UPBox e DataNotes: um ambiente de suporte à gestão colaborativa de dados científicos

UPBox and DataNotes : an environment for the collaborative management of research data

João Rocha da Silva[1]
Doutorando do Programa Doutoral em Engenharia Informática e Computação da Universidade do Porto / INESC TEC, Porto, Portugal
Email : joaorosilva@gmail.com

Cristina Ribeiro
Professora Auxiliar do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto / INESC TEC, Porto, Portugal
Email : mcr@fe.up.pt

João Correia Lopes
Professor Auxiliar do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto / INESC TEC, Porto, Portugal
Email : jlopes@fe.up.pt

Resumo

A ciência depende cada vez mais de dados e é reconhecida a dificuldade em armazenar, descrever e partilhar conjuntos de dados produzidos em ambiente de investigação. Em muitos casos, os conjuntos de dados de investigação apenas são sujeitos a processos de gestão após a publicação dos resultados da sua análise. Esta abordagem *a posteriori* funciona bem para recursos estáticos, como por exemplo as publicações, mas não resulta tão bem para recursos com contextos de produção dinâmicos, como é o caso dos conjuntos de dados. No caso destes tipo de recursos, o processo de gestão deverá começar no início das atividades de investigação, tornando-se uma parte integrante do próprio processo científico. Verifica-se neste momento que os repositórios de dados científicos dependem completamente dos curadores para efetuar a anotação dos conjuntos de dados. Para facilitar a preparação dos dados para o seu depósito, propomos um ambiente colaborativo de gestão de dados de investigação, desenhado para ajudar pequenas equipas de investigação a armazenar e descrever os seus conjuntos de dados, em preparação para o seu posterior depósito. É suportado por dois componentes integrados: o UPBox—uma "nuvem" privada desenhada para suportar o armazenamento dos ficheiros—e o DataNotes—uma solução "wiki" desenhada para os investigadores descreverem os seus ficheiros de forma colaborativa, construída sobre a Semantic MediaWiki. Os resultados preliminares de testes com utilizadores produziram respostas positivas às funcionalidades implementadas, o que permite concluir que estas respondem a necessidades reais dos investigadores.

Palavras-chave: Gestão de dados de investigação, dados de investigação, web semântica, repositórios de dados

1 Trabalho suportado pela bolsa de doutoramento com o código SFRH/BD/77092/2011 da FCT (Fundação para a Ciência e Tecnologia).

Abstract

Research datasets are, in most cases, only targeted for appropriate management after the results of their analysis are published. This *a posteriori* approach works well for static resources such as publications, but does not suit research datasets because of their rapidly changing context and content. For these resources, the data management process should be present from the start of the research, effectively becoming a part of the research workflow itself. However, research data repositories rely completely on curators for the description of the deposited datasets; this can create problems, as the curators can become a bottleneck in the process. In this paper we present a collaborative data management environment designed to help a small research group store and describe their datasets in preparation for later deposit in a data repository. The environment is made up of two integrated components: UPBox—a private cloud and web-based file storage environment—and DataNotes—a wiki solution tailored for researchers to collaboratively describe their data, based on Semantic MediaWiki. Preliminary user tests have yielded positive responses to the implemented features, as they correspond to some of the needs on the targeted research groups.

Keywords: Research data management, research data, semantic web, data repositories

Introdução

A gestão dos dados de investigação tem vindo a assumir um papel cada vez mais relevante no *workflow* de investigação. A adoção de práticas adequadas de gestão de dados de investigação apresenta diversas vantagens para as instituições de investigação, entre as quais se destaca o reconhecimento externo dos resultados dos projetos por elas desenvolvidos. No entanto e em última análise, são os investigadores que mais devem estar conscientes das potenciais melhorias para o seu trabalho que podem advir da adopção de práticas consistentes de gestão dos dados que produzem ou reutilizam. Estas vantagens, já amplamente discutidas na comunidade de preservação e gestão de dados, passam por um aumento no número de citações para os artigos que fornecem acesso aos dados base, pela reprodutibilidade dos resultados da investigação, pelas oportunidades de formulação de novas questões de investigação (PIWOWAR; DAY; FRIDSMA, 2007; GIL et al, 2007; LORD et al., 2008) e também por um objetivo mais amplo: o avanço mais rápido da ciência (BORGMAN, 2012). Estas vantagens são claras para a gestão das entidades financiadoras e para os responsáveis das instituições de investigação, mas o mesmo não se passa junto dos investigadores; para estes últimos, essas vantagens são muitas vezes vistas como benefícios obscuros a longo prazo, com a desvantagem de exigirem um esforço muito concreto no imediato. Com efeito, o esforço de anotação dos dados por parte dos investigadores é muitas vezes encarado como um trabalho extra, de um certo cariz burocrático, a somar a todas as tarefas necessárias à produção científica.

As descrições, ou *metadados*, são informação acerca dos próprios conjuntos de dados, que torna possível a sua interpretação por terceiros; só com a presença desta informação se torna possível a esses outros investigadores reutilizar os conjuntos de dados em atividades de investigação subsequentes, algo que fomenta a citação de dados em paralelo com a citação das publicações. Para que a citação de dados se torne uma realidade, contudo, é crucial combater a falta de motivação expressa por diversos investigadores para participar no depósito e partilha dos seus dados; para tal, acreditamos que a eliminação da barreira entre os sistemas de suporte às atividades diárias de investigação e os chamados *repositórios de dados* é um passo chave para fazer com que cada vez mais conjuntos de dados façam a transição dos sistemas de suporte à investigação (do inglês "Current Research Information Systems") para os repositórios de dados, onde poderão estar acessíveis e, obviamente, ser citados.

