OR2021 Presentation Submission
16th International Open Repositories Conference, June 7th -10th

On building a tool for finding datasets based on a list of researchers or publications

Washington L. R. Carvalho-Segundo, IBICT, washingtonsegundo@ibict.br;

Thiago M. R. Dias, CEFET-MG, thiagomagela@cefetmg.br .

Abstract: *This proposal presents a tool developed in the Python language used to find related datasets of a list of researchers or publications. This tool was applied to a list of articles that a specific group of researchers had declared in their CVs. The target group was chosen based on the highest level that these researchers had obtained in a research productivity grant (1A). As a result, form a list of 1,227 researchers and more than 225 thousand deduplicated publications, it was possible to find 12,030 related datasets, were the most frequent access type is OPEN and the five most frequent related areas of research are Zoology; Chemistry; Genetics; Physics; and Agronomy. The proposed tool will be applied to facilitate populating the research data repository of the national funding agency in Brazil, but it can also be used in other more general contexts, extracting information from open databases, such as ORCID and Wikidata.*

Keywords: Open Science, Scientific Data Repositories, Scientific Publications, Open Data.

Audience: The availability and accessibility of scientific data in open access has been the subject of research that starts to circulate and reach a wide audience inside and outside the academy. Researchers, Editors of scientific journals, Managers of journals portals and Institutional Repositories, in addition to Librarians, are definitely an audience for the presented proposal.

Introduction: Open Science proposes free access to scientific information, as well as research data, providing greater transparency in scientific methods and promoting scientific collaboration. In this context, Open Science has as one of its objectives to facilitate the sharing and reuse of data by the scientific community.

The research data, which cover various stages of investigations such as experiments, empirical studies or observations of natural phenomena, as well as the publication of the results obtained, are characterized as an important element in the context of Open Science. Such data, in general, are arranged in different levels of aggregation, different formats and documentation in different layouts, which makes it difficult to access and especially to reuse. Currently, several efforts have been made in order to make research data more accessible and, consequently, reusable, which will definitely provide significant advances in scientific research.

In line with the guidelines of Open Science, several initiatives have been employed in order for researchers to guarantee accessibility to the data generated within the scope of their research projects. Several research funding agencies have been demanding from their financiers that data originating from their research be published within a reasonable time, in accessible formats and within widely adopted standards, especially those defined by public repositories. In this context, an important initiative of a research data repository directory is the re3data.org[1] portal, which indexes the information of more than 2,620 data repositories.

Therefore, understanding how researchers have been publishing the data resulting from their research is essential to obtain an overview of the current scenario regarding data publication, especially those researchers who are references in their areas of expertise. Due to the representativeness of their research, the identification of this panorama can provide incentives for other researchers to also publish their data sets, as well as to reuse the data published in other research.

[1] https://www.re3data.org

In Brazil, a group of researchers with high scientific production are recognized for excellence in their research, receiving assistance for the maintenance of their scientific production. The CNPq research productivity grant (National Council for Scientific and Technological Development) is intended for researchers who stand out among their peers in carrying out research in the scientific and technological areas. Thus, CNPq awards researchers according to their scientific production and training of human resources with a monthly scholarship that varies according to the category / level of the same. The Research Productivity Scholarship has three categories, 1, 2 and Senior, and category 1 has four levels, in ascending order: 1D, 1C, 1B and 1A. Each level provides an increasing monthly fee, with a greater jump between category 2 and level 1D.

Level A ("1A Productivity Research") is reserved for candidates who have shown continued excellence in scientific production and training of human resources, and who lead consolidated research groups. The profile of this level of researcher should, in most cases, extrapolate only the aspects of productivity to include additional aspects that show significant leadership within his research area in Brazil and the ability to explore new scientific frontiers in risky projects. Therefore, in this work, the entire set of scholarship holders in CNPq research productivity in modality 1A was selected because they are considered the elite of Brazilian researchers.

Methodology[2]: Initially were collected a relation of 1,227 research identifiers in the Brazilian National CV Platform, the Lattes Platform, using the filter "1A Productivity Research". From this research id list it was used the "**Publications Extractor**" module for obtaining a deduplicated list of publications that these researchers declare in their respective CVs in the Lattes Platform.

In the module "**Finding Datasets Tool**", each publication in the list provided by the Publication Extractor module is queried in the publications API of OpenAIRE[3] in two different ways: 1) if the publication has a DOI, the search is from this number. If the publication with the referred DOI is not found in the OpenAIRE database, the tool performs a second query that is described in the next sentence; 2) if the publication does not have a DOI or the search for the provided DOI does not return results, it is performed a query by title in OpenAIRE publications API. For each obtained result it is checked if the publication year matches with the year provided in the input publications list. For each publication that has the year matched it is calculated the *Levenshtein distance*[4] between the result title and title that is provided by the input publications list. If the distance is less than 10, it is considered that one has a publication match. From this, if the matched result has a persistent identifier it is stored in the corresponding place in the publications list.

From the list of publications that were found in the OpenAIRE database and that has a persistent identifier, one performs a query in the OpenAIRE Scholix API[5], with the parameter targetPid=$pid, where $pid is replaced by the persistent identifier in the publications list. Is necessary to add an extra parameter (&sourceType=dataset) in the query to bring only the relations that the publication specifically has with datasets. The relationship of type "Reference", that means that the publication only cites the dataset, is filtered and discharged. Also datasets that do not have declared authors are not considered.

