cadernos BAD

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

Projeto TAIL—Gestão de dados de investigação da produção ao depósito e à partilha (resultados preliminares)

Cristina Ribeiro
INESC TEC—Faculdade de Engenharia da Universidade do Porto
mcr@fe.up.pt

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

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

Ricardo Carvalho Amorim
INESC TEC—Faculdade de Engenharia da Universidade do Porto
ricardo.amorim3@gmail.com

João Correia Lopes
INESC TEC—Faculdade de Engenharia da Universidade do Porto
jlopes@fe.up.pt

Resumo

A gestão dos dados de investigação preocupa neste momento tanto os investigadores como os responsáveis por gestão de ciência e as agências de financiamento. Os investigadores têm consciência do valor dos dados de investigação, enquanto as agências de financiamento estabelecem mandatos para planos de curadoria e partilha de dados como parte dos seus regulamentos. Os responsáveis por políticas de ciência querem também garantir que os resultados obtidos com os seus planos de investimento têm o maior impacto possível. O projeto TAIL, a decorrer de 2016 a 2019, vai construir um portfólio de exemplos de gestão de dados em diversos

domínios que poderão ser usados pelos investigadores para avaliar o esforço requerido e as compensações a obter com esta atividade. O projeto tem como base o trabalho realizado no estudo dos fluxos de trabalho dos investigadores usando a plataforma Dendro e as interfaces móveis para a recolha de dados e metadados de que é exemplo o LabTablet. Estes resultados preliminares informam os processos a usar na publicação de conjuntos de dados existentes em repositórios nacionais e internacionais, no desenho de modelos de metadados para a descrição pormenorizada dos domínios e no alinhamento com as infraestruturas europeias e nacionais.

Palavras-chave: Gestão de dados de investigação; repositórios de dados; modelos de metadados; infraestruturas de investigação; cadernos de laboratório eletrónicos

The TAIL project: Research data management from creation to deposit and sharing (preliminary results)

Abstract

Research data management is currently a concern for researchers, science managers and funding agencies. Researchers are aware of the value of research data, while funding agencies have data curation and data sharing plans as part of their mandates. Policy agents are committed to obtain the best possible impact for the results from their funding. The TAIL project, running from 2016 to 2019, will build a portfolio of data management cases in several domains, to be used by researchers to assess the effort required by this activity and the rewards it may bring. The project is based on the work in research data management workflows using the Dendro platform and mobile platforms for data and metadata collection such as the LabTablet. These preliminary results anticipate the processes to use in the publication of existing datasets in national and international repositories, in the design of metadata models for detailed domain-specific description, and in the alignment with national and European research infrastructures.

Keywords: Research data management, data repositories, metadata models, research infrastructures; electronic laboratory notebooks

Introdução

A gestão dos dados de investigação (designação internacional: Research Data Management, ou RDM) preocupa neste momento tanto os investigadores como as agências de financiamento e os responsáveis por políticas de ciência. Do lado dos investigadores, há a consciência do valor dos dados criados ou reunidos em contexto de investigação, especialmente quando começa a ser notado que os artigos com dados associados têm mais citações e que a divulgação de dados interessantes atrai a colaboração. As agências de financiamento estão a estabelecer os mandatos para planos de curadoria e partilha de dados como parte dos seus regulamentos e, tanto a nível nacional como internacional, os

responsáveis por políticas de ciência querem garantir que os resultados obtidos com os seus planos de investimento têm o maior impacto possível.

A gestão dos dados ainda não está estabelecida como parte integrante dos projetos de investigação. Este processo está mais adiantado em áreas como as ciências da vida, em que projetos como o do genoma humano foram construídos sobre bases de dados dedicadas desde o início. Os grupos pequenos na chamada cauda longa da ciência não têm ainda a tecnologia para gerir os seus dados, mas têm potencial para gerar grandes quantidades de dados valiosos e únicos.

Os resultados em RDM requerem a consideração de questões tecnológicas e conceptuais. Os problemas tecnológicos têm a ver com armazenamento, mas também com as interfaces para organizar, descrever e depositar os dados. Os problemas conceptuais têm a ver com a descrição. Os dados armazenados numa infraestrutura, por muito sólida que seja, serão inúteis sem a organização e descrição que só os criadores podem fornecer e que permitirá a outros especialistas do domínio reutilizá-los.

