melhor é construir o seu próprio vocabulário controlado (Warner 2002).

L. Bermudez, E. Montgomery, S. Miller et al. (2011) classificam os vocabulários controlados pela sua:

- finalidade: *discovery vocabulary*, que ajuda os utilizadores a encontrar os dados; *usage vocabulary*, que auxilia na interpretação dos dados; *semantic vocabulary*, que fornece o significado mais compreensível para humanos; *syntactic vocabulary*, que traduz a informação para formato mais legível por máquina.

- forma: *flat*, que fornece um conjunto de conceitos necessários, que podem ser utilizados; *multilevel*, que é baseado em *flat*, atribuindo a cada conceito uma categoria; *relational*, que fornece um conjunto de conceitos e relação entre eles.

- funcionalidade: a Tabela 2 resume as categorias e tipos de vocabulários controlados com base nas suas funcionalidades; por exemplo, as taxonomias estão relacionados com o *Multilevel Controlled Vocabulary*.

Tabela 2 - Relações entre as grandes categorias, com base em suas funcionalidades (Bermudez et al. 2011).

Broad, Form-based Category	Functionality-based Type	Description
Flat Controlled Vocabulary	Authority File	List of terms
	Glossary	List of terms and definitions within a specific domain
	Dictionary	List of terms, definitions, and additional information
	Gazetteer	List of place names
	Code List	List of codes (e.g., abbreviations) and definitions
Multilevel Controlled Vocabulary	Taxonomy	Terms classified into categories
	Subject Heading	Terms classified into categories, which may be broad classes
Relational Controlled Vocabulary	Thesaurus	Set of terms and relationships among individual values
	Semantic Network	Set of terms/concepts and directed relationships
	Ontology	Set of terms and relationships among terms, enhanced by additional information provided by rules and axioms.

Projetar os elementos de metadados para um repositório digital e projetar o vocabulário controlado são processos integrados. Cada vocabulário controlado, taxonomia ou ficheiro de autoridade corresponderá a um elemento diferente de metadados. As decisões iniciais no desenvolvimento dos vocabulários devem ser tomadas após decidir quais os elementos na descrição que deverão ter vocabulários. Há elementos de metadados que não precisam de ter vocabulário controlado. Por exemplo, elementos como título ou nome do ficheiro devem permitir texto livre; os campos numéricos tais como tamanho e data, também não usam vocabulários controlados, embora possa haver políticas relativas ao formato de entrada (Hedden 2010). Tal como o estudo efetuado por Zhang, Ogletree e Greenberg mostra, a maioria dos utilizadores têm um interesse enorme em ter e usar ferramentas avançadas para selecionar conceitos definidos em vocabulários controlados, mas também querem manter a possibilidade de introdução do texto livre (Y. Zhang et al. 2015).

1.3.2 Estudos sobre vocabulários controlados

A existência de vocabulário controlado pode melhorar a qualidade de descrição de dados e seguidamente o acesso aos mesmos. Pode ajudar a estabelecer a navegação, servir como base para personalização dos recursos e preparação de projetos ligados à gestão do conhecimento e dos dados, uma vez que muitos deles exigem este tipo de estrutura. Em geral o vocabulário controlado é um "mapa conceptual" e quando bem concebido e atualizado pode ajudar a diminuir os erros humanos (Leise, Fast, e Steckel 2002).

Muitas iniciativas fornecem exemplos de vocabulários controlados:

- *Marine Metadata Interoperability* no seu site[9] fornece referências para vocabulários controlados específicos, assim como: *List of sensors developed by the Alliance for Coastal Tecnhologies; Instrument code list from Argo; Collection of vocabularies and ontologies from TDWG(Taxonomic Databases Working Group - Biodiversity Information Standards) for biodiversity;*
- *ISO* 639 codifica os nomes das línguas naturais. Versão *ISO* 639-1 fornece designação de duas letras e *ISO* 639-2 – de três. Pode ser visto na Tabela 3;
- *Darwin Core* recomenda utilizar os vocabulários controlados da lista definida em *ISO 3166-1-alpha-2* para descritor *countryCode* (Wieczorek et al. 2012).

Tabela 3 - Os códigos para representação de nomes de idiomas (excerto)[10]

ISO 639-2 Code	ISO 639-1 Code	English name of Language
aar	aa	Afar
abk	ab	Abkhazian
ace		Achinese
ach		Acoli
ada		Adangme
ady		Adyghe; Adygei

Um dos vocabulários controlados mais utilizados é LCSH (Library of Congress Subject Headings), que é a única lista de assuntos que foi aceite como um padrão mundial. Este vocabulário é composto por vários vocabulários de domínios específicos e está disponível através de *Classification Web*, sendo atualizado diariamente (Y. Zhang et al. 2015).

Além disto Znahg, Ogletree e Greenberg falam sobre as tecnologias de *Simple Knowledge Organization System* (SKOS) e *Helping Interdisciplinary Vocabulary Engineering* (HIVE): o SKOS é um modelo desenvolvido pelo consórcio W3C, que facilita a interação entre diferentes sistemas de informação devido à padronização de sistemas de organização do conhecimento, tais como tesauros, esquemas de classificação, taxonomias e outros tipos de normalização de

[9] Marine Metadata Interoperability. Vocabularies References. https://marinemetadata.org/conventions/vocabularies

[10] ISSO 639.2 - Codes for the Representation of Names of Languages. http://www.loc.gov/standards/iso639-2/php/code_list.php

vocabulário no âmbito de Web Semântica[11]. O HIVE é tecnologia que permite criação automática de metadados, que integra dinamicamente vocabulários codificados com SKOS, é uma tecnologia *Linked Open Data* alinhada com as *Linked Open Vocabularies activities*[12]. A ideia geral disto é promover a interoperabilidade entre repositório de dados, bibliotecas e arquivos, para indexação de dados de investigação de várias áreas específicas.

Outros exemplos de vocabulários controlados são o *Thesaurus of Geographic Names* (TGN) e *Union List of Artist Names* (ULAN). O TGN atualmente contém 1 115 000 nomes geográficos. O foco de cada registo de TGN é um lugar. Existem aproximadamente 895 000 locais representados no banco de dados de TGN, que são identificados por um identificador numérico único. O ULAN tem aproximadamente 293 mil nomes e informação sobre artistas e criadores de obras culturais. Os registos de ULAN podem incluir nomes próprios, apelidos e pseudónimos e são geralmente apresentados como uma lista em conformidade com as normas *ISO* e *NISO* de construção de tesauros (Harpring 2010).

Enquanto os vocabulários controlados podem funcionar como normas para valores de dados eles mesmos deveriam ser construídos de acordo com normas, tal como:

1. *ANSI/NISO Z39.19 – 2004: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies*
2. *BS 8723 -1:2005, BS 8723 – 2:2005, BS 8723 – 3:2007, BS 8723 – 4:2007: Structured Vocabularies for Information Retrieval*
3. *ISO 2788:1986: Documentation – Guidelines for the Establishment and Development of Monolingual Thesauri*
4. *ISO 5964:1985: Documentation – Guidelines for the Establishment and Development of Multilingual Thesauri* (Harpring 2010).

Durante a sua elaboração deve-se ter em mente que os vocabulários controlados podem ou não conter "sinónimos" – *non-preferred terms*, que servem como pontos de entrada adicional e fazem um vocabulário controlado mais eficaz, mas ao mesmo tempo mais complexo para desenvolver, implementar e manter (Hedden 2010). Além disto, podemos utilizar a taxonomia que geralmente significa um vocabulário controlado com uma estrutura hierárquica. Uma pequena taxonomia tem o espaço para incluir todos os conceitos e também é mais simples de implementar em uma interface para a seleção de metadados. Uma grande taxonomia seria demasiado longa para percorrer todos os níveis da hierarquia, sendo mais complicada de implementar. Mas pode ser criada em forma de Tópicos e Subtópicos, como pode ser visto na Figura 3, o que é muito desejável para os utilizadores finais:

[11] W3C Semantic Web. Introduction to SKOS. http://www.w3.org/2004/02/skos/intro

[12] Helping Interdisciplinary Vocabulary Engineering. https://code.google.com/p/hive-mrc/

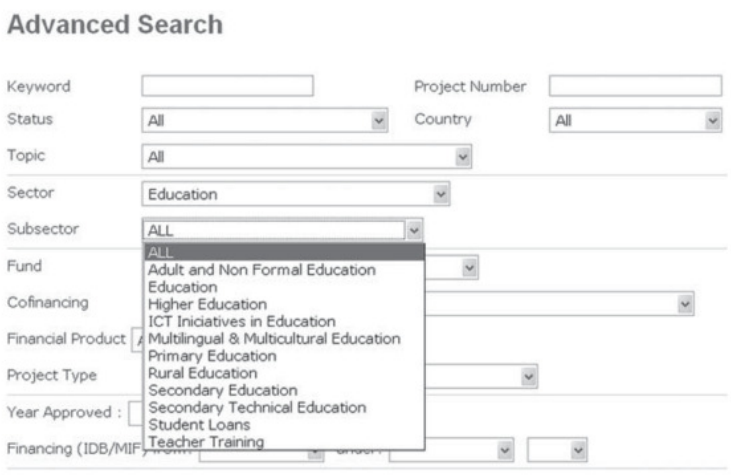


Figura 3 - Uma taxonomia hierárquica dos sectores com dois campos de metadados em pesquisa avançada para projeto do *Inter-American Development Bank* (Hedden 2010).

P. Harping (2010) propõe as questões básicas relacionadas com desenvolvimento de um novo vocabulário controlado. Quando possível, devem ser:

- utilizadas as normas oficiais para escolher os conceitos para vocabulário controlado;

- estabelecidos o foco lógico de cada registo de vocabulário e a estrutura de dados e relacionamento entre vários tipos de dados;

- utilizados ambos tipos de campos – controlado e texto livre;

- estabelecida a informação mínima necessária para cada registo, identificando que informação é necessária e qual é opcional;

- identificar e adotar as regras para construção do vocabulários controlados (Harpring 2010).

Algumas implementações de vocabulários controlados usam listas de seleção (*pick-lists*) para ajudar o utilizador a escolher um termo num conjunto de conceitos pré-definidos. Eles são implementados através de *drop-down lists,* que permite mostrar todas as opções existentes e escolher o necessário. Normalmente estas listas não incluem sinónimos e são mais fáceis de implementar, mas não deixam de ser uma boa sugestão para melhoria de qualidade na descrição de dados. Neste trabalho optou-se pela criação de vocabulário controlado deste tipo.

Um dos objetivos deste trabalho é agilizar o processo de descrição de dados e ao mesmo tempo obter metadados com maior detalhe. Neste contexto os vocabulários controlados apresentam-se como uma boa ferramenta para simplificar o processo de descrição de dados no Dendro e, provavelmente obter-se uma melhoria da qualidade dos metadados.

1.4 Expressões regulares

Mais uma ferramenta que pode ser utilizada na simplificação do processo de descrição de dados no Dendro são as expressões regulares. As expressões regulares (*regex*) permitem trabalhar com texto, criando padrões de texto; são um mecanismo que permite definir o padrão de formato de entrada do texto. Além disto, elas podem ser utilizadas para validação a estrutura de campo de dados tipo *string* e correspondência do mesmo com as regras definidos para este campo (Skoglund 2011; Grimalovskii 2013; Standen 2010).

Em geral, as expressões regulares são um meio poderoso, flexível e eficiente de processamento de texto. As templates universais de expressões regulares são muito parecidas com uma linguagem de programação que serve para descrição e análise do texto. Com o apoio adicional das ferramentas específicas de programação elas podem inserir, excluir, selecionar e executar as operações com dados de texto (Friedl 2006; Goyavaerts e Levithan 2009).

Aqui estão alguns exemplos de utilização de expressões regulares:

1) Testar se o número de telefone tem os dígitos corretos;
2) Verificar se endereço de e-mail está em formato válido;
3) Procurar um documento e substituir todas as ocorrências de Bob, Bobby para Robert;
4) Definir o formato de entrada de Data (YYYY-MM-DD) (Skoglund 2011).

Para entender o funcionamento das expressões regulares, pode-se ver o exemplo da validação de uma entrada de data em formato DD.MM.YYYY. A expressão regular:

(0?[1-9]|[12][0-9]|3[01])\.(0?[1-9]|1[012])\.((19|20)\d\d), onde

DD são os dois dígitos para o Dia, MM são os dois dígitos para o Mês e YYYY – são os quatro dígitos para o Ano.

De acordo com a expressão regular definida, os valores de DD podem ser os seguintes: 01-09 ou 1-9; 10-19 ou 20-29; 30 ou 31. Da mesma maneira os valores de MM podem ser: 01-09 ou 1-9; 10,11 ou 12 e do YYYY: qualquer valor do intervalo 1900 – 2099. Desta forma, a data 23.08.2010 está validada com esta expressão regular.

Juntamente com outras ferramentas de programação, consegue-se efetuar o acompanhamento de todo o processo com o auxílio de dicas de contexto (*Tooltip´s*), que conseguem ajudar os utilizadores no preenchimento correto dos dados. Na Figura 4 é apresentado Tooltip com o formato sugerido para preenchimento de descritor Authors - Last Name, *First Names.*

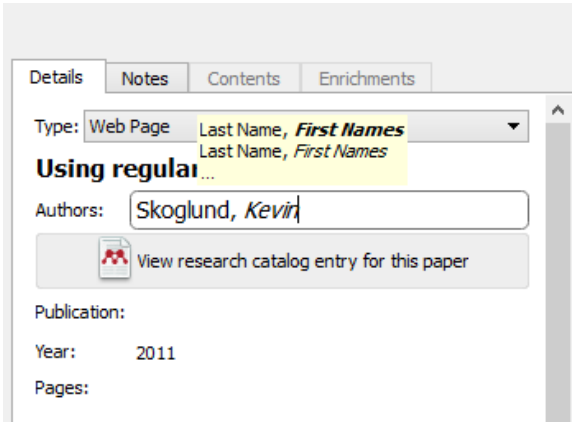


Figura 4 - *Tooltip* na aplicação Mendeley Desktop

De acordo com a literatura, as expressões regulares podem reduzir o esforço manual na introdução da informação e ajudar na qualidade de dados. As expressões regulares complexas aquando da sua boa configuração e funcionamento trazem sempre benefícios da sua utilização (Friedl 2006).

Existem várias aplicações *on-line* criadas para verificação expressões regulares, que podem ajudar na criação e validação da sintaxe de expressões (Goyavaerts e Levithan 2009). Por exemplo:

No site Regexpal.com podemos verificar a expressão regular, criada para formato de Data mm/dd/yyyy (Figura 5).

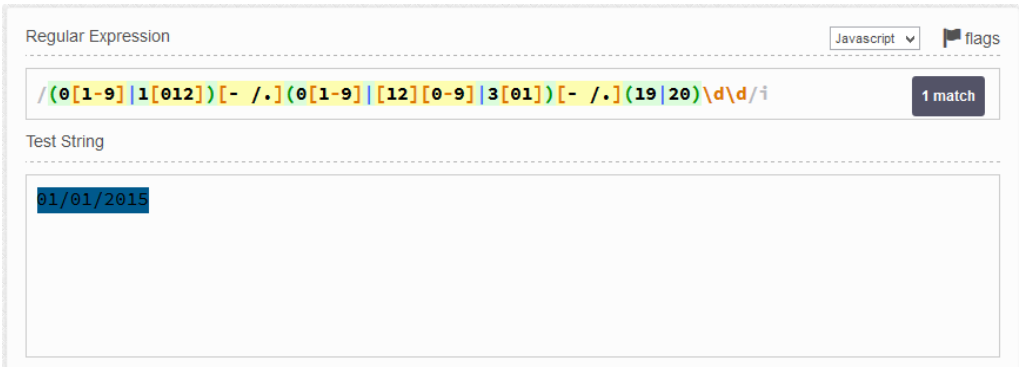


Figura 5 - Comparação de formato de Data mm/dd/yyyy no Regexpal.com

Como já foi dito, nem todos os elementos de metadados precisam de ter um vocabulário controlado. Alguns campos podem sugerir o formato da entrada (Hedden 2010). Após efetuada a análise de funcionamento e vantagens das expressões regulares, ficou evidente que estas podem ser adaptadas para os descritores existentes no Dendro, com o objetivo de dinamizar a introdução manual da informação e desta forma contribuir para a melhoria da qualidade dos metadados (Friedl 2006).

1.5 Dendro: repositório de gestão de dados de investigação

Atualmente verifica-se um incremento na quantidade de repositórios digitais contendo objetos partilhados. No meio disto, a necessidade de avaliar a qualidade de metadados tornou-se ainda mais crítica. De acordo com a literatura, há poucos repositórios que utilizam práticas de controlo de qualidade de metadados (Chassanoff 2009).

A plataforma Dendro tem vindo a ser desenvolvida na Universidade do Porto, de maneira a satisfazer as necessidades existentes nos processos de descrição de dados. Através do Dendro pretende-se apoiar os investigadores na preparação dos dados de investigação, nomeadamente na captura atempada de metadados para que os dados sejam transferidos para repositórios externos em condições de serem preservados. O Dendro utiliza os conjuntos de perfis de aplicação, baseados em ontologias, para descrever dados de investigação de domínios específicos, tais como Produção de Hidrogénio ou Oceanografia Biológica (Silva et al. 2014).

O Dendro consiste numa interface web, destinada aos investigadores no geral, tendo algum cuidado em criar interfaces adequadas para aqueles que não tenham experiência na gestão de dados. Tal como indicado na Figura 6 a interface inclui quatro áreas principais: a área de utilizador (1), o gestor de ficheiros (2), a zona de descrição de dados (3) e a zona de seleção de descritores (4).

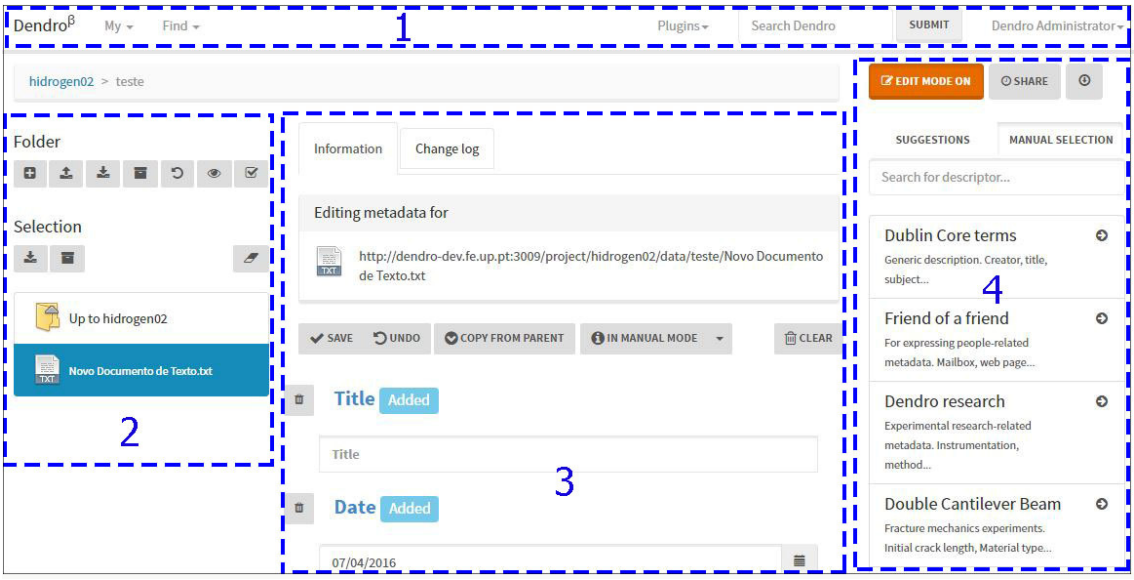


Figura 6 – A interface da plataforma Dendro

Na primeira área os utilizadores podem consultar os seus projetos e as respetivas descrições assim como criar novos. Além disso, podem também verificar os dados do seu perfil, efectuar correções se necessário e encontrar a informação sobre outros utilizadores registados na plataforma.

A segunda área é responsável pela gestão dos ficheiros que estejam depositados na plataforma. Neste módulo os investigadores podem criar e remover pastas e ficheiros, consoante as suas necessidades. Igualmente nesta área, os investigadores podem realizar cópias de segurança dos dados que poderão ser restauradas em qualquer altura.

O processo de descrição dos dados é efetuado conjugando ambas as áreas 3 e 4. Na quarta área existem vários domínios com os seus respetivos descritores que podem ser selecionados. A seleção de um descritor faz com que este seja adicionado à área de descrição para ser posteriormente preenchido pelo investigador. Cada descritor tem a informação sobre o seu propósito, apresentando uma descrição breve que facilite o seu objetivo. Na Figura 7 é apresentado o processo de escolha do descritor.