The obtained datasets list is confronted with the list of related publications, checking if at least one of the names of the identified authors in the publication is in the list of authors of the corresponding dataset. This is also performed using the Levenshtein distance and the tolerance is a value at most

[2] The tool described in this section was developed in Python 3.9 and it can be downloaded from https://zenodo.org/record/4513945, https://doi.org/10.5281/zenodo.4506826, or https://github.com/brcris0/findingdataset .
[3] http://api.openaire.eu/search/publications
[4] https://en.wikipedia.org/wiki/Levenshtein_distance
[5] http://api.scholexplorer.openaire.eu/v2/Links?targetPid=$pid&sourceType=dataset

equal to 3. The datasets that do not have an author match are discharged. Also is captured a DOI of each resulted dataset in the list, if it exists.

Finally, having the datasets list in the belt one performs a query against the OpenAIRE datasets API[6] to check the *best access type* of the dataset. With this, one obtains the output of the tool. One final step is performed with a backtrack search of the main research areas that authors declare in their CV. This is useful to try estimating the research area in which the dataset was produced. The diagram of Figure 1 illustrates all the described processes.

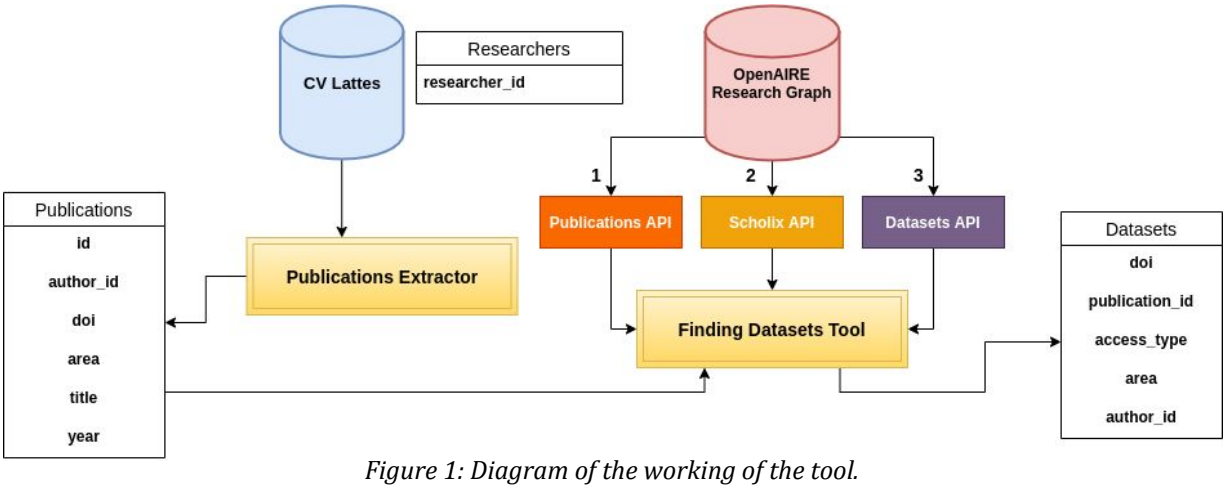

Figure 1: Diagram of the working of the tool.

Results: The Publications Extractor provided a list of **227,640 publications**. **141,790 of these were found in the OpenAIRE Research Graph**[7]. From the last set, **12,030 related datasets** were identified whose at least one of the authors of the observed dataset matches with an author of the related publication. From this, it was possible to verify in the OpenAIRE API, for **11.542 datasets**, that the access type is **OPEN in 66.80% of the cases**, while it is **UNKNOWN**, **EMBARGOED** and **CLOSE in**, respectively, **33.15%, 0.03% and 0.02%** of the cases. Regarding the identified main research area of the authors of the datasets, the top ten occurrences were: **Zoology (17,22%)**; **Chemistry (10,01%)**; **Genetics (9,42%)**; **Physics (9,31%)**; **Agronomy (6,47%)**; **Medicine (6,29%)**; **Ecology (4,72%)**; **Biochemistry (4,65%)**; **Veterinary Medicine (2,53%)**; and **Collective Health (2,46%)**.

Conclusion: This *finding datasets tool* was developed in the scope of two in-progress projects, a project for the construction of the scientific data repository of CNPq, named Lattes Data, and of the project for the construction of a CRIS, Current Research Information System, at the national level, in Brazil. This last is named BrCris. The tool can be applied not only to the Brazilian context but also to an arbitrary list of scientific publications. It is only necessary to have a publications list with four columns: 1) title; 2) publication date; 3) authors; and 4) persistent identifier (optional), such as DOI. It also can be adapted to be applied directly to a list of researchers that are present in an open search database, such as ORCID, Microsoft Academic Graph, Open Citations, Wikidata, among others. Future work is considered to build an application module that gets the output dataset list and performs a download of the datasets and their corresponding metadata to be automatically uploaded to the Lattes Data Repository. Moreover, the entities present in the processed metadata, such as Person, Publication, Dataset and Organization, and their relationships will be aggregated in the BrCris database.

References:

MATAS, Lautaro J.; DIAS, Thiago M. R.; CARVALHO-SEGUNDO, Washington L. R. *Improving LA Referencia metadata by linking research profiles to repositories: the case of the Brazilian Digital Library of Thesis and Dissertations (BDTD) and the Lattes CV Platform*. Open Repositories 2019.

DIAS, Thiago Magela Rodrigues; MOITA, Gray Farias. *A method for the identification of collaboration in large scientific databases*. Em Questão, v. 21, n. 2, p. 140-161, 2015.

[6] http://api.openaire.eu/search/datasets
[7] https://graph.openaire.eu