cadernos BAD

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

Social Dendro: Aplicação de técnicas das redes sociais à gestão de dados de investigação

João Rocha da Silva
Faculdade de Engenharia da Universidade do Porto / INESC – TEC
joaorosilva@gmail.com

Nelson Pereira
Faculdade de Engenharia da Universidade do Porto
ei09025@fe.up.pt

Resumo

A gestão de dados de investigação é atualmente um grande desafio para as instituições de investigação. Os pequenos grupos de investigação, em particular, necessitam de apoio na gestão dos seus dados para serem capazes de corresponder aos requisitos de dados abertos que começam a ser impostos pelas entidades financiadoras. Dado que o trabalho de investigação é tipicamente desenvolvido em ambiente colaborativo, é importante que seja suportado por ferramentas desenhadas para suportar essa colaboração. É neste sentido que tem vindo a ser desenvolvida a plataforma Dendro para gestão de dados de investigação, uma ferramenta open-source, fácil de instalar e usar por pequenos grupos de investigação, que pretende facilitar o armazenamento e descrição dos dados de investigação em preparação para o seu depósito em praticamente qualquer repositório final. Atualmente está a ser desenvolvida uma extensão à plataforma Dendro, que pretende incorporar diversos conceitos das redes sociais na plataforma (incluindo gostos, comentários e partilhas), com o objetivo de tornar a descrição de dados mais dinâmica e fácil de realizar dentro do grupo de investigação. Após o desenho de esboços de interações, o Social Dendro está neste momento em fase de implementação e testes junto de alguns grupos de investigação. Nesta comunicação será apresentado o modelo de dados da plataforma, alguns exemplos de interação e o estado atual de desenvolvimento da

aplicação; terminaremos com algumas breves conclusões e algumas perspetivas de trabalho futuro.

Palavras-chave: Gestão de dados de investigação; repositórios de dados; redes sociais; interação; interfaces com o utilizador

Social Dendro: Applying social network techniques to research data management

Abstract

Research data management is currently a large challenge for research institutions. Small research groups, in particular, need support in the management of their data in order to comply with the open data requirements that start to be imposed by funding institutions. Since research activities are typically carried out in a collaborative environment, it is important that it is supported by tools designed to support that collaboration. It is with this in mind that we have been developing the Dendro research data management platform. It is open-source, easy to install and use even by small research groups, which facilitates the storage and description of research data in preparation for their deposit in practically any repository. We are presently developing an extension that intends to incorporate several concepts from social networks into the platform (including likes, comments and shares) with the goal of making data description more dynamic and easy to carry out within the research group. After the initial outline of the interactions with the extension, Social Dendro is currently in development and testing stage with several research groups. In this paper, we will present the platform's data model, some interaction examples and its current state of development. We will finish with some conclusions and perspectives of future work.

Keywords: Research data management; data repositories; social networks; interaction; user interfaces

Introdução

A gestão de dados de investigação é cada vez mais uma preocupação dos investigadores, talvez devido à presença de cada vez mais recomendações (COMMISSION, 2013) ou até mesmo imposições (FOUNDATION, 2011) por parte das entidades financiadoras no sentido da disponibilização dos conjuntos de dados que servem e suporte às publicações resultantes dos projetos por elas financiados. Muito devido a estas novas políticas, tem sido possível a assistir a uma mudança na atitude dos investigadores relativamente à gestão dos seus dados de investigação. Em particular, a partilha de dados tem aumentado em grande medida (TENOPIR et al., 2011, TENOPIR et al., 2015).

Apesar destes sinais muito positivos, a descrição e partilha de dados de investigação ainda está longe de estar perfeitamente enraizada nos processos de investigação. Diversos estudos indicam, contudo, que as melhorias devem passar não só pelas ferramentas de

gestão e partilha como também pelo melhor reconhecimento do trabalho dos investigadores que partilham dados (TENOPIR et al., 2015, WALLISROLANDO e BORGMAN, 2013). O problema da gestão de dados é um tema muito complexo, que inclui tanto questões técnicas como políticas, sociais e de propriedade intelectual (BORGMAN, 2012).

