M. Gouveia

Master in Informatics and Computing Engineering
DataNotes— Um sistema colaborativo para anotação de estruturas de diretórios
Mariana Freitas de Gouveia
—

Abstract

Scientific research is increasingly based on the collection and use of significant amounts of data, which has led researchers to consider depositing them in data repositories. The goal of these repositories is the storage and preservation of datasets. One of the problems that usually arise is the difficulty experienced in interpreting these data. To overcome this issue, it is necessary not only to save the data but also to describe it. Descriptions can be specialised to varying degrees: they can be based only on generic descriptors such as title, date and creator, or include domain features.

The creation of rich descriptions for the datasets in a repository requires the collaboration of an information-management specialist who is responsible for creating those descriptions. In this scenario, as the researcher has little control over data descriptions, the process tends to be time-consuming. In the case of the storage and description of data in universities, there are some projects where the uploading process depends on these specialized staff, called “curators”.

This project aims to design and develop a collaborative annotation system to be used by researchers at the University of Porto. Using this system, they will be able to upload and describe their datasets themselves, using a set of tools to assist them in this process. They will also be able to describe, in free text, other observed facts that may be hard to fit into the available descriptors. Researchers are able not only to describe and update their own data but also data belonging to other researchers, provided they are authorised to do so.

Based on the Semantic MediaWiki platform and one of its extensions, Semantic Forms, this new extension was developed to ensure the functionalities of DataNotes. The platform was also integrated with another project called UPBox so that they can both be part of a complete research data curation process.

The initial objectives have been met, but there is plenty of opportunity for future expansion. Possible improvements include the development of a new extension on Semantic MediaWiki which would allow metadata schemas to be automatically imported into DataNotes, as well as semi-automatic annotation of data using the content of the datasets to be annotated.

Resumo

A investigação científica está cada vez mais baseada na recolha e exploração de quantidades apreciáveis de dados, o que leva os investigadores a considerar o seu depósito em repositórios. O objetivo destes repositórios é o armazenamento e preservação dos conjuntos de dados submetidos. Um problema que surge habitualmente é a dificuldade de interpretação destes dados. Para o ultrapassar, é necessário não só preservar os dados, mas também descrevê-los. As descrições podem

ser mais ou menos especializadas, ou seja, podem apenas basear-se em descritores genéricos, como o título, a data e o criador, ou incluir características próprias de um domínio.

Atualmente, os repositórios que possuem as descrições mais ricas supõem a colaboração de um técnico — o curador de dados — que é responsável pela descrição dos conjuntos de dados. Neste cenário, como o investigador tem pouco controlo sobre a descrição dos dados, o processo tende a ser demorado. No caso do depósito e descrição dos dados em universidades, existem alguns projetos em que o processo de depósito recorre à figura do curador.

Este trabalho tem por objetivos a conceção e desenvolvimento de um sistema de anotação colaborativa, DataNotes, destinado ao uso por parte dos investigadores da Universidade do Porto. Os investigadores terão a possibilidade de depositar e descrever os seus conjuntos dados autonomamente, recorrendo a algumas ferramentas que os auxiliam nesse processo. Poderão ainda descrever, em texto livre, outros factos observados que, através dos descritores disponíveis, não se conseguem anotar. Além dos investigadores terem a possibilidade de descrever os seus dados, podem também fazê-lo nos dados de outro investigador, desde que para tal estejam autorizados.

Com base na plataforma Semantic MediaWiki e uma das suas extensões, o Semantic Forms, foi desenvolvida uma nova extensão capaz de assegurar as funcionalidades do DataNotes. A plataforma foi também integrada com outro projeto denominado UPBox, para que possam ambos fazer parte de um processo de curadoria de dados científicos.

Os objetivos propostos foram cumpridos, embora haja algumas perspetivas de evolução futura. Possíveis melhorias incluem o desenvolvimento de uma nova extensão sobre a Semantic MediaWiki que permita que os esquemas de metadados sejam importados de forma automática para o DataNotes, assim como obter uma anotação semiautomática dos dados, através dos conteúdos dos conjuntos de dados a ser anotados.

Jury

- Chair: Rui Carlos Camacho de Sousa Ferreira da Silva
- External Examiner: José Luís Brinquete Borbinha
- Supervisor: João António Correia Lopes
- Date: 8/2/2013