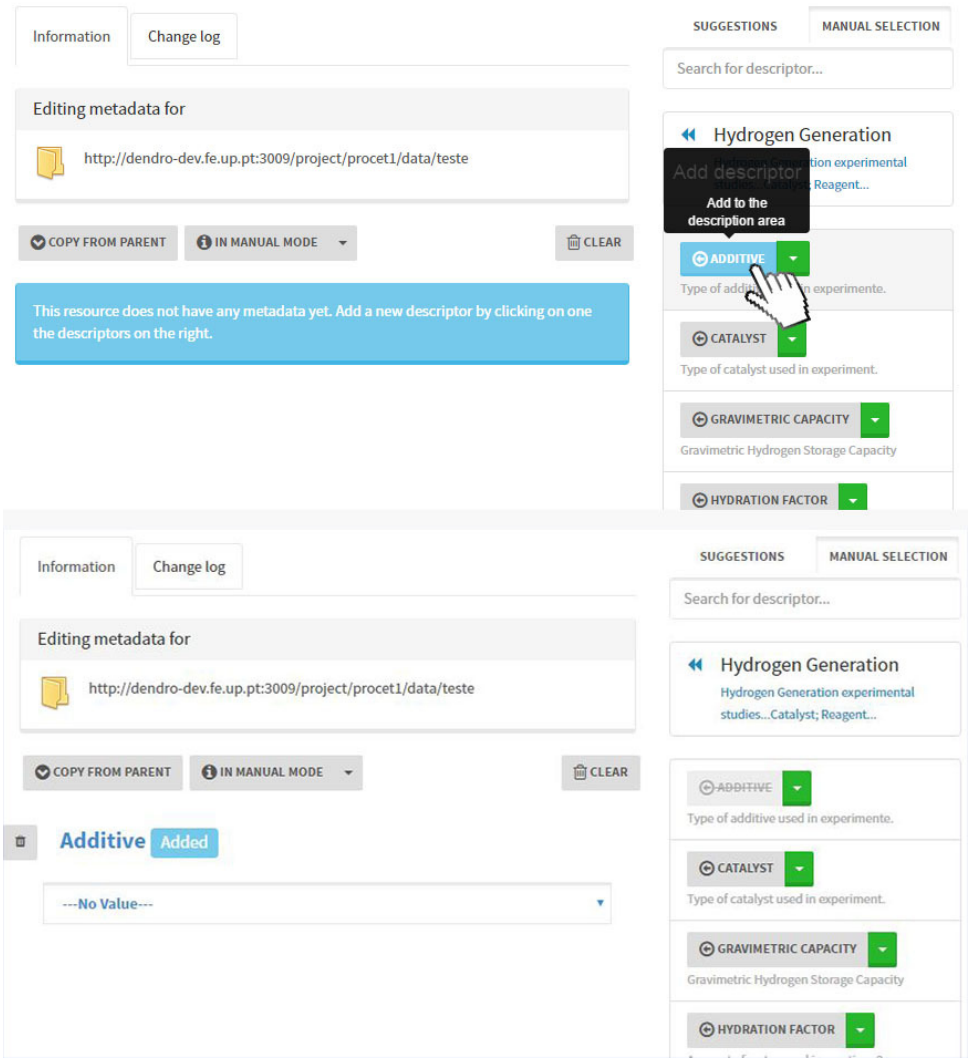


Figura 7 – Escolha do descritor

Além disto, os dados descritos podem ser transferidos para repositórios de dados, tais como Eudat, DSpace, entre outros. Na Figura 8 é demostrado o processo de transferência de dados descritos para repositório B2Share.

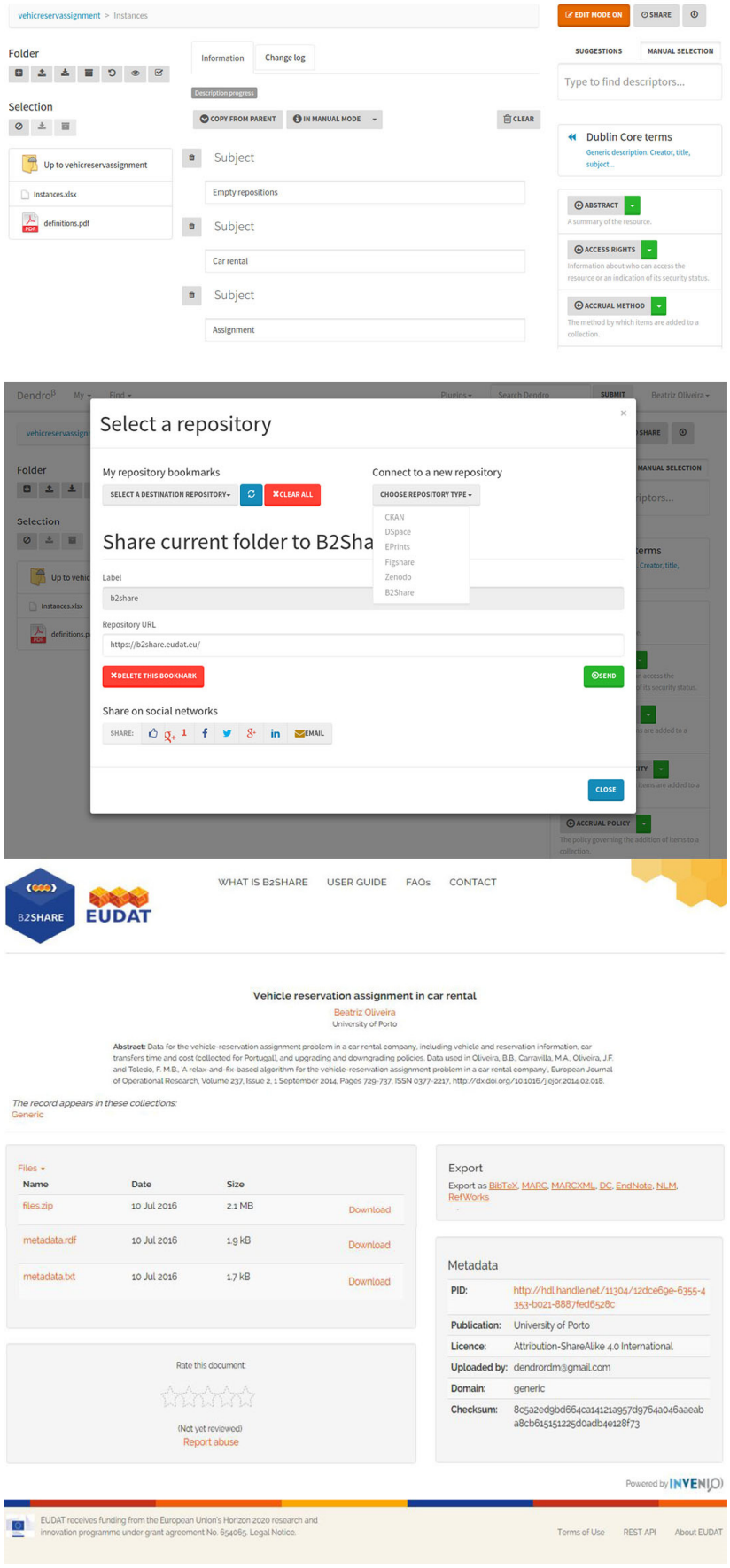


Figura 8 - Transferência de dados descritos para repositório B2Share

No âmbito do projeto Dendro, para apoiar os investigadores na descrição dos seus conjuntos de dados de domínios específicos, os investigadores do grupo InfoLab criaram várias ontologias para satisfazer a necessidade de descrição de dados de investigação em cada domínio. Os primeiros modelos de ontologias para Dendro foram: Fratura de Materiais, Química Analítica e Biodiversidade. Estas ontologias adicionaram novos descritores ao Dendro, desta forma permitindo ao investigador combinar os descritores específicos destas ontologias com descritores com origem em normas e esquemas de metadados genéricos carregados no Dendro: Dublin Core, Friend of a friend e CERIF (Castro, Silva, e Ribeiro 2014; Castro et al. 2015). Seguidamente, foram criadas ontologias para domínios como a Simulação de Veículos, Oceanografia Biológica, Produção de Hidrogénio e Ciências Socias.

Após das suas incorporações na plataforma, os investigadores participantes daqueles domínios em que foram criados as ontologias realizaram uma séria de experiências, onde testaram o depósito e descrição dos seus dados de investigação, avaliaram o funcionamento do Dendro, verificaram a compatibilidade dos descritores com o seu domínio de investigação, fizeram comentários e responderam um questionário sobre as funcionalidades de Dendro.

Durante as experiências realizadas com investigadores, verificou-se que no processo da criação de ontologias por vezes apareciam tentativas de criação de vocabulários controlados. Por isso, o foco desta dissertação está na proposta de vocabulários controlados para a descrição de dados em domínios científicos específicos. Estes vocabulários controlados irão disponibilizar uma lista de entradas permitidas para determinado descritor, facilitando o uso do Dendro e assim diminuindo os erros na descrição e contribuindo para a melhoria da qualidade de metadados.

Capítulo 2. Seleção de caso de estudo e definição de métricas de qualidade de metadados

Este capítulo inclui as seguintes secções:

- A primeira secção apresenta o processo da seleção do caso de estudo; mais especificamente, o trabalho realizado com investigadores e informação geral sobre o grupo de investigadores escolhido para nosso estudo;
- A segunda seção inclui a recolha de registos de metadados, existentes no Dendro e os comentários sobre a qualidade da descrição dos dados;
- Na terceira seção vão ser definidos os descritores mais utilizados por investigadores do domínio escolhido;
- Na quarta secção são definidas as métricas de qualidade de descrição de dados de investigação e descrito o processo da experiência realizada com o investigador do grupo escolhido.

No geral, este capítulo descreve o processo preparativo da análise de qualidade da descrição dos dados.

2.1 Seleção do caso de estudo

Envolvimento dos investigadores é importante para fácil compreensão de domínios científicos e diminuição dos erros na modelação de ontologias. A colaboração com grupo de investigadores de Produção de Hidrogénio e criação da ontologia para este domínio foi feita numa fase anterior a esta dissertação.

O grupo de investigadores da área de produção de hidrogénio é um grupo do CEFT (Centro de Estudos de Fenómenos de Transporte) – Grupo Energia. O principal objetivo da investigação deste grupo é a produção instantânea de hidrogénio através da hidrólise catalítica do borohidreto de sódio (NaBH4) de forma a alimentar células de combustível do tipo PEM com aplicação em 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. Na partilha dos dados, a equipa de investigação utiliza ferramentas de comunicação tradicionais como o email. Em outros casos, os dados são copiados do computador para discos externos.

O trabalho neste domínio é dividido em três fases principais: Na fase inicial é efetuada a descrição das medidas de dados que antecedem a experiência, nomeadamente de matérias-primas, e a recolha de dados brutos, depois da experiência no reator; A segunda fase contém os cálculos de verificação e construções gráficas onde se afere se existem erros na recolha de dados, de acordo com os dados teóricos previstos; A última fase é baseada em processamento dos dados brutos. Os dados serão refinados de maneira a poder-se efetuar uma análise de resultados e assim obter conclusões sobre o desempenho do processo e a qualidade das saídas. Do ponto de vista da preservação, os investigadores identificaram as saídas da primeira fase como os dados mais importantes a serem depositados e preservados.

O trabalho realizado com este grupo de investigadores inclui várias entrevistas e experiências de descrição de dados efetuados por 2 investigadores – Utilizador 1 e Utilizador 2. Durante as experiências eles depositaram e descreveram seus conjuntos de dados com o intuito de avaliar o desempenho da plataforma Dendro. Utilizador 1 depositou e descreveu três conjuntos de dados durante a experiência, Utilizador 2 – um conjunto de dados.

A Tabela 4 mostra todos os descritores específicos que foram criados para descrição de dados de investigação de Produção de Hidrogénio na plataforma Dendro.

Tabela 4 – Lista de descritores específicos, criados para domínio Produção de Hidrogénio

Descritor	Descrição
Catalyst	Type of catalyst used in experiment
Gravimetric Capacity	Gravimetric Hydrogen Storage Capacity
Hydration Factor	Amount of water used in reaction=2+x
Hydrogen Generation Rate	Amount of Hydrogen per minute
Hydrolysis	Type of hydrolysis reaction
Number of reutilization	Number of catalyst reutilizations
Reactor Type	Type of reactor used in experiment
Reagent	Type of reagent used in experiment

Additive[13]	Type of additive used in experiment

Durante as entrevistas os investigadores afirmaram que seria uma grande vantagem ter-se os dados descritos com qualidade e serem disponíveis para partilha, reutilização e publicação. A criação de metadados com qualidade para descritores acima indicados facilita a localização de conjuntos de dados, a compreensão e reutilização dos mesmos.

Depois de criação e implementação de vocabulários controlados para este domínio e obtendo-se os resultados positivos, ou seja provando a hipótese – vocabulários controlados facilitam o processo da descrição e melhoram a qualidade da descrição de dados de investigação, espera-se que a mesma abordagem possa ser alargada aos outros domínios.

2.2 Recolha de dados existentes no Dendro no domínio escolhido

Para recolher os dados existentes foram analisados as descrições efetuadas pelos Utilizador 1 e Utilizador 2. Para este fim foi criada uma tabela no ficheiro Excel, com a lista de descritores utilizados por ambos, que posteriormente foi preenchida manualmente de acordo com as descrições efetuadas na plataforma Dendro (Anexo 1). A Tabela 5 contém uma breve análise de descrição de dados efetuado por ambos utilizadores, que mostra a existência de diferenças na descrição de dados:

Tabela 5 – Descrição de dados efetuada por ambos utilizadores e sua breve análise

Descritor utilizado por ambos utilizadores	Valor de descritor do utilizador 1	Valor de descritor do utilizador 2	Breve análise
Temperature	25ºC	24	Uma vez indicado a unidade ºC, outra não
Hydration Factor	16	16	Não existe diferença
Number of reutilization	45	49	Não existe diferença

[13] Additive é descritor incluido na ontologia, mas não incorporado no Dendro

Reactor Type	RG/RM	ovoid	Maneiras diferentes de escrita de tipos de reactor (RG, RM, ovoid)
Reagent	NaBH4	Sodium Borohydride	Maneiras diferentes de escrita dos mesmos reagentes (NaBH4, Sodium Borohydride)
Date Issued	06/08/2015	06/23/2015	Não existe diferença
Contributor	R*** C	He*** Nu***	Maneiras diferentes da escrita (Fe*** MJF, Ra*** C, He*** Nu***)
Coverage	Faculdade de engenharia universidade porto	CEFT-Lab206	Maneiras diferentes da escrita (CEFT-Lab206, faculdade de engenharia universidade porto)
Creator	Fe*** MJF	Al*** Pi***	Maneiras diferentes da escrita (Fe*** MJF, Al*** Pi***)
Catalyst	Ni-Ru	Nickel-ruthenium	Maneiras diferentes da escrita (Ni-Ru, Nickel-ruthenium)
Gravimetric Capacity	<5wt%	2,3	Uma vez indicado a unidade wt%, outra vez não
Hydrolysis	Hidrolise clássica	alkali	Maneiras diferentes da escrita (classic, Hidrolise clássica, procedimento habitual, alkali)

*- omitido por revelar a identidade do utilizador.

Após a recolha de dados e análise dos mesmos, é possível afirmar que existem diferenças sintáticas e gramaticais na linguagem usada na descrição de dados de investigação. E por isso necessário identificar os descritores mais utilizados pelos investigadores deste domínio, que

podem estar associados com vocabulários controlados e expressões regulares, com o objetivo de facilitar a descrição e melhorar a qualidade de descrição de dados de investigação.

2.3 Definição dos descritores mais utilizados na plataforma Dendro no domínio escolhido

Analisando os dados recolhidos (Anexo 1), antes de se proceder à identificação dos descritores mais utilizados no domínio Produção de Hidrogénio, foram criadas as tabelas que visavam os descritores utilizados de domínios diferentes de Produção de Hidrogénio (Tabela 4). A Tabela 6 inclui os descritores genéricos, com a identificação da norma. A Tabela 7 inclui os descritores que foram criados para outros domínios.

Tabela 6 – Descritores genéricos, utilizados na descrição

Descritor	Norma
Access Rights	Dublin Core
Contributor	Dublin Core
Coverage	Dublin Core
Alternative Title	Dublin Core
Title	Dublin Core
Creator	Dublin Core
Date Issued	Dublin Core
Subject	Dublin Core
Instrumentation	Dendro research – ecological metadata language
Software	Dendro research – ecological metadata language

Tabela 7 – Descritores específicos de outros domínios, utilizados na descrição

Descritor	Domínio
Temperature	Double Cantilever Beam (Fracture mechanics)
Compound	Pollutant analysis

Além disto, durante a análise de descritores do domínio Produção de Hidrogénio, verificou-se que o descritor *Additive* foi incluído na ontologia de domínio Produção de Hidrogénio, mas não foi incorporado na plataforma Dendro. Contudo, com base nas experiências realizadas com investigadores este descritor é importante neste domínio e vai ser incluído na ontologia do domínio na plataforma Dendro.

A análise de dados mostrou que um dos investigadores selecionou o descritor *Method* para descrições do tipo de hidrólise, pois não estava familiarizado com o funcionamento do sistema Dendro e não sabia que descritores existiam no sistema. O descritor *Hydrolysis* é tido como adequado à descrição, pois o mesmo foi criado para as descrições do tipo de hidrólise no domínio Produção de Hidrogénio.

De acordo com a análise quantitativa, identificam-se os descritores mais utilizados no domínio Produção de Hidrogénio (Tabela 8):

Tabela 8 – Os descritores mais utilizados na plataforma Dendro para o domínio Produção de Hidrogénio

Descritor	Numero de utilização de cada descritor
Alternative Title	4
Contributor	4
Creator	4
Date Issued	4
Hydration Factor	4
Reactor Type	4
Temperature	4
Hydrolysis	3
Number of Reutilization	3
Reagent	3
Catalyst	2
Coverage	2
Gravimetric Capacity	2
Access Rights	1
Compound	1
Hydrogen Generation Rate	1
Instrumentation	1
Software	1
Subject	1
Title	1

A análise geral de descritores mostra que os descritores com valor 4 e 3 são mais utilizados, contudo os descritores com valor 2 e 1 também vão ser analisados, porque a qualidade de descrição de dados depende de metadados em todos os descritores preenchidos.

2.4 Definição de métricas de qualidade de descrição de dados de investigação

Um dos objetivos específicos deste trabalho é definir as métricas para avaliação de qualidade de descrição de dados de investigação no Dendro. Na Tabela 1 foram apresentadas as métricas de qualidade mais comuns na visão de autores diferentes.

Tendo em conta o domínio, foram escolhidas as seguintes métricas para avaliar a qualidade dos metadados:

1. ***Correctness*** – grau em que a linguagem utilizada nos metadados é correta sintaticamente e gramaticalmente;
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 requisites de uma determinada comunidade de utilizadores;
4. ***Overall Rating*** – pontuação geral do registo de metadados, tendo em conta todas as métricas acima.

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.

De forma a poder de efetuar uma análise à qualidade da descrição de dados de investigação no domínio Produção de Hidrogénio, aplicando os critérios e medidas elaboradas, no dia 18.02.2016 realizou-se mais uma experiência da descrição de dados na plataforma Dendro com a duração de 30 minutos, com o Utilizador 3, que não fez parte dos testes anteriores.

Antes da realização da experiência foi elaborado um guião (Anexo 2) e criado o login para utilizador. Este utilizador foi adicionado ao projeto "Experiência da descrição de dados no Produção de Hidrogénio" na plataforma Dendro e descreveu três conjuntos de dados durante esta experiência.

Após a realização da descrição de dados de investigação, foi enviado o inquérito de usabilidade da plataforma Dendro (Anexo 2). Além disto, as perguntas foram elaboradas de modo a que as respostas ajudem na análise dos resultados aplicando as métricas definidas.

Os dados desta experiência foram recolhidos manualmente (Anexo 4) e comparados com os dados existentes no Anexo 1 e Tabela 5. A Tabela 9 descreve as diferenças existentes na descrição de dados de investigação levada a cabo pelos três utilizadores do mesmo grupo de investigação e a sua breve análise:

Tabela 9 – Descrição de dados efetuada por 3 utilizadores e comentários

Descritor utilizado por ambos utilizadores	**Valor de descritor do utilizador 1**	**Valor de descritor do utilizador 2**	**Valor de descritor do utilizador 3**	**Comentário**
Temperature	25ºC	24	28ºC	Uma vez indicado a unidade ºC, outra não
Hydration Factor	16	16	15	Não existe diferença
Number of reutilization	45	49	50	Não existe diferença
Reactor Type	RG/RM	ovoid	Egg Reactor / Conical Small Reactor	Maneiras diferentes de escrita de tipos de reactor (RG, RM, ovoid, Egg Reactor, Conical Small Reactor)
Reagent	NaBH4	Sodium Borohydride	NaBH4	Maneiras diferentes de escrita (NaBH4, Sodium Borohydride)
Date Issued	06/08/2015	06/23/2015	07/13/2015	Não existe diferença
Creator	Fe*** MJF	Al*** Pi***	HXN-CEFT	Maneiras diferentes da escrita (Fe*** MJF, Al*** Pi***)
Catalyst	Ni-Ru	Nickel-ruthenium	NiRu	Maneiras diferentes da escrita (Ni-Ru, Nickel-ruthenium, NiRu)
Gravimetric Capacity	<5wt%	2,3	1.9wt%	Uma vez indicado a unidade wt%, outra vez não
Hydrolysis	Hidrolise clássica	alkali	Classic hydrolysis	Maneiras diferentes da escrita (Classic hydrolysis, Hidrolise clássica, procedimento habitual, alkali)

* - omitido por revelar a identidade do utilizador

Após a experiência com o Utilizador 3 identificaram-se as seguintes limitações:

1) A diferença na descrição de dados de investigação;

2) Os erros na descrição de dados de investigação;

3) A descrição incompleta em vários descritores específicos.

