Artigo
de Dados

Encontros Bibli

PYTHON SCRIPTS FOR WEB SCRAPING METADATA FROM DESCRIPTIONS ABOUT THE DATASETS OF THE INTERNATIONAL SCENARIO OF RESEARCH DATA REPOSITORIES

Python scripts para o web scraping de metadados das descrições sobre os conjuntos de dados do cenário internacional de repositórios de dados de pesquisa

Alexandre Ribas Semeler
Universidade Federal do Rio Grande do Sul,
Porto Alegre, RS, Brazil
Alexandre.semeler@ufrgs.br
https://orcid.org/0000-0002-8036-4271

Arthur Longoni Oliveira
Universidade Federal do Rio Grande do Sul,
Porto Alegre, RS, Brazil
arthur.holiver@gmail.com
https://orcid.org/0000-0001-6916-8934

Policarpo Camilo Silvestre Matiquite
Universidade Federal de Santa Catarina,
Santa Catarina, SC, Brazil
Universidade Eduardo Mondlane, Maputo,
Republica Popular de Mocambique
cmatiquite@gmail.com
https://orcid.org/0000-0001-5374-1900

Fabiana Andrade Pereira
Universidade de São Paulo,
São Paulo, SP, Brazil
Fundação de Amparo à Pesquisa de São Paulo,
São Paulo, SP, Brazil
fpereira@fapesp.br
https://orcid.org/0000-0002-3779-222X

A lista completa com informações dos autores está no final do artigo

ABSTRACT

Objective: Research data repositories are an evolution of document repositories that aim to access and preserve all materials used before, during, and after scientific research. In this context, this study aims to conduct an exploratory and descriptive investigation of the international scenario of data repositories by monitoring the descriptive metadata of the international register of this type of repositories in the Registry of Research Data Repositories (re3data.org).
Methods: The process requires applying knowledge inherent to the techniques and technologies used for descriptive data analysis, information retrieval, manipulation, analysis, and data visualization. Consequently, three scripts in Python 3.11 are provided for collecting metadata from re3data and scripts and converting the metadata to enable visualization in software such as VOSviewer, a dataset with metadata descriptions of repositories and conversions for visualization of networks. The datasets produced in this study can be found in the ZENODO Data Repository (https://doi.org/10.5281/zenodo.7903109). In a collection on (05/05/2023), 3108 links to the repository descriptions were retrieved. Data and scripts were created for this methodological experiment and shared at (DOI: doi.org/10.5281/zenodo.7903109). The dataset contains a root directory with three subdirectories: (scripts) with (.py) Python codes, another directory called (data) with textual files containing tab-separated values (.TSV), and the file (Information Systems Research, RIS). The third directory (env) contains the Python libraries required to run the scripts.
Potential for reuse: The research method applied to manipulate this dataset is based on automated re3data metadata extraction and network visualization; after the data collection and analysis process, it is possible to trigger a study based on the descriptions extracted from the Registry of Research Data Repositories (re3data), researchers can visualize the international scenario of research data repositories, verified by re3data, which allows ethical monitoring of the number of research data repositories that are registered in re3data, what are their areas, institutions, countries, the language of research data, the typology of repositories and deposited data, their themes, areas of knowledge, types of access, licenses and software used. In addition, other issues can be raised while interpreting the data. The community of Librarianship and Information Science professionals need to share data and the extraction technique these research data. Finally, it can be concluded whether information about research data repositories allows us to state that they are heterogeneous data sources that enable access and preservation of a wide range of research data types.
KEYWORDS: Data Repository. Research Data. Geosciences. Re3data. Python. Scripts. Web Scraping.

RESUMO

Objetivo: Os repositórios de dados de pesquisa são a evolução dos repositórios de documentos e visam acessar e preservar todos os materiais usados antes, durante e depois da realização pesquisa científica. Nesse contexto, o objetivo deste estudo é realizar uma abordagem exploratória e descritiva do cenário internacional de repositórios de dados de pesquisa, por meio do monitoramento dos metadados descritivos do registro internacional desse tipo de repositórios no Registry of Research Data Repositories (re3data.org).