Durante 3 anos, o projeto TAIL irá fazer a ponte entre os mandatos de dados com que os grupos de investigação se confrontam e as ferramentas, fluxos de trabalho e modelos de descrição disponíveis. No final teremos um conjunto de grupos de investigação que passaram pela experiência de gerir com sucesso os dados que criaram e que também colheram os benefícios de terem os seus dados publicados, pesquisáveis e citados. Estas histórias de sucesso faltam na comunidade e as suas lições são essenciais para o progresso na gestão dos dados de investigação. Ao longo do projeto, haverá ações promovidas em colaboração pelas universidades participantes, para acompanhar o teste das ferramentas adotadas tendo em vista o seu uso em serviços estáveis de suporte aos investigadores.

Nesta apresentação mostraremos os resultados preliminares do projeto TAIL, nomeadamente os resultados já obtidos nas seguintes tarefas:

1. Fluxos de trabalhos dos investigadores usando a plataforma Dendro: ao alargar o número de grupos usando o Dendro e ao observá-los nas suas tarefas habituais, estamos a recolher informação importante para o aperfeiçoamento das tarefas de organização e descrição de dados.

2. Interfaces móveis para a recolha de dados e metadados: o LabTablet é uma aplicação móvel que permite a recolha de metadados de forma expedita e a geração automática de valores para alguns descritores usando sensores disponíveis no dispositivo móvel. Estamos a expandir o âmbito do LabTablet, usando-o também para adquirir dados que são colecionados e agregados no Dendro.

3. Publicação de conjuntos de dados existentes: nos grupos de investigação existem dados cujo valor e potencial para reutilização são reconhecidos. Para estes dados

estamos a desenvolver ações de publicação em repositórios internacionais, tendo em vista a avaliação do seu percurso em termos de pesquisas, descargas e reutilização.

4. Modelos de metadados: a atividade de descrição é tanto mais efetiva quanto mais os metadados criados tenham um bom reconhecimento nas comunidades. O método que propomos, criando ontologias no Dendro, vai ser testado em domínios com normas bem estabelecidas e em outros que não as têm ainda.

5. Alinhamento com as infraestruturas europeias: nos parceiros do TAIL temos grupos diretamente envolvidos nas iniciativas para infraestruturas especializadas por domínios. Os requisitos identificados por esses grupos são uma fonte importante para o projeto.

Estes são resultados de trabalho em curso e a sua evolução nos primeiros meses deste projeto permite uma reflexão preliminar sobre o estado das tecnologias que podem facilitar o avanço da gestão de dados de investigação. Para além disso vai permitir aferir tanto a recetividade dos grupos de investigação às novas tarefas com que são confrontados como a sua perceção sobre os ganhos a obter com a conformidade a uma política de gestão de dados centrada nos investigadores.

Gestão de dados de investigação

Os dados de investigação são criados e usados em contextos diversos. Podem ser gerados especificamente para um projeto de investigação, como dados de sensores captados numa experiência, ou entrevistas para uma análise. Mas também incluem dados recolhidos sistematicamente para algum fim e que também são usados em investigação, como dados meteorológicos ou os registos de acesso a um serviço de computação. Os dados podem ser documentos comuns, como um conjunto de páginas web, colecionadas ad-hoc para avaliar o desempenho de uma ferramenta de pesquisa. Esta diversidade torna a gestão de dados de investigação uma tarefa difícil de definir, para a qual os investigadores não têm processos bem estabelecidos nem uma visão clara sobre a sua utilidade.

Muitos dados produzidos em grandes projetos são curados em infraestruturas disciplinares. O NCBI nas ciências da vida e o ICPSR nas ciências sociais são exemplos de infraestruturas maduras que suportam as tarefas de curadoria de dados de comunidades bem estabelecidas. Estas infraestruturas são usadas pelos investigadores tanto para pesquisar e obter dados como para contribuir com novos resultados de investigação.

Na chamada "cauda longa" da ciência, onde muitos grupos de pequena dimensão desenvolvem uma parte substancial do trabalho da comunidade, não existe ainda uma solução bem estabelecida para a tornar os dados visíveis, descritos de forma satisfatória, depositados e pesquisáveis. Os esforços nesta área estão ainda ao nível de projeto, como ilustrado por duas iniciativas com financiamento europeu, o EUDAT (GENTZSCHLECARPENTIER e WITTENBURG, 2014, LECARPENTIERMICHELINI e WITTENBURG,

2013) e o OpenAIRE (MANGHI et al., 2012). Ambos têm trabalho substancial no desenho de serviços e no envolvimento com grupos da cauda longa, mas as infraestruturas que propõem são ainda ligadas a projetos.