Por isso acredita-se, que a criação e implementação de vocabulários controlados e expressões regulares pode facilitar a descrição de dados de investigação no Dendro e assim normalizar o formato da descrição, sintaxe e semântica, melhorando consequentemente a qualidade de descrição dos mesmos.

A Tabela 10 mostra quais os descritores que podem estar associados com vocabulários controlados e respetiva adequação a expressões regulares:

Tabela 10 – Análise de descritores para criação de vocabulários controlados ou expressões regulares

Descritor	Numero de utilização de cada descritor	Comentários
Alternative Title	4	Texto-livre
Contributor	4	Expressão regular, para sugestão de formato
Creator	4	Expressão regular, para sugestão de formato
Date Issued	4	Expressão regular, para sugestão de formato
Hydration Factor	4	Texto-livre
Reactor Type	4	Vocabulário controlado
Temperature	4	Texto-livre
Hydrolysis	3	Vocabulário controlado
Number of Reutilization	3	Texto-livre
Reagent	3	Vocabulário controlado
Catalyst	2	Vocabulário controlado
Coverage	2	Expressão regular, para sugestão de formato
Gravimetric Capacity	2	Dividir este descritor por dois. Para valor - texto livre, para unidade - vocabulário controlado
Access Rights	1	Texto-livre
Compound	1	Texto-livre
Hydrogen Generation Rate	1	Dividir este descritor por dois. Para valor - texto livre, para unidade - vocabulario controlado
Instrumentation	1	Texto-livre
Software	1	Texto-livre
Subject	1	Texto-livre
Title	1	Texto-livre
*Additivo		Vocabulário controlado

Obtém-se assim dois grupos de descritores onde é possível melhorar a qualidade de descrição utilizando vocabulários controlados ou expressões regulares:

1) Descritores com vocabulários controlados – *Reactor Type, Hydrolysis, Reagent, Catalyst, Additivo*

2) Descritores com expressões regulares[14] – *Contributor, Creator, Date Issued, Coverage*.

Antes de se proceder a realização da análise de qualidade da descrição de dados de investigação, serão criados os vocabulários controlados, ou seja, serão definidos e aprovados os conceitos para vocabulários controlado em conjunto com os investigadores. A utilização destes conceitos presentes nos vocabulários controlados corresponde a uma descrição com 100% de qualidade. Esta percentagem é tida em consideração para a análise da qualidade das descrições existentes.

[14] As expressões regulares não são o foco deste trabalho. Por isso não vai ser realizada análise de qualidade da descrição dos descritores indicados.

Capítulo 3. Elaboração de vocabulários controlados para domínio escolhido

No contexto de elaboração de vocabulários controlados, este capítulo inclui as seguintes secções:

- A primeira secção apresenta o processo de criação de vocabulários controlados juntamente com os investigadores de grupo CEFT (Centro de Estudos de Fenómenos de Transporte); mais especificamente, a concretização de conceitos pré-definidos em vocabulários controlados no domínio Produção de Hidrogénio, e vantagens que trazem a implementação dos mesmos;
- A segunda seção mostra como foram implementados os vocabulários controlados na plataforma Dendro; explica-se a razão de utilização de *Annotation Property* na modelação de ontologia com vocabulários controlados; é apresentada a interface do Dendro após a implementação dos mesmos;
- A terceira seção contém a informação sobre as experiências realizadas por investigadores; recolha dos registos de metadados para análise de qualidade de descrição de dados de investigação após a implementação de vocabulários controlados.

No geral, este capítulo descreve todas as ações ligadas a criação e implementação dos vocabulários controlados na plataforma Dendro no domínio Produção de Hidrogénio.

3.1 Criação de vocabulários controlados

Um dos objetivos específicos definidos neste trabalho para melhoria da qualidade da descrição de dados de investigação no Dendro é elaboração de vocabulários controlados. Para atingir este objetivo, após a escolha de descritores que podem estar associados com vocabulários controlados, juntamente com os investigadores do grupo CEFT (Centro de Estudos de Fenómenos de Transporte) foi realizada a reunião sobre concretização de conceitos pré-definidos em vocabulários controlados.

No início da reunião foi explicado que um dos nossos objetivos principais na criação de vocabulários controlados é a tentativa de melhoria da qualidade de descrição de dados, diminuição dos erros e facilitação do processo da descrição no geral. Além disto, foi explicado

o que é um vocabulário controlados e o que são expressões regulares, através de exemplos dos mesmos. Em seguida, foram discutidos alguns aspetos importantes no processo de descrição de dados de investigação, assim como a importância de qualidade de descrição, a existência de diferenças na utilização de expressões lexicais, a ocorrência dos erros, a exatidão da informação na descrição, etc. Reconhecendo esta importância o grupo CEFT achou muito útil o processo de criação de vocabulários controlados e expressões regulares.

Para todos os investigadores do grupo CEFT foi apresentada uma lista com conceitos pré-definidos para os descritores *Reactor Type*, *Catalyst*, *Hydrolysis*, *Reagent* e *Additive* com a finalidade de concretizar e aprovar as propostas sugeridas ou, sendo necessário durante a discussão, proceder ao ajuste dos mesmos.

Com a validação dos termos sugeridos para os vocabulários controlados, concordou-se com a implementação dos mesmos na plataforma Dendro. Na Tabela 11 são apresentados os conceitos pré-definidos para vocabulários controlados.

Tabela 11 – Descritores com conceitos pré-definidos para vocabulários controlados

<table>
<tr><th colspan="2">Reactor Type</th><th colspan="2">Hydrolysis</th></tr>
<tr><td colspan="2">○ EggR – ovoid mini reactor
○ LR – large reactor
○ MR_c – conical medium reactor
○ MR_f – flat medium reactor
○ SR_c – conical small reactor
○ SR_f – flat small reactor</td><td colspan="2">○ Acid hydrolysis
○ Alkali-free hydrolysis
○ Classic hydrolysis</td></tr>
<tr><th>Catalyst</th><th>Reagent</th><th colspan="2">Additive</th></tr>
<tr><td>○ Ni-Ru
○ Pt/C
○ Co-B
○ Co-Mn-B
○ Co-B/Ni</td><td>○ $NaBH_4$
○ NH_3BH_3
○ $LiAlH_4$
○ $LiBH_4$
○ KBH_4</td><td colspan="2">○ SDS
○ CMC</td></tr>
</table>

O Anexo 5 apresenta o relatório completo da reunião acima indicado. Durante a reflexão, além da definição e aprovação dos conceitos para vocabulários controlados, foram discutidos vários aspetos importantes no processo de descrição de dados de investigação no geral. Foram vistos, entre outros, a definição de vocabulários controlados abertos ou fechados. O grupo CEFT sugeriu deixar os vocabulários controlados abertos para todos os descritores, exceto para o descritor *Hydrolysis*. De acordo com esta sugestão foi proposto o seguinte:

Um dos investigadores fica responsável para administrar os projetos deste grupo no Dendro, de forma a decidir se adiciona os conceitos sugeridos por utilizadores do grupo à lista de vocabulário controlado. Ou seja, se um utilizador não encontra um conceito no vocabulário

controlado, escreve-o e adiciona-o ao vocabulário. Contudo este conceito só aparecerá disponível na lista após aprovação e validação por parte do responsável.

3.2 Implementação de vocabulários controlados no Dendro

A formalização do vocabulário controlado tirou partido da ontologia desenvolvida para o domínio Produção de Hidrogénio[15]. A escolha da maneira da adaptação da ontologia existente baseia-se na análise de estudos relacionados. De acordo com literatura, existem várias maneiras de modelar vocabulários controlados numa ontologia.

Segundo a W3C Recommendation - OWL Web Ontology Language Reference[16] os vocabulários controlados podem ser criados na ontologia como *Annotation Property*. 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 disto, no mesmo documento afirma-se que é possível especificar o tipo de valor de um literal numa indicação da *Annotation Property*. Existem 5 *Annotation Property* pré-definidas: *VersionInfo*, *label*, *comment*, *seeAlso*, *isDefineBy*, que podem ser utilizados para anotação de *DataProperties* como exemplificado na Figura 9[17].

[15] A modelação da ontologia é efetuada utilizando o *software Protégé*

[16] W3C: OWL Web Ontology Language Reference - https://www.w3.org/TR/owl-ref/#Header

[17] Dublin Core in OWL 2 - http://bloody-byte.net/rdf/dc_owl2dl/

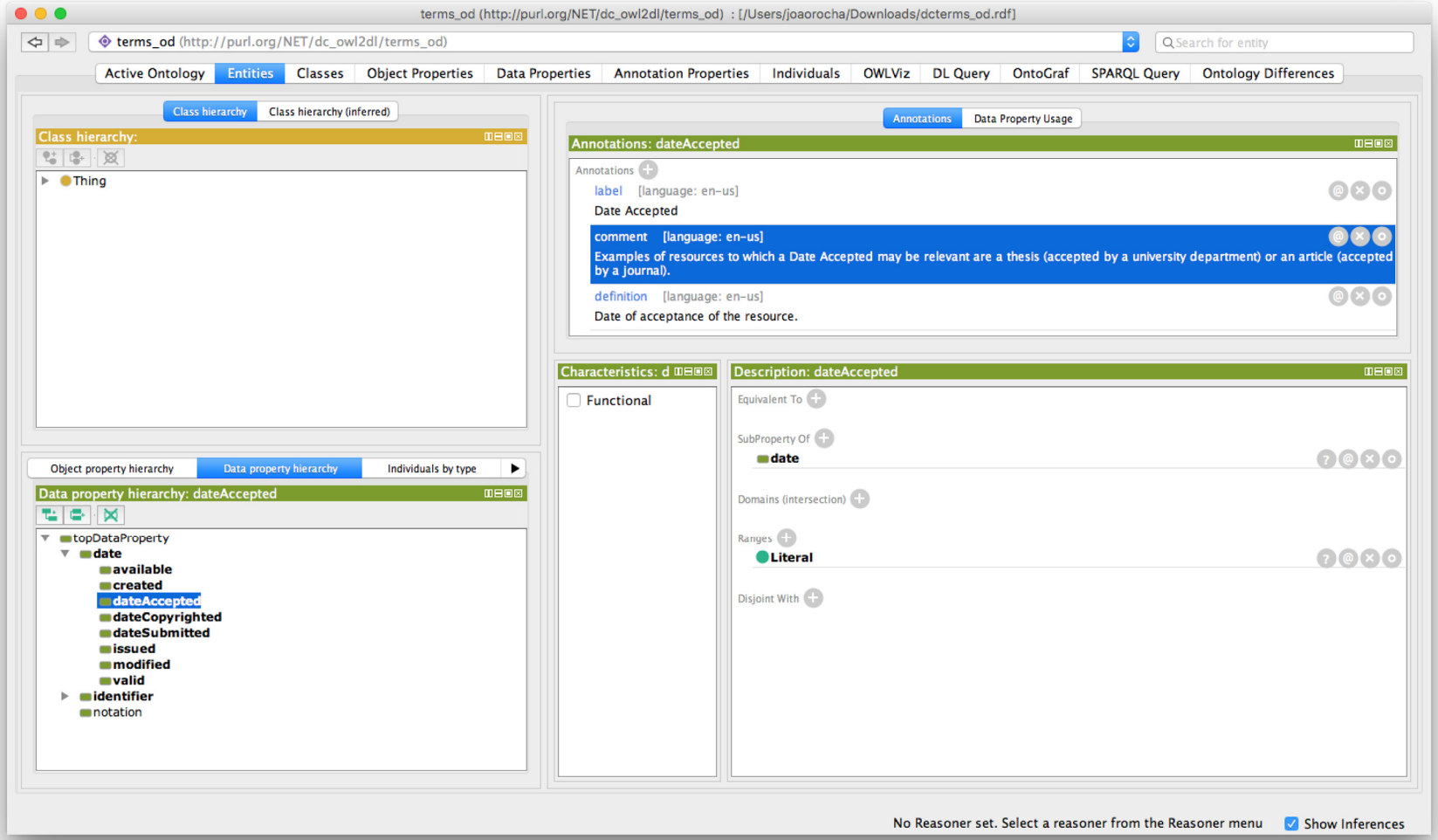


Figura 9 – Exemplo de anotação de *Data Property*, usando a *Annotation Property*

Além da recomendação na linguagem OWL, Liyang Yu (Yu 2011) afirma que as anotações fornecem maneiras de associar as informações adicionais a classes e propriedades e devem ser usadas para a informação de entidades que não faz parte do domínio e não deve contribuir para as consequências lógicas da ontologia principal. Tal como no W3C Recommendation, o autor confirma que o OWL permite que classes, propriedades, indivíduos e cabeçalhos de ontologia sejam anotados com a informação útil, como sejam *labels, comments, authors, creation date*, etc. Esta informação adicional pode ser muito importante para a reutilização de ontologias. A seleção de nomes significativos e o uso de anotações é especialmente importante para documentação, manutenção e rastreabilidade. As anotações são frequentemente utilizadas em ferramentas para fornecer a expressão em linguagem natural para ser exibida em *help windows* (Yu 2011). Além disto, podem ser utilizados em vocabulários controlados, para especificar a semântica e relações dos conceitos (Qin e Paling 2001).

Mais um exemplo de utilização de *Annotation Properties* é encontrado no *Foundational Model of Anatomy* (Golbreich, Zhang, e Bodenreider 2006), que tem como base o caso do *NCI-Thesaurus de National Cancer Institute* (de Coronado e Fragoso 2004). O FMA é uma fonte de conhecimento (ontologia) em ciências biomédicas sobre anatomia humana, desenvolvido para representação simbólica da estrutura fenotípica do corpo humano de uma forma compreensível, navegável, analisável e interpretável. Este modelo contém aproximadamente 75.000 classes e mais de 120.000 termos, foi implementado no *Protégé* com utilização de *Annotation Property* para descrição de classes, restrições ou valores. Na Figura 10 é demonstrado a definição de *Annotation Property* assim como: *Preferred_name, Synonyms*, entre ouros. O exemplo da utilização do *Synonyms* é apresentado na Figura 11.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:owl="http://www.w3.org/2002/07/owl#" xmlns:rdfs="http://www.w3.org/2000/01/
  <owl:Ontology rdf:about=""/>
  <owl:AnnotationProperty rdf:ID="Preferred_name"/>
  <owl:AnnotationProperty rdf:ID="Synonyms"/>
  <owl:AnnotationProperty rdf:ID="UWDAID"/>
  <owl:AnnotationProperty rdf:ID="author"/>
  <owl:AnnotationProperty rdf:ID="authority"/>
  <owl:AnnotationProperty rdf:ID="modification"/>
  <owl:AnnotationProperty rdf:ID="name"/>
  <owl:AnnotationProperty rdf:ID="Date_entered_modified"/>
  <owl:ObjectProperty rdf:ID="dimension">
    <rdfs:domain rdf:resource="#Physical_anatomical_entity"/>
    <rdfs:range>
      <owl:Class>
      <owl:oneOf rdf:parseType="Collection">
        <owl:Thing rdf:about="#individual_d1-dimension"/>
        <owl:Thing rdf:about="#individual_d0-dimension"/>
        <owl:Thing rdf:about="#individual_d2-dimension"/>
        <owl:Thing rdf:about="#individual_d3-dimension"/>
      </owl:oneOf>
      </owl:Class>
    </rdfs:range>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#FunctionalProperty"/>
  </owl:ObjectProperty>
  <owl:AnnotationProperty rdf:ID="TA_ID"/>
  <owl:AnnotationProperty rdf:ID="definition"/>
```

Figura 10 - Definição de *Annotation Property* em FMA-constitutionalPartofNS.owl[18]

[18] https://mor.nlm.nih.gov//pubs/supp/2005-owled-cg/FMA-constitutionalPartForNS.owl

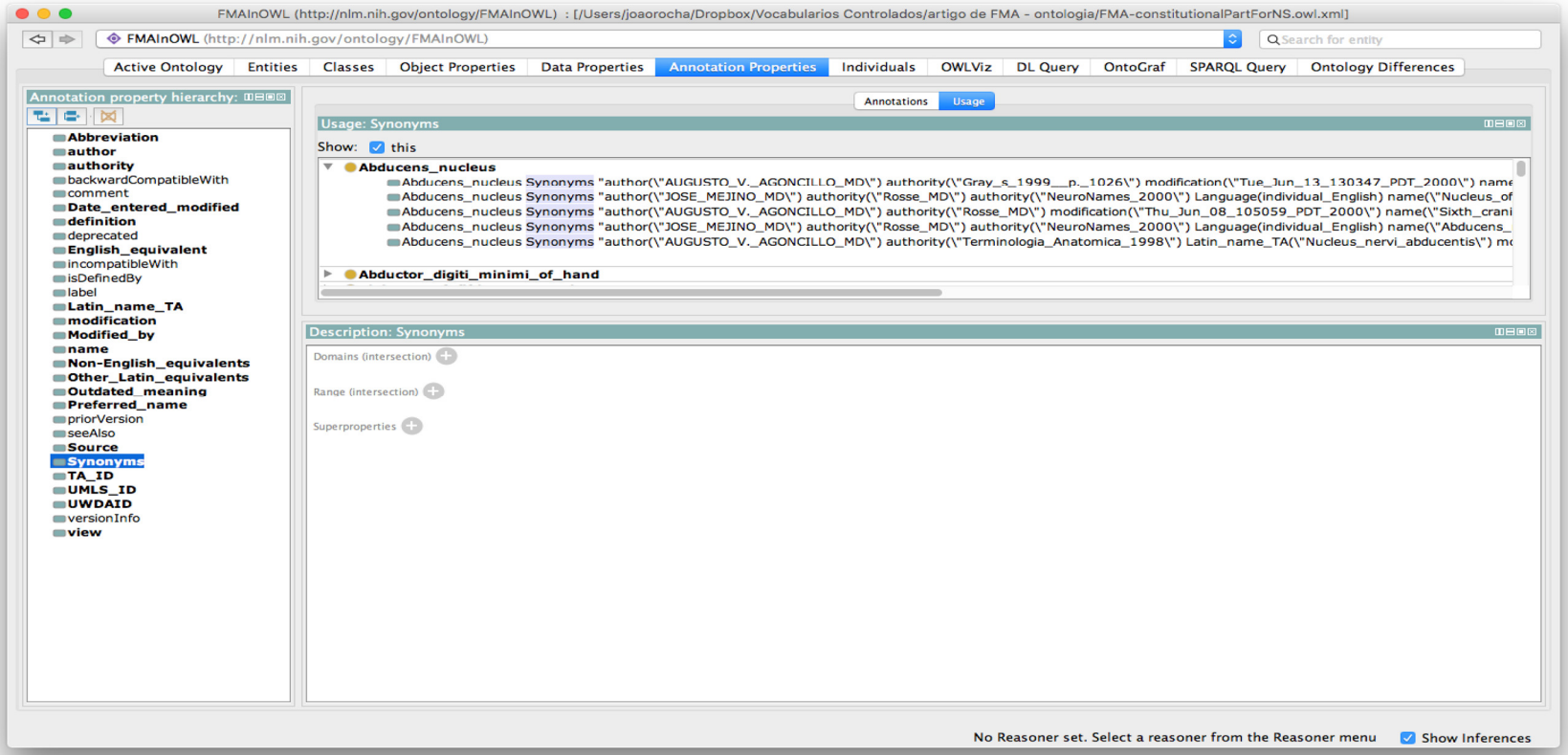


Figura 11 - O exemplo de utilização de *Annotation Property Synonyms* na ontologia FMA no *Protégé*[19]

[19] https://mor.nlm.nih.gov//pubs/supp/2005-owled-cg/FMA-constitutionalPartForNS.owl

Pode-se então afirmar que, em geral, as anotações podem servir para adicionar metadados às classes, aos indivíduos, às *Object Properties* e às *Data Properties*. Do ponto de vista de modelação pode-se então criar uma *Annotation Poroperty* própria que irá descrever as *DataProperty* existentes no Dendro.

Adaptando esta metodologia foram criadas as seguintes *Annotation Property* que definem as relações entre os *Data Properties* existentes na nossa ontologia (Figura 12). Os seus significados são:

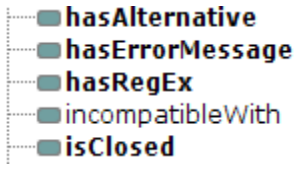


Figura 12 - *Annotation Property* no *Protégé*

hasAlternative - é uma das alternativas possíveis para o valor de um descritor, ou seja é valor de vocabulário controlado;

isClosed – tem valor *True*, quando o vocabulário controlado é fechado, significando que quando não existe possibilidade de adicionar novo termo para a lista dos conceitos definidos, e valor *False* quando é aberto, ou seja novos conceitos podem ser adicionados ao longo do tempo. É utilizado em conjunto com *hasAlternative;*

hasRegEx – é uma expressão regular usada para verificar se o valor de descritor é válido; é usado para validação de uma expressão regular;

hasErrorMessage - é mensagem de erro quando o valor do descritor especifico não está de acordo com o formato sugerido. É utilizado em conjunto com *hasRegEx*.

Após a criação dos *Annotation Property* foram definidos os valores para os vocabulários controlados, aprovados pelo Grupo CEFT. A Figura 13 mostra um exemplo de vocabulário controlado de descritor ***Reactor Type*** em OWL e a Figura 14 o mesmo no *Protégé*.