Nas instituições de investigação onde existem recursos para tal, a gestão dos dados de investigação pode ser suportada por especialistas denominados curadores de dados (PENG et al., 2016); em outros casos, os investigadores acabam por realizar eles próprios a descrição e depósito dos seus conjuntos de dados. Para obter os melhores resultados, contudo, tanto curadores com investigadores devem participar no processo de descrição e depósito, pois desempenham funções complementares. Os curadores são peritos na área de gestão de dados, mas não propriamente nas áreas científicas de que provêm esses mesmos dados; por outro lado, os investigadores têm profundo conhecimento das áreas científicas em questão, mas não sabem necessariamente como formalizar esses metadados de acordo com as melhores práticas de gestão de dados.

Para que seja possível criar metadados de elevada qualidade, capazes de fomentar a descoberta e reutilização dos conjuntos de dados é, portanto, necessário construir *workflows* que integrem o trabalho dos curadores e dos investigadores, desde o momento da produção dos dados. É portanto necessário motivar a participação dos investigadores através da oferta de soluções que lhes proporcionem vantagens imediatas no seu dia-a-dia, que podem passar pelo simples armazenamento seguro dos dados e sua descoberta mais fácil dentro do grupo de investigação (POSCHEN et al., 2012). Outras vantagens como o indiscutível aumento das citações originada pela partilha de dados base (PIWOWAR e VISION, 2013) são muito importantes mas ainda pouco palpáveis para os investigadores—especialmente quando consideramos que a produção de registos de metadados para partilha com elementos externos implica um investimento muito maior na sua descrição (BORGMAN, 2012).

Como já demonstrado em projetos como o ADMIRAL (HODSON, 2011), o DataStaR (STEINHART, 2011) ou o MaDAM (POSCHEN et al., 2012) a gestão de dados deverá estar presente em todo o processo de investigação, focando-se na organização dos dados desde a sua criação até à sua disseminação, passando pelo seu armazenamento e publicação. A fase de publicação representa o passo final do trabalho dos investigadores, onde os dados produzidos são depositados em repositórios reconhecidos pela comunidade científica, tornando possível a revisão por pares.

O processo de investigação implica a comunicação entre os elementos do grupo, assim como a partilha de ficheiros entre eles. Esta colaboração é tipicamente levada a cabo por e-mail ou conversa online, enquanto a partilha de ficheiros é feita com recurso, por exemplo, a unidades de armazenamento USB, partilha na rede ou soluções de armazenamento na nuvem, como é caso da Dropbox. Apesar de serem rápidos e simples de usar, estes processos podem fazer com que os dados fiquem dispersos e separados das

eventuais notas relativas ao seu contexto de produção. Por outras palavras, os dados armazenados ficam separados dos seus metadados, que acabam por se perder em cadeias de e-mail trocados entre elementos dos grupos de investigação ou outros contatos informais. Ao analisar o tempo gasto na escrita de todas essas mensagens e emails conclui-se que o trabalho de descrição acaba por ser feito, mas não de forma centralizada, sistemática ou interoperável.

O Dendro[i] é uma plataforma de gestão de dados de investigação atualmente em desenvolvimento, que permite aos grupos de investigação armazenar, descrever e partilhar os seus dados de investigação nos principais repositórios (AUTOR, 2014). A plataforma permite aos utilizadores criar projetos, que são semelhantes às pastas partilhadas na Dropbox. Dentro de cada projeto poderão ser depositados ficheiros e criadas pastas. Contrariamente à Dropbox, contudo, há também um forte foco na produção de metadados de qualidade. Para tal, a plataforma usa ontologias genéricas e específicas dos domínios de investigação no seu modelo de dados (AUTOR, 2014). A plataforma também inclui um sistema que ajuda os investigadores a escolher os descritores mais adequados para os seus ficheiros e pastas, tendo por base as suas interações com o mesmo (descritores mais usados no projeto, por exemplo) (AUTOR, 2016).