Grande parte dos *workflows* de gestão de dados de investigação atuais dependem de processos de descrição realizados por curadores profissionais. Embora este processo seja eficaz para a produção de metadados genéricos de alta qualidade, a inclusão de metadados de domínios específicos nessas descrições só pode ser feita com a estreita colaboração entre os curadores e os criadores dos conjuntos de dados. Por outras palavras, só um investigador de um determinado domínio sabe o que os seus pares precisam de saber acerca de um conjunto de dados para considerarem ou não a sua reutilização nas suas atividades. No entanto, os investigadores muito dificilmente possuem as capacidades ou conhecimentos de gestão de dados necessárias à produção de descrições abrangentes dos seus conjuntos de dados (SWAN; BROWN, 2008). Assim sendo, só com a colaboração entre curadores e investigadores se torna possível produzir descrições ricas e completas conjuntos de dados de investigação (JONES; ROSS; RUUSALEPP, 2012). Esta abordagem tende, contudo, a exigir muito tempo dos investigadores ao mesmo tempo que cria uma grande pressão sobre os curadores, que correm o risco de se tornarem o *bottleneck* no processo de preservação. O resultado final é um processo que corre o risco de ficar reduzido a uma série de contactos esporádicos; ao mesmo tempo, vão-se perdendo oportunidades para descrever os conjuntos de dados à medida que os seus autores mudam de projeto ou decidem perseguir outras questões de investigação.

Recentemente, têm vindo a surgir diretórios de repositórios de dados suportados por comunidades, como por exemplo o DataBib, um diretório ou agregador de repositórios de dados de investigação (WITT; GIARLO, 2012). No entanto, as iniciativas para a criação de verdadeiros ambientes de depósito e descrição continuada e colaborativa para curadores e investigadores são ainda em número reduzido. Em 2013, o projeto DataUP (STRASSER;

CRUSE, 2012) mostrou como uma ferramenta de auto-depósito embutida diretamente no Microsoft Excel pode ajudar os investigadores no depósito de folhas de cálculo a partir dessa mesma aplicação. Um aspeto interessante do projeto é o facto de este procurar orientar os investigadores ao longo da descrição das folhas de cálculo em questão, apontando possíveis erros na sua formatação e organização e, ao mesmo tempo, tornando mais fácil descrevê-los utilizando metadados genéricos, ou seja não específicos para domínios de investigação dos conjuntos de dados em causa. Outro exemplo relevante é o projeto ADMIRAL, uma iniciativa que decorreu de 2009 a 2011 com o objetivo de criar uma infraestrutura federada de dois níveis de gestão de dados para investigadores das ciências da vida (HODSON, 2011). As figuras 1 e 2 mostram as diferenças introduzidas pela solução desenvolvida no *workflow* de gestão de dados em vigor até então. Na Figura 1 (antes de terem sido introduzidas as mudanças), os repositórios institucionais encontram-se dissociados das atividades de gestão de dados desenvolvidas pelos investigadores no seu dia a dia; na Figura 2 é possível ver a presença de uma plataforma de partilha de armazenamento de dados que os investigadores podem usar (componente "Management"), e que por seu turno está ligada diretamente aos repositórios para depósito de conjuntos de dados que foram sendo geridos ao longo do processo de investigação. O resultado final é a disponibilização desses conjuntos como *Open Data* na Web, para serem reutilizados na formulação de novas questões de investigação, por exemplo. No projeto ADMIRAL é patente a tentativa de eliminar as barreiras iniciais ao depósito de conjuntos de dados, mantendo no entanto uma etapa explícita de depósito num repositório, com vista a tornar mais fácil a associação das publicações aos dados base que lhes dão origem. Para tal, as diretrizes do projeto ditam que cada conjunto deve ser disponibilizado sob licença CC Zero[2] e identificado através de um DataCite DOI[3]; só desta forma se torna possível citar os conjuntos de dados mencionados e oferecer acesso livre aos mesmos.

2 Creative Commons Zero http://creativecommons.org/publicdomain/zero/1.0/

3 Digital Object Identifier (http://www.doi.org/)

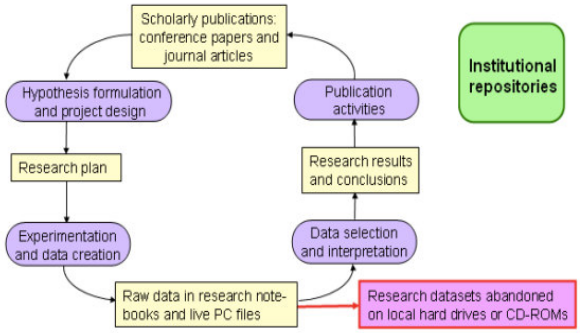

Figura 1: "The conventional research data lifecycle" (imagem obtida a partir de (HODSON, 2011))

Na abordagem que aqui propomos, a gestão de dados científicos é vista como um processo contínuo de apoio às atividades diárias de cada investigador no seio do seu grupo de investigação, uma necessidade já expressa recentemente (JAHNKE; ASHER; KERALIS, 2012).