```
<!-- http://dendro.fe.up.pt/ontology/hydrogen#reactorType -->

<owl:DatatypeProperty rdf:about="&ontology;hydrogen#reactorType">
    <rdfs:label>Reactor Type</rdfs:label>
    <ddr:isClosed rdf:datatype="&xsd;boolean">true</ddr:isClosed>
    <ddr:hasAlternative>SRc - conical small reactor</ddr:hasAlternative>
    <rdfs:comment>Type of reactor used in experiment.</rdfs:comment>
    <ddr:hasAlternative>LR - large reactor</ddr:hasAlternative>
    <ddr:hasAlternative>EggR - ovoid  mini reactor</ddr:hasAlternative>
    <ddr:hasAlternative>SRf - flat small reactor</ddr:hasAlternative>
    <ddr:hasAlternative>MRc - conical medium reactor</ddr:hasAlternative>
    <ddr:hasAlternative>MRf - flat medium reactor</ddr:hasAlternative>
    <rdfs:domain rdf:resource="&ontology;hydrogen#HydrogenGeneration"/>
</owl:DatatypeProperty>
```

Figura 13 - Vocabulário Controlado de descritor *Reactor Type*

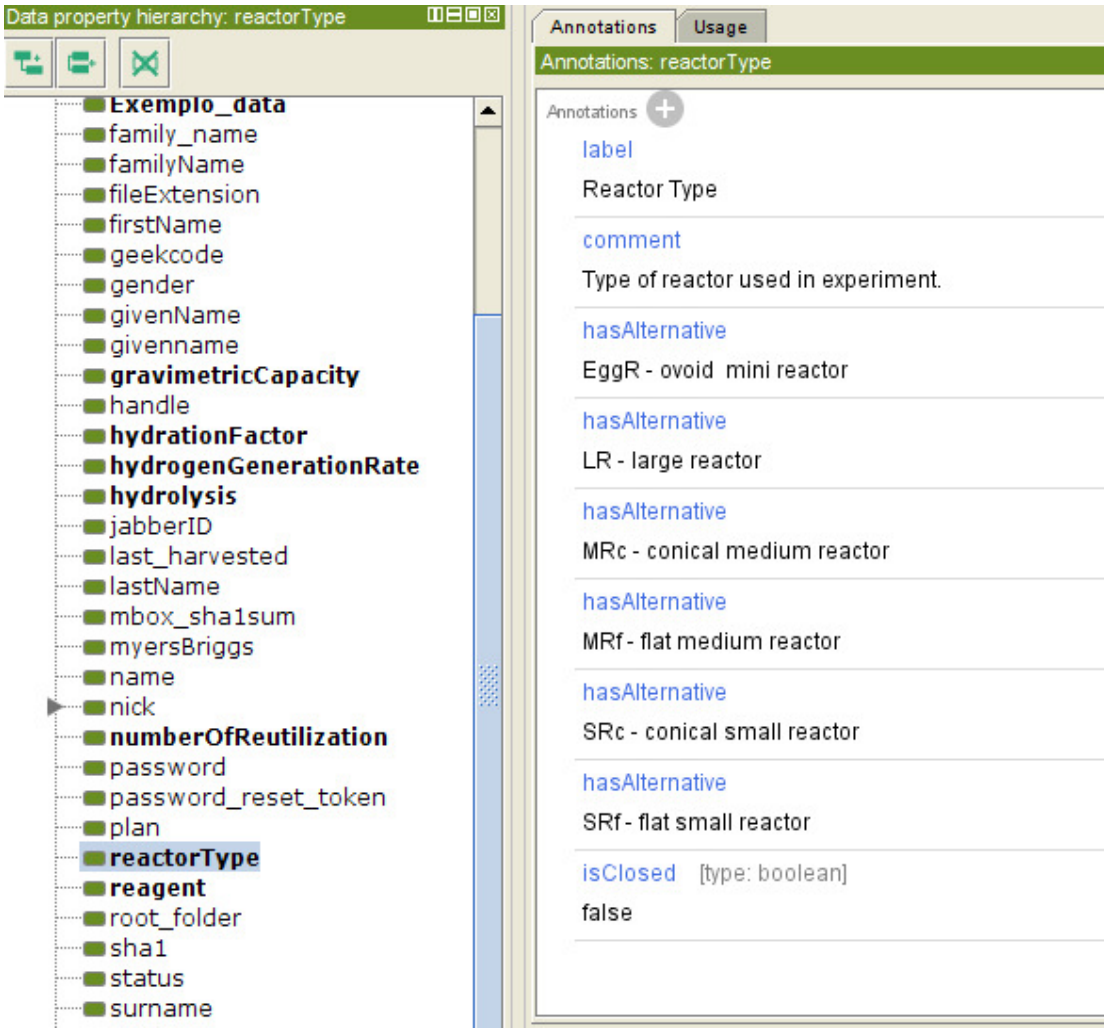


Figura 14 – *Reactor Type* no *Protégé*

A implementação dos vocabulários abertos e fechados depende de desenvolvimentos adicionais na plataforma Dendro, pelo que não está ainda realizada.

Após a modelação da ontologia e a implementação da mesma, os descritores com vocabulários controlados apareçam no interface Dendro tal como ilustrado nas Figuras 15 a 19.

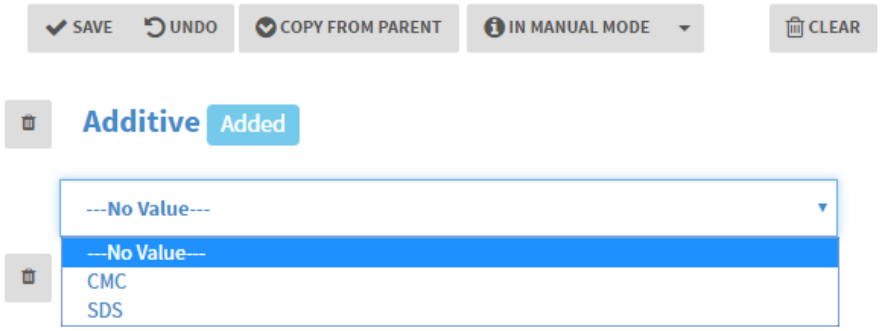


Figura 15 – Descritor *Additive*

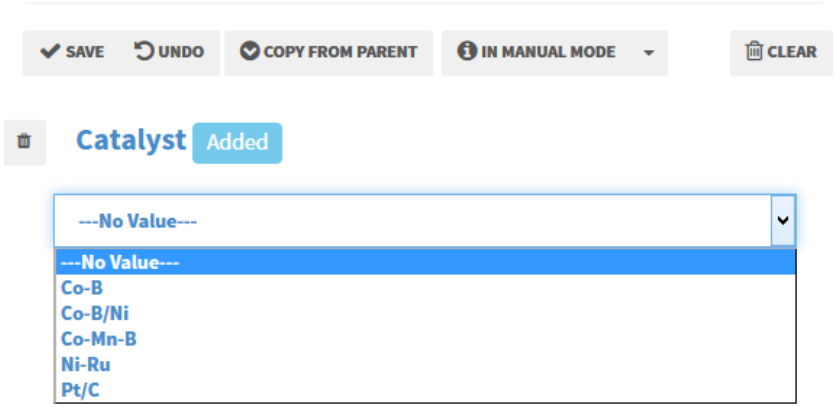


Figura 16 – Descritor *Catalyst*

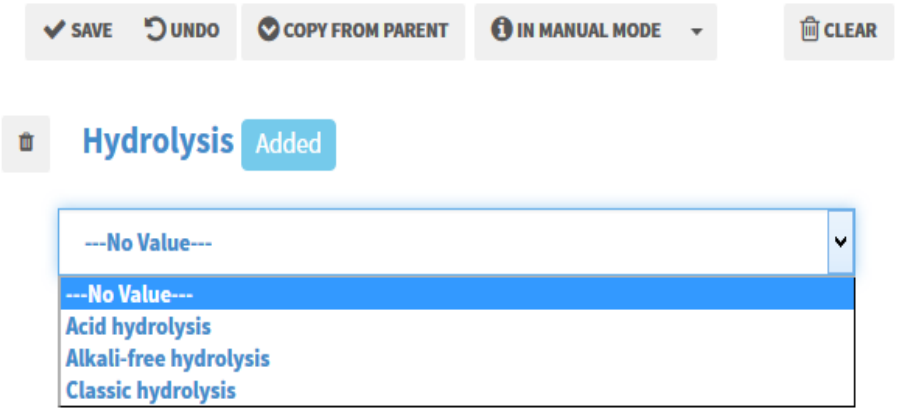


Figura 17 – Descritor *Hydrolysis*

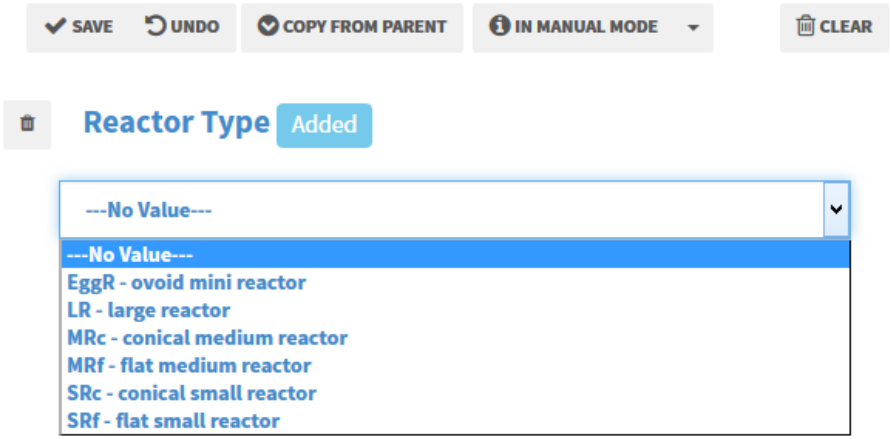


Figura 18 – Descritor *Reactor Type*

Figura 19 – Descritor *Reagent*

Após a implementação de vocabulários controlados na plataforma Dendro, pode-se então prosseguir com as experiências de descrição de dados de investigação do domínio Produção de Hidrogénio. Estas experiências permitiram obter dados das descrições com e sem recurso a vocabulários controlados para comparar a qualidade das mesmas.

3.3 Experiências da descrição de dados de investigação no Dendro após a implementação de vocabulários controlados e recolha de dados para respetiva análise

Para conseguir analisar a qualidade de descrição de dados de investigação no domínio Produção de Hidrogénio, após a implementação de vocabulários controlados e chegar a conclusões sobre a utilidade dos vocabulários controlados na melhoria de qualidade de descrição e na facilidade da utilização da plataforma Dendro, realizamos as experiências da descrição de dados com 3 utilizadores do domínio escolhido.

As experiências foram realizadas no dia 21.06.2016. Cada uma delas tem a duração de 40 minutos aproximadamente. O guião de processo da experiência e o inquérito, que foi enviado após a realização da experiência para cada utilizador, encontra-se no Anexo 6. O inquérito ajuda a avaliar a usabilidade da plataforma Dendro, bem como a experiência no geral.

Para recolha e análise dos dados destas experiências, utilizamos a metodologia utilizada no Capítulo 2. Da mesma maneira foi criado o ficheiro Excel e os dados foram recolhidos manualmente de acordo com as descrições efetuadas na plataforma Dendro (Anexo 8). A Tabela 12 contém uma breve análise da descrição de dados mostrando as diferenças na descrição de dados de investigação entre os três utilizadores do mesmo grupo de investigação.

Tabela 12 – Descrição de dados efetuada por três utilizadores, utilizando vocabulários controlados, e comentários

Descritor utilizado por cada utilizador	Valor de descritor do utilizador 1	Valor de descritor do utilizador 2	Valor de descritor do utilizador 3	Comentário
Gravimetric Capacity	1.7 wt.% / 1.7 wt%	2 wt% / 1.8 wt%	1.8 wt.% / 2 wt%	Maneiras diferentes de escrita (uma vez existe sinal de pontuação ".", outra vez não)

Number of Reutilization	50 / 53	50 / 49	50 / 52	Não existe a diferença
Hydratation Factor	16	15 /11	16	Não existe a diferença
Additive	SDS	CMC	CMC	Conceitos de vocabulário controlado
Hydrolysis	Classic hydrolysis	Classic hydrolysis	Classic hydrolysis	Conceitos de vocabulário controlado
Reactor Type	EggR - ovoid mini reactor	EggR - ovoid mini reactor	EggR - ovoid mini reactor / SRc - conical small reactor	Conceitos de vocabulário controlado
Catalyst	Ni-Ru	Ni-Ru	Ni-Ru	Conceitos de vocabulário controlado
Hydrogen Generation Rate	0.8 L/min.gcat / 0.2 L/min.gcat	0.2 L/min.gcat / 0.1 L/mi.gcat	0.8 L/min.gcat / 0.1 L/min.gcat	Não existe a diferença, mas existe erro na introdução de unidade (mi.gcat em vez de min.gcat)
Reagent	NaBH4	NaBH4	NaBH4	Conceitos de vocabulário controlado
Creator	CEFT_Energy group - H2 / CEFT- Energy Group -H2	Da*** Fa***	H. X. Nu***	Maneiras diferentes de escrita

Após as experiências realizadas e respetiva análise é possível afirmar o seguinte:

1) A descrição de dados de investigação em descritores *Reactor Type, Hydrolysis, Catayst, Reagent, Additivo* foi normalizada com utilização de vocabulários controlados e agora não existe a diferença na descrição, tal como erros na introdução da descrição;
2) O preenchimento destes descritores está completo e correto;
3) O preenchimento de outros descritores continua a ter a diferença na descrição e erros na introdução do texto.

Antes de se efetuar a análise da qualidade de descrição detalhadamente, com valores concretos, pode-se constatar que a utilização de vocabulários controlados na descrição de dados melhorou a qualidade de descrição no geral. Por isso, acredita-se que a implementação de expressões regulares, em descritores sem vocabulários controlados, vai continuar a aumentar a qualidade de descrição e ao mesmo tempo facilitar utilização do sistema Dendro, tal como aconteceu com a implementação de vocabulários controlados.

De maneira a provar esta hipótese e obter dados concretos sobre a melhoria de qualidade dos metadados, continuou-se com a análise dos dados das experiências.

Capítulo 4. Análise de qualidade da descrição de dados de investigação

No contexto da análise de qualidade da descrição de dados de investigação, este capítulo inclui as seguintes secções:

- A primeira secção descreve o processo da análise de qualidade de descrição de dados de investigação antes de implementação de vocabulários controlados, ou seja, a análise de metadados criados por investigadores sem auxílio de vocabulários controlados;
- A segunda seção mostra o processo da análise de qualidade de descrição de dados após a implementação de vocabulários controlados, ou seja, a análise de registos de metadados, criados com uso de vocabulários controlados;
- Na terceira seção é realizada comparação dos resultados de dados obtidos após análises efetuadas.

Ao longo deste capítulo, é feita uma análise da qualidade das descrições dos dados de investigação, e avaliado o contributo de vocabulários controlados para este aspeto.

4.1 Antes da implementação de vocabulários controlados

A primeira análise ligada à qualidade de descrição de dados foi realizada com metadados obtidos manualmente nas 3 experiências de descrição de dados efetuadas por utilizadores do domínio Produção de Hidrogénio. Estas experiências ocorreram antes da implementação dos vocabulários controlados. Os dados recolhidos podem ser vistos nos Anexos 1 e 4 e a definição das métricas *Correctness*, *Completeness*, *Conformance to expectations*, *Overall Rating*, *Satisfaction e Task time* escolhidas para avaliação de qualidades de descrição estão especificadas no Capítulo 2.

Palavitsinis (2013), tal como Ochoa (2006) num dos seus trabalhos mostrou como quantificar a qualidade de metadados. Ochoa afirmou que a relevância de elemento é depende do contexto, o texto livre, por exemplo, exige um cálculo mais complexo (Ochoa e Duval 2006; Palavitsinis 2013). Analisando várias fórmulas existentes para avaliação da qualidade dos dados, não foi possível adaptar nenhuma ao presente caso, desta forma, ficou decidido propor as próprias regras e fórmulas de avaliação da qualidade da descrição dos dados de investigação no Dendro.

Correctness

Começando com avaliação dos dados de descrição aplicando a métrica *Correctness* criamos as seguintes regras de avaliação de cada registo de metadados e fórmula de cálculo:

100% - registo de metadados está 100% alinhado, sintaticamente e gramaticalmente, com o valor definido pelo grupo de investigadores de domínio Produção de Hidrogénio, ou seja, com o conceito correspondente no vocabulário controlado;

0 % - registo de metadados não está alinhado, nem sintaticamente nem gramaticalmente com o conceito correspondente no vocabulário controlado;

$$Cor_{ra} = \frac{N_{ra}}{N_{100}} * 100\% \text{ , onde}$$

Cor_{ra} - percentagem de exatidão de registo analisado *(ra),*

N_{ra} - número de palavras de registo analisado,

N_{100} - numero de palavras de registo 100 % correto, definido no vocabulário controlado.

Exemplo de análise de um registo do descritor *Reactor Type*:

$$Cor_{ra} = \frac{3}{4} * 100\% = 75\% \text{ , onde}$$

SR_c – conical small reactor	registo 100% correto, definido no vocabulário controlado	Cor_{100} = 100%
SR_c; conical; small; reactor	4 palavras que contém o conceito definido no vocabulário controlado	N_{100} = 4
Conical Small Reactor	3 palavras	*Nra* = 3

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. Estas discrepâncias não comprometem o significado da palavra, sendo negligenciáveis neste cálculo.

A análise deve ser efetuada manualmente com muito cuidado, para não perder o peso das palavras contidas num conceito. De maneira a entender-se melhor o que é entendido por "peso das palavras", usa-se então o seguinte exemplo:

1) Registo 100% correto – Conical Small Reactor
2) Registo analisado –Small Modular Reactor

Neste caso, Small Modular Reactor e Conical Small Reactor são entidades diferentes, e apesar de partilharem duas palavras iguais o valor de *Correctness* é 0%. Ou seja, se um termo

(uma palavra) influencia e muda o significado de conceito por completo, não se pode então efetuar o cálculo dos restantes conceitos e compará-los com o conceito definido no vocabulário controlado. Esta comparação, atualmente não pode ser efetuada automaticamente, por isso a análise deve ser realizada com precaução.

A Tabela 13 apresenta os valores das descrições que recolhemos de todas as experiências realizadas por investigadores antes da implementação de vocabulários controlados, aplicando a métrica *Correctness* e regras definidas. Numa das colunas são indicados os conceitos de vocabulários controlados para descritores, que foram definidas junto com grupo CEFT com valor atribuído de 100% de qualidade. Outra coluna contém os registos de metadados dos descritores preenchidos por investigadores durante as experiências. As diferentes descrições de utilizadores estão separados com símbolo "/", ou seja, "Egg Reactor / ovoid" significa que um utilizador criou registo Egg Reactor, e outro utilizador – ovoid. Neste caso, o valor calculado é uma média de dois valores Cor_{ra}. Na terceira coluna é o valor obtido após cálculo realizado.

Tabela 13 – Avaliação de registos de metadados, aplicando métrica *Correctness*

	Linguagem usada nos metadados sintatiamente e gramaticamente correta?		
Correctnes	**A descrição 100% coreta sintaticamente e gramaticamente**	**Os registos de metadados criados por Utilizadores 1-3**	**0-100%**
Reactor Type	EggR – ovoid mini reactor	Egg Reactor/ovoid	38%
	LR – large reactor	RG	0%
	MR_c – conical medium reactor	RM	0%
	MR_f – flat medium reactor		
	SR_c – conical small reactor	Conical Small Reactor	75%
	SR_f – flat small reactor		
Hydrolysis	Acid hydrolysis		
	Alkali-free hydrolysis	alkali	34%
	Classic hydrolysis	Classic hydrolysis/Classic/ classic	67%
Catalyst	Ni-Ru	NiRu/ Nickel - ruthenium	50%
	Pt/C		
	Co-B		
	Co-Mn-B		
	Co-B/Ni		
Reagent	$NaBH_4$	NaBH4 /Sodium Borohydride	50%
	NH_3BH_3		
	$LiAlH_4$		
	$LiBH_4$		
	KBH_4		
Additive	SDS		
	CMC		

Understanding

Antes de avaliar a qualidade de descrição de dados de investigação aplicando a métrica *Conformance to expectations,* prosseguiu-se com avaliação de compreensão destes desrições.

U_{ra} é compreensão da descrição (percentagem), que foi obtida através de inquérito com o Utilizador 4 do domínio Produção de Hidrogénio. O inquérito pode ser consultado no Anexo 9 e os valores obtidos após a análise estão apresentados na Tabela 14. Na coluna, intitulada "Os registos de metadados criados pelos Utilizadores 1-3" estão indicadas todas as descrições de dados efetuadas por utilizadores durante as experiências de descrição antes da implementação de vocabulários controlados. O Utilizador 4 devia indicar o nível de compreensão destas descrições (0% atribuído se o registo não é compreensível, 50% se o registo é compreensível, mas o utilizador tem algumas dúvidas sobre esta descrição e 100% se o registo é compreensível). De acordo com as respostas do Utilizador 4, na coluna 0-100%, temos os valores de U_{ra}. Na outra coluna são apresentados as descrições, que na opinião deste investigador, completava o registo de metadados.

Tabela 14 – Avaliação de compreensão da descrição existente no Dendro

	Percentagem de compreensão da descrição existente no Dendro (até implementaçao de vocabularios controlados)		
Ura	**Os registos de metadados criados por Utilizadores 1-3**	**Descrição completa proposta por Utilizador 4**	**0-100%**
Reactor Type	Egg Reactor	Reactor com forma de ovo	100%
	ovoid	-	50%
	RG	-	0%
	RM	-	0%
	Conical Small Reactor	Reactor com forma cónica	100%
Hydrolysis	alkali	Hidrólise em meio alcalino	100%
	classic	-	50%
	Classic hydrolysis	Hidrólise classica	100%
Catalyst	NiRu	Catalisador com base de Niquel-Ruténio	100%
	Nickel - ruthenium	Catalisador com base de Niquel-Ruténio	100%
Reagent	NaBH4	Tetra borohidreto de sódio (ou borohidreto de sódio)	100%
	Sodium Borohydride	Borohidreto de sódio	100%

Conformance to expectations

De forma a avaliar os dados aplicando a métrica *Conformance to expectations* foram criadas as seguintes regras:

100% - o registo de metadados cumpre na totalidade os requisitos de uma determinada comunidade de utilizadores, neste caso o grupo de investigadores de domínio Produção de Hidrogénio. Assim, o valor 100% atribuído ao registo se é compreensível pelos utilizadores deste domínio (U_{ra} = *100%)* e está com uma descrição exata (Cor_{ra} = *100%*) em comparação com o conceito correspondente no vocabulário controlado.

0 % - registo de metadados não preenche os requisitos da comunidade de utilizadores, não é compreensível e não tem descrição exata (Cor_{ra}=0%) em comparação com o conceito correspondente no vocabulário controlado;

$$Conf_{ra} = \frac{Cor_{ra}+U_{ra}}{2}$$, onde

$Conf_{ra}$ – conformidade de registo analisado,

Cor_{ra} – valor de exatidão de registo analisado,

U_{ra} – compreensão da descrição (percentagem).

Exemplo de análise de um registo do descritor *Reactor Type*:

Registo "Conical Small Reactor" – 100% compreensível na comunidade (U_{ra}=*100%,* ver Tabela 14*)*, mas não está correto (Cor_{ra}=*75%*, ver Tabela 13):

$$Conf_{ra} = \frac{75\% + 100\%}{2} = 88\%$$

A Tabela 15 mostra os valores das descrições analisadas, aplicando a métrica *Conformance to expectations* e avaliação destas descrições de acordo com as regras definidas. A primeira coluna, intitulada "A descrição 100% compreensível, correta sintaticamente e gramaticamente" contém os conceitos de vocabulários controlados para os descritores *Reactor Type, Hydrolysis, Catalyst, Reagent, Additive*, que foram definidos junto com grupo CEFT com valor atribuído de 100% de qualidade. Outra coluna contém os registos de metadados criados pelos Utilizadores 1-3 durante as experiências realizadas antes de implementação de vocabulários controlados.

Tabela 15 - Avaliação de registos de metadados, aplicando métrica *Conformance to expectations*

	Os registos de metadados preenchem os requisites de uma determinada comunidade de utilizadores?		
Conformance to expectations	**A descrição 100% compreensivel, coreta sintaticamente e gramaticamente**	**Os registos de metadados criados por Utilizadores 1-3**	**0-100%**
Reactor Type	EggR – ovoid mini reactor	Egg Reactor /ovoid	56%
	LR – large reactor	RG	0%
	MR_c – conical medium reactor	RM	0%
	MR_f – flat medium reactor		
	SR_c – conical small reactor	Conical Small Reactor	88%
	SR_f – flat small reactor		
Hydrolysis	Acid hydrolysis		
	Alkali-free hydrolysis	alkali	67%
	Classic hydrolysis	Classic hydrolysis/Classic/ classic	71%
Catalyst	Ni-Ru	NiRu/ Nickel - ruthenium	75%
	Pt/C		
	Co-B		
	Co-Mn-B		
	Co-B/Ni		
Reagent	$NaBH_4$	NaBH4 /Sodium Borohydride	75%
	NH_3BH_3		
	$LiAlH_4$		
	$LiBH_4$		
	KBH_4		
Additive	SDS		
	CMC		

Completeness

Seguidamente foi realizada a avaliação de qualidade de dados de descrição, aplicando a métrica *Completeness* – número de descritores preenchidos em comparação com o número total de descritores. Para realizar o cálculo foi obtida a totalidade dos descritores preenchidos pelos investigadores em três experiências, não contabilizando a repetição dos mesmos. Por sua vez obteve-se a totalidade dos descritores preenchidos por cada investigador nas mesmas experiências, não contabilizando a repetição dos mesmos. Para a comparação dos valores adquiridos, foi criada a seguinte fórmula de cálculo:

