5. OPEN DATA OF THE LATTES PLATFORM ACCORDING TO FAIR PRINCIPLES: EXAMPLES OF EXTRACTOR AND UFSC INFORMATION OBSERVATORY

Adilson Luiz Pinto[120]
Thiago Magela Rodrigues Dias[121]
Fábio Lorensi do Canto[122]
Washington Luís Ribeiro de Carvalho Segundo[123]

5.1 INTRODUCTION

The term *Open Access* (OA) is a concept related to the free access to scientific information on the Internet, especially peer reviewed scientific papers or papers published in a specialized scientific magazine. Open access is based on the premise that the scientific research is mostly financed with public resources; therefore, its results should be available and accessible with no cost to society. It considers that researchers do not write for financial reasons, but to maximize the visibility, use, and impact of their research results. The open access movement also argues that, although the process of editing and disseminating an article involves costs, these must be incorporated into the general costs of the research and not passed on to readers. (Shavell, 2010; Freire, 2011).

In the late 1990s, there were several demonstrations in favor of open access. Among the reasons that drove the creation of this movement, the scientific journal's price crises stands out, a phenomenon that limited or prevented the access to scientific information from countries and institutions lacking resources to pay for subscriptions and licenses. Alternatives were sought to provide a broader access, forming consortia to acquire content to be available in portals and databases (Fladung, 2007).

With the development of tools for building repositories and databases, the open access model gains consistency. Therefore, several declarations in favor of this model are published, intensifying the implementation of a basic infrastructure of open access in national and international levels (Kuramoto, 2006).

Open access drives the return of efforts put on researches with public investment, making the results more accessible. Regardless of the meanings the term contains, open access must be discussed based on different aspects, among which access to literature or the knowledge in it stands out (SUBER, 2007). It is important to note

120 Author details: PhD in Documentation, Professor and Researcher at the Federal University of Santa Catarina (PGCIN/UFSC), adilson.pinto@ufsc.br

121 Author details: PhD in Mathematical and Computational Modeling, Professor and Researcher at the Federal Center for Technological Education of Minas Gerais, thiagomagela@cefetmg.br

122 Author details: PhD student in Information Science at PGCIN/UFSC, fabio.lc@ufsc.br

123 Author details: PhD student in Computer Science (UnB), Researcher at the Brazilian Institute of Information in Science and Technology, washingtonsegundo@ibict.br

that open access to scientific knowledge refers to both formal and informal aspects of the scientific communication process (Leite, 2016).

Recently, open access guidelines have also been applied to data management plans, since those are instruments that guide practices to promote accessibility and reuse of research data. To make data more findable, accessible, interoperable and reusable, FAIR principles are used, an acronym for *'Findable', 'Accessible', 'Interoperable' and 'Reusable'* (Veiga *et al.*, 2019).

Currently, FAIR principles are considered the guiding elements for good practices in the whole research data management process. They aim at implementing a metadata set defined to be used by both automated computational mechanisms and by people. If properly adopted, FAIR principles facilitates the interoperability among different data environments (Henning *et al.*, 2019).

FAIR principles have been part of discussions and contemporaneous practices of data science since the beginning of 2014. They had their application consolidated in 2017, when the European Commission required that data management plans be adopted based on those principles in projects funded by its resources. Since then, the principles have been used to guide the discovery, access, interoperability, sharing, and reuse of research data (Henning *et al.*, 2018).

In Veiga *et al.* (2019) it is possible to find a summary of the FAIR principles:

Findable: (a) data and metadata need to have a single persistent identifier; (b) data should be described with rich metadata; (c) have the persistent identifier for the dataset described in the metadata, and; (d) metadata and data must be retrievable through trustworthy repositories;

Accessible: (a) data and metadata must be retrieved by its identifier using standard communication protocols; (b) the protocols must be free, open and support authentication and authorization, and; (c) metadata must be accessible even when data is not available anymore.

Interoperable: (a) data and metadata must be coded using agreed standards of representation, and; (b) data and metadata must use vocabulary aligned to FAIR principles and include relevant references.