O trabalho que desenvolvemos tem como objetivo promover a publicação de dados em instituições de investigação, partindo de dois pressupostos: 1) os investigadores são os atores principais, pelo que os fluxos de trabalho devem ser simplificados do seu ponto de vista; e 2) as ferramentas multidisciplinares podem ser desenhadas e adaptadas para áreas específicas com um esforço razoável. Começámos com grupos pequenos com dados interessantes e sem muito tempo ou financiamento para curadoria de dados. Nos casos tratados os investigadores reconhecem os benefícios da gestão de dados, envolvem-se no processo de preparação de dados para publicação e obtêm resultados verificáveis da sua investigação. Procuramos o equilíbrio entre o uso de ferramentas genéricas e a satisfação dos requisitos de cada grupo de investigação.

As ferramentas disponíveis para os investigadores são decisivas para a sua motivação para a gestão de dados. É de esperar que as ferramentas que simplificam o trabalho necessário na gestão de dados e produzem resultados claros sejam mais facilmente adotadas (RIBEIRO et al., 2015). Compreendendo isso, demos especial atenção à preparação dos dados, nomeadamente à organização dos conjuntos de dados e à recolha de metadados. O nosso objetivo foi reduzir o atraso entre a captura dos dados e a sua organização e descrição, que é determinante na perda de dados valiosos.

As experiências que realizámos foram apoiadas num conjunto de grupos em diversas áreas cujos investigadores se comprometeram a realizar algumas tarefas de gestão de dados. Os contactos preliminares em cada grupo identificaram a natureza e objetivos dos seus dados, enquanto se apresentaram aos investigadores os conceitos da gestão de dados. O método usado a seguir adotou as recomendações da literatura (MAYERNIK, 2011, TENOPIR et al., 2011). Fez-se a seleção de modelos de dados específicos dos domínios, a identificação de conjuntos de dados relevantes para publicação e a avaliação de ferramentas de gestão de dados no contexto das atividades regulares dos grupos (DAF, 2016).

O trabalho que descrevemos adotou uma estratégia minimalista para a cauda longa da ciência partindo de alguns pressupostos e tendo em vista responder a uma questão. As suposições foram que os investigadores na cauda longa estão geralmente pouco informados sobre gestão de dados e não têm serviços que os apoiem na preparação de dados para depósito e publicação. A questão é se as ferramentas podem ajudar a integrar a gestão de dados no processo de investigação, resultando em mais dados chegarem à fase de depósito.

Para estes grupos na cauda longa, os nossos resultados são uma plataforma *open source* para gestão de dados de investigação (Dendro), um fluxo de trabalho que inclui a análise de conceitos do domínio e a sua organização em ontologias, a criação automática de

metadados com dispositivos móveis e a avaliação das ferramentas e dos fluxos de trabalho com investigadores de vários domínios.

É importante reunir evidência relativamente a esta proposta de fluxos de trabalho e ferramentas genéricos. Para esse efeito estamos a expandir as experiências para um público mais vasto, tendo em conta as políticas e prática nas instituições de investigação e a possibilidade de ligação a repositórios internacionais com capacidade de preservação. Considerando que os conceitos de gestão de dados estão a evoluir, bem como as recomendações em várias disciplinas, temos também em vista a ligação entre as soluções para a cauda longa e as que estão a aparecer em ambientes disciplinares. Soluções testadas em muitos domínios podem ser valiosas em projetos disciplinares, identificando conceitos e requisitos comuns. Inversamente, os requisitos de projetos de curadoria disciplinares podem revelar-se suficientemente genéricos para serem incorporados nas plataformas mais ligeiras desenvolvidas para a cauda longa.

Atores na gestão de dados de investigação

A gestão de dados ainda procura o modelo certo para se embeber nos fluxos de trabalho de investigação. Vários atores estão identificados, mas para alguns deles o caminho a seguir não é claro e as relações com outros precisam de ser consolidadas.

O nosso contexto é o de uma unidade de investigação grande, com um leque grande de disciplinas. O primeiro ator é a própria universidade, que tem de lidar com um grande número de problemas relacionados com dados, nomeadamente direitos, ética, armazenamento e auditoria.