$$Compl = \frac{N_{dpu} * 100\%}{N_{dt}}, \text{ onde}$$

Compl – completude,

Ndpu – número de descritores preenchidos por um utilizador que não são repetidos numa experiência, ou seja, se o utilizador está a preencher três vezes o mesmo descritor, então este descritor será contado uma única vez,

Ndt – número total de descritores não repetidos numa experiência. O número total é a soma de todos os descritores, sem repetições, utilizados por todos os utilizadores em todas as experiências.

Exemplo,

N_{dt} = 20 – total de descritores, utilizados por todos utilizadores (não repetidos)

N_{dpu} = 14 – número de descritores, utilizados por Utilizador 1 (não repetidos)

$$Compl = \frac{14 * 100\%}{20} = 70\%.$$

De acordo com as regras definidas, obtém-se os seguintes resultados, que podem ser vistos na Tabela 16.

Tabela 16 - Avaliação de registos de metadados, aplicando métrica *Completeness*

	Numero de descritores preenchidos em comparação com o numero total de descritores		
Utilizador	**Total de descritores, utilizados por todos utilizadores (não repetidos)**	**Quantidade de descritores preenchidos por Utilizador (não repetidos)**	**Completeness 0-100%**
Utilizador 1	20	14	70%
Utilizador 2	20	19	95%
Utilizador 3	20	12	60%
Média			75%

Overall Rating

Overall Rating é uma agregação de todas as métricas acima aplicadas, sendo calculada como a média dos valores das métricas propostas, calculada para cada descritor (Tabela 17):

$$OvR_{ra} = \frac{Cor_{ra}+Conf_{ra}+Compl}{3}.$$

Tabela 17 – Avaliação de registos de metadados, aplicando métrica *Overall Rating*

	A pontuação geral do registo de metadados, tendo em conta todos os critérios acima
Overall Rating	**0-100%**
Reactor Type	46%
Hydrolysis	65%
Catalyst	67%
Reagent	67%
Additive	

A Tabela 18 agrupa os resultados obtidos durante a análise:

Tabela 18 – Avaliação de qualidade de metadados de descritores com vocabulários controlados

AVALIAÇÃO DE QUALIDADE DE METADADOS		Avaliador		Yulia Karimova		
		→				
DESCRITOR	MÉTRICAS	0%-24%	25%-49%	50%-74%	75%-99%	100%
Reactor Type	Correctness		28%			
Hydrolysis	Correctness			50,50%		
Catalyst	Correctness			50%		
Reagent	Correctness			50%		
Additive	Correctness					
		→				
		0%-24%	25%-49%	50%-74%	75%-99%	100%
Reactor Type	Conformance to expectations		36%			
Hydrolysis	Conformance to expectations			69%		
Catalyst	Conformance to expectations				75%	
Reagent	Conformance to expectations				75%	
Additive	Conformance to expectations					
		→				
		0%-24%	25%-49%	50%-74%	75%-99%	100%
Média	Completeness				75%	
		Pontuação geral do registo de metadados tendo em conta todos os criterios acima				
		0%-24%	25%-49%	50%-74%	75%-99%	100%
Reactor Type	Overall Rating		46%			
Hydrolysis	Overall Rating			65%		
Catalyst	Overall Rating			67%		
Reagent	Overall Rating			67%		
Additive	Overall Rating					

Satisfaction e Task Time

Após efetuada a análise de qualidade da descrição de dados de investigação, procedeu-se com a avaliação de usabilidade, aplicando as métricas *Satisfaction* e *Task Time*. Essas métricas podem ajudar a avaliar a usabilidade, a simplicidade e o desempenho do Dendro do ponto de vista de utilizador. Além disto, estas métricas podem ajudar a comprovar ou refutar a hipótese de que os vocabulários controlados facilitam o processo de descrição de dados de investigação e diminuem o tempo necessário na criação de metadados.

A avaliação da satisfação dos utilizadores é baseada nas respostas ao inquérito (ver Anexo 3) preenchido após a experiência de descrição de dados na plataforma Dendro. A avaliação tem uma escala compreendida entre 1 e 5, sendo 1 muito insatisfeito e 5 muito satisfeito.

O cálculo de *Task Time* numa tarefa é baseado na seguinte fórmula:

$$TT = \frac{\sum_{i=1}^{n} T_i}{n}, \text{ onde}$$

TT – média de tempo gasto por cada utilizador numa tarefa. Uma tarefa é a descrição de um conjunto de dados por um utilizador. Uma experiência contém várias tarefas, por exemplo um utilizador numa experiência pode efetuar a descrição de três conjuntos de dados.

T_i – Tempo gasto na descrição de um conjunto de dados (uma tarefa).

i - número de tarefas.

O *Task Time* por descritor é calculado na seguinte maneira:

$$TT_d = \frac{\sum_{i=1}^{n} T_i}{N_d}. \text{ onde}$$

TT_d - média de tempo gasto para preenchimento de um descritor. Cada experiência tem sua quantidade de descritores preenchidos,

N_d – quantidade de descritores preenchidos numa experiência. Neste caso a soma de descritores inclui todos os descritores, e estes podem ser repetidos.

Exemplo:

Uma experiência inclui duas tarefas da descrição (dois conjuntos de dados). O tempo de descrição de primeiro conjunto de dados é de 7 minutos, o tempo de descrição do segundo conjunto de dados é de 8 minutos. Esta experiência contou com um total de 24 descritores preenchidos.

Neste caso obtêm-se os seguintes cálculos para o tempo por tarefa e o tempo por descritor:

$$TT_t = \frac{7min + 8min}{2} = 7.5min$$

$$TT_d = \frac{7min + 8min}{24} = \frac{420seg + 480seg}{24} = 38seg$$

Os valores do tempo que cada utilizador precisava para realizar a descrição de cada conjunto de dados de investigação, e o número de descritores preenchidos por eles numa experiência são mostrados na Tabela 19.

Tabela 19 – Número de descritores de cada utilizador e tempo gasto

	O tempo utilizado na descrição: Tarefa 1	O tempo utilizado na descrição: Tarefa 2	Quantidade de descritores, preenchidos numa experiência
Utilizador 1	27min	15min	32
Utilizador 2	23min		20
Utilizador 3	7min	8min	24

Após os cálculos realizados os resultados para o tempo médio que cada utilizador usou para efetuar a descrição dos seus conjuntos de dados de investigação podem ser vistos na Tabela 20.

Tabela 20 – Avaliação de usabilidade da plataforma Dendro sem utilização de vocabulários controlados

AVALIAÇÃO DE USABILIDADE	Avaliador Yulia Karimova		
Métricas	Utilizador 1	Utilizador 2	Utilizador 3
Task Time de tarefa	21 min	23 min	7,5 min
Task Time por descritor	79 seg	69 seg	38 seg
	Média		
Satifaction (1 ⟶ 5)	4		

Resultados da análise

A análise realizada mostrou que a qualidade de descrição sem utilização de vocabulários controlados está muito abaixo do desejável. Por exemplo, o valor de descritor *Reactor Type* ficou abaixo dos 50%, o que indica que as diferenças na descrição de dados de investigação, erros sintáticos e gramáticos, bem como a descrição incompleta estão a influenciar a qualidade dos metadados.

Apesar disto, os utilizadores estão satisfeitos com a utilização da plataforma Dendro e não necessitam de muito tempo para a criação de metadados para os seus dados de investigação.

A base deste trabalho tinha a hipótese de que a implementação de vocabulários controlados ajuda a aumentar a qualidade de descrição de dados de investigação e facilita o processo de descrição no geral. Na próxima secção prosseguimos com análise dos dados das experiências realizadas após a implementação de vocabulários controlados.

4.2 Após a implementação de vocabulários controlados

Após a criação e implementação de vocabulários controlados foram realizadas três experiências pelos utilizadores do domínio Produção de Hidrogénio. Seguidamente procedeu-se a análise da qualidade de descrição de dados de investigação depositados na plataforma Dendro. Os metadados recolhidos podem ser vistos no Anexo 8.

Para que seja possível comparar os resultados da qualidade de descrição antes e após a implementação de vocabulários controlados, a análise corrente baseou-se nas mesmas métricas e regras definidas na análise anterior.

Correctness

Para descrever os dados de investigação, os utilizadores de domínio Produção de Hidrogénio durante as experiências realizadas utilizaram-se vocabulários controlados fechados, ou seja, para descritores *Reactor Type, Hydrolysis, Catalyst, Reagent* e *Additive* escolheram-se os conceitos pré-definidos. Sendo assim todas as descrições, aplicando a métrica *Correctness,* têm 100% da qualidade.

Conformance to expectation

Em seguida, avaliamos as descrições obtidos após de experiencias realizados pelos Utilizadores 1-3 após a implementação de vocabulários controlados, aplicando a métrica *Conformance to expectation*. Antes de avançar com esta análise, prosseguiu-se com a análise de inquérito de forma a obter os valores de compreensão da descrição U_{ra} pelo Utilizador 4.

O inquérito pode ser consultado no Anexo 10 e os valores obtidos após a análise são apresentados na Tabela 21. Na coluna, intitulada "Os conceitos vocabulários controlados" são apresentadas as descrições que foram definidas em conjunto com grupo CEFT. Na próxima coluna são apresentadas as descrições propostas pelo Utilizador 4, que podem ajudar a completar descrição de conceitos nos vocabulários controlados definidos.

Tabela 21 – Avaliação de compreensão dos conceitos escolhidos para vocabulários controlados

	Percentagem de compreensão da descrição existente no Dendro (após implementaçao de vocabularios controlados)		
Ura	**Os conceitos de vocabularios controlados**	**Descrição completa proposta por Utilizador 4**	**0-100%**
Reactor Type	EggR – ovoid mini reactor	mini reactor com formato de ovo	100%
	LR – large reactor	reactor grande	100%
	MR_c – conical medium reactor	reator médico com formato cónico	100%
	MR_f – flat medium reactor	reactor médio com formato plano	100%
	SR_c – conical small reactor	reactor pequeno com formato cónico	100%
	SR_f – flat small reactor	reactor pequendo com formato plano	100%
Hydrolysis	Acid hydrolysis	hidrólise em meio ácido	100%
	Alkali-free hydrolysis	hidrólise com ausência de inibidor alcalino	100%
	Classic hydrolysis	hidrólise clássica	100%
Catalyst	Ni-Ru	catalisador à base de Níquel-Ruténio	100%
	Pt/C	catalisado à base de Platina suportada em carvão	100%
	Co-B	catalisador à base de Cobalto-Boro	100%
	Co-Mn-B	catalisador à base de Cobalto-Manganês-Boro	100%
	Co-B/Ni	catalisador à base de Cobalto-Boro/Níquel	100%
Reagent	$NaBH_4$	Borohidreto de sódio	100%
	NH_3BH_3		0%
	$LiAlH_4$		0%
	$LiBH_4$		0%
	KBH_4	Borohidreto de potássio	100%
Additive	SDS		0%
	CMC		0%

Seguidamente, na Tabela 22 são apresentados os valores das descrições analisadas, aplicando a métrica *Conformance to expectations* obtidos após a análise efetuada. Numa das colunas encontram-se os conceitos dos vocabulários controlados para descritores *Reactor Type, Hydrolysis, Catalyst, Reagent, Additive*, que foram definidos juntamente com o grupo CEFT com valor atribuído de 100% de qualidade. Outra coluna contém registos de metadados criados pelos Utilizadores 1-3 durante as experiências realizadas após a implementação de vocabulários controlados.

Tabela 22 - Avaliação de registos de metadados, aplicando a métrica *Conformance to expectations*

Conformance to expectations	Os registos de metadados preenchem os requisites de uma determinada comunidade de utilizadores?		
	A descrição 100% compreensivel, coreta sintaticamente e gramaticamente	**Os registos de metadados criados por Utilizadores 1-3**	**0-100%**
Reactor Type	EggR – ovoid mini reactor	EggR – ovoid mini reactor	100%
	LR – large reactor		
	MR_c – conical medium reactor		
	MR_f – flat medium reactor		
	SR_c – conical small reactor	SR_c – conical small reactor	100%
	SR_f – flat small reactor		
Hydrolysis	Acid hydrolysis		
	Alkali-free hydrolysis		
	Classic hydrolysis	Classic hydrolysis	100%
Catalyst	Ni-Ru	Ni-Ru	100%
	Pt/C		
	Co-B		
	Co-Mn-B		
	Co-B/Ni		
Reagent	$NaBH_4$	$NaBH_4$	100%
	NH_3BH_3		
	$LiAlH_4$		
	$LiBH_4$		
	KBH_4		
Additive	SDS	SDS	50%
	CMC	CMC	50%

Completeness

A próxima avaliação é realizada aplicando-se a métrica *Completeness*. Utilizando as regras e fórmulas definidas na seção anterior, obtemos os seguintes resultados (Tabela 23).

Tabela 23 - Avaliação de registos de metadados, aplicando a métrica *Completeness*

Utilizador	Numero de descritores preenchidos em comparação com o numero total de descritores		
	Total de descritores, utilizados por todos utilizadores (não repetidos)	**Quantidade de descritores preenchidos por Utilizador (não repetidos)**	**Completeness 0-100%**
Utilizador 1	20	19	95%
Utilizador 2	20	13	65%
Utilizador 3	20	17	85%
Média			82%

Overall Rating

Após aplicação da métrica *Overall Rating*, obtiveram-se os resultados apresentados na Tabela 24.

Tabela 24 – Avaliação de registos de metadados, aplicando a métrica *Overall Rating*

	A pontuação geral do registo de metadados, tendo em conta todos os critérios acima
Overall Rating	**0-100%**
Reactor Type	94%
Hydrolysis	94%
Catalyst	94%
Reagent	94%
Additive	77%

De maneira a resumir e facilitar a visualização e compreensão dos dados obtidos foi criada a Tabela 25.

Tabela 25 – Avaliação de qualidade de metadados de descritores com vocabulários controlados

AVALIAÇÃO DE QUALIDADE DE METADADOS		Avaliador		Yulia Karimova		
		→				
DESCRITOR	MÉTRICAS	0%-24%	25%-49%	50%-74%	75%-99%	100%
Reactor Type	Correctness					100%
Hydrolysis	Correctness					100%
Catalyst	Correctness					100%
Reagent	Correctness					100%
Additive	Correctness					100%
		→				
		0%-24%	25%-49%	50%-74%	75%-99%	100%
Reactor Type	Conformance to expectations					100%
Hydrolysis	Conformance to expectations					100%
Catalyst	Conformance to expectations					100%
Reagent	Conformance to expectations					100%
Additive	Conformance to expectations			50%		
		→				
		0%-24%	25%-49%	50%-74%	75%-99%	100%
Média	Completeness				82%	
		Pontuação geral do registo de metadados tendo em conta todos os criterios acima				
		0%-24%	25%-49%	50%-74%	75%-99%	100%
Reactor Type	Overall Rating				94%	
Hydrolysis	Overall Rating				94%	
Catalyst	Overall Rating				94%	
Reagent	Overall Rating				94%	
Additive	Overall Rating				77%	

Satisfaction e Task Time

Em seguida, procedeu-se com a avaliação de usabilidade, aplicando as métricas *Satisfaction* e *Task Time*.

Os valores do tempo que cada utilizador precisava para realizar a descrição de cada conjunto de dados de investigação, e a número de descritores preenchidos por ele numa experiência são demonstrados na Tabela 26:

Tabela 26 – Número de descritores de cada utilizador e tempo gasto

	O tempo utilizado na descrição: Tarefa 1	O tempo utilizado na descrição: Tarefa 2	Quantidade de descritores, preenchidos numa experiência
Utilizador 1	25	13	35
Utilizador 2	10	4	25
Utilizador 3	5	3	33

Com base nas respostas do inquérito (Anexo 7), tempo gasto durante das experiências de descrição (Tabela 26) e utilização da mesma metodologia e as mesmas métricas da análise da usabilidade antes de criação de vocabulários controlados, foram obtidos os seguintes resultados (Tabela 27):

Tabela 27 – Avaliação de usabilidade de plataforma Dendro com utilização de vocabulários controlados

AVALIAÇÃO DE USABILIDADE	Avaliador Yulia Karimova		
Métricas	Utilizador 1	Utilizador 2	Utilizador 3
Task Time de tarefa	19 min	7 min	4 min
Task Time por descritor	65 seg	34 seg	15 seg
	Média		
Satifaction (1 ⟶ 5)	4,5		

Resultados da análise

Os resultados da análise realizada mostram que a qualidade de descrição de dados de investigação com utilização de vocabulários controlados aumentou. Aliás, todos os valores ficam muito próximos de 100%. Isto indica que a utilização de vocabulários controlados está a diminuir as diferenças na descrição, erros sintáticos e gramaticais, e ajuda a obter uma descrição mais completa.

Contudo, a qualidade dos metadados do descritor *Additive* ainda se encontra com valor baixo. A razão disto é a descrição dos conceitos escolhidos para vocabulário controlado. De acordo com resultados de inquérito de avaliação de compreensão (Anexo 10), a descrição é

incompleta e pouco percetível para o grupo de investigadores do domínio Produção de Hidrogénio.

As experiências da descrição de dados de investigação realizadas no Dendro com utilização de vocabulários controlados deixaram os utilizadores satisfeitos com a usabilidade da plataforma. O tempo que cada investigador precisava para realização das tarefas diminuiu.

De modo a comprovar ou refutar a hipótese desta dissertação, que a implementação de vocabulários controlados ajuda a aumentar a qualidade de descrição de dados de investigação e facilita o processo de descrição no geral, a comparação dos resultados é feita na próxima secção.

4.3 Comparação dos resultados antes e após de implementação de vocabulários controlados

Os resultados de análises efetuados mostram 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 28, a qualidade da descrição após a implementação dos vocabulários controlados melhorou.

