João Rocha

Doctoral Programme in Informatics Engineering
Usage driven Application Profile generation using ontologies
João Miguel Rocha da Silva
—

Abstract

Nowadays, research is driven by ever increasing amounts of datasets and supported by more and more accessible computation power. This data is expensive to produce and very diverse due to its domain-specific nature. Such domain specificity calls for the participation of data creators in the description of their datasets, since they are the most knowledgeable regarding the meaning of their data.

Capturing each dataset's production context is thus necessary to enable their reuse, either by the creators or by other researchers within their research group, as well as from external groups. Researchers are not data management experts, though, so they need assistance in the production of their metadata records. The most common approach is to adopt a fixed set of descriptors for all research domains, but those descriptors are often not enough.

This research focuses on supporting the description of research datasets by the researchers themselves, using appropriate metadata. We started by gathering requirements in collaboration with a panel of researchers. The requirements, combined with our vision of an integrated data management workflow have led to the development of several solutions which have evolved towards a platform where data can be organised and described.

The first approach towards a research data management environment was a prototype repository designed for the deposit of datasets and based on DSpace (UPData). Shifting to an earlier moment in the data production workflow, we followed it up with two complementary solutions (UPBox and DataNotes) designed to support researchers in their regular data management tasks. These solutions introduced the two aspects identified as the most important during our requirements gathering: ease of use by non-experts and high-quality, appropriate metadata.

From the lessons learned from these developments we could outline, design and build the Dendro platform, a user-friendly research data management platform that allows researchers to deposit and describe their datasets. Behind the scenes is a graph-based data model supported by Linked Open Data. It is fully built on ontologies that can be directly drawn from the web or designed according to the metadata requirements of each research domain. We have also have developed a modeling process for these lightweight ontologies, having instantiated it several times during this work.

As more of these ontologies are introduced in Dendro, researchers can be easily overwhelmed by the increasing number of descriptors available for dataset description. To cope with this information overload and assist them in building their own application profiles, we implemented a recommendation module in Dendro. This module recommends descriptors suitable for different research domains and is driven by the usage patterns of users, taking different interactions into account to select the most appropriate descriptors. Our approach was tested with a panel of researchers from 11 different domains. Results show an increase in the usage of domains specific

descriptors, as well as an easier adaptation to the process of data description by these non-expert users, while maintaining the quality of the metadata records.

Resumo

O trabalho de investigação é atualmente baseado em conjuntos de dados cada vez mais numerosos e é suportado por capacidades computacionais cada vez maiores e mais acessíveis. A produção de tais dados é dispendiosa e a sua natureza é muito diversa, dependendo do domínio de investigação de onde esses dados provêm. Esta especificidade torna necessária a participação dos criadores na descrição dos seus conjuntos de dados, já que estes possuem um conhecimento profundo do significado dos dados que produzem.

A captura do contexto de produção de cada conjunto de dados é necessária para tornar possível a sua reutilização, tanto pelos criadores como por outros investigadores dentro e fora do grupo de investigação. Contudo, os investigadores não são peritos em gestão de dados, portanto necessitam de suporte à produção dos seus registos de metadados. A abordagem mais comum passa pela adoção de um conjunto fixo de descritores aplicado a todos os domínios de investigação, mas esses descritores são frequentemente insuficientes.

Este trabalho de investigação foca-se no suporte à descrição de conjuntos de dados de investigação por parte dos próprios investigadores, utilizando metadados adequados. Começou com uma recolha de requisitos em estreita colaboração com um painel de investigadores; estes requisitos, combinados com a nossa visão para um *workflow* integrado de gestão de dados, levaram ao desenvolvimento de diversas soluções que evoluíram no sentido de uma plataforma onde os dados podem ser organizados e descritos.

A primeira abordagem no sentido da criação de um ambiente de gestão de dados científicos consistiu num protótipo de repositório desenhado para depósito de dados e baseado na ferramenta DSpace (UPData). No sentido de introduzir a descrição de dados mais a montante no processo de criação de dados, duas soluções integradas foram também implementadas (UPBox e DataNotes). Estas soluções introduziram dois aspetos considerados muito importantes durante o processo de recolha de requisitos: a facilidade de utilização por parte de utilizadores sem conhecimento especializado e a produção de metadados apropriados e de qualidade.

As lições retiradas destes desenvolvimentos suportaram o desenho e construção da plataforma Dendro, um ambiente amigável para gestão de dados de investigação que permite aos investigadores depositar e descrever os seus conjuntos de dados. Em segundo plano existe um modelo de dados em grafo, assente em *Linked Open Data*. É completamente construído sobre ontologias que podem ser retiradas da web, ou desenhadas de acordo com os requisitos de metadados de cada grupo de investigação. Um processo de modelação destas ontologias *lightweight* foi também desenvolvido e instanciado por diversas vezes durante este trabalho.

Á medida que mais ontologias deste tipo são introduzidas na plataforma Dendro, mais difícil se torna para os investigadores escolher os descritores mais adequados para a descrição dos seus conjuntos de dados. Por forma a lidar com esta sobrecarga de informação e ajudar os investigadores a construir os seus próprios perfis de aplicação, foi implementado um módulo de recomendação na plataforma Dendro. Este módulo recomenda descritores adequados para a descrição de conjunto de dados de diferentes domínios, e baseia-se nos padrões de utilização dos diferentes utilizadores, tendo em conta diversos tipos de interações na seleção dos descritores mais apropriados. A abordagem foi testada com um painel de investigadores provenientes de 11 domínios de investigação distintos. Os

resultados obtidos mostram um aumento da utilização de descritores específicos do domínio, assim como uma adaptação mais fácil ao processo de descrição por parte destes utilizadores sem conhecimentos profundos de descrição e metadados, ao mesmo tempo que mantém a qualidade dos registos de metadados produzidos

Jury

- Chair: Prof. Eugénio da Costa Oliveira (FEUP, DEI)
- Members: Prof. Siegfried Handschuh (U. Passau), Prof. José Luís Brinquete Borbinha (U. Lisboa), Prof. Ana Alice Rodrigues Pereira Baptista (U. Minho), Prof. Álvaro Pedro de Barros Borges Reis Figueira (FCUP, DEI), Prof. Gabriel de Sousa Torcato David (FEUP, DEI), Prof. Maria Cristina de Carvalho Alves Ribeiro (FEUP, DEI)
- Supervisor: Prof. Cristina Ribeiro (FEUP, DEI) and Prof. João Correia Lopes (FEUP, DEI)
- Date: 18/05/2016