No nível imediatamente a seguir estão os departamentos e grupos de investigação, que partilham as preocupações com os dados e na prática têm de fornecer as soluções operacionais para armazenamento, acesso e auditoria.

Os investigadores vêm a seguir e são os atores principais, com os papéis de criadores e utilizadores. Um investigador pode gerar ou recolher um conjunto de dados, partilhá-lo com outro investigador num projeto colaborativo e espera-se que prossiga para a publicação de dados em repositórios e sua associação a artigos publicados.

Os atores seguintes são os gestores de investigação. O seu trabalho pode ser distanciado dos investigadores se as suas preocupações principais forem as oportunidades de financiamento e a apresentação de resultados de projetos. Cada vez mais, porém, vemos gestores de investigação que são especialistas do domínio a participar na gestão de todos os resultados de investigação, incluindo os dados.

As agências de financiamento são um ator muito influente. Adotando a visão da ciência aberta, em particular para a ciência financiada com fundos públicos, as agências de

financiamento estão cada vez mais a requerer que os resultados de investigação, incluindo os dados, estejam disponíveis publicamente para outros grupos os usarem(EOSC, 2016). Esta atitude tem como consequência tornar os resultados publicados mais facilmente verificáveis e reutilizáveis.

Dado que estamos ainda longe de estar generalizado o acesso a infraestruturas de gestão de dados, os desenvolvedores de aplicações são também atores importantes. Precisam de familiaridade com os conceitos de gestão de dados, eles próprios ainda a evoluir, e criar boas metáforas para a ligação entre pessoas e tecnologia.

Os resultados de investigação tradicionais, como artigos de revista e relatórios de projetos, têm uma prática bem estabelecida de disseminação, envolvendo revistas, editoras e mais recentemente repositórios institucionais, que fornecem visibilidade e possibilidades de referência cruzada. Tudo isto é incipiente na publicação de dados. Os agregadores de dados e os coletores de metadados são também chamados como atores.

A preservação é uma dimensão importante na gestão de dados, pelo que as capacidades e caraterísticas dos arquivos têm de estar de acordo com os requisitos de preservação (CAPLAN, 2009, COUNCIL OF THE CONSULTATIVE COMMITTEE FOR SPACE DATA, 2002). Os arquivos, institucionais ou disciplinares, são por isso atores importantes, estando o seu papel ainda mal estabelecido em muitas áreas. Os arquivos de dados podem ser muito volumosos, pelo que a seleção e a descrição apropriada se tornam essenciais. Estas funções são consumidoras de tempo, mas está provado que se tornam mais comportáveis quando realizadas ao tempo da geração dos dados, e com ferramentas apropriadas, do que mais tarde no processo. A seleção e descrição passam aqui a ser tarefas embebidas no processo de gestão de dados desde o início, em vez de serem tarefas à posteriori.