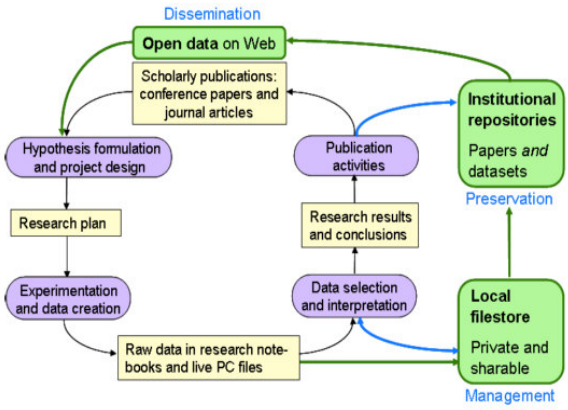

Figura 2: "The ADMIRAL enhanced research data lifecycle" (imagem obtida de (HODSON, 2011)

Este ambiente mais dinâmico relaxa alguns requisitos de interoperabilidade associados à produção de metadados no *workflow* de investigação e curadoria, em favor de uma captura mais imediata e ágil do contexto de produção dos dados tão cedo quanto possível no desenrolar das atividades de investigação—idealmente logo após a sua produção ou processamento. Para o conseguir, oferece aos investigadores um conjunto de recursos desenhados para resolver problemas com que se deparam diariamente na gestão dos seus dados de investigação. A solução é composta por dois módulos distintos mas integrados: o UPBox, uma solução de software semelhante à conhecida "Dropbox", desenhada para o

armazenamento e partilha de conjuntos de dados e o DataNotes, uma solução *wiki* que permite aos investigadores anotar os dados depositados no UPBox de forma colaborativa. Ao participar no processo de gestão de dados científicos, os investigadores ganham imediatamente acesso a uma área de armazenamento simples e segura para os seus conjuntos de dados, completamente controlada pela instituição a que pertencem. Este é um ponto importante em contexto de investigação, pois o armazenamento dos conjuntos de dados produzidos pelos seus investigadores em serviços na nuvem externos à instituição pode representar potenciais perigos de segurança ou quebras de sigilo.

Estas duas soluções são desenhadas para estabelecer uma *área de preparação* desenhada para trabalhar sobre os conjuntos de dados com vista à sua ingestão num repositório num momento posterior. Contudo, esta abordagem dá um passo para além do desenvolvido no projeto ADMIRAL devido ao seu foco principal, que é a produção de metadados de forma colaborativa pelos próprios investigadores. Desta forma, aquando da finalização de um conjunto de dados ou da escrita da publicação, uma parte importante do trabalho de descrição de dados já terá sido executada; consegue-se tornar o passo de depósito torna-se uma tarefa mais fácil e será mais simples incentivar os investigadores a concluir o processo de depósito dos seus dados com a assistência de curadores. Convém no entanto referir que este trabalho é orientado para a *long-tail* dos dados de investigação, não se assumindo portanto como uma solução para os grandes conjuntos de dados criados em algumas áreas de investigação—que são aliás muitas vezes já depositados em infraestruturas desenhadas à medida para satisfazer esses requisitos. Destina-se por isso a satisfazer as necessidades de grupos de investigação que produzem uma miríade de pequenos conjuntos de dados (PALMER; CRAGIN, 2007), dados esses que tendem a correr mais risco de serem perdidos devido à falta de recursos para sua gestão nos projetos em que foram produzidos.

Uma nuvem privada combinada com uma wiki semântica

O ambiente de gestão de dados de investigação proposto neste trabalho é composto por dois sistemas interligados por um conjunto de pontos de comunicação baseados na Web, normalmente chamados *webservices* (ver Figura 3). O UPBox (1) utiliza o armazenamento