Métodos: O desenvolvimento do método exigiu a aplicação de conhecimentos inerentes às técnicas e tecnologias utilizadas para análise descritiva de dados, recuperação de informações, manipulação, análise e visualização de dados. A aplicado ao método resulta em três scripts em Python 3.11 para coleta de metadados do re3data, scripts para conversão de metadados e scripts para visualização dos metadados em softwares como o VOSviewer. Os conjuntos de dados produzidos pela pesquisa pode ser encontrados no repositório de dados ZENODO (https://doi.org/10.5281/zenodo.7903109), em uma coleção de software depositada em (05/05/2023), nela foram recuperados 3108 registros de links para descrições de repositórios distribuídos internacionalmente. Conforme o experimento metodológico o conjunto de dados contém um diretório raiz com 3 subdiretórios, um chamado (scripts) com os códigos Pyhton (.py), outro diretório chamado (data) com os arquivos textuais **(***Tab-separated values,TSV*) contidos e o arquivo *(Information Systems Research, RIS*). O terceiro diretório (env) é onde estão as bibliotecas Python necessárias para executar os scripts.
Potencial de reutilização: O método de pesquisa aplicado para manipular este conjunto de dados é baseado na extração automatizada de metadados do re3data e na visualização de redes; após o processo de coleta e análise dos dados é possível desencadear um estudo exploratório e descritivo sobre o cenário internacional dos repositórios de dados de pesquisa, verificados pelo re3data, o que permite o monitoramento ético da quantidade de repositórios de dados de pesquisa que estão cadastrados no re3data, quais são suas áreas, as instituições, os países o idioma o idiomas dos dados da pesquisa, a tipologia dos repositórios e dos dados depositados, suas os temáticas, áreas do conhecimento, tipos de acessos, licenças e softwares utilizados. Além disso, outras questões podem ser levantadas durante a interpretação dos dados. O que reforça a necessidade desse conjunto de dados para a comunidade de profissionais da Biblioteconomia e da Ciência da Informação, o compartilhamento de dados e a técnica de extração podem colaborar com o reaproveitamento desses dados de pesquisa.
PALAVRAS-CHAVE: Repositório de Dados. Dados de Pesquisa. Geociências. Re3data. Python. Scripts. Web Scraping.

1 INTRODUCTION

The emergence of data repositories, which are technological and organizational information systems that assist researchers in managing and curating their research data, and the proliferation of collections of digital research data have recently drawn considerable interest. Research data repositories have become a current trend in librarianship and information sciences research. These repositories are an evolution of document repositories and exhibit the elementary function of facilitating the search for and access to research data. They constitute an essential part of scientific research cyber-infrastructure and aim to preserve long-term visas and reuse research data. Therefore, these strategies must be planned and structured from the beginning of their implementation.