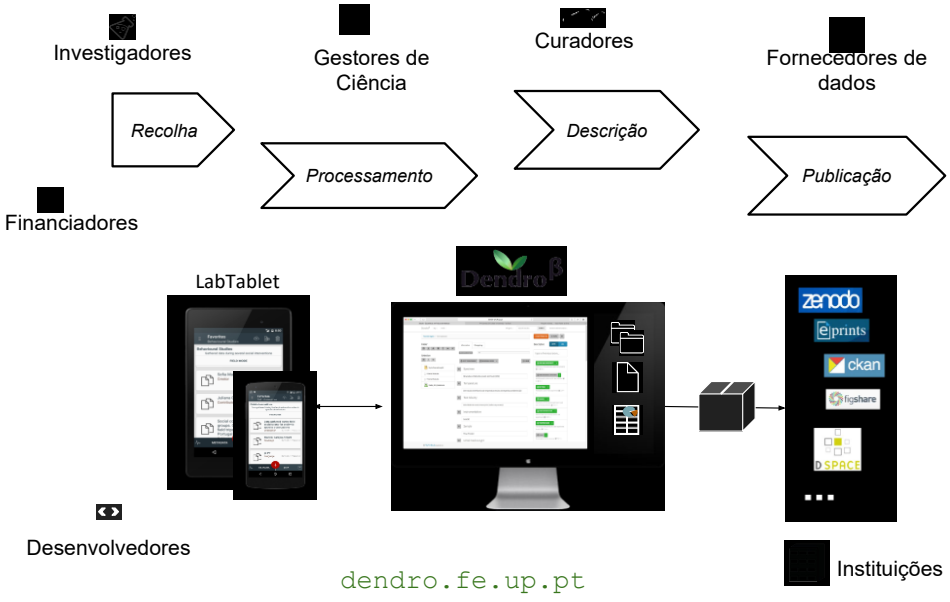

FIGURA 1. ATORES NA GESTÃO DE DADOS DE INVESTIGAÇÃO

Modelos adaptados à realidade dos grupos

A gestão de dados inclui atividades com diferentes âmbitos num grupo de investigação. Num extremo, a gestão de dados de investigação está fortemente ligada com a publicação: os investigadores colocam um pouco de esforço extra na preparação de materiais adicionais—dados recolhidos, módulos de software, versões digitais de dados apresentados num artigo—para atrair mais atenção sobre os seus resultados, para conformar ao mandato de uma agência de financiamento ou para satisfazer os requisitos de uma revista. No outro extremo, a gestão de dados pode fazer parte de um ambiente de *e-science* muito completo, em que as atividades do dia a dia são registadas, os dados são gerados e explorados de forma colaborativa e os resultados de investigação são preparados.

O trabalho realizado começou com um estudo de avaliação na Universidade do Porto, com a colaboração de 8 grupos de investigação, usando recomendações existentes e cobrindo aspetos como a consciência da necessidade de curadoria de dados, as necessidades prementes relativas a dados existentes, as soluções correntes para armazenamento, o valor atribuído a dados legados e as necessidades de apoio em ações de gestão de dados (RIBEIRO e FERNANDES, 2011). Como resultado deste estudo, o trabalho começou com base na seguinte hipótese. Os repositórios institucionais cresceram e criaram uma grande comunidade, que promoveu o auto-depósito e o acesso aberto, devido à disponibilidade de plataformas de repositórios muito sólidas, como DSpace, Fedora e EPrints. Para a gestão de dados de investigação se tornar parte do fluxo de trabalho num grupo de investigação, precisamos também de ferramentas convenientes. Para fazer a comunidade adotar estas ferramentas, elas devem permitir um benefício imediato aos investigadores, e ainda criar uma expectativa de melhoria na eficácia da publicação e disseminação da sua investigação (AMORIM et al., 2016).

A preocupação com a gestão de dados na cauda longa é relativamente recente e diversas entidades fornecem apoio ao depósito de dados multidisciplinares (ANDS, 2016, DANS, 2016, DASH, 2016, DATAONE, 2016, DCC, 2016). No entanto, nas áreas em que conjuntos de dados grandes estão no centro da investigação há um registo mais longo de iniciativas de gestão de dados, como sejam bases de dados que identificam contribuições individuais e têm uma forte ligação às publicações. O NCBI e o ICPSR são exemplos disto nas ciências de vida e nas ciências sociais, respetivamente(ICPSR, 2016, NCBI RESOURCE COORDINATORS, 2013). Noutras áreas, à medida que crescem os grupos internacionais, a partilha de dados se torna mais importante e a avaliação da investigação requer a ligação a fontes de dados, há novas infraestruturas internacionais a serem montadas. Um bom exemplo são os projetos promovidos pelo ESFRI, o *European Strategy Forum on Research*

Infrastructures, que está a apoiar uma rede de infraestruturas de investigação de interesse europeu a longo prazo. O ESFRI suporta correntemente 29 projetos em fase de implementação e 21 em projeto, nas grandes áreas da Energia, Ambiente, Saúde e Alimentação, Ciências Físicas e Engenharia, Inovação Cultural e Social e e-infraestruturas.