Reusable: (a) data and metadata need to be associated to relevant attributes; (b) data and metadata must be released with use license clearly defined; (c) metadata and data must be associated to their origin in a detailed way, and; (d) data and metadata must meet the community standards.

From the consolidation of these four principles in the context of open data and open science, it is presented one of the forms through which the Federal University of Santa Catarina (UFSC) have used data available in Lattes Platform in its internal management systems.

5.2 METHODOLOGY

The data source defined for this work was the UFSC faculty résumé database registered in Lattes Platform (2,581 UFSC permanent faculty in May 2020). Résumé in the Lattes system were chosen because they have a lot of information. It is a resource that allows the integration of scientific, professional and academic data, and data update can be done permed by research. Among the main information in the résumé, academic qualification, research field, professional performance and academic orientation are those that stand out, in addition to technical and scientific productions.

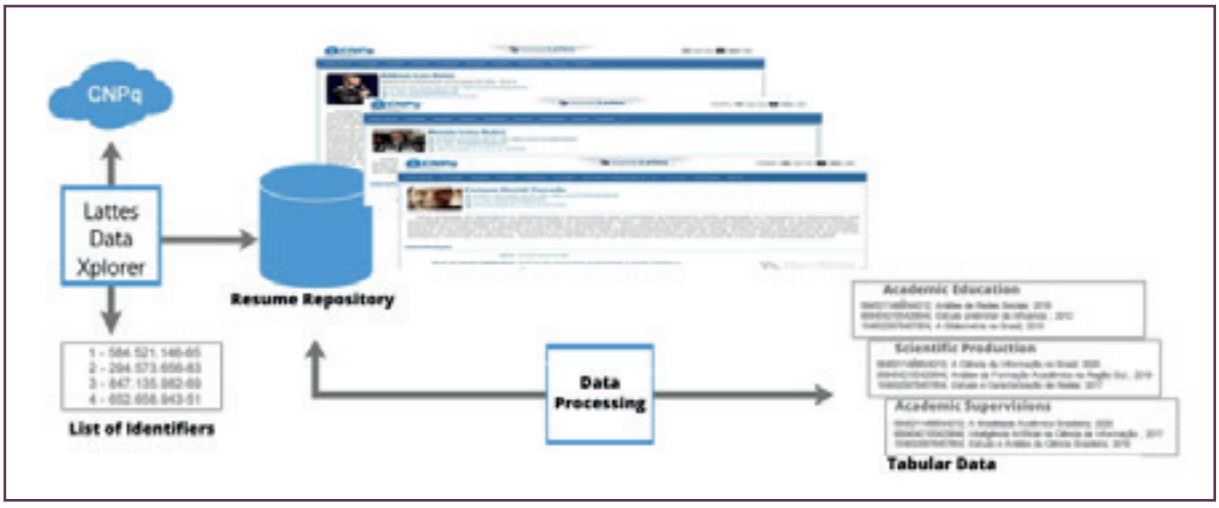
The Résumé in Lattes system became a national standard used for individual evaluation of scientific and academic activities. They add data from researchers from all fields of knowledge, making the platform a relevant source for analysis and understanding of research groups behavior (Digiampietri *et al.*, 2012).

Although the data from the résumé is freely available, they can only be viewed individually through a query interface provided by the CNPq. However, this interface is limited with no possibility of grouping, analysis and comparisons with other résumé. That said, techniques and tools for extracting data are necessary for analyzing large sets of curriculum data.

LattesDataXplorer framework (Dias, 2016) was used for extracting and treating data. It is a tool developed with the aim of collecting and treating curriculum data from Lattes Platform, with low computational cost.

LattesDataXplorer is responsible for encompassing the whole set of techniques and methods for collecting, treating and analyzing data used herein. The extraction is performed by a component that searches and retrieves each faculty résumé from the single indicator in their résumé in Lattes Platform. Consequently, with all résumé stored in XML format, the institution handles its data in UFSC information Observatory (https://observatoriodainformacao.ufsc.br/indicadores-cnpq/ufsc/) making data management possible for the Dean of Research and Graduate Studies of the Institution.

Figure 1 - Model of the system used by UFSC for extracting and treating data coming from the Lattes Platform



Source: Research data (2020).