Para tentar direcionar os esforços dos investigadores para a adequada descrição dos seus dados e incorporá-la nas suas atividades diárias, começaram recentemente a ser propostas algumas abordagens colaborativas assentes em paradigmas das redes sociais. O conceito "Science 2.0 Repositories" traz para o processo de descrição e partilha de dados de investigação algumas metáforas de interação normalmente presentes nas redes sociais. Estas incluem, por exemplo os "gostos", comentários e partilhas, mas associados a eventos como a criação ou descrição de conjuntos de dados de investigação (ASSANTE et al., 2015). Esta visão de repositório colaborativo e social alinha-se com um importante foco de desenvolvimento do Dendro: a produção de metadados flexíveis e com histórico completo de edição (AUTOR, 2014). Dado que esta informação de auditoria sobre os metadados é parte integral do Dendro, uma extensão social deste tipo pode tirar grande partido da estrutura e modelo de dados já existente.

A extensão Social Dendro

A extensão Social Dendro tem como objetivo trazer para a plataforma Dendro alguns conceitos das redes sociais para melhor suportar a gestão de dados de investigação. A extensão social irá atuar sobre as alterações aos metadados de ficheiros (edição, adição e remoção) bem como a adição, edição e remoção de ficheiros e pastas de um projeto. A cada alteração realizada no contexto de um projeto, será gerado um Post com a informação descrevendo a alteração em questão, e identificando o autor da mesma. Os restantes contribuidores do projeto terão a possibilidade de fornecer feedback sobre cada Post, nomeadamente sob a forma de Gostos, Comentários e Partilhas.

Os Posts são apresentados numa linha temporal, permitindo aos intervenientes no projecto ter uma visão clara do decorrer das atividades de investigação. A possibilidade de comentar esses Posts ou fazer "Gosto" poderá ajudar também a determinar quais os projetos mais ativos, e ao mesmo tempo passar de certa forma pelas etapas da fase de publicação em cada alteração ao projeto, já que o uso dos "Gostos" e Comentários podem ser vistos como uma forma informal de revisão por pares. As Partilhas são também uma forma de disseminação de dados dentro do grupo de investigação. Também se pode assumir como uma ferramenta de auditoria, ajudando a manter registos de autoria de trabalhos e colaborações, complementando o papel do e-mail como ferramenta de documentação das interações dentro do grupo de investigação.

Funcionalidades principais da componente social

As principais funcionalidades implementadas relativamente à componente social do Dendro assentam sobre princípios introduzidos pelas redes sociais, bem como ideias apresentadas no conceito Science 2.0 Repositories. Por outro lado, dado que o modelo de dados do Dendro é totalmente construído sobre ontologias, foi necessário modelá-lo com recurso a estas especificações. O reaproveitamento de conceitos e ontologias já definidas é ponto essencial para a sua correta modelação, pois só assim é possível assegurar a interoperabilidade entre os modelos de dados de diferentes sistemas que caracteriza a web semântica.

Na Figura 1 está representado o modelo de dados da extensão e as suas relações com conceitos do esquema schema.org, que resulta de uma colaboração entre as principais empresas de motor de busca (Google, Bing, Yandex e Yahoo!). Após a sua análise, foram reaproveitadas as classes CommentAction, LikeAction e ShareAction para a definição dos comentários, gostos e partilhas presentes no Social Dendro. O significado da classe SocialMediaPosting está alinhado com a nossa definição para os posts registados no Social Dendro e apresentados na timeline social.