Organização de dados e criação de metadados com Dendro e LabTablet

A adoção da gestão de dados depende da existência de processos claros para os investigadores relativamente a recolha de dados, organização, descrição e publicação ou depósito. Ao entrevistarmos investigadores acerca das práticas de gestão de dados, tornou-se claro o fosso entre os processos usados na preparação de artigos e comunicações (embora dependentes do domínio) e as rotinas necessárias para organizar dados e para os tratar como resultados de investigação. O nosso trabalho concentrou-se por isso em duas frentes: o contacto com os investigadores, identificando tarefas que os possam levar a comprometer-se com a gestão de dados, e o desenho e implementação de ferramentas para suportar as suas atividades. As ferramentas de gestão de dados têm um campo de aplicação largo, incluindo a recolha, limpeza, processamento, organização, descrição, depósito, pesquisa. À medida que a gestão de dados se estabelece, espera-se que as plataformas de repositórios, disciplinares ou genéricas, se tornem comuns. As ferramentas de recolha e processamento serão provavelmente dependentes do domínio. Escolhemos por isso investir nas tarefas de organização e descrição de dados, na interface entre os investigadores e os gestores de repositórios (CASTROROCHA DA SILVA e RIBEIRO, 2014, CASTRO et al., 2015, CASTROROCHA DA SILVA e RIBEIRO, 2013, DA SILVA et al., 2014). As ferramentas de preparação e descrição de conjuntos de dados contribuem para um discurso consistente sobre gestão de dados com os investigadores, e uma mesma ferramenta pode fazer a ligação a diversos repositórios. A descrição completa e apropriada ao domínio é essencial para a pesquisa, uma vez que muitos dados têm pouco ou nenhum conteúdo indexável.

O Dendro é uma plataforma de gestão de dados baseada em ontologias e desenvolvimento na Universidade do Porto e recentemente publicada em código aberto (ROCHA DA SILVA, 2016, ROCHA DA SILVA et al., 2014). Os investigadores são o seu público, e o objetivo é ajudá-los a depositar e partilhar dados tanto dentro do seu grupo de investigação como com elementos externos, de forma controlada. O Dendro supõe a disponibilidade de repositórios para a preservação a longo prazo e concentra-se na organização e descrição dos dados, usando a metáfora da *Dropbox* e oferecendo possibilidades de descrição sofisticadas. No Dendro um grupo de investigação cria "projetos", na forma de pastas partilhadas em que cada colaborador deposita ficheiros, cria pastas e descreve os seus recursos com descritores genéricos ou dependentes do domínio. A

orientação para a publicação revela-se na exportação de dados e metadados para as principais plataformas de repositórios nos momento escolhidos pelos autores.

Além de suportar a criação colaborativa de metadados, o Dendro ajuda os investigadores a escolher os conjuntos de descritores que são mais relevantes para os dados do seu domínio. Duas componentes contribuem para isso: um modelo de dados flexível baseado em ontologias que cresce à medida que se juntam ontologias específicas dos domínios e um sistema de ordenação de descritores que aprende com as interações passadas na plataforma. O modelo de dados permite exportar registos como dados abertos ligados (*Linked Open Data*) para interoperabilidade, mas vai para além das soluções correntes como o OAI-PMH nas facilidades de interrogação e recuperação. O Dendro permite ligar a produção de dados com a sua preservação a longo prazo de acordo com um plano de gestão de dados, por exemplo.

Os dispositivos móveis têm evoluído para incluir funcionalidades avançadas que os tornam apropriados para diversas tarefas no âmbito da investigação. Além da capacidade de armazenamento, estes dispositivos podem estar ligados à Internet em permanência e incluem sensores capazes de fornecer informação sobre o contexto do utilizador.

O LabTablet é uma ferramenta na linha dos cadernos de laboratório electrónicos, aplicações que tiram partido dos sensores no dispositivo móvel para ajudar os investigadores a descrever os seus dados (AMORIM et al., 2015, AMORIM et al., 2014). Nalguns casos a descrição com o dispositivo móvel substitui o processo realizado em papel, com uma maior garantia da associação entre dados e metadados. O dispositivo móvel contribui para a integração de dados e metadados desde o início de um projeto, distribuindo o esforço de criação de metadados pela sua duração. Isto contribui para evitar que a descrição seja uma tarefa consumidora de tempo e realizada numa altura em que nem todos os pormenores do processo estão presentes.

Os metadados também são valiosos antes do depósito dos dados: os parceiros de um projeto colaboram e precisam de trocar dados; se estes têm descrições associadas, é menos provável que sejam mal interpretados e a colaboração fica facilitada. O resultado desta abordagem é que os conjuntos de dados têm um registo de metadados quando o projeto acaba. Isto garante que os investigadores são envolvidos desde o início e que não será necessário muito esforço para depositar ou publicar dados no final do projeto.