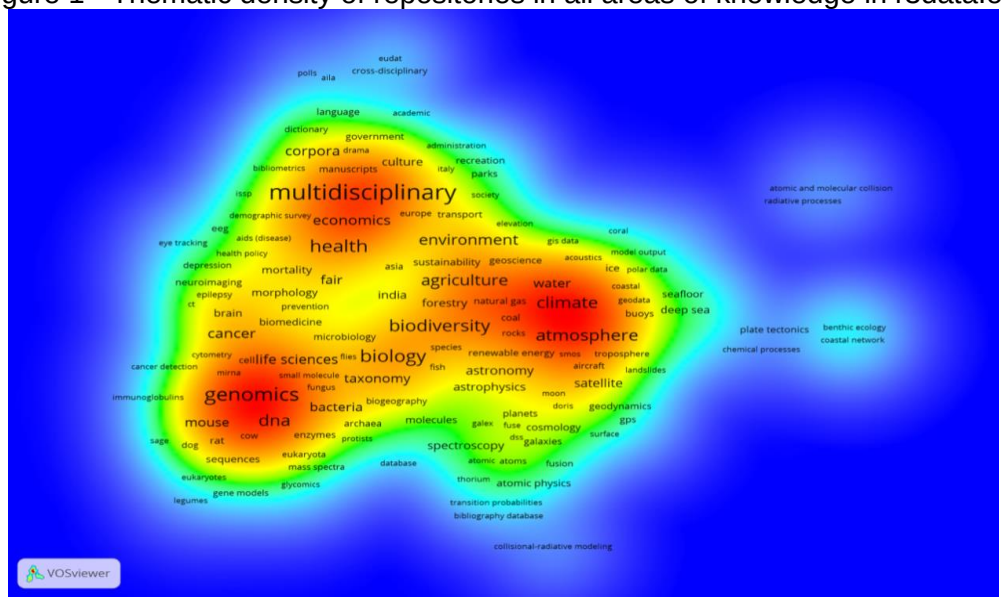
The global distribution of research data repositories is cataloged by the Registry of Research Data Repositories (re3data), an international registry of research data repositories. Founded by the German Research Foundation (DFG), re3data brings together initiatives from the Library and Information Services (LIS) of the GFZ German Research Center for Geosciences, Library of the Karlsruhe Institute of Technology (KIT), School of Library and Information Science (BSLIS) at Humboldt-Universität in Berlin, and Libraries of the Purdue University in Germany. re3data.org indexes approximately (3108, May. 2023) research data repositories internationally as of the fall of 2012. It provides researchers, funding organizations, libraries, and publishers with a systematic overview of the heterogeneous landscape of research data repositories.

Re3data spans different academic disciplines and presents a list of repositories that provide permanent storage and access to research datasets, in which funding bodies,

publishers, and educational institutions promote a culture of sharing, access, and visibility of research data (RE3DATA, 2012).

Based on the context above, a consolidated environment exists where these research data repositories are registered in the most diverse areas of knowledge and globally distributed. Regarding the thematic scope, the holdings are multidisciplinary. Figure 1 shows the main subjects of the research data repositories.

Figure 1 - Thematic density of repositories in all areas of knowledge in r3data.org



Source: Prepared by the authors (2023)

In this context, the data are valuable because they can, as Figure 1 shows, reveal the three main areas of the thematic density of research data available on the international scene. The first explores a cluster linked to biological and genetic issues, the second relates to the exact and Earth sciences, and the third deals with multidisciplinary problems. The density zones have four primary colors (blue, green, yellow, and red). The themes in blue are less recurrent, and those in the green range are more recurrent issues than the previous ones (prospecting areas). The yellow zone indicates the central theme linked to the cluster core (red). Research data on water, climate, and astrophysics are available in the thematic heart. In this context, this study aims to describe the metadata describing the distribution of these repositories.

3108 research data repositories were registered in (May. 2023). The international composition of these repositories can be mapped using re3data. Another factor that demonstrates this data set's usefulness is its ability to answer questions such as: How many research data repositories are registered in re3data, and what are their areas? What are the leading institutions? In which countries and languages are the survey data available? What

is the typology of a repository? What types of data were deposited? What are the themes and areas of knowledge? What types of accesses and licenses are used? What kind of software has been used? In addition, other questions may be raised during data interpretation. These issues underscore the complexity and originality of this study in terms of analyzing research data repositories using scripts and techniques to retrieve network information. In brief, a group of Python 3.11 scripts is proposed for web scraping and visualization of metadata collected from the re3data API interface.

In this context, the Librarianship and Information Science professional community must share data and the extraction technique to collaborate on reusing these research data. The potential for reuse to manipulate this dataset is based on automated re3data metadata extraction and network visualization; after the data collection and analysis process, it is possible to trigger a study based on the descriptions extracted from the Registry of Research Data Repositories (re3data), researchers can visualize the international scenario of research data repositories, verified by re3data, which allows ethical monitoring of the number of research data repositories that are registered in re3data, what are their areas, institutions, countries, the language of research data, the typology of repositories and deposited data, their themes, areas of knowledge, types of access, licenses, and software used. In addition, other issues can be raised while interpreting the data.

Finally, the additional potential of this dataset reuse is to create investigative services for monitoring and prospecting the technological environment and organizations that offer research data, mapping in the global scenario that provides services related to research data management by libraries; and developing a system for monitoring and retrieving scientific information (research data) within the global scenario data repositories.