Durante a realização da segunda experiência da descrição os investigadores utilizaram praticamente a mesma quantidade de descritores em comparação com a primeira e não tiveram grandes dificuldades na utilização da plataforma. O nível da satisfação manteve-se igual com um ligeiro aumento. Contudo, afirmaram que a utilização de vocabulários controlados facilita o processo da descrição. Analisando as respostas de inquéritos e comentários de utilizadores (Anexo 9) após a segunda experiência conclui-se que os investigadores preferem criar os metadados com auxílio de vocabulários controlados, pois a comparação dos valores na Tabela 28 obtidos aplicando a métrica *Task Time*, apontam para que o tempo médio necessário para realização de uma tarefa diminuiu. Para além disso os vocabulários controlados ajudaram o Utilizador 3 a diminuir o tempo gasto na descrição de dados de investigação e ao mesmo tempo aumentar a quantidade de descritores preenchidos.

Sendo assim, a comparação dos resultados demostra que a utilização de vocabulários controlados na criação dos metadados melhora a qualidade de descrição dos dados de investigação e facilita o processo de descrição em geral, o que confirma a hipótese desta dissertação.

Tabela 28 – Comparação de resultados de qualidade de dados antes e após a implementação de vocabulários controlados

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%	
Descritor	**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%	
Descritor	**Qualidade antes de implementação de vocabulários controlados (*Overall Rating*)**		**Qualidade após de implementação de vocabulários controlados (*Overall Rating*)**	
Reactor Type	46%		94%	
Hydrolsysis	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 1	**Task-time de tarefa 2**	**Task-time de tarefa 1**	**Task-time de tarefa 2**
Utilizador 1	27 min	15 min	25 min	13 min
Utilizador 2	23 min		10 min	4 min
Utilizador 3	7 min	8 min	5 min	3 min
	Quantidade de descritores preenchidos na experiência			
Utilizador 1	32		35	
Utilizador 2	20		25	
Utilizador 3	24		33	
	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	

Conclusões e perspetivas futuras

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 eles precisam. O interesse e a motivação de investigadores tanto na descrição de dados como na escolha de um sistema para criação dos metadados depende de vários fatores, tais como: usabilidade, simplicidade, facilidade de compreensão dos descritores e possibilidade de descrever os seus dados com qualidade sem grande esforço, entre outros. Os resultados do trabalho efetuado nesta dissertação mostraram a presença de todos os fatores acima indicados, mas mais concretamente a existência de problemas ligados a erros ocorridos durante a descrição dos dados, de ordem sintática ou gramatical.

O trabalho desenvolvido nesta dissertação refletiu estes problemas e enquadrou-se em geral no processo de desenvolvimento da plataforma de gestão de dados de investigação Dendro da Universidade do Porto. Pretendeu-se encontrar uma solução para agilizar todo o processo de descrição de dados de investigação e assim contribuir para a melhoria da qualidade dos metadados criados no Dendro.

Resumindo, os objetivos desta dissertação foram segmentados em três vertentes. Em primeiro lugar procedeu-se à escolha do domínio Produção de Hidrogénio como caso de estudo, definiram-se as métricas para a avaliação de qualidade de descrição de dados de investigação e foi feita a recolha de metadados existentes no Dendro. Em segundo lugar foram elaborados e implementados os vocabulários controlados para o domínio escolhido e realizaram-se as experiências de descrição de dados de investigação no contexto do domínio Produção de Hidrogénio. Em terceiro lugar 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. Nomeadamente a descrição efetuada com uso de vocabulários controlados melhorou a qualidade de descrição de dados e simplificou todo o processo de criação de metadados no

Dendro, normalizando a descrição, diminuído os erros sintáticos e gramaticais, obtendo-se uma descrição mais completa e correta sem aumento do tempo necessário[20].

Durante o trabalho realizado foram enfrentadas várias dificuldades. Uma dessas dificuldades é a natureza do grupo CEFT, que é um grupo pequeno com um número restrito de investigadores. Devido a este fator os dados obtidos para análise através das experiências de descrição e preenchimento de inquéritos não contém um número elevado de dados para análise. Outra dificuldade está relacionada com a necessidade de desenvolvimentos adicionais na plataforma Dendro, para a implementação de todas as ideias que surgiram no desenrolar desta dissertação.

Como perspetiva de trabalho futuro, a Tabela 29 apresentada todas as sugestões que foram apuradas para a facilitação do processo de descrição e continuação de melhoria de qualidade de descrição de dados de investigação no Dendro.

Tabela 29 – Sugestões de melhoria e perspetivas futuras

	Vocabulários controlados
1	Aplicar a metodologia elaborada nesta dissertação em outros domínios de investigação existentes no Dendro
2	Criação de vocabulários controlados para descritores genéricos, assim como *Language, Format, etc.*
3	Para os descritores *Gravimetric Capacity, Hydrogen Generation Rate* e *Temperature* criar vocabulários controlados com conceitos pré-definidos de unidades utilizados pelos investigadores do domínio Produção de Hidrogénio, tais como: *wt.% (Gravimetric Capacity), L/min.gcat (Hydrogen Generation Rate)* e *Celsius (Temperature*). Para estes descritores está prevista a existência de um campo para inserção de texto-livre. Este campo será dividido em duas partes: uma para valor (texto-livre) e outra para vocabulário controlado com lista de unidades deste valor.
4	Criar ferramentas no Dendro, para o responsável do grupo de investigadores administrar os projetos deste grupo, de forma a decidir se adiciona os conceitos sugeridos por utilizadores do grupo à lista de vocabulários controlados. Ou seja, implementar vocabulários controlados abertos e fechados, utilizando *Annotation Property - isClosed* criado para este propósito

[20] Com base nestes resultados está a ser preparada a escrita de uma comunicação para 7ª Conferencia Luso-Brasileira sobre Acesso Aberto, que decorrerá em novembro de 2016.

5	Para facilitar a compreensão dos conceitos de descritor *Additive*, adicionar a descrição mais completa para os conceitos pré-definidos de vocabulários controlados: *SDS - Sodium dodecyl sulfate* e *CMC - Carboxymethyl celulose*
6	Adicionar mais um descritor com vocabulário controlado aberto para domínio Produção de Hidrogénio – *Inhibitor - Type of inhibitor used in the experiment*. Adicionar os conceitos pré-definidos para vocabulário controlado, tal como: NaOH, entre outros.
	Expressões regulares
1	Criação e implementação de expressões regulares. Realização de análise da qualidade de descrição em descritores – *Contributor, Creator, Date, Coverage, etc.* Na Figura 20 é apresentado o exemplo da expressão regular para descritor *Date*, criado no *Protégé* através de *Annotation Property - hasRegEx* e *hasErrorMessage* criados para este propósito
	Sugestões gerais
1	Adicionar mais um descritor (texto livre) para domínio Produção de Hidrogénio – *Successive loads - Indicates the corresponding injection*
2	Adicionar mais um descritor (texto livre) para domínio Produção de Hidrogénio – *Catalyst / Reagent ratio - Ratio between the mass of catalyst and the mass of reactant*
3	Adicionar mais um descritor (texto livre) para domínio Produção de Hidrogénio – *Hydrogen yield - Percentage of hydrogen obtained in the reaction*

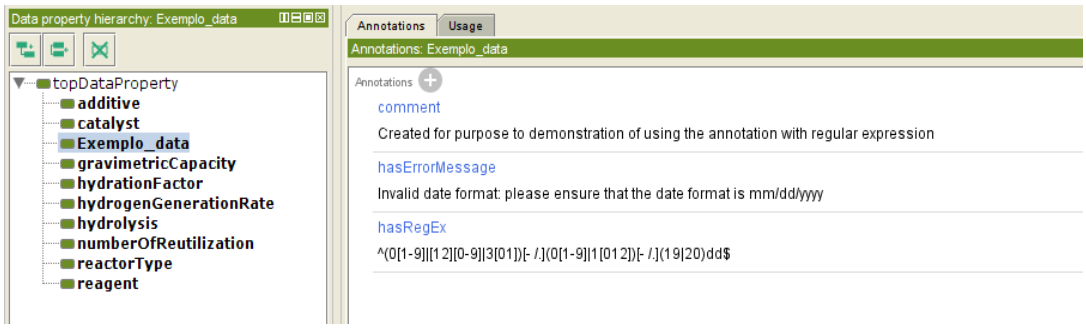


Figura 20 - Expressão Regular com mensagem de erro para descritor *Data*

Para terminar, pode-se afirmar que a continuação do desenvolvimento da plataforma Dendro, a implementação de sugestões de melhoria de qualidade de descrição de dados de investigação e a continuação de interação e elaboração com os investigadores de diferentes domínios científicos vai continuar trazer benefícios, tais como o aumento da usabilidade de

plataforma Dendro e simplicidade da mesma, de forma a motivar os investigadores para utilização do Dendro e melhorar a qualidade na descrição de dados de investigação.

Referências

Akmon, D, A Zimmerman, M Daniels, e M Hedstrom. 2011. «The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs». *Archival Science*, 1–22.

Alkhattabi, M, D Neagu, e A Cullen. 2010. «Information quality framework for e-learning systems». *Knowledge Management & E-Learning: An International Journal* 2 (4): 340–62. http://www.kmel-journal.org/ojs/index.php/online-publication/article/view/21/62.

Arrison, Thomas, Deborah D. Stine, Steve Olson, Neeraj P. Gorkhaly, Albert Swiston, e Sage Arbor. 2009. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in Digital Age*. Committe on Ensuring the Utility and Integrity of Research Data in a Digital Age; National Academy of Sciences. http://www.nap.edu/catalog.php?record_id=12615.

Arzberger, Peter, P Schroeder, Anne Beaulieu, G Bowker, K Casey, L Laaksonen, e D Moorman. 2004. «Promoting Access to Public Research Data for Scientific , Economic , and Social Development». *Data Science Journal* 3: 135–52. http://www.ics.uci.edu/~gbowker/promoting access.pdf.

Baca, Murtha. 2003. «Practical Issues in Applying Metadata Schemas and Controlled Vocabularies to Cultural Heritage Information». *Cataloging & Classification Quarterly* 36 (3-4): 47–55. http://polaris.gseis.ucla.edu/gleazer/260_readings/Baca.pdf.

Bargmeyer, Bruce E., e Daniel W. Gillman. 2000. «Metadata standards and Metadata registries: an overview». *Internation Conference on Establishment Surveys II, Buffalo, New York*. http://www.bls.gov/ore/pdf/st000010.pdf.

Barton, Jane, Sarah Currier, e Jessie Hey. 2003. «Building Quality Assurance into Metadata Creation: an Analysis based on the Learning Objects and e-Prints Communities of Practice». http://eprints.rclis.org/5237/.

Beall, Jeffrey. 2006. «Metadata and Data Quality Problems in the Digital Library». *Journal of Digital Information* 6 (3). http://www.cndwebzine.hcp.ma/img/pdf/beall.pdf.

Bermudez, L., E. Montgomery, S.P. Miller, C. Neiswender, e A. Isenor. 2011. «The Importance of Controlled Vocabularies». *The MMI Guides: Navigating the World of Marine Metadata*. http://marinemetadata.org/guides/vocabs/vocimportance.

Bruce, Thomas R., e Diane I. Hillmann. 2004. «The Continuum of Metadata Quality: Defining, Expressing, Exploiting». *Metadata in Practice*2. https://ecommons.cornell.edu/handle/1813/7895.

Candela, Leonardo. 2011. «Virtual Research Environments». *GRDI2020*. http://www.grdi2020.eu/Repository/FileScaricati/eb0e8fea-c496-45b7-a0c5-831b90fe0045.pdf.

Caplan, Priscilla. 1995. «You Call It Corn, We Call It Syntax-Independent Metadata for Document -Like Objects.» *The Public-Access Computer Systems Review* 6 (4): 19–23. http://xml.coverpages.org/caplan.html.

———. 2003. «Metadata fundamentals for all librarians.» American Library Association.

Castro, João Aguiar, Deborah Perrotta, Ricardo Amorim, João Rocha da Silva, e Cristina Ribeiro. 2015. «Ontologies for research data description: a design process applied to vehicle simulation». *Proceedings of the 9th Metadata and Semantics Research Conference (MTSR 2015)*.

Castro, João Aguiar, João Rocha da Silva, e Cristina Ribeiro. 2014. «Creating lightweight ontologies for data asset description Practical applications in a cross-domain research

data management workflow». *Joint Conference on Digital Libraries, Londres, 8 a 12 de Setembro de 2014*. http://dendro.fe.up.pt/pdf/papers/dl2014.pdf.

Chassanoff, A.M. 2009. «Metadata Quality Evaluation in Institutional Repositories: A Survey of Current Practices.» *A Master´s Paper for the M.S. in I.S. degree.*, 48. https://cdr.lib.unc.edu/indexablecontent/uuid:1e82a108-6404-4382-b775-d66c0a9711e3.

Corrado, E.M., e H.L. Moulaison. 2014. *Digital preservation for libraries, archives, and museums*. Rowman & Littlefield.

de Coronado, Sherri., e Gilberto. Fragoso. 2004. Enterprise Vocabulary Development in Protege/OWL: Workflow and Concept History Requirements NCI Center for Bioinformatics.

Duval, Erik, Wayne Hodgins, Stuart Sutton, e Stuart L. Weibel. 2002. «Metadata Principles and Practicalities». *D-Lib Magazine* 8 (4). http://www.dlib.org/dlib/april02/weibel/04weibel.html.

Ercegovac, Z. 1999. «Introduction, special topic issue: Integrating multiple overlapping metadata standards.» *Journal of the American Society for Information Science* 50: 1165–68.

«ESS Qulity Glossary». 2010. *Developed by Unit B1 «Qulity; Classifications», Eurostat*. http://unstats.un.org/unsd/dnss/docs-nqaf/ESS Quality Glossary 2010.pdf.

Fidel, Raya. 1992. «Who needs controlled vocabulary?» *Special Libraries* 83 (1): 1–9. http://faculty.washington.edu/fidelr/RayaPubs/WhoNeedsControlledVocabulary.pdf.

Fienberg, Stephen E., Margaret E. Martin, e Miron L. Straf. 1985. *Sharing research data. Science (New York, N.Y.)*. Vol. 229. Committe on National Statistics, National Research Council. http://www.nap.edu/catalog/2033/sharing-research-data.

Friedl, Jeffrey E F. 2006. *Mastering Regular Expressions*. O'Reilly Media, Inc.

Gattelli, Rúbia Tatiana. 2015. «Gestão de dados de investigação no domínio da oceanografia biológica: criação e avaliação de um perfil de aplicação baseado em ontologia». *FEUP*. https://repositorio-aberto.up.pt/bitstream/10216/79336/2/117377.pdf.

Golbreich, Christine, Songmao Zhang, e Olivier Bodenreider. 2006. «The Foundational Model of Anatomy in OWL: Experience and Perspectives». *Web Semantics: Science, Services and Agents on the World Wide Web* 4 (3): 181–95. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.9032&rep=rep1&type=pdf.

Goyavaerts, Jan, e Steven Levithan. 2009. *Regular expressions Cookbook*. O'Reilly Media, Inc.

Greenberg, Jane. 2005. «Understanding Metadata and Metadata Schemes». *Cataloging & Classification Quarterly* 40 (3-4): 17–36. http://www.columbia.edu/cu/libraries/inside/units/bibcontrol/osmc/greenberg.pdf.

Grimalovskii, Alexandr. 2013. «Expressões regulares.» *ProviderZ.ru*. http://www.codenet.ru/webmast/php/regexps.php.

Guenther, Rebecca, e Jaqueline Radebaugh. 2004. «Understanding Metadata». *National Information Standards Organization. Bethesda, MD: NISO Press* 20. http://www.niso.org/publications/press/UnderstandingMetadata.pdf.

Harpring, Patricia. 2010. *Introduction to controlled Vocabularies : Terminology for Art, Architecture, and Other Cultural Works*. Getty Publications. http://d2aohiyo3d3idm.cloudfront.net/publications/virtuallibrary/160606018X.pdf.

Harvey, Ross. 2005. *Preserving digital materials*.

Hedden, Heather. 2010. «Taxonomies and Controlled Vocabularies Best Practices for Metadata». *Journal of Digital Asset Management* 6 (5). Palgrave Macmillan: 279–84. http://www.palgrave-journals.com/doifinder/10.1057/dam.2010.29.

Heery, R., e M. Patel. 2000. «Application profiles: mixing and matching metadata schemas.» *Ariadne* 25. http://www.ariadne.ac.uk/issue25/app-profiles/.

Heldman, Kim. 2005. *PMP: Project Management Professional Study Guide*. *Imprint*. Wiley Publishing, Inc. http://bbu.yolasite.com/resources/Project Mgt.pdf.

Higgins, Sarah. 2007. «What are Metadata Standards». http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards.

———. 2008. «The DCC Curation Lifecycle Model». *International Journal of Digital Curation* 3 (1): 134–40. http://www.ijdc.net/index.php/ijdc/article/viewFile/69/48.

Juan, Angel A. 2012. *Collaborative and Distributed E-Research: Innovations in Technologies, Strategies and applications*. IGI Global.

Lee, Yang W., Diane M. Strong, Beverly K. Kahn, e Richard Y. Wang. 2002. «AIMQ: a methodology for information quality assessment». *Information & Management* 40 (2): 133–46. http://www.sciencedirect.com/science/article/pii/S0378720602000435.

Leise, Fred, Karl Fast, e Mike Steckel. 2002. «What is a controlled vocabulary?» *Boxes and Arrows*, 1–19. http://boxesandarrows.com/what-is-a-controlled-vocabulary/.

MANTRA. 2014. «Research Data Managemenet Training. Research data explaining.» *Edimburgo: University of Edinburgh*. http://datalib.edina.ac.uk/mantra/researchdataexplained/.

Mayernik, Matthew S. 2012. «Data citation initiatives and issues». *Bulletin of the American Society for Information Science and Technology* 38 (5): 23–28. doi:10.1002/bult.2012.1720380508.

McGilvary, D. 2008. «Data Quality Dimensions». Em *Executing data quality projects: ten steps to quality data and trusted information (TM)*. Elsevier. http://www.gfalls.com/storage/book/individual-downloads-quick-ref/10steps_DQDimen.pdf.

Mitchell, E. 2015. *Metadata Standards and Web Services in Libraries, Archives, and Museums: An Active Learning Resource: An Active Learning Resource*. ABC-CLIO.

Moreira, Bárbara L., Marcos André Gonçalves, Alberto H F Laender, e Edward A. Fox. 2009. «Automatic evaluation of digital libraries with 5SQual». *Journal of Informetrics* 3 (2): 102–23.