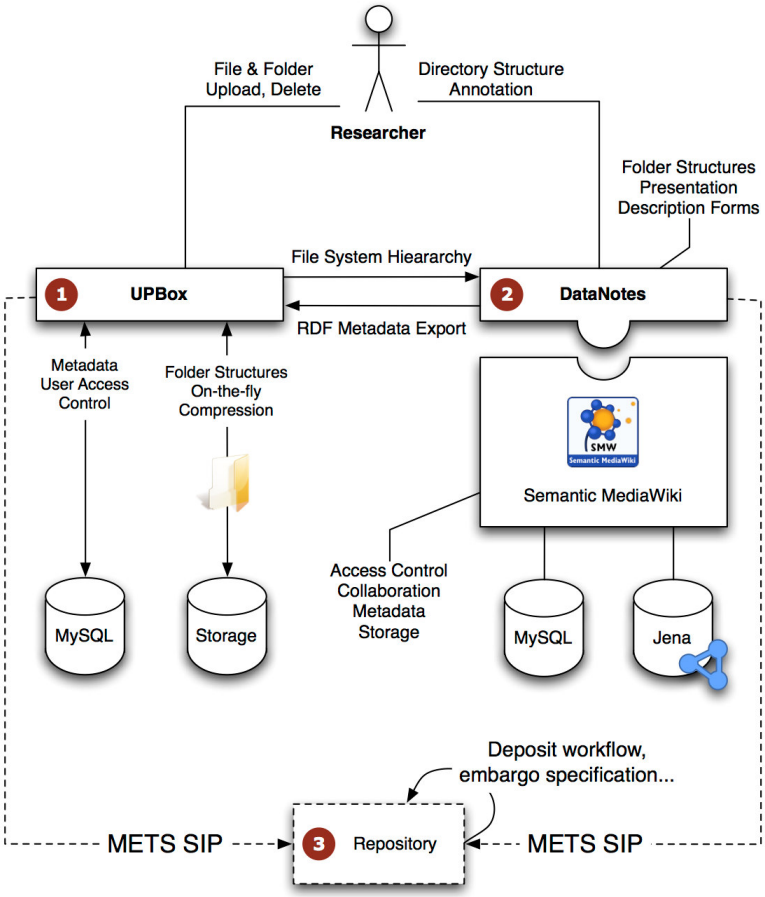

Figura 3: Arquitectura do sistema UPBox + DataNotes

local do servidor, que pode ser mapeado num ponto de armazenamento baseado em RAID (a solução adotada nesta fase), ou um volume de armazenamento distribuído. Uma alternativa possível para assegurar a escalabilidade horizontal do sistema seria a montagem do volume de armazenamento do UPBox num sistema de ficheiros Hadoop (HDFS), o que proporcionaria a abstração necessária à existência de uma nuvem privada. Uma base de dados MySQL é usada para manter os dados necessários à gestão de utilizadores, controlo de acessos e metadados internos acerca das estruturas de diretórios. Todos os ficheiros são compactados ao serem enviados para o servidor e descompactados ao serem pedidos pelos investigadores para minimizar o espaço de armazenamento necessário para sustentar o sistema. No que diz

respeito à gestão de acessos, o UPBox está ligado ao sistema de informação da U.Porto (SIGARRA) através do protocolo LDAP (Lightweight Directory Access Protocol), o que permite aos elementos da U.Porto autenticarem-se no sistema recorrendo às suas credenciais do SIGARRA. Os utilizadores externos podem também registar-se no sistema, facilitando a colaboração entre investigadores da U.Porto e investigadores de outras instituições. A plataforma permite aos investigadores criar *projetos*; dentro de um projeto é possível criar pastas, onde ficheiros podem ser depositados de uma forma muito semelhante ao que acontece com a "Dropbox". Um projeto pode ser partilhado entre os colaboradores da equipa e novos membros podem ser adicionados, sendo que o sistema oferece sugestões automáticas a partir da lista de utilizadores já registados. Os membros de um projeto podem fazer carregamentos de ficheiros, criar pastas e também apagar tanto pastas como ficheiros. O sistema também permite aos investigadores carregar diversos ficheiros simultaneamente para facilitar a migração de grandes quantidades de ficheiros já existentes.

O DataNotes (número 2 na Figura 3) é uma plataforma de anotação de estruturas de diretórios baseada numa wiki, assente na tecnologia Semantic MediaWiki[4] e permite aos utilizadores produzir rapidamente páginas wiki com os metadados dos seus conjuntos de dados. O DataNotes tem como objetivos:

1. Proporcionar um ambiente de colaboração para descrever estruturas de diretórios, com suporte para controlo de versões de anotações, controlo de bloqueios das páginas, gestão de edição concorrente e controlo de acessos.

2. Ajudar os investigadores do grupo a encontrar conjuntos de dados através de pesquisa sobre o texto livre contido nos metadados.

3. Oferecer uma interface com o utilizador amigável, embora com capacidades sofisticadas, capazes de capturar as relações entre as partes do conjunto de dados e também a semântica dessas mesmas ligações, nos casos em que é necessário tal nível de detalhe.

4. Facilitar a partilha de descrições de conjuntos de dados (incluindo a possibilidade de enviar um *link* direto para uma pasta ou ficheiro depositado no sistema).

5. Eliminar dependências de software em código fechado. O facto de uma solução de

4 http://semantic-mediawiki.org/

preservação depender de módulos ou bibliotecas fechadas pode comprometer o acesso aos dados armazenados na solução à medida que essas dependências ficam obsoletas; quando isso acontece, essas soluções não podem ser atualizadas nem é possível alterar a sua lógica de negócio para suportar mudanças nos requisitos da plataforma no seu todo.

6. Facilitar a instalação de um sistema de preservação: o DataNotes pode ser instalado facilmente em qualquer máquina com um servidor Web Apache e o interpretador de PHP instalados, ambos soluções *open-source* e presentes em praticamente em qualquer servidor. Assim, qualquer instituição de investigação poderá hospedar a sua própria instância do DataNotes para apoiar o trabalho dos seus grupos de investigação.

7. Preparar os conjuntos de dados para a preservação a longo prazo ao facilitar a exportação de metadados sobre esses conjunto em formatos normalizados (RDF é um exemplo). Desta forma, a sobrevivência dos dados está assegurada, mesmo no caso do DataNotes precisar de ser substituído por outra plataforma.

8. Oferecer interfaces programáticas ao exterior, que permitam a sistemas externos recuperar os conteúdos depositados na wiki, de maneira a poderem integrá-los nas suas próprias lógicas de negócio.

Dado que o DataNotes é baseado numa plataforma wiki, a gestão de *namespaces* e os controlos de acesso já existem, assim como as capacidades de edição simultânea e controlo de versões. A pesquisa sobre texto livre também está presente, o que permite aos utilizadores recuperar conjuntos de dados através de uma função de pesquisa global (que cobre todos os metadados introduzidos no sistema pelos investigadores). A interface já tem provas dadas de ser considerada amigável por utilizadores sem conhecimentos profundos de informática; o *look-and-feel* da MediaWiki é mantido na Semantic MediaWiki, mantendo a curva de aprendizagem simples que tem permitido à Wikipedia crescer e manter-se em funcionamento. O sistema também permite aos seus utilizadores partilhar facilmente as descrições dos seus conjuntos de dados, uma vez que cada descrição é uma página wiki, acessível através de um URL. Dado que essas URLs são mostradas no navegador web durante a visualização da página, qualquer endereço pode ser simplesmente copiado e colado numa mensagem de email para ser partilhada com outros utilizadores com permissões para aceder ao recurso.