Fluxo de trabalho dos curadores usando ontologias ligeiras

Na história recente da curadoria e publicação de dados, vemos duas linhas de evolução nos modelos de metadados. O primeiro considera que os dados de investigação

são basicamente um resultado de investigação tal como os artigos, e que as descrições têm de seguir os modelos bem estabelecidos para publicações. Esta vista tem o mérito de facilitar a abordagem aos investigadores, tornando o depósito de dados semelhante ao das publicações, e promovendo repositórios em que as publicações e os dados são apenas distinguidos pelo tipo de recurso. A desvantagem é que não se investe na descrição de dados e em tornar os dados compreensíveis fora do grupo de investigação onde foram criados; um investigador que pretenda reutilizar os dados terá provavelmente de contactar os criadores para perceber os dados. Esta abordagem é seguida em iniciativas tais como o projeto OpenAIRE, promovendo o repositório Zenodo para recolher e interligar artigos, conjuntos de dados, software e os projetos em que tiveram origem (OPENAIRE, 2016).

A segunda linha, onde se enquadra este trabalho, parte da hipótese de que os grupos de investigação podem construir competências para descrever os seus dados com descritores mais expressivos. Isto requer ferramentas, tal como as que desenvolvemos, mas também um fluxo de curadoria com algum suporte especializado. Na nossa abordagem, isto é capturado como o fluxo do curador e envolve a análise dos conceitos essenciais para o domínio, a sua representação numa ontologia, a incorporação da ontologia na plataforma de gestão de dados e a publicação da própria ontologia. A evolução da descrição baseada nas ontologias também favorece a identificação, dentro de um domínio, de descritores apropriados que podem evoluir no sentido da adoção na comunidade, ou mesmo da normalização.

Conclusões e trabalho em curso

O projeto TAIL vai desenvolver-se numa época em que a gestão de dados de investigação está na ordem do dia, com muitas iniciativas internacionais e uma crescente consciência da sua importância nas várias comunidades científicas. A equipa do projeto parte de um ponto em que existe trabalho feito na identificação dos requisitos dos investigadores, no desenvolvimento de ferramentas e no seu teste em condições reais. O trabalho em curso vai desenvolver-se em três linhas. Na primeira vamos continuar o trabalho junto dos investigadores, aumentando o número de grupos envolvidos, experimentando com repositórios internos às instituições e com outros externos e construindo uma base de casos que possam servir de referência a outros. Na segunda vamos trabalhar junto de algumas infraestruturas do roteiro nacional, identificando as necessidades de comunidades muito estruturadas e estudando soluções adaptadas. Na terceira vamos continuar o trabalho sobre modelos de metadados e processos para facilitar a sua definição, além de desenvolver métricas para avaliar a qualidade dos metadados produzidos.

Referências bibliográficas

AMORIM, Ricardo Carvalho [et al.] (2014) - LabTablet: Semantic Metadata Collection on a Multi-domain Laboratory Notebook - Springer Communications in Computer and Information Science [Em linha]. Vol: 478, nº (2014), p. 193-205. ISSN: 978-3-319-13673-8

AMORIM, Ricardo Carvalho [et al.] (2016) - A comparison of research data management platforms: architecture, flexible metadata and interoperability - Universal Access in the Information Society [Em linha]. (2016), p. 1-12. Disponível em WWW: <http://dx.doi.org/10.1007/s10209-016-0475-y>. ISSN: 1615-5297

AMORIM, Ricardo [et al.] (2015) - Engaging researchers in data management with LabTablet, an electronic laboratory notebook. In Symposium on Languages, Applications and Technologies, SLATE'2015. 2015.

ANDS - ANDS- Australian National Data Service [Em linha]. Disponível em WWW: <URL: http://ands.org.au/>.

CAPLAN, Priscilla (2009) - Understanding PREMIS: an overview of the PREMIS Data Dictionary for Preservation Metadata 2009.

CASTRO, J. A. ; ROCHA DA SILVA, J. ; RIBEIRO, Cristina (2013) - Designing an Application Profile Using Qualified Dublin Core: A Case Study with Fracture Mechanics Datasets. In International Conference on Dublin Core and Metadata Applications. 2013. 2013. p. 47--52.

CASTRO, J. ; ROCHA DA SILVA, J. ; RIBEIRO, C. (2014) - Creating lightweight ontologies for dataset description: Practical applications in a cross-domain research data management workflow. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL and TPDL) DL 2014. 2014. ACM Press, 2014.