National Information Standards Organization. 2005. *Z39.19-2005: Guidelines for the Construction , Format , and Management of Monolingual Controlled Vocabularies*. *ANSI/NISO*. http://www.niso.org/apps/group_public/download.php/12591/z39-19-2005r2010.pdf.

Ochoa, Xavier, e Erik Duval. 2006. «Quality Metrics for Learning Object Metadata». *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, 1004–11.

———. 2009. «Automatic evaluation of metadata quality in digital repositories». *International Journal on Digital Libraries* 10 (2-3): 67–91. http://link.springer.com/10.1007/s00799-009-0054-4.

Palavitsinis, Nikos. 2013. «Metadata Quality Issues in Learning Repositories». http://dspace.uah.es/dspace/bitstream/handle/10017/20664/Thesis Palavitsinis.pdf?sequence=1&isAllowed=y.

Pennock, Maureen. 2007. «Digital Curation: A Life-Cycle Approach to Managing and Preserving Usable Digital Information». *Library and Archives Journal* 1: 1–3. http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf.

Qin, Jian, Alex Ball, e Jane Greenberg. 2012. «Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data». *Twelfth International Conference on Dublin Core and Metadata Applications*, 62–71. http://opus.bath.ac.uk/34849/1/107_396_1_PB.pdf.

Qin, Jian, e Stephen Paling. 2001. «Converting a controlled vocabulary into an ontology: the case of GEM». *Information Research* 6 (2): 6–2. http://www.citeulike.org/group/560/article/365347.

Sanz-Rodriguez, J., Juan Manuel Dodero, e Salvador Sánchez-Alonso. 2010. «Ranking Learning Objects trough Integration of Different Qulity Indicators». *Learning Technologies, IEEE Transactions on* 3 (4): 358–63. http://www.computer.org/csdl/trans/lt/2010/04/tlt2010040358.pdf.

Sayogo, Djoko Sigit, e Theresa A. Pardo. 2013. «Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data». *Government Information Quarterly* 30: 19–31. http://isiarticles.com/bundles/Article/pre/pdf/5106.pdf.

Seffah, Ahmed, Mohammad Donyaee, Rex B. Kline, e Harkirat K. Padda. 2006. «Usability measurement and metrics: A consolidated model». *Software Quality Journal* 14 (2): 159–78.

Shankaranarayanan, G., e A. Even. 2006. «The metadata enigma.» *Communications of the ACM* 49 (2): 88–94.

Shreeves, S.L., J. Riley, e L. Milewicz. 2006. «Moving towards shareable metadata.» *First Monday* 11 (8). http://firstmonday.org/ojs/index.php/fm/article/view/1386/1304.

Silva, João Rocha da, João Aguiar Castro, Cristina Ribeiro, e João Correia Lopes. 2014. «Dendro: Collaborative Research Data Management Built on Linked Open Data». *The Semantic Web: ESWC 2014 Satellite Events*, 483–87. http://2014.eswc-conferences.org/sites/default/files/eswc2014pd_submission_54.pdf.

Simberloff, D., B.C. Barish, K.K. Droegemeier, Etter D.M., N.V. Fedoroff, K.M. Ford, ..., e Jr.J.A. White. 2005. *Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century*. Technical Report NSB-05-40, National Science Foundation, Washington DC, USA. http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf.

Skoglund, Kevin. 2011. «Using regular expressions». http://www.lynda.com/Regular-Expressions-tutorials/Using-Regular-Expressions/85870-2.html.

Smit, Johanna Wilhelmina, e Nair Yumiko Kobashi. 2003. *Como elaborar vocabulário controlado para aplicação em arquivos. Como fazer*. Vol. 10. Arquivo do Estado. http://www.arqsp.org.br/arquivos/oficinas_colecao_como_fazer/cf10.pdf.

Smith, K., L. Seligman, e V. Swarup. 2008. «Everybody Share: The challenge of data-sharing systems.» *ComputerComputer* 41: 54–61.

Standen, James. 2010. «Using regular expressions to check data quality. Part 2». http://www.datamartist.com/how-to-use-regular-expressions-to-check-data-quality-part-2.

Strasser, Carly, Robert Cook, William Michener, e Amber Budden. 2012. «Primer on Data Management : What you always wanted to know». http://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf.

Stvilia, Besiki, Les Gasser, Michael B Twidale, Sarah L Shreeves, e Tim W Cole. 2004. «Metadata quality for federated collections». *Proceedings of the Ninth International Conference on Information Quality (ICIQ-04)*, 111–25. doi:10.4018/978-1-59904-420-0.ch008.

Stvilia, Besiki, Les Gasser, Michael B Twidale, e Linda C Smith. 2007. «A framework for Information Quality Assessment». *Journal of the American Society for Information Science and Technology* 58 (12): 1720–33.

Swan, Alma, e Sheridan Brown. 2008. «To share or not to share: Publication and quality assurance of research data outputs. Report commissioned by the Research Information Network», n. June: 56. http://eprints.soton.ac.uk/266742/.

Treloar, Andrew, e Ross Wilkinson. 2008. «Rethinking Metadata Creation and Management in a Data-Driven Research World». *IEEE Fourth International Conference on eScience*. Ieee,

782–89. http://andrew.treloar.net/research/publications/escience/2008_3535a782.pdf.

Van de Eynden, Veerle, Louise Corti, Matthew Woollard, Libby Bishop, e Laurence Horton. 2013. *Managing and Sharing Data. Best Practice For Researchers*. UK Data Archive. http://www.data-archive.ac.uk/media/2894/managingsharing.pdf.

Vickers, A. 2006. «Whose data set is it anyway? Sharing raw data from randomized trials». *Trials* 7 (1): 15. http://www.trialsjournal.com/content/7/1/15.

Walters, Tyler, e Katherine Skinner. 2011. *New Roles for New Times : Digital Curation for Preservation*. Association of Research Libraries. https://vtechworks.lib.vt.edu/bitstream/handle/10919/10183/nrnt_digital_curation17mar11.pdf?sequence=1&isAllowed=y.

Warner, A.J. 2002. «A taxonomy primer». *Lexonomy*. https://www.ischool.utexas.edu/~i385e/readings/Warner-aTaxonomyPrimer.html.

Wieczorek, John, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, e David Vieglais. 2012. «Darwin Core: An Evolving Community-Developed Biodiversity Data Standard». *PLoS ONE* 7 (1). http://dx.plos.org/10.1371/journal.pone.0029715.

Willis, Craig, Jane Greenberg, e Hollie White. 2012. «Analysis and Synthesis of Metadata Goals for Scientific Data». *Journal of the American Society for Information Science and Technology* 63 (8): 1505–20. http://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=5391&context=faculty_scholarship.

Woodley, M.S., G. Clement, e P. Winn. 2003. «DCMI Glossary». http://dublincore.org/documents/2003/08/26/usageguide/glossary.shtml.

Yu, Liyang. 2011. *A Developer's Guide to the Semantic Web*. Springer Science & Business Media. http://link.springer.com/10.1007/978-3-642-15970-1.

Zhang, Tao, Deborah J. Maron, e Christopher C. Charles. 2013. «Usability Evaluation of a Research Repository and Collaboration Web Site». *Journal of Web Librarianship* 7 (1): 58–82. http://www.tandfonline.com/doi/abs/10.1080/19322909.2013.739041.

Zhang, Yue, Adrian Ogletree, Jane Greenberg, e Chelcie Rowell. 2015. «Controlled Vocabularies for Scientific Data : Users and Desired Functionalities». Em *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, 54. American Society for Information Science.

ANEXOS

Anexo 1 – Os dados da experiência da descrição de dados de investigação, efetuados pelos Utilizador 1 e Utilizador 2 na plataforma Dendro antes de implementação de vocabulários controlados

Utilizador 1	Nome de ficheiro com dados de investigação	Descritores utilizados	Valor	Utilizador 2	Nome de ficheiro com dados de investigação	Descritores utilizados	Valor
		Temperature	25ºC		**teste AS**	Temperature	24
	3 classic hydrolysis RG 10NaBH4 040g T25 10 ml varias wt NaOH - TESE	Hydration Factor	16			Catalyst	Nickel - ruthenium
		Number of Reutilization	45			Gravimetric Capacity	2,3
		Reactor Type	RG			Hydration Factor	16
		Reagent	NaBH4, H2O			Hydrogen Generation Rate	0,5 liter/ (min.g)
		Date Issued	06/08/2015			Hydrolysis	alkali
		Access Rights	CEFT-Grupo de energia			Number of Reutilization	49
		Contributor	Ra*** C			Reactor Type	ovoid
		Coverage	faculdade de engenharia universidade porto			Reagent	Sodium Borohydride
		Alternative Title	RG-10ml-varias conc NaOH			Title	Catalytic Hydrolysis of Sodium Borohydride - 8 Successive Loads
		Creator	Fe*** MJF			Alternative Title	Number of loads - 8
		Descritores utilizados				Contributor	He*** Nu***
	1 classic hydrolysis RM 10NaBH4 3NaOH 040g 20ml varias temperaturas - TESE	Temperature	gama temperatura (20,25, ...50ºC)			Creator	Al*** P***
		Catalyst	Ni-Ru			Date Issued	06/23/2015
		Gravimetric Capacity	<5wt%			Subject	Catalytic Hydrolysis of Sodium Borohydride
		Hydration Factor	16			Instrumentation	Reactor, syringe, temperature and pressure probes
		Hydrolysis	classic			Instrumentation	acquisition data system
		Number of Reutilization	46			Software	Labview
		Reactor Type	RM			Compound	Hydrogen
		Reagent	NaBH4 (10wt%), NaOH (3wt%), resto água			Method (Hydrolysis)	Loads 1-7 Classic Hydrolysis; Load 8 with agitation
		Date Issued	06/08/2015				
		Alternative Title	classic-RM 10-3-040g/g-varias T				
		Contributor	Fe*** MJF				
		Coverage	CEFT-LabE206				
		Creator	Fe*** MJF				
		Descritores utilizados					
	2 classic hydrolysis RG 3NaOH 040g T27 10ml varias wt NaBH4 - TESE	Temperature	27ºC				
		Hydration Factor	<5 wt%				
		Reactor Type	RG				
		Date Issued	06/08/2015				
		Alternative Title	Classic-RG-3NaOH-varias con NaBH4				
		Contributor	Ra*** CM				
		Creator	Fe*** MJF				
		Method (Hydrolysis)	Hidrolise classica, procedimento habitual				

*** - omitido por revelar a identidade do utilizador

Anexo 2 – Guião de processo da experiência de descrição de dados no Dendro antes de implementação de vocabulários controlados (Produção de Hidrogénio)

Localização da máquina da experiência	Credenciais do utilizador
http://dendro-prd.fe.up.pt:3007	Login: *he***** Password: *he*****

Demonstração do funcionamento do Dendro

Vai ser explicado o objetivo da descrição de dados de investigação e demonstradas as funcionalidades gerais do Dendro, na base de experiencias realizadas por colegas de grupo CEFT.

Registo de utilizadores

Todos os utilizadores deverão ser criados e verificados antecipadamente e o *login* e *password* dos mesmos ter facultado em papel a hora da experiência.

Criação de projeto

O projeto Experiencia em Produção de Hidrogénio foi criado em Junho de 2015 para experiencias iniciais de funcionamento de Dendro. O utilizador 3 foi adicionado ao mesmo projeto do mesmo grupo CEFT. Nenhum dos investigadores poderá administrar, para já, os seus projetos, porque a plataforma ainda está em modo desenvolvimento.

Guião

1. Dar noção do conceito de descritor, caso necessário;
2. Apresentar os objetivos da avaliação;
3. Apresentar o funcionamento da plataforma Dendro - navegação, seleção e preenchimento de descritores;
4. Apresentar os descritores de domínios Dublin Core, Friend of Friend e Produção de Hidrogénio;
5. Se não houver dúvidas, pedir ao investigador para efetuar o depósito do conjunto de dados e realizar a descrição dos mesmos na plataforma Dendro;
6. Deixar o investigador à vontade durante a interação;
7. Apontar todas as dúvidas sugeridas durante a experiência;
8. Efetuar a gravação, com a indicação do tempo gasto para cada tarefa;
9. Em caso de ocorrerem bugs – registar os erros com mais detalhe possível;

10. Enviar o inquérito após a experiência realizada.

Inquérito após experiência

(https://docs.google.com/forms/d/15OjjEdvJUt46kdqueMNCNMeGU9upVSEZv9ThwhLg8P8/viewform):

1 – O que achou da experiencia de descrição em geral?

2 - Sentiu algumas dificuldades durante de descrição? Se sim quais?

3 – De 1 a 5 qual foi o grau de esforço para descrever os dados? (1 – pouco esforço; 5 – muito esforço)

4 – Acha que o processo de descrição de dados que exigiu muito tempo? Se sim, como acha onde e como podemos diminuir o tempo gasto?

5 – Acha que pode incluir esta tarefa para o seu dia-a-dia de investigador? (Sim; Não; Sim, mas com menos tempo gasto)

6 – Considera a descrição de dados como um processo importante para o seu trabalho? Se sim, porquê?

7 – Acha que toda a descrição que fez está adequada aos descritores utilizados? (Sim; Não; Outra)

8 – Existem mais descritores que gostava e acha importante de preencher? Se sim, não conseguiu encontrar ou?

9 – De 1 a 5, qual o seu nível de satisfação após de experimentar a plataforma Dendro? (1 – muito insatisfeito; 5 – muito satisfeito).

Anexo 3 – Respostas de inquérito após a experiência de descrição de dados antes de implementação de vocabulários controlados para avaliação dos dados aplicando métricas *Satisfaction* e *Task Time*

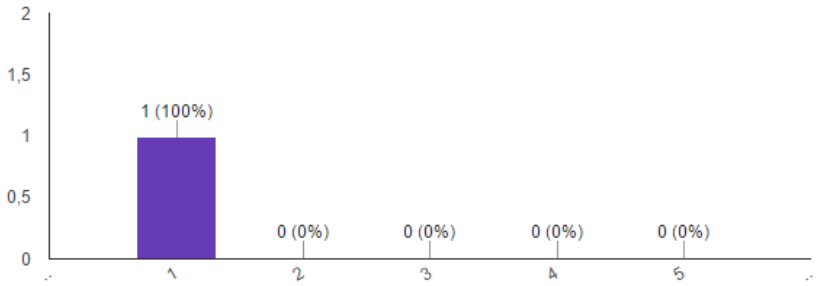
O que achou da experiencia de descrição no geral? (1 resposta)

Muito interessante, fácil e intuitiva

Sentiu alguma dificuldade durante de descrição? Se sim quais? (1 resposta)

Encontrar um descritor comum em várias áreas - Temperatura

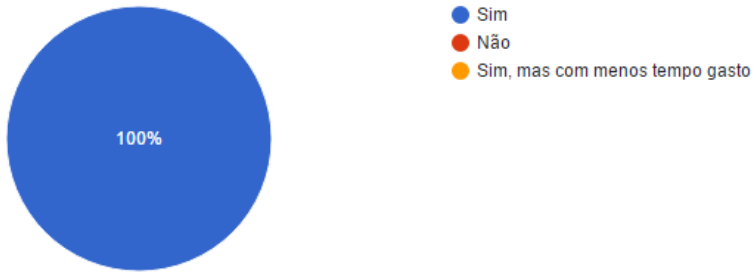
De 1 a 5 qual foi o grau de esforço para descrever os dados? (1 resposta)



Acha que o processo de descrição de dados exigiu muito tempo? Se sim, como acha onde e como podemos diminuir o tempo gasto?
(1 resposta)

Não.

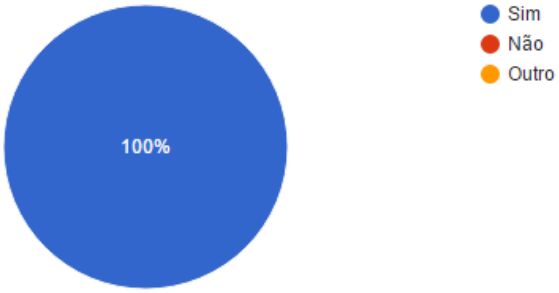
Acha que pode incluir esta tarefa para o seu dia – o- dia de investigador?
(1 resposta)



Considera a descrição dos dados como processo importante para o seu trabalho? Se sim, porque?
(1 resposta)

Sim. Permite uma fácil partilha e organização de dados sem que haja perdas de informação relevantes.

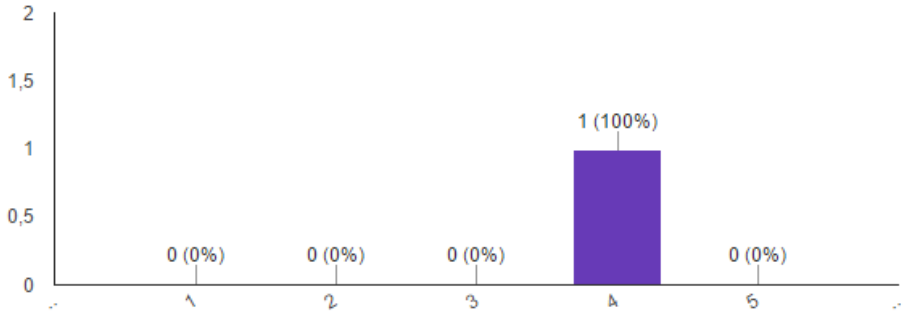
Acha que toda a descrição que fez está adequada aos descritores utilizados?
(1 resposta)



Existe mais descritores que gostava e acha importante de preencher? Se sim, não conseguiu encontrar ou?
(1 resposta)

Importante a adição de mais dois descritores (Additives e Successive loads)

De 1 a 5, qual o seu nível de satisfação após de experimentar a plataforma Dendro?
(1 resposta)



Anexo 4 - Os dados da experiência da descrição de dados de investigação, efetuados pelo Utilizador 3 na plataforma Dendro antes de implementação de vocabulários controlados

Utilizador 3	Nome de ficheiro com dados de investigação	Descritores utilizados	Valor
	xxx_Calssic_EggR_NiRu	Temperature	28ºC
		Catalyst	NiRu
		Gravimetric Capacity	1.9 wt%
		Hydration Factor	15
		Hydrogen Generation Rate	0.345 L/min.gcat
		Hydrolysis	Classic hydrolysis
		Number of Reutilization	50
		Reactor Type	Egg Reactor
		Reagent	NaBH4
		Title	Hydrogen generation from hydrolysis of NaBH4
		Creator	HXN-CEFT
		Date Issued	07/13/2015
	xxx_Calssic_SRc_NiRu	Temperature	22ºC
		Catalyst	NiRu
		Gravimetric Capacity	2.25 wt%
		Hydration Factor	16
		Hydrogen Generation Rate	0.48 L/min.gcat
		Hydrolysis	Classic
		Number of Reutilization	51
		Reactor Type	Conical Small Reactor
		Reagent	NaBH4
		Title	Hydrogen generation from hydrolysis of NaBH4
		Creator	HXN-CEFT
		Date	12/17/2015

Anexo 5 – Relatório da reunião com grupo CEFT sobre concretização de conceitos pré-definidos em Vocabulários Controlados

Data: 25/02/2016

Local: Laboratório IE206

Participantes: Utilizador 1, Utilizador 2, Utilizador 3, Yulia Karimova

Resumo

O presente relatório apresenta os resultados da Reunião do grupo CEFT juntamente com o autor desta dissertação Yulia Karimova, com a finalização de aprovar os conceitos pré-definidas, para Vocabulários Controlados, que foram elaborados com apoio do Utilizador 3.

1 – Objetivos da criação de Vocabulários Controlados

No início da reunião foi explicado que um dos nossos objetivos principais da criação de vocabulários controlados é a tentativa de melhoria da qualidade de descrição de dados, diminuição dos erros e facilitação do processo de descrição no geral. Além disto, foi explicado o que é um vocabulário controlado e expressões regulares, através de exemplos dos mesmos.

2 – Pontos a destacar

Como os participantes desta reunião, iá tiveram a oportunidade de conhecer a plataforma Dendro, foi-lhes perguntado se existem dúvidas sobre o funcionamento da plataforma Dendro, ou se há algum ponto que seja necessário salientar ou ajuda sobre o mesmo. Como não tinha nenhumas dúvidas prosseguisse para a discussão do ponto principal da reunião.

3- Discussão sobre conceitos; Comentários

Foi apresentada uma lista com conceitos pré-definidos para os descritores Reactor Type, Catalyst, Hydrolysis, Reagent e Additive com a finalização de concretizar e aprovar e no caso se necessário alterar, adicionar ou eliminar alguns dos conceitos ou seja proceder o ajuste. Existiam vários comentários sobre a lista de conceitos, mas não eram relevantes para vocabulários controlados em si, mas sim relevantes para outros pontos existentes no Dendro, tal como: um comentário sobre adição de descritor Sucesive loads, uns comentários sobre a definição de padrão para criação de formato sugerido através de expressão regular para descritores Creator, Contributor e um comentário sobre as normas existentes, ligados ao descritor Hydrogen Generation Rate, mais concretamente podem ser utilizados normas: **NPT** – normal pressure temperature 0ºC com 1 Atm, e **SPT** – standard pressure temperature 20º C com 1 Atm e por isso é relevante indicar com qual norma as unidades de valores estão descritas.