O módulo “Repository” (número 3 na Figura 3) representa uma plataforma de

repositório existente na instituição (software como o DSpace, Fedora, ou ePrints). Após os conjuntos de dado serem depositados na UPBox e as suas descrições terem sido produzidas no DataNotes, os investigadores devem ser capazes de empacotar automaticamente o estado atual de uma pasta (por exemplo) e enviá-la para o repositório, onde um novo *workflow* de depósito e ingestão será iniciado. Os metadados que serão adicionados ao novo item de depósito terão que ser sujeitos às validações necessárias por parte de um curador (incluindo especificações de embargo). Após o processo ser concluído, e devido à natureza "estática" do item de repositório que é produzido, este poderá ser citado de forma segura em publicações através de um identificador (URL) persistente.

Tecnologias para preservação

O facto de a plataforma depender única e exclusivamente de componentes *open-source* incentiva o desenvolvimento sobre a mesma e elimina as dependências "caixa-preta". Estas podem contribuir para a rápida obsolescência da plataforma, acelerando a necessidade de migrar o conteúdo para novas plataformas, muitas vezes incorrendo em custos e colocando em risco o acesso aos dados sempre que tais operações se tornarem inviáveis. A plataforma é fácil de instalar tanto em sistemas Linux como Windows e é totalmente suportada por software de código aberto que está disponível gratuitamente. Estas características tornam-na atrativa para os administradores de TI, que a podem instalar mesmo em máquinas relativamente modestas.

Para proporcionar funcionalidades de exportação de metadados em formatos normalizados, a Semantic MediaWiki possibilita a exportação de páginas em formato RDF[5], o que facilita a migração dos dados no wiki para outra solução. O ponto de acesso SPARQL fornecido permite a agentes externos executar consultas complexas sobre o conteúdo da wiki e recuperar conjuntos de dados nela contidos. De forma a reduzir as alterações ao núcleo da Semantic MediaWiki, o DataNotes foi escrito como um módulo totalmente autossuficiente. O código da extensão está contido numa única pasta, tornando possível instalar a extensão em qualquer instância Semantic MediaWiki, sem ser necessário modificar a instalação existente.

5 http://semantic-mediawiki.org/wiki/Help:RDF_export

Demonstração de um workflow

Após a descrição da solução em causa, passa-se a demonstrar um *workflow* de interação com os sistemas mencionados. Estas atividades começam com a criação de um *projeto* na plataforma UPBox (ver número 4 da Figura 4).

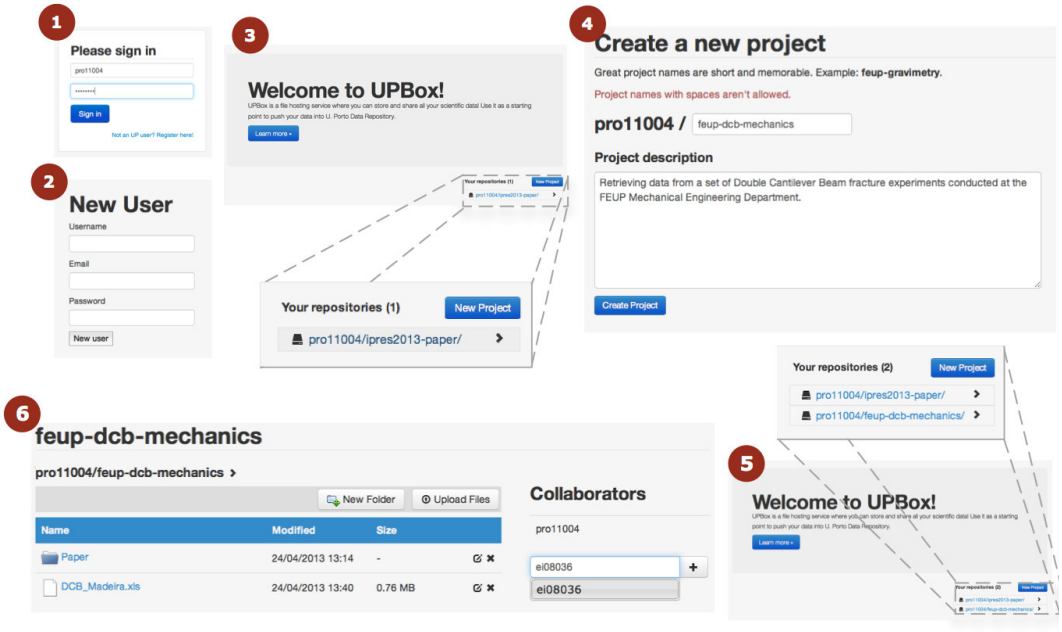

Figura 4: Workflow de depósito (UPBox)