Neste diagrama, a classe MetadataChange representa uma alteração aos metadados de uma pasta ou ficheiro dentro de um projeto. Uma MetadataChange possui um tipo (adição, edição ou remoção de uma instância de um descritor)
e o valor anterior e o novo valor caso existam. Está também associada a um Descriptor: no projeto podem haver diversas modificações feitas ao descritor "Descrição" (Dublin Core) ao longo do tempo, por exemplo-cada uma será uma instância de MetadataChange, associada a esse Descriptor.

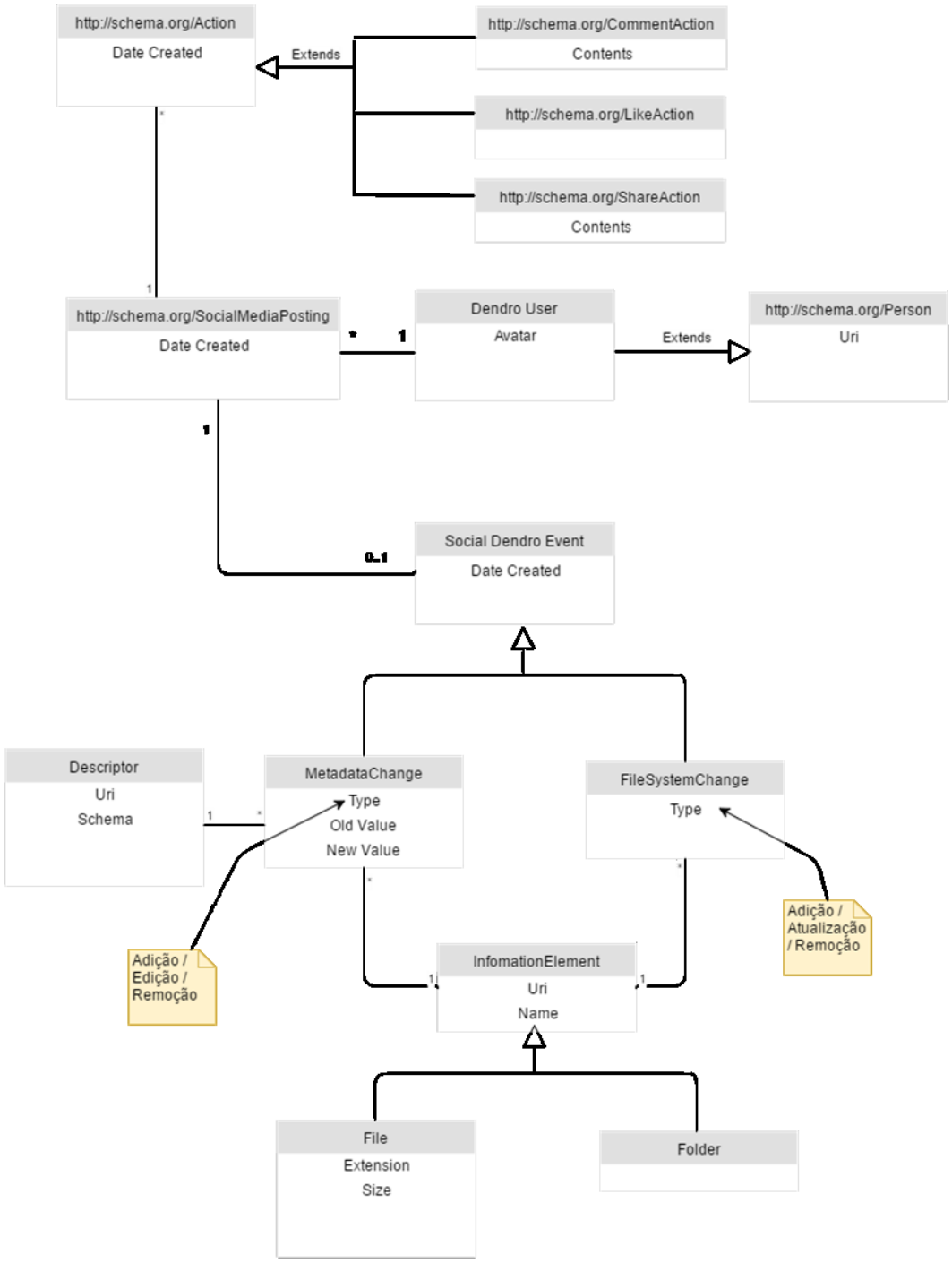

FIGURA 1 – MODELO DE DADOS

Outro exemplo de reaproveitamento de conceitos pré-existentes neste modelo de dados é a representação da hierarquia de pastas e ficheiros, que reutiliza classes definidas na Nepomuk File Ontology[ii]. Desta forma, os ficheiros e pastas de um projeto são representados pela classe InformationElement, as pastas pela classe Folder e os ficheiros pela classe File definidas nessa ontologia. Quando alguma alteração ocorre ao nível do sistema de ficheiros (criação, atualização ou remoção) é registada uma FileSystemChange e

gerado um SocialDendroEvent, que representa todos os eventos de manipulação quer de metadados quer de ficheiros. O sistema da timeline regista também a informação do que aconteceu nesse evento na forma de uma instância de SocialMediaPosting. Todos os utilizadores associados ao projeto da qual foi originada a alteração ao ficheiro ou pasta têm a possibilidade de realizar ações sociais sobre esse post, e o mesmo aparecerá na sua timeline.

Interface com o utilizador

A extensão foi desenvolvida como uma timeline que apresenta, sobre a forma de posts, informação proveniente das alterações aos conjuntos de dados que foram sujeitos ao processo de curadoria. Como se mostra na Figura 2, o utilizador tem primeiro que aceder à sua área de projetos (1,2 na interface A), sendo-lhe apresentada a interface B. Seguidamente, poderá visualizar a timeline representativa das alterações aos seus conjuntos de dados (4) ou então aceder a uma listagem convencional (3).

A interface gráfica da extensão social foi desenvolvida de forma a ser o mais simples possível, a par com a orientação visual introduzida por aplicações de redes sociais bastante conhecidas. O objetivo principal foi o de tornar o processo de aprendizagem da extensão o mais rápido possível.