CASTRO, João Aguiar [et al.] (2015) - Ontologies for Research Data Description: A Design Process Applied to Vehicle Simulation. In GAROUFALLOU, Emmanouel, HARTLEY, Richard J. & GAITANOU, Panorea - Metadata and Semantics Research: 9th Research Conference, MTSR 2015. Springer, 2015. p. 348-354. ISBN: 978-3-319-24129-6.

COUNCIL OF THE CONSULTATIVE COMMITTEE FOR SPACE DATA, Systems - Reference Model for an Open Archival Information System (OAIS) 2002.

DA SILVA, João Rocha [et al.] (2014)- The Dendro research data management platform- Applying ontologies to long-term preservation in a collaborative environment. In 11th International Conference on Digital Preservation iPRES 2014 [Em linha]. iPRES, 2014, Disponível em WWW: <URL: http://dcpapers.dublincore.org/pubs/issue/view/165>.

DAF – Data Asset Framework [Em linha]. Disponível em WWW: <URL: http://www.data-audit.eu/>.

DANS – Data Archiving and Networked Services [Em linha]. Disponível em WWW: <URL: http://www.dans.knaw.nl/en>.

DASH – Dash– Data Sharing made easy [Em linha]. Disponível em WWW: <URL: https://dash.cdlib.org/>.

DATAONE – DataONE [Em linha]. Disponível em WWW: <URL: https://www.dataone.org/>.
DCC – DCC– Digital Curation Centre [Em linha]. Disponível em WWW: <URL: http://www.dcc.ac.uk/>.

EOSC – European Open Science Cloud [Em linha]. Disponível em WWW: <URL: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud>.

GENTZSCH, W. ; LECARPENTIER, D. ; WITTENBURG, P. (2014) – Big Data in Science and the EUDAT Project. In 2014 Annual SRII Global Conference. 2014. p. 191–194.

ICPSR – Interuniversity Consortium for Political and Social Research [Em linha]. Disponível em WWW: <URL: https://www.icpsr.umich.edu/icpsrweb/>.

LECARPENTIER, D. ; MICHELINI, A. ; WITTENBURG, P. (2013) – The building of the EUDAT Cross-Disciplinary Data Infrastructure. In EGU General Assembly Conference Abstracts. 2013.
MANGHI, Paolo [et al.] (2012)– OpenAIREplus: the European Scholarly Communication Data Infrastructure – D-Lib Magazine [Em linha]. Vol: 18, nº 9/10 (2012), Disponível em WWW: <http://www.dlib.org/dlib/september12/manghi/09manghi.html>. ISSN: 1082-9873

MAYERNIK, Matthew Stephen (2011) – Metadata realities for cyberinfrastructure: Data authors as metadata creators – ProQuest Dissertations and Theses [Em linha]. (2011), p. 338.

NCBI RESOURCE COORDINATORS – Database resources of the National Center for Biotechnology Information – Nucleic Acids Research [Em linha]. Vol: 41, nº Database issue (2013), p. D8–D20. Disponível em WWW: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531099/>. ISSN: 0305-1048

OPENAIRE – The OpenAIRE project. 2016. Disponível em WWW: <https://www.openaire.eu/>.

RIBEIRO, Cristina [et al.] (2015) – Motivators and Deterrents for Data Description and Publication: Preliminary Results (Short Paper). In On the Move to Meaningful Internet Systems: OTM 2015 Workshops 2015. Disponível em WWW: <URL: http://dx.doi.org/10.1007/978-3-319-26138-6_55>. p. 512--516.

RIBEIRO, Cristina ; FERNANDES, Maria Eugénia Matos (2011) – Data Curation at U.Porto: Identifying current practices across disciplinary domains – IASSIST Quarterly [Em linha]. Vol: 35, nº 4 (2011), p. 14–17.

ROCHA DA SILVA, J. (2016) – The Dendro RDM platform. 2016. Disponível em WWW: <https://github.com/feup-infolab-rdm/dendro>.

ROCHA DA SILVA, João [et al.] (2014) – Dendro: Collaborative Research Data Management Built on Linked Open Data. Springer International Publishing, 2014. Disponível em WWW: <http://dx.doi.org/10.1007/978-3-319-11955-7_71>.

TENOPIR, Carol [et al.] (2011)– Data Sharing by Scientists: Practices and Perceptions – PLoS ONE [Em linha]. Vol: 6, nº 6. Disponível em WWW: <http://dx.plos.org/10.1371/journal.pone.0021101>.