Para terem acesso à plataforma, os utilizadores que sejam elementos da U.Porto podem autenticar-se com as suas credenciais SIGARRA (1), enquanto os utilizadores externos à comunidade da U.Porto poderão executar uma etapa de registo simples para obter um conjunto de credenciais UPBox (2). Após uma autenticação bem-sucedida, os utilizadores podem aceder aos projetos existentes ou criar um novo projeto (3), especificando o título e uma breve descrição do projeto (4). O projeto é então adicionado à lista de projetos aos quais o utilizador tem acesso (5). Ao aceder a um dos projetos dessa lista, é apresentada a principal interface com o utilizador da plataforma UPBox(6), que exibe um *look and feel* conhecido de qualquer utilizador do "Dropbox"; os ficheiros e pastas são exibidos como uma lista, permitindo ao utilizador criar novas pastas ou fazer *upload* de novos ficheiros para a pasta aberta em dado instante. O utilizador pode fazer clique no nome da pasta para a abrir ou fazer download de ficheiros ao fazer clique nos seus nomes ou ícones. Ao lado direito é visível uma área de "Colaboradores", na qual o utilizador poderá adicionar outros utilizadores da UPBox

à lista de colaboradores do projeto atual, e é também possível remover um colaborador a qualquer momento. Após depositar um ficheiro, o utilizado pode descrevê-lo, selecionando o botão "Anotar" (ícone com uma folha e lápis à esquerda da cruz presente no item da lista para esse ficheiro). Ao selecionar essa opção, o utilizador é redirecionado para a plataforma DataNotes—mais especificamente para a página que contém os metadados mais recentes para esse ficheiro caso tenha sido previamente anotado, ou para uma página em branco onde a anotação pode ser realizada. Tanto ficheiros como pastas podem ser apagadas ao selecionar a cruz à direita de cada item da lista de ficheiros da pasta.

A Figura 5 mostra o *workflow* de descrição do conjunto de dados na plaaforma wiki, usando o projeto que foi criado na UPBox .

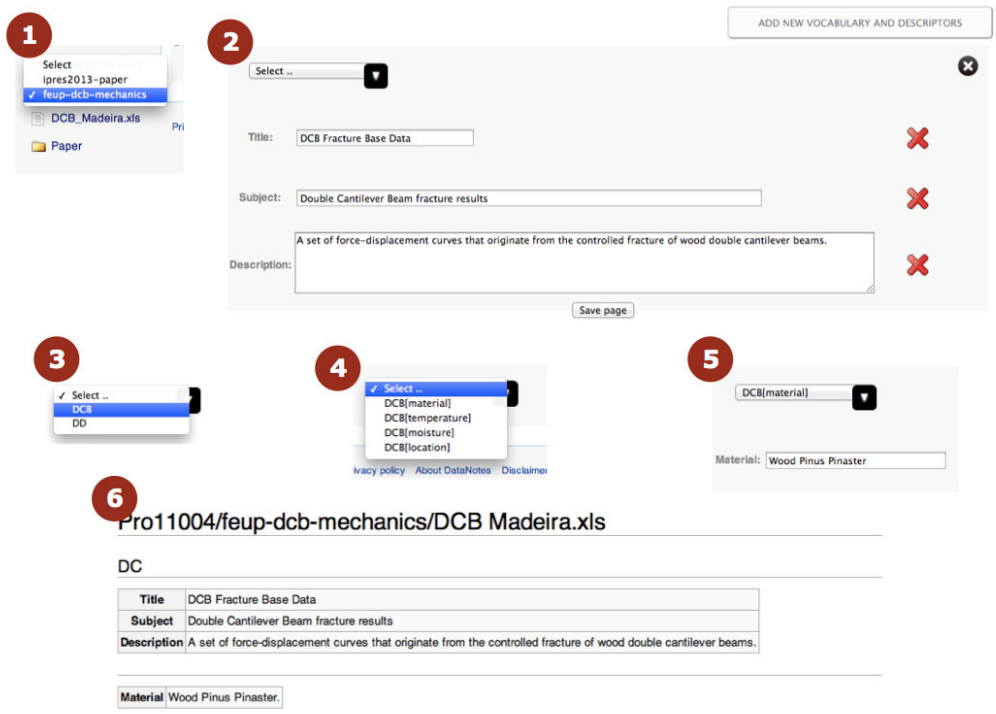

Figura 5: Descrição de conjuntos de dados depositados (DataNotes)

Após efetuar a sua autenticação no DataNotes, o utilizador acede a uma interface semelhante à de uma página da Wikipedia, mas incluindo com um elemento adicional: uma área de seleção de projetos na barra lateral esquerda. Esta área inclui uma caixa de combinação (1) que permite ao utilizado selecionar um dos seus projetos da UPBox. Ao selecionar um projeto, a sua estrutura de pastas e ficheiros é apresentada. Ao selecionar um dos ficheiros ou pastas, o interface apresentará uma página com os metadados relativos a esse nó na área de visualização de páginas (2). O utilizador pode também expandir pastas (sem

realmente selecionar um arquivo ou pasta para anotar) para explorar os níveis mais profundos da estrutura de diretórios. Neste exemplo proceder-se-á à anotação do ficheiro "DCB_Madeira.xls" mostrado em (1); a área principal de exibição de página (2) mostra as caixas de texto usadas para entrada de valores de metadados. Dependendo do tipo de descritor, as caixas de texto podem ser substituídas por outros controlos. Por exemplo, para inserir o valor de um descritor que consista numa data, a interface do utilizador irá mostrar um seletor de data em vez de uma caixa de texto. A Semantic MediaWiki oferece a possibilidade de especificar descritores de metadados tipados como URL's, tornando possível selecionar uma URL de um recurso interno ou externo à *wiki* como o valor para um descritor. Isto torna possível, por exemplo, especificar um descritor "Autor", cujas instâncias podem ser ligações para as páginas Web dos autores dos recursos em questão, ao invés de sequências de caracteres contendo o nome de uma pessoa. Esta pequena diferença é muito importante do ponto de vista da qualidade dos metadados, pois reduz a ambiguidade e evita duplicados devido a convenções de nomenclatura, que são muitas vezes inconsistentes. Ao mesmo tempo, é um primeiro passo para a plena integração dos conjuntos de dados depositados na Web semântica. Os utilizadores podem remover um descritor (selecionando a cruz vermelha à direita do mesmo) ou todos os descritores de um determinado esquema, selecionando o círculo preto com uma cruz dentro que é mostrado no canto superior direito (ambos os elementos são visíveis na Figura 5).