In summary, the re3data platform can be used to map the international composition of these repositories, moreover, used to answer questions about the leading institutions, countries, languages, typology of repositories, types of data deposited, and more. The re3data platform can be used. Librarianship and Information Science professionals must collaborate to share data and extraction techniques for reusing research data. The data can be used to create investigative services for monitoring and prospecting the technological environment and organizations offering research data.

2 METHODOLOGY AND RESULTS

The research method applied to manipulate this dataset is based on automated re3data metadata extraction and network visualization; after the data collection and analysis

process, it is possible to trigger a study based on the descriptions extracted from the Registry of Research Data Repositories (re3data), researchers can visualize the international scenario of research data repositories, verified by re3data, which allows ethical monitoring of the number of research data repositories that are registered in re3data.org.

We utilized the API (http://www.re3data.org/api/<apiidentifier>) provided by re3data. It is crucial to note that re3data offers complete or partial content recovery via API. Currently, the platform implements a simple open search and a representational state transfer (REST) interface for data extraction. Upon querying re3data, a list of links was generated, each containing the address of the repository descriptions. It is essential to highlight that the schema employed by re3data to describe repositories is in version 3.1 of May 2023, which provides the structure for data analysis. This pattern is accessible in XML and features metadata encompassing the overall scope, content, infrastructure, technical, and quality standards for data repositories metadata descriptions.

The metadata available in this type of link represents the structure of the collected data and is used to analyze the global scenario of data stores. This standard is open to XML and contains metadata properties regarding research data repositories' general scope, content, infrastructure, technical, and quality standards.

Once collected, the data must be manipulated to meet the preparation requirements for analysis. It is necessary to incorporate data refinement software that performs cleaning, recoding, and merging of datasets and imports and exports data from one type of software to another without losing content. Analytical datasets can be generated using OpenRefine, which loads, cleans, reconciles, categorizes, and converts data from one format to another. It is not a web service but a desktop application used to process data. It is applied to ensure consistency in the data.

The use of OpenRefine follows these steps: a) import the data set, b) model the columns for analysis, c) verify the consistency of the data, d) apply a clustering algorithm to look for words that are spelled differently, such as “New York” and “New York.” OpenRefine allows the use of the key-collision method and fingerprint function, e) generates datasets with the frequency of occurrence of the contents contained in the dataset, and f) converts the dataset for analysis and visualization. Summarizing the data and creating different tables and views after refinement is possible using OpenRefine. Notably, the refinement process allows summarization based on the description of frequencies and mining of the texts that constitute the dataset collected in re3data. Typical data-mining tasks include classification (e.g., constructing a decision tree), clustering, regression, summarization, and association

rule learning. These are all based on simple tabular data techniques, where rows correspond to instances and columns correspond to variables. After conducting such research procedures, it was possible to summarize the international scenario of research data repositories. The collected data and scripts were organized according to the Research Data Management Plan (PGD) and are available under the CC BY 4.0 license in the ZENODO repository (https://doi.org/10.5281/zenodo.7903109).

3 SPECIFICATION TABLE

The dataset, tab.1, contains all dataset scripts description.

Table 1: dataset scripts descriptions

Area of knowledge	B. Information use and sociology of information
Specific subject área	Research data repository
Language	English
Data type	Number, Text, Scripts
How the data was acquired	Metadata web scraping of descriptions re3data.org Schema 3.1 XML Schema DOI: http://doi.org/10.48440/re3.011. API links: https://www.re3data.org/api/v1/repository/r3d100000002
Data state	Secondary
Parameters for data collection	Name, Description, URL, Software Name, License, Data Access, Type of Repository, Type of Content, Type of Repository Language, Institution, Country, Institution Type, Keyword, and Subject
Description of data collection	Data collection selects information about the international composition of these repositories through Python 3.11 scripts. The API (http://www.re3data.org/api/) available on re3data was used as a source. re3data offers recovery of all or part of content via API. Currently, the platform offers a simple open search implementation and representational state transfer (REST) interface for extracting data.
Data source location	Record_base = 'https://www.re3data.org/api/v1/repository/' Repository_list = 'https://www.re3data.org/api/v1/repositories'
Data accessibility	Repository name: ZENODO Data identification number: https://doi.org/10.5281/zenodo.7903109

Source: Prepared by the authors (2023).

