Repositório de dados na U.PORTO: um fluxo de curadoria suportado numa extensão ao DSpace

Cristina Ribeiro
DEI-Faculdade de Engenharia da Universidade do Porto/INESC-TEC,
Rua Dr Roberto Frias, s/n, Porto, Portugal, mcr@fe.up.pt

João Rocha da Silva
DEI-Faculdade de Engenharia da Universidade do Porto/INESC-TEC,
Rua Dr Roberto Frias, s/n, Porto, Portugal, joaorosilva@gmail.com

Maria Eugénia Matos Fernandes
Reitoria da Universidade do Porto, Universidade Digital
Praça Gomes Teixeira, Porto, Portugal, efernand@reit.up.pt

João Correia Lopes
DEI-Faculdade de Engenharia da Universidade do Porto/INESC-TEC,
Rua Dr Roberto Frias, s/n, Porto, Portugal, jlopes@fe.up.pt

Resumo. É reconhecida a complexidade dos processos de produção de dados de investigação, e o interesse de os armazenar e descrever para que possam ser preservados e eventualmente reutilizados. Na Universidade do Porto, que gera uma parte significativa da produção científica nacional, este problema começa a ser tratado considerando tanto a experiência na comunidade internacional como as necessidades concretas dos investigadores. Com base numa iniciativa de auditoria de dados que lidou com uma amostra de grupos de investigação em vários domínios, foi proposto um fluxo de curadoria e produzido um protótipo de repositório para o suportar. O protótipo pretende responder a algumas das necessidade identificadas junto dos investigadores e oferece a possibilidade de os investigadores registarem conjuntos de dados, pesquisarem tanto sobre a sua descrição como sobre os conteúdos e de gerarem subconjuntos dos dados. Aproveitando o envolvimento dos investigadores com o repositório, os próximos passos serão de avaliação da abordagem, observando a utilização do repositório pelos investigadores, e o desenvolvimento de novas formas de interação com os investigadores.

Palavras-chave: Curadoria de dados, gestão de dados científicos, repositórios de dados.

1 Requisitos para a gestão de dados científicos

Os investigadores atribuem grande valor aos dados que recolhem ou que usam nos seus trabalhos, mas são em geral demasiado optimistas quanto à sua persistência e à possibilidade de os utilizarem no futuro. Ao contrário dos artigos publicados, em que eventuais restrições de acesso não partem dos autores, a divulgação dos dados nem sempre é favorecida pelos investigadores. As restrições podem ser devidas a questões de ética, de privacidade ou decorrerem de contratos. Por outro lado os investigadores sentem que a publicação dos dados lhes pode retirar vantagem na publicação de resultados. É por isso reconhecido que qualquer iniciativa de constituição de repositórios de dados que não esteja voltada para uma comunidade de investigadores estará fortemente limitada à partida.

Na Universidade do Porto (U.PORTO) a procura de soluções para a gestão dos dados científicos começou com a seleção de um grupos de investigadores de diversas áreas científicas cujos procedimentos de gestão de dados foram observados e descritos [1,2]. A auditoria de dados seguiu as recomendações internacionais, nomeadamente a metodologia proposta no Data Asset Framework [3]. Observaram-se práticas de gestão dos dados muito diferentes nas diversas disciplinas e foi possível identificar as funcionalidades mais importantes para um serviço de curadoria de dados [4]. Os investigadores mostraram interesse num serviço que lhes permitisse divulgar dados de forma seletiva, fazer pesquisas sobre os dados armazenados e exportar subconjuntos de dados.

2 Fluxo de curadoria e protótipo de repositório

Para satisfazer os requisitos identificados, foi desenhado um fluxo de curadoria de dados e desenvolvido um protótipo de repositório como uma extensão à plataforma DSpace. O fluxo de curadoria proposto inclui o depósito dos dados pelos investigadores, a intervenção de um curador que colabora com o investigador na descrição dos dados, a organização dos dados na forma de um conjunto de tabelas e a geração de um formato de preservação para o armazenamento do conjunto de dados no repositório [5].

Este processo é suportado num protótipo de repositório de dados. Algumas das funções necessárias são as existentes na plataforma DSpace, por exemplo o depósito de ficheiros e a criação de metadados ao nível do item. Um conjunto de dados é visto como um "Item" DSpace, com metadados gerais a este nível. A extensão desenvolvida para o repositório de dados fornece uma visão dos dados ao nível da tabela. Um conjunto de dados curado aparece no repositório como uma sequência de tabelas. Cada tabela tem metadados próprios e pode ser navegada com uma interface especial dentro do respetivo item. Para além da navegação o repositório tem funcionalidade de pesquisa e de exportação de subconjuntos de dados [6].

3 Conclusões e trabalho futuro

Na sequência do trabalho de auditoria de dados realizado na U.PORTO, foram identificados requisitos para um repositório de dados e proposto um fluxo de curadoria. Este fluxo pode ser a base de um serviço de curadoria de dados de investigação para a universidade. Para suportar as operações de transformação de dados e de criação de metadados foi desenvolvido um protótipo de repositório de dados que fornece algumas das funcionalidade identificadas junto dos investigadores. Uma experiência preliminar de curadoria permitiu povoar o repositório com alguns conjuntos de dados reais.

O trabalho realizado requer validação pelos investigadores, e essa é a primeira linha de trabalho futuro. O conjunto de investigadores que colaborou no projeto irá ser envolvido numa tarefa de avaliação incluindo mais ações de curadoria e a recolha de comentários dos investigadores ao fluxo proposto. Para facilitar a interação dos investigadores e lhes dar um maior controlo sobre os dados depositados, estão a ser estudadas novas interfaces para o carregamento de dados e para a sua anotação.

Referências

1. Cristina Ribeiro, Maria Eugénia Matos Fernandes. Data Curation at U.Porto: Identifying current practices across disciplinary domains. IASSIST Quarterly, 35(4):14–17, 2011.
2. Cristina Ribeiro, Maria Eugénia Matos Fernandes. Curadoria de Dados na Universidade do Porto: Identificação de práticas em diversas áreas disciplinares. 2ª Conferência Luso-Brasileira sobre Acesso Aberto, CONFOA 2011.
3. DAF Team: The Data Asset Framework Implementation Guide. http://www.data-audit.eu/
4. UPData Team: Project UPData. http://sciencedata.up.pt/doc/
5. João Rocha, Cristina Ribeiro, João Correia Lopes. UPData- a data curation experiment at U.Porto using DSpace. In 8th International Conference on Preservation of Digital Objects (IPRES 2011). iPRES, 2011.
6. UPData Team: Repositório de dados da U.PORTO (protótipo). http://sciencedata.up.pt/