O DataNotes permite aos seus utilizadores selecionar descritores de diferentes esquemas de metadados, combinando-os livremente para criar os seus próprios perfis de aplicação. Esta abordagem contrasta com a distribuição padrão da Semantic MediaWiki, que apenas permite aos utilizadores preencher descritores de um único esquema de metadados (normalmente o esquema Dublin Core, que é definido como padrão). A decisão de proporcionar esta flexibilidade na anotação é motivada pela necessidade de diversificar a descrição de dados, incentivando o surgimento de perfis de aplicação à medida que os utilizadores adicionam ou removem descritores provenientes de diferentes esquemas de metadados. Da mesma forma, os conjuntos de descritores selecionados por determinados grupos poderão ser reutilizados por outros grupos de investigação que se foquem nos mesmos domínios.

Para adicionar descritores de um esquema de metadados diferente do esquema Dublin Core, o utilizador pode selecionar a opção "Add new vocabulary and descriptors". Ao selecionar essa opção, uma caixa de combinação será adicionada na parte inferior da lista de

descritores, que poderá ser usada selecionar descritores do esquema metadados recém-adicionado(3); neste caso, são disponibilizados dois esquemas personalizados (DCB e DD). Após selecionar o esquema de metadados pretendido, o utilizador pode adicionar descritores deste através de outra caixa de combinação, que é apresentada em (4). O descritor recém-adicionado do esquema DCB é adicionado à página de descrição e, em seguida, é possível ver este novo descritor já preenchido (5). O resultado do *workflow* exemplo de descrição do conjunto de dados, contendo já todos os descritores e seus respetivos valores, é mostrado em (6). Nesta área é possível ver todos valores de metadados, corretamente guardados e divididos em secções de acordo com os esquemas de metadados a que pertencem. Ao selecionar o botão "Save", a informação é guardada, sendo apresentada como uma página wiki, que pode ser vista imediatamente por todos os elementos do grupo de investigação e editada por aqueles que têm a permissão para tal. Tanto a gestão de versões como a edição simultânea são tratadas de forma transparente pelo Semantic MediaWiki.

Conclusões e perspetivas de trabalho futuro

Os investigadores expressam geralmente alguma relutância quanto à produção de descrições para os seus conjuntos de dados. No entanto, é claro que registam o contexto de produção dos seus dados através de meios convencionais, tais como arquivos "LEIAME", notas escritas ou mesmo mensagens de correio eletrónico trocadas entre elementos dos grupos de investigação. Uma das grandes motivações por detrás da criação destas notas prende-se com a necessidade de passar o conhecimento necessário para interpretar os conjuntos de dados a outros investigadores que integram o grupo.

A análise das práticas de gestão de dados atuais sugere que a gestão de dados deve acompanhar as atividades diárias dos investigadores, ao invés de ser realizada após toda a atividade de investigação ter sido concluída e os seus resultados terem sido publicados. Desta forma, maximizam-se as oportunidades de recolha do contexto de produção dos conjuntos de dados, facilitando a ingestão posterior num repositório capaz de lidar com a sua preservação a longo prazo. Outra vantagem desta abordagem é o facto de proporcionar às instituições de investigação uma visão dos dados que são produzidos pelos seus investigadores, ao mesmo tempo que mantêm o controlo completo sobre os mesmos.

Para atender às necessidades dos investigadores, foi apresentado um ambiente

colaborativo totalmente *open-source* (tanto no que diz respeito à solução em si como no seu ecossistema de dependências), desenhado para a partilha de dados e sua descrição por parte dos próprios elementos dos grupos de investigação. O sistema UPBox + DataNotes permite que os investigadores depositem os seus conjuntos de dados em pastas, procedendo depois à sua anotação através de uma interface wiki integrada.

No estado atual de desenvolvimento, o UPBox não proporciona funcionalidades de controlo de versões de ficheiros. O DataNotes, por sua vez, já oferece recursos de controlo de versões para as páginas de metadados de cada arquivo ou pasta, uma vez que é construído sobre a Semantic MediaWiki. A possibilidade de especificar quotas de armazenamento em disco para os utilizadores do UPBox está também na lista de possíveis melhorias para fornecer o controlo sobre o espaço de armazenamento do servidor. Um sistema de controlo de acesso mais sofisticado também poderia ser implementado do lado do UPBox, permitindo aos administradores do projeto especificar as ações que podem serem realizadas por cada membro da equipa para cada diretório (e respetivos subdiretórios) presentes no projeto.

No que diz respeito às potenciais melhorias a realizar sobre o UPBox, pensa-se que um cliente de desktop para a UPBox seria uma excelente maneira de manter a sincronização constante e transparente com o sistema de armazenamento remoto (de forma semelhante ao cliente desktop da Dropbox). Uma funcionalidade de carregamento de pastas completas também facilitaria a migração de dados armazenados em sistemas convencionais.