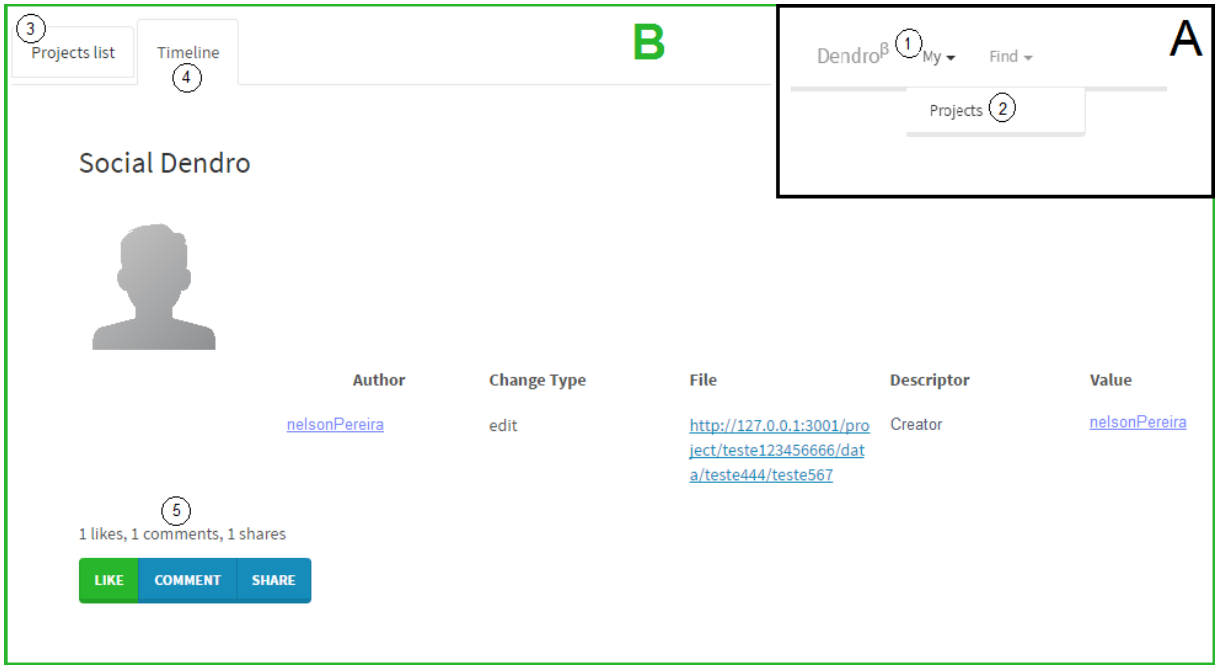

FIGURA 2 – ACESSO À EXTENSÃO SOCIAL DENDRO

Selecionando o separador "Timeline" (4) é possível observar todas as atualizações recentes relativas aos projetos de um utilizador, por ordem cronológica. Cada alteração gera um post respetivo, contendo informação sobre o autor da mesma, o tipo de alteração, bem como o ficheiro, descritor de metadados alterado e o valor do mesmo.

A Figura 3 representa as três possíveis interações a realizar perante um post numa timeline do Social Dendro. É possível interagir com cada post da timeline realizando um comentário (2), um gosto (1), ou uma partilha sobre o mesmo (3), fornecendo feedback ao autor perante as alterações introduzidas ao conjunto de dados de um projeto.

Estado atual do desenvolvimento

Presentemente, na fase de protótipo, já são criados *posts* a partir das alterações aos conjuntos de dados, estando também implementada toda a componente da interação social com os mesmos (*gostos*, *comentários* e *partilhas*). O foco nas fases seguintes de implementação estará no processo de otimização bem como a realização de testes à usabilidade de forma a avaliar todo o processo de interação. O código fonte encontra-se disponível em regime open-source em [[Ligação github omitida para processo de revisão]].

Conclusões e trabalho futuro

No estado atual de desenvolvimento, as principais funcionalidades do Social Dendro estão já implementadas. O resultado final encontra-se alinhado com os objetivos iniciais introduzidos no conceito de Science 2.0 Repository.

Serão efetuados testes de usabilidade com dois grupos de utilizadores: o primeiro grupo será constituído por utilizadores que nunca usaram o Dendro e um segundo com utilizadores experientes no seu uso. O objetivo será o de comparar a facilidade com que os dois grupos conseguem estar a par das alterações aos seus projetos, e a comunicação entre