4 DESCRIPTION OF THE DATASET

The datasets produced in this study can be found in the ZENODO Data Repository (https://doi.org/10.5281/zenodo.7903109). In a collection on (05/05/2023), 3108 links to the repository descriptions were retrieved.

The dataset contains, in the root directory, a readme.txt file with explanations of use, versioning, contact with the authors of the data, and technical descriptions for performing the analysis. The root folder contains three directories: (scripts), with (.py) Python codes for the parser, data collection, and conversion. Another guide, called (data), with textual files (.tsv) containing a list of API links to the re3data repositories, descriptions, and files (.ris), with the nodes and frequencies of keywords and institutions related to the repositories described by the re3data. The third directory (env) was the Python library needed to run scripts.

In the folder (scripts), the scrapingRe3Data.py file collects the links and names of the re3data repositories and follows the following logic: lists the repositories to be extracted, copies all the HTML codes of the page, open a file where it will be written, copies the headers of the copy file of the name and metadata links of the metadata descriptions of the re3data repository, traverses the tree of these data in their respective XML in re3data, and writes the file in (.tsv), and closes the file.

The second script, ScrapMetadataRe3dataGen.py, extracts the data cataloged by re3data (Name, Description, URL, Software Name, License, Data Access, Type of Repository, Type of Content, Type of Repository Language, Institution, Country, Institution Type, Keyword, and Subject). The script parserRedeTSVRISRe3dataGen.py converts the Re3Data_repositories.tsv file from the (.tsv) to (.ris).

The Re3Data_repositories.tsv file is the database describing the 3,108 repositories. The data repository, type of content, language of the data, responsible institution, country, and subjects of the data are made available by the repository. In the data folder (Data), the RepoListAPI.tsv file contains the decisions for a list of links to the metadata and names of the repositories. Occasionally, this directory may generate a log file of errors if they occur.

In summary, to reuse the dataset described in this data paper, it is necessary to download the dataset from: (https://doi.org/10.5281/zenodo.7903109) and run the scripts in the following order:

1 - **scrapingRe3Data.py** (this script generates a .tsv file (RepoListAPI.tsv with links to each record on re3data.org);

2 - **scrapMetadataRe3dataGen.py** (this script generates a .tsv file (Re3Data_repositories.tsv) with descriptions of each record in re3data.org);

4 - **parserRedeTSVRISRe3dataGen.py** (this script generates a .RIS file (rede.ris) with the graph for visualization in VOSviewer software.

REFERENCES

RE3DATA. Retrieve from: http://www.re3data.org/about (Accessed: may 2023).

NOTAS

CONTRIBUIÇÃO DE AUTORIA
Concepção e elaboração do manuscrito: Semeler, A. R., Oliveira, A. L. Matiquite, P.C.S, Pereira, F. A.
Coleta de dados: Semeler, A. R., Oliveira, A. L.
Análise de dados: Matiquite, P.C.S, Pereira, F. A.
Discussão dos resultados: Semeler, A. R., Oliveira, A. L. Matiquite, P.C.S, Pereira, F. A.
Revisão e aprovação: Semeler, A. R.,

FINANCIAMENTO
Não se aplica.

CONSENTIMENTO DE USO DE IMAGEM
Não se aplica

APROVAÇÃO DE COMITÊ DE ÉTICA EM PESQUISA
não se aplica.

CONFLITO DE INTERESSES
Não se aplica

PUBLISHER
Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no Portal de Periódicos UFSC. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES
Edgar Bisset Alvarez, Ana Clara Cândido, Patrícia Neubert, Genilson Geraldo, Mayara Madeira Trevisol, Jônatas Edison da Silva, Camila Letícia Melo Furtado e Beatriz Tarré Alonso.

HISTÓRICO
Recebido em: 10-06-2023 – Aprovado em: 18-07-2023 - Publicado em: 04-08-2023.