As melhorias possíveis do lado do DataNotes estão relacionadas com a descrição do conjunto de dados, particularmente na automatização de tarefas repetitivas. Ao permitir que os investigadores reutilizem conjuntos de descritores de uma pasta para anotar outra, será possível incentivar a criação de perfis de aplicação para cada domínio, tirando partido da reutilização dos mesmos por parte da comunidade. Outra importante funcionalidade necessária para completar o *workflow* de gestão de dados descritos neste ambiente é o depósito dos dados em repositório de uma forma transparente, num momento escolhido pelos responsáveis do projeto. Para o conseguir, é necessário implementar uma conexão a um repositório de dados, como por exemplo o software DSpace[6]. No futuro, será possível iniciar *workflows* de depósito DSpace a partir do UPBox ou do DataNotes; para que tal seja possível, será necessário implementar funcionalidades de exportação de conjuntos de dados num

6 Mais informação disponível em : http://www.dspace.org

formato aceite pelas soluções de repositório, como por exemplo pacotes METS[7] SIP[8]. O objetivo é tornar esse processo suficientemente rápido para os investigadores conseguirem citar os dados base aquando da publicação dos seus resultados, tornando mais fácil para os leitores das publicações encontrarem os dados base correspondentes.

Uma pequena experiência de validação foi realizada junto de um grupo de investigadores do Departamento de Engenharia Mecânica da FEUP (Faculdade de Engenharia da Universidade do Porto), proporcionando *feedback* essencial acerca de potenciais melhorias introduzidas por esta plataforma. A resposta foi positiva, tendo oferecido importantes *inputs* para guiar os desenvolvimentos futuros. Por exemplo, a decisão de permitir que utilizadores externos se registassem no sistema foi tomada após se constatar que este grupo de investigação incluía membros de uma outra universidade. Os investigadores da FEUP colocaram a hipótese de usar o UPBox para partilhar conjuntos de dados entre todos os elementos do grupo de investigação e a funcionalidade foi implementada dado que outros grupos muito provavelmente teriam as mesmas necessidades.

À medida que as ferramentas forem usadas por diferentes grupos de investigação, será também importante avaliar se estas ferramentas devem agir apenas como suporte às atividades diárias dos investigadores ou se devem ser estendidas de forma a também satisfazerem requisitos de preservação a longo prazo, tomando para si o papel atualmente desempenhado pelos repositórios de dados.

Referências

ALAN, T.; PETER, M. Project SPECTRa-T Submission, Preservation and Exposure of Chemistry Teaching and Research Data in Theses. **Technical Report**, 2008.

BORGMAN, C. The conundrum of sharing research data. **Journal of the American Society for Information Science and Technology**, v. 63, n. 6, 2012.

GIL, Y. et al. Examining the challenges of scientific workflows. **Computer (Long Beach California)**, v. 40, n. 12, p. 24–32, Dec. 2007.

HEERY, R.; PATEL, M. Application profiles: mixing and matching metadata schemas. **Ariadne**, n. 25, 2000.

HODSON, S. ADMIRAL: a data management infrastructure for research activities in the life sciences. **Technical Report**, p. 18, 2011

7 METS: Metadata Encoding and Transmission Standard.
8 SIP : Submission Information Package

JAHNKE, L.; ASHER, A.; KERALIS, S. D. C. The problem of data. **Council on Library and Information Resources**, p. 1-43, Aug. 2012.

JONES, S.; ROSS, S.; RUUSALEPP, R. **Data audit framework methodology**. HATII: University of Glasgow, 2009.

LORD, P. et al. From data deluge to data curation. **eScience All Hands Meeting,** p. 371–375, 2008.

MARTINEZ-URIBE, L.; MACDONALD, S. User engagement in research data curation. In: EUROPEAN CONFERENCE ON RESEARCH AND ADVANCED TECHNOLOGY FOR DIGITAL LIBRARIES, 13., 2009, Berlin. **Proceedings**... Berlin: Springer-Verlag. 2009. v. 5714, p. 309–314.

PALMER, C.; CRAGIN, M. Data curation for the long tail of science: the case of environmental sciences. In: INTERNATIONAL DIGITAL CURATION CONFERENCE, 3., 2007, Renaissance. **Proceedings...** Renaissance, 2007.

PIWOWAR, H. A.; DAY, R. B.; FRIDSMA, D. S. Sharing detailed research data is associated with increased citation rate. **PLoS One,** v. 2, n. 3, 2007.

SILVA, J. R.; RIBEIRO, C.; LOPES, J. C. Managing multidisciplinary research data: Extending DSpace to enable long-term preservation of tabular datasets. In: IPRES CONFERENCE, 2012, Toronto. **Proceedings...** Toronto: University of Toronto, 2012. p. 105–108.

STRASSER, C.; CRUSE, P. The DMPTool and dataup: helping researchers manage, archive, and share their data. In: RESEARCH DATA MANAGEMENT IMPLEMENTATIONS WORKSHOP, 2013, Arlington. **Experience & Position Papers**. Arlington: University of Chicago, 2013. p. 5–8

SWAN, A.; BROWN, S. **The skills, role and career structure of data scientists and curators: an assessment of current practice and future needs**: report to the JISC. Key Perspectives, 2008.

WITT, M.; GIARLO, M. Databib: an online bibliography of research data repositories. In: ALA ANNUAL CONFERENCE, 2012, Anaheim. **Paper 2**. Anaheim: ALA, 2012.

DOI: 10.11606/issn.2178-2075.v4i2p95-111