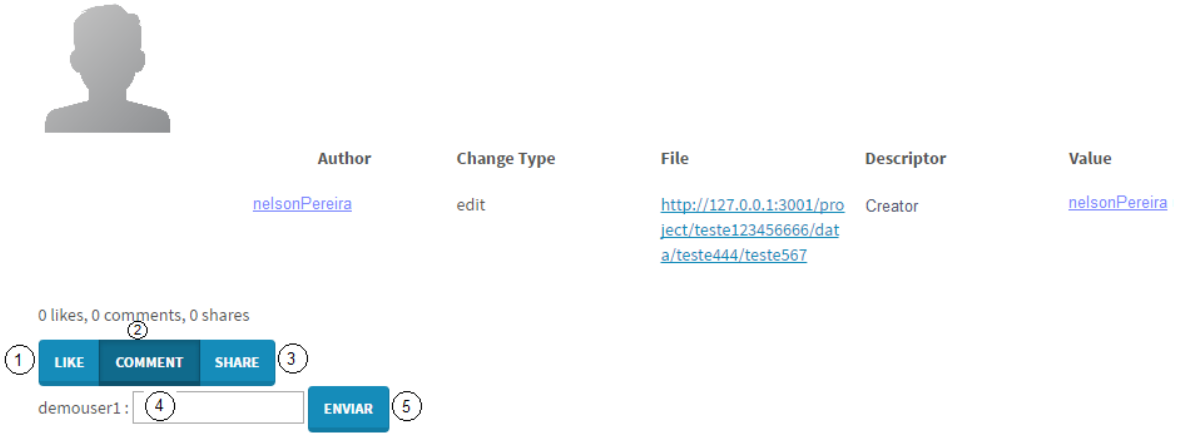

FIGURA 3 – OPÇÕES DE INTERAÇÃO

os diferentes elementos (investigadores) de um projeto, bem como a temporização das diferentes interações com a interface gráfica e observação e contagem do número de erros cometidos pelos utilizadores. Um inquérito com métricas bem estabelecidas de qualidade da experiência do utilizador (MACK e NIELSEN, 1993) será também levado a cabo.

Referências bibliográficas

ASSANTE, Massimiliano [et al.] - Science 2.0 Repositories: Time for a Change in Scholarly Communication - D-Lib Magazine [Em linha]. Vol: 21, nº (2015),

BORGMAN, C.L. - The conundrum of sharing research data - Journal of the American Society for Information Science and Technology [Em linha]. Vol: 63, nº (2012),

COMMISSION, European - Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 - [Em linha]. (2013), p. 1-14.

FOUNDATION, National Science - Grants.Gov Application Guide A Guide for Preparation and Submission of NSF Applications via Grants.gov. 2011. ISBN: ISBN.

HODSON, Simon - ADMIRAL: A Data Management Infrastructure for Research Activities in the Life sciences.In [Em linha]. 2011,

MACK, Robert ; NIELSEN, Jakob - Usability inspection methods - ACM SIGCHI Bulletin [Em linha]. Vol: 25, nº (1993), p. 28-33. ISSN: 0897916514

PENG, Ge [et al.] - Scientific Stewardship in the Open Data and Big Data Era -- Roles and Responsibilities of Stewards and Other Major Product Stakeholders - D-Lib Magazine [Em linha]. Vol: 22, nº (2016),

PIWOWAR, Heather ; VISION, Todd - Data reuse and the open data citation advantage - PeerJ [Em linha]. Vol: 1, nº (2013), p. e175. ISSN: 2167-8359

POSCHEN, Meik [et al.] - Development of a Pilot Data Management Infrastructure for Biomedical Researchers at University of Manchester - Approach, Findings, Challenges and Outlook of the MaDAM Project - International Journal of Digital Curation [Em linha]. Vol: 7, nº (2012), p. 110-122.

STEINHART, Gail - DataStaR: A Data Sharing and Publication Infrastructure to Support Research - Agricultural Information Worldwide [Em linha]. (2011), p. 26-29.

TENOPIR, C. [et al.] - Data sharing by scientists: practices and perceptions - PLoS One [Em linha]. Vol: 6, nº 6 (2011), p. e21101. Disponível em WWW: <https://www.ncbi.nlm.nih.gov/pubmed/21738610>. ISSN: 1932-6203 (Electronic)

TENOPIR, C. [et al.] - Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide - PLoS One [Em linha]. Vol: 10, nº 8 (2015), p. e0134826. Disponível em WWW: <https://www.ncbi.nlm.nih.gov/pubmed/26308551>. ISSN: 1932-6203 (Electronic)

WALLIS, Jillian C. ; ROLANDO, Elizabeth ; BORGMAN, Christine L. – If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology – PLoS ONE [Em linha]. (2013), ISSN: 19326203 (Electronic)

[i] Disponível em código aberto em: [[link para repositório GitHub omitido para revisão]]

[ii] Mais informações em http://www.semanticdesktop.org/ontologies/2007/03/22/nfo/