Mais um ponto levantado durante esta reunião é a possibilidade de adicionar outros valores para estas listas, ou seja deixar alguns descritores com vocabulários abertos.

4- Resultados

Ficou então acordado entre todos os membros do grupo CEFT, que os conceitos pré-definidos apresentados, não precisam de alteração e que os mesmos podem ser avançados com a sua implementação na plataforma.

5 – Comentários adicionais

Após a análise efetuada dos resultados da reunião, foram definidos descritores, que podem ter Vocabulários controlados abertos. Descritores Reactor Type, Hydrolysis – vocabulários fechados, Catalyst, Reagent e Additive – vocabulários abertos.

Gravação da reunião: contactar Yulia Karimova

Anexo 6 – Guião de processo da experiência de descrição de dados no Dendro após a implementação de vocabulários controlados (Produção de Hidrogénio)

Localização da máquina da experiência	Credenciais dos utilizadores
http://dendro-dev.fe.up.pt:3009	**Login: *h**** Password: ******* **Login: *d**** Password: ******* **Login: *j**** Password: *******

Demonstração do funcionamento do Dendro

Vai ser explicado o objetivo da descrição de dados de investigação e demonstradas as funcionalidades gerais do Dendro.

Registo de utilizadores

Todos os utilizadores deverão ser criados e verificados antecipadamente e o *login* e *password* dos mesmos ter facultado em papel a hora da experiência.

Criação de projeto

Todos os utilizadores serão adicionados para um novo projeto. Como a plataforma encontra-se em fase de desenvolvimento, nenhum dos investigadores pode administrar os seus projetos.

Guião

1. Apresentar o objetivo da experiência;
2. Relembrar a informação sobre descritores e funcionamento da plataforma Dendro no geral;
3. Caso não haver dúvidas, começar o processo de descrição de dados de investigação;
4. Deixar o investigador à vontade durante a experiência;
5. Apontar todas as dúvidas sugeridos durante a descrição;
6. Realizar a gravação, com a indicação do tempo gasto para cada tarefa de descrição de dados;
7. Analisar a utilização de vocabulários controlados;
8. Caso ocorrência dos bugs – registar os erros com mais detalhe possível;
9. Enviar o inquérito apás a experiência realizada.

Questionário após experiência

(https://docs.google.com/forms/d/1KXjuYlFDy3g1LewFQQ4FyiVYxxL39SvKwZUAw4g7By0/viewform):

1 – O que achou da experiência em geral?

2 - Sentiu algumas dificuldades durante a descrição? Se sim quais?

3 – De 1 a 5 qual foi o grau de esforço necessário para descrever os dados? (1 – pouco esforço; 5 – muito esforço)

4 – Acha que o processo de descrição de dados exigiu muito tempo? Se sim, como podemos diminuir o tempo gasto?

5 – Acha que utilização de vocabulários controlados na descrição de dados de investigação facilita ou complica o processo? (Facilita; Complica)

6 – Preferia de criar os metadados com ou sem utilização de vocabulários controlados? Se responder sem vocabulários controlados, indique por favor o motivo.

7 – Acha que o uso de vocabulários controlados diminui o tempo gasto para a descrição? (Diminui o tempo; Aumentou o tempo; Permaneceu igual; Outro)

8 – Gostaria de adicionar a descrição de dados de investigação na plataforma Dendro para o seu dia-a-dia de investigador? (Sim; Não; Sim, mas com menos tempo gasto; Outro)

6 – Considera a descrição de dados um processo importante para o seu trabalho? Se sim, porquê?

7 – Acha que todos os descritores estão adequados a sua descrição? (Sim; Não; Outra)

8 – Os conceitos de vocabulários controlados que utilizou causaram dúvidas? Se sim, indique quais.

9 – De 1 a 5, qual o seu nível de satisfação da plataforma Dendro? (1 – muito insatisfeito; 5 – muito satisfeito).

Resultados de inquérito: contactar Yulia Karimova

Anexo 7 – Respostas de inquérito após a experiência de descrição de dados no Dendro após de implementação de vocabulários controlados (Produção de Hidrogénio)

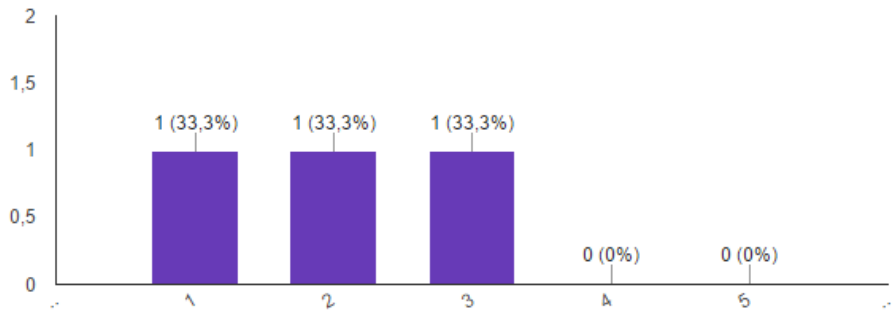
O que achou da experiência em geral? (3 respostas)

Interessante e produtiva
Boa
Boa.

Sentiu algumas dificuldades durante a descrição? Se sim quais? (3 respostas)

Não
Não senti dificuldades
Talvez no início; na escolha dos descritores mais adequados para o 'depósito'/registo dos metadados.

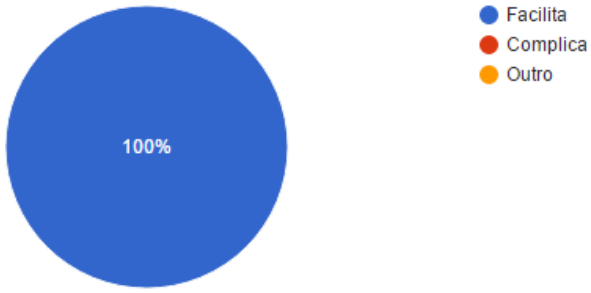
De 1 a 5 qual foi o grau de esforço necessário para descrever os dados?
(3 respostas)



Acha que o processo de descrição de dados exigiu muito tempo? Se sim, como podemos diminuir o tempo gasto?
(3 respostas)

Não
Acho que não mas ainda poderão diminuir mais o tempo dando opção das unidades com vocabulários controlados.
Algum tempo ...

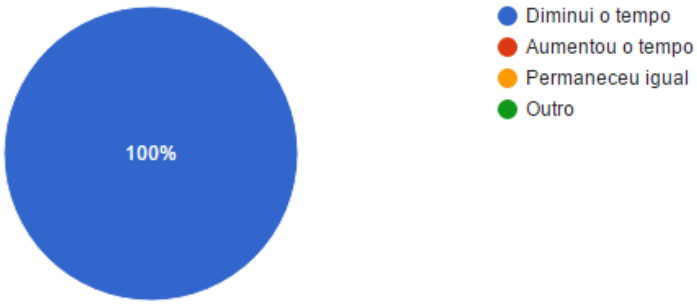
Acha que utilização de vocabulários controlados na descrição dos dados de investigação facilita ou complica o processo?
(3 respostas)



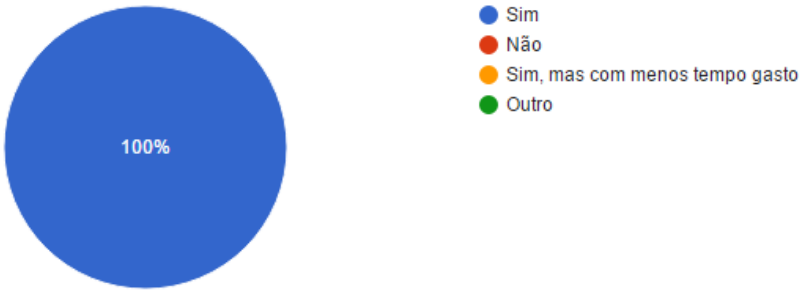
Preferia de criar os metadados com ou sem utilização de vocabulários controlados? Se responder sem vocabulários controlados, indique por favor o motivo.
(3 respostas)

Com vocabulários controlados
Com.
Com e sem vocabulários controlados.

Acha que o uso de vocabulários controlados diminui o tempo gasto para a descrição?
(3 respostas)



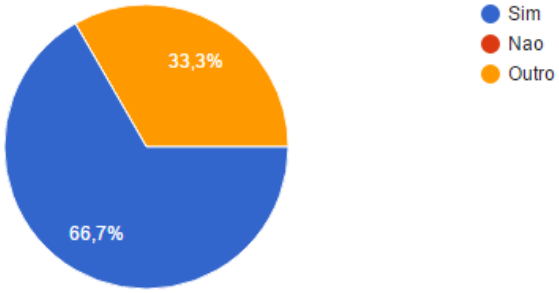
Gostaria de adicionar a descrição de dados de investigação na plataforma Dendro para o seu dia-a-dia de investigador?
(3 respostas)



Considera a descrição dos dados um processo importante para o seu trabalho? Se sim, porquê?
(3 respostas)

Sim. Facilita a partilha (interna e externa) de informação e trabalho realizado
Sim para ter tudo bem organizado.
Sim, porque facilita a pesquisa/acesso e promove a partilha dos mesmos.

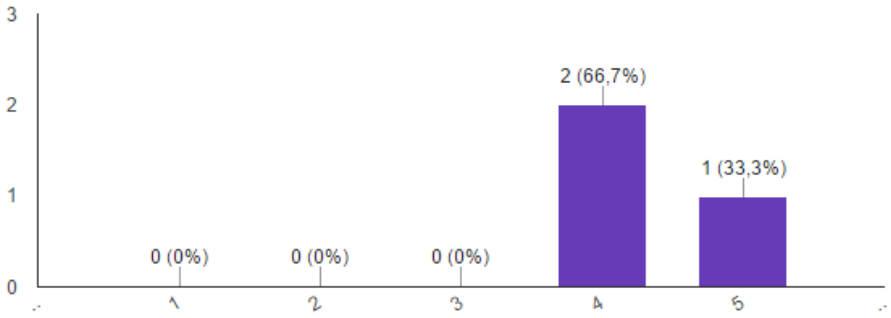
Acha que todos os descritores estão adequados a sua descrição? (3 respostas)



Os conceitos de vocabulários controlados que utilizou causaram dúvidas? Se sim, indique quais.
(3 respostas)

Não
Não
Não.

De 1 a 5, qual o seu nível de satisfação da plataforma Dendro? (3 respostas)



Anexo 8 - Os dados da experiência da descrição de dados de investigação, efetuados pelo Utilizador 1,2,3 na plataforma Dendro após implementação de vocabulários controlados

Utilizador 1	Nome de ficheiro com dados de investigação	Descritores utilizados	Valor	Utilizador 2	Nome de ficheiro com dados de investigação	Descritores utilizados	Valor
	xxxx_Classic_EggR_NiRu(IV)_50x_10_7_0,25SDS_Single_03jul2015.xlsm	Gravimetric Capacity	1.7 wt.%		xxx2_Classic_EggR_NiRu(IV)_50x_10_7_1CMC_Single_13jul2015.xlsm	Gravimetric Capacity	2 wt%
		Number of Reutilization	50			Number of Reutilization	50
		Temperature	27 ºC			Hydration Factor	15
		Hydration Factor	16			Additive	CMC
		Additive	SDS			Hydrolysis	Classic hydrolysis
		Hydrolysis	Classic hydrolysis			Reactor Type	EggR - ovoid mini reactor
		Reactor Type	EggR - ovoid mini reactor			Catalyst	Ni-Ru
		Catalyst	Ni-Ru			Hydrogen Generation Rate	0.2 L/min.gcat
		Hydrogen Generation Rate	0.8 L/min.gcat			Reagent	NaBH4
		Reagent	NaBH4			Description	Foi realizada uma experiência usando um catalisador de NiRu
		Title	Classic_EggR_NiRu(IV)_50x_10_7_0,25SDS_Single_03jul2015			Title	Experiência com mini reactor ovoide
		Access Rights	CEFT_Energy group - H2			Creator	Da*** Fa***
		Contributor	Hé*** Xa*** Nu***			Date	06/21/2016
		Coverage	UP-FEUP-DEQ			**Descritores utilizados**	
		Creator	CEFT_Energy group - H2		xxxx_Classic_EggR_NiRu(IV)_49x_10_30_0_8Feed_29jun2015.xlsm	Gravimetric Capacity	1.8 wt%
		Date	07/03/2015			Number of Reutilization	49
		Format	Excel			Hydration Factor	11
		Language	English			Hydrolysis	Classic hydrolysis
						Reactor Type	EggR - ovoid mini reactor
		Descritores utilizados				Catalyst	Ni-Ru
	xxxx_Classic_EggR_NiRu(IV)_53x_10_7_0_3Feed_18mai2016.xlsm	Gravimetric Capacity	1.7 wt%			Hydrogen Generation Rate	0.1 L/mi.gcat
		Number of Reutilization	53			Reagent	NaBH4
		Temperature	25 ºC			Description	Foi realizada uma experiência com catalisador de NiRu
		Hydration Factor	16			Title	Experiência com mini reactor ovoide
		Hydrolysis	Classic hydrolysis			Creator	Da*** Fa***
		Reactor Type	EggR - ovoid mini reactor			Date	06/21/2016
		Catalyst	Ni-Ru				
		Hydrogen Generation Rate	0.2 L/min.gcat				
		Reagent	NaBH4				
		Title	Classic_EggR_NiRu(IV)_53x_10_7_0_3Feed_18mai2016				
		Contributor	He*** Xa*** Nu***				
		Coverage	Faculdade de Engenharia Universidade do Porto				
		Creator	CEFT- Energy Group -H2				
		Date	05/18/2016				
		Format	Excel				
		Language	English				
		Subject	NaBH4 classic hydrolysis_suc loads				

*** - omitido por revelar a identidade do utilizador

Utilizador 3	Nome de ficheiro com dados de investigação	Descritores utilizados	Valor
		Gravimetric Capacity	1.8 wt.%
		Number of Reutilization	50
	xxxx_Classic_EggR_NiRu(IV)_50x_10_7_0,25CMC_1Feed_08jul2015.xlsm	Temperature	28 ºC
		Hydration Factor	16
		Additive	CMC
		Hydrolysis	Classic hydrolysis
		Reactor Type	EggR - ovoid mini reactor
		Catalyst	Ni-Ru
		Hydrogen Generation Rate	0.8 L/min.gcat
		Reagent	NaBH4
		Title	Classic_EggR_NiRu(IV)_50x_10_7_0,25CMC_1Feed_08jul2015
		Access Rights	CEFT - Energy group
		Coverage	Faculdade Engenharia Universidade do Porto - DEQ
		Creator	H. X. Nu***
		Date	07/08/2015
		Format	Excel
		Language	English
		Subject	NaBH4 classic hydrolysis; Successive loads
		Descritores utilizados	
	xxxx_Classic_SRc_NiRu(IV)_52x_10_7_0_Single_08jan2016.xlsm	Gravimetric Capacity	2 wt.%
		Number of Reutilization	52
		Hydration Factor	16
		Hydrolysis	Classic hydrolysis
		Reactor Type	SRc - conical small reactor
		Catalyst	Ni-Ru
		Hydrogen Generation Rate	0.1 L/min.gcat
		Reagent	NaBH4
		Title	Classic_SRc_NiRu(IV)_52x_10_7_0_Single_08jan2016
		Coverage	Faculdade Engenharia da Universidade do Porto - DEQ
		Creator	H. X. Nu***
		Date	01/08/2016
		Format	Excel
		Language	English
		Subject	NaBH4 classic hydrolysis

Anexo 9 – Inquérito de avaliação de compreensão da descrição existente por grupo de investigadores, ligados ao domínio Produção de Hidrogénio até implementação de vocabulários controlados

Data: 01.06.2016

Local: Laboratório IE206

Participantes: Utilizador 4

Resumo

1 – Objetivos de inquérito

O presente inquérito serve para obter o valor de U_{ra} – percentagem de compreensão da descrição existente no Dendro por utilizadores de grupo Produção de Hidrogénio, utilizado na fórmula de cálculo de *Conformance to expectations* (Tabela 12). Utilizador tem de ver os dados demonstrados e indicar se consegue os interpretar ou não.

2 – Respostas que podem ser escolhidos

Sim – 100%

Não – 0%

Outro – 50%

3 - Lista de perguntas para utilizador

1. Consegue entender que tipo de reator é utilizado na experiência:

   Egg Reactor - ?
   Se responder Sim, escreve a descrição completa
   _____

   Ovoid -?
   Se responder Sim, escreve a descrição completa
   _____

   RG - ?
   Se responder Sim, escreve a descrição completa
   _____

   RM - ?
   Se responder Sim, escreve a descrição completa
   _____

Conical Small Reactor - ?

Se responder Sim, escreve a descrição completa

_____

2. Consegue entender que tipo de hidrólise é utilizado na experiência:

Alkali - ?

Se responder Sim, escreve a descrição completa

_____

Classic - ?

Se responder Sim, escreve a descrição completa

_____

Classic hydrolysis - ?

Se responder Sim, escreve a descrição completa

_____

3. Consegue entender que tipo de catalizador é utilizado na experiência:

NiRu - ?

Se responder Sim, escreve a descrição completa

_____

Nickel – ruthenium - ?

Se responder Sim, escreve a descrição completa

_____

4. Consegue entender que tipo de reagente é utilizado na experiência:

NaBH4 - ?

Se responder Sim, escreve a descrição completa

_____

Sodium Borohydride – ?

Se responder Sim, escreve a descrição completo

_____

4- Resultados

	Percentagem de compreensão da descrição existente no Dendro (até implementaçao de vocabularios controlados)		
Ura	**Os registos de Utilizadores 1-3**	**Descrição completa proposta por Utilizador 4**	0-100%
Reactor Type	Egg Reactor	Reactor com forma de ovo	100%
	ovoid	-	50%
	RG	-	0%
	RM	-	0%
	Conical Small Reactor	Reactor com forma cónica	100%
Hydrolysis	alkali	Hidrólise em meio alcalino	100%
	classic	-	50%
	Classic hydrolysis	Hidrólise classica	100%
Catalyst	NiRu	Catalisador com base de Niquel-Ruténio	100%
	Nickel - ruthenium	Catalisador com base de Niquel-Ruténio	100%
Reagent	NaBH4	Tetra borohidreto de sódio (ou borohidreto de sódio)	100%
	Sodium Borohydride	Borohidreto de sódio	100%

5 – Comentários adicionais

Após análise de resultados deste inquérito mais uma vez provou-se que a descrição completa, proposta pelo Utilizador 4 é diferente da descrição, elaborada e aprovada pelo grupo de utilizadores do domínio Produção de Hidrogénio para vocabulários controlados.

Resultados de inquérito: contactar Yulia Karimova

Anexo 10 - Inquérito de avaliação de compreensão da descrição após a implementação de vocabulários controlados por grupo de investigadores, ligados ao domínio Produção de Hidrogénio

Data: 20.06.2016

Local: Laboratório IE206

Participantes: Utilizador 4

Resumo

1 – Objetivos de inquérito

O presente inquérito serve para obter o valor de U_{ra} – percentagem de compreensão da descrição realizado após a implementação de vocabulários controlados no Dendro. Utilizador tem de ver os dados demonstrados e indicar se consegue os interpretar ou não.

2 – Respostas que podem ser escolhidos

Sim – 100%

Não – 0%

Outro – 50%

3 - Lista de perguntas para utilizador

1. Consegue entender que tipo de reator é utilizado na experiência:

EggR- ovoid mini reactor - ?
Se responder Sim, escreve a descrição completa
_____

LR – large reactor -?
Se responder Sim, escreve a descrição completa
_____

MR_c – conical médium reactor - ?
Se responder Sim, escreve a descrição completa
_____

MR_f – flat médium reacto - ?
Se responder Sim, escreve a descrição completa
_____

SR_c – conical small reactor - ?

Se responder Sim, escreve a descrição completa

_____

SR_f – flat small reactor - ?

Se responder Sim, escreve a descrição completa

_____

2. Consegue entender que tipo de hidrólise é utilizado na experiência:

Acid hydrolysis- ?

Se responder Sim, escreve a descrição completa

_____

Alkali-free hydrolysis - ?

Se responder Sim, escreve a descrição completa

_____

Classic hydrolysis - ?

Se responder Sim, escreve a descrição completa

_____

3. Consegue entender que tipo de catalizador é utilizado na experiência:

Ni-Ru - ?

Se responder Sim, escreve a descrição completa

_____

Pt/C - ?

Se responder Sim, escreve a descrição completa

_____

Co-B - ?

Se responder Sim, escreve a descrição completa

_____

Co-Mn-B - ?

Se responder Sim, escreve a descrição completa

_____

Co-B/Ni - ?

Se responder Sim, escreve a descrição completa

_____

4. Consegue entender que tipo de reagente é utilizado na experiência:

$NaBH_4$ - ?

Se responder Sim, escreve a descrição completa

_____

NH_3BH_3 – ?

Se responder Sim, escreve a descrição completo

_____

$LiAlH_4$ – ?

Se responder Sim, escreve a descrição completo

_____

$LiBH_4$ – ?

Se responder Sim, escreve a descrição completo

_____

KBH_4 – ?

Se responder Sim, escreve a descrição completo

_____

5. Consegue entender que tipo de aditivo é utilizado na experiência:

SDS - ?

Se responder Sim, escreve a descrição completa

_____

CMC - ?

Se responder Sim, escreve a descrição completa

_____

4- Resultados

Ura	Percentagem de compreensão da descrição existente no Dendro (após implementaçao de vocabularios controlados)		
	Os conceitos de vocabularios controlados	**Descrição completa proposta por Utilizador 4**	0-100%
Reactor Type	EggR – ovoid mini reactor	mini reactor com formato de ovo	100%
	LR – large reactor	reactor grande	100%
	MR_c – conical medium reactor	reator médico com formato cónico	100%
	MR_f – flat medium reactor	reactor médio com formato plano	100%
	SR_c – conical small reactor	reactor pequeno com formato cónico	100%
	SR_f – flat small reactor	reactor pequendo com formato plano	100%
Hydrolysis	Acid hydrolysis	hidrólise em meio ácido	100%
	Alkali-free hydrolysis	hidrólise com ausência de inibidor alcalino	100%
	Classic hydrolysis	hidrólise clássica	100%
Catalyst	Ni-Ru	catalisador à base de Níquel-Ruténio	100%
	Pt/C	catalisado à base de Platina suportada em carvão	100%
	Co-B	catalisador à base de Cobalto-Boro	100%
	Co-Mn-B	catalisador à base de Cobalto-Manganês-Boro	100%
	Co-B/Ni	catalisador à base de Cobalto-Boro/Níquel	100%
Reagent	$NaBH_4$	Borohidreto de sódio	100%
	NH_3BH_3		0%
	$LiAlH_4$		0%
	$LiBH_4$		0%
	KBH_4	Borohidreto de potássio	100%
Additive	SDS		0%
	CMC		0%

5 - Comentários adicionais

A análise dos resultados deste inquérito mais uma vez mostra que a descrição sugerida pelo Utilizador 4 e introduzida sem uso de vocabulários controlados contém erros ortográficos, tais como:

reator médico com formato cónico

catalisado à base de Platina suportada em carvão

Figura 21 - Os erros ortográficos, causados or introdução manual, sem auxílio de vocabulários controlados.

Resultados de inquérito: contactar Yulia Karimova