With *LattesDataXplorer* it is possible to group a set of resumes based on predefined parameters. In the process of selecting résumé based on parameters, regardless of the section in which they are found, they are selected and grouped for analysis. Data is organized in a list of selected resumes, which would not be possible without the strategy adopted.

Subsequently, having as input a specific list of resumes obtained through a specific query or even through using a global listing with the entire local resume repository, it is possible to process data with specific computational routines for each type of analysis. This processing aims at extracting relevant information from the resumes and grouping them in preprocessed data files. Such a strategy aims at generating sets of specific data for application of metric analysis, making it no longer necessary to access the entire set of resumes in each new analysis. Thus, resumes that have a large quantity of data to be processed are accessed and treated only once. As examples of preprocessed data files, there are the ones that gather information about orientations, academic qualification, collaboration and scientific production, as well as about research projects and technical production. All data is available in tabular form.

5.3 OUTCOMES

The main results can be seen at UFSC Information Observatory - https://observatoriodainformacao.ufsc.br/indicadores-cnpq/ufsc/[124]-, an environment that presents historical data from on the institution, which are: (a) indicators that are cross-referenced with data from CNPq, such as scholarship holders of Research Productivity, Technological Development, Research Groups, Research Technical Support and Technical Support Extension; (b) UFSC general indicators (2012-2018), such as Distribution of scientific/technical/artistic productivity and orientation, Scientific production by typology, Technical production by typology, Artistic production and Orientations; (c) Technological indicators (2000-2018), such as patents in general, Thematic patents, Patents by large areas, Patents by inventor, Patents by repository country, Collaborative patents by areas, Patents by institutions collaboration and Patents by areas; (d) indicators by departments per person between 2012-2018; (e) indicators of visibility in Web of Science database (2012-2018), aiming at Typologies of publication, Areas of publication, partner institutions; Main Sponsors, partner countries and most notable authors.

However, this study aims at the feasibility of this entire set of information in open access based on FAIR principles, in which it is intended to determine in which way the Extractor and UFSC Information Observatory datasets are represented

Meeting the ***Findable*** principle, it was possible to identify a context of unique data, the Unique resume identifier (http://lattes.cnpq.br/4767432940301118). Another aspect to be pointed out is that these data have several resources and features, such as tracking system, training levels and work functions, levels of production typologies, either scientific, technical or artistic, as well as tutorials in all training levels. The permanent identification of the dataset can be retrieved

124 The Observatory is managed by SETIC/UFSC and because it is in constant maintenance during COVID-19 pandemic, it is suggested to right-click to open the link in a new tab to access the spreadsheet. Thus, any maintenance problem of the system is solved, and the content can be accessed.

any moment, since the character sets are unique and constantly surveyed, in which no researcher can have two ways of input. The key point is that the system is updated by this entry point Every weekend.

For the ***Accessible*** principle, the system developed by the Federal University of Santa Catarina from the Extractor and the Information Observatory shows that even if data is retrieved, a standard must be kept for identifying protocols in the resumes in the Lattes system (4767432940301118), which is also open access. Access to data indexing is exclusive to researches through their resumes; however, data extraction is free, especially as it is a government record, maintained by the Ministry of Science and Technology. Finally, data are available even if researches do not update them, including in the case of death.

Regarding ***Interoperable***, data is a set of metadata that each researcher adds to his/her resume. Some fields may use filters. In addition, as they are formatted using the XML standard, the interoperability with other datasets is facilitated. Several other systems were developed for similar treatments in Brazil, such as *Script Lattes* (Mena-Chalco; Cesar Junior, 2009).

For ***Reusable*** data, UFSC provides a system for data gathering in XML format, which goes to a systematic extractor. This system organizes the information according to each professor's functional data, not worrying about overlapping of departments at first. Later, institutional overlapping is possible. Data using license is from the government, so as they are data of national need, it is required to be updated for possible evaluations (Productivity Scholarship, Public notice projects, among others).

5.4 DATA FUNCTION FOR INSTITUTIONAL MANAGEMENT CONTROL

The Federal University of Santa Catarina, as a higher education teaching, research, and extension institution, needs to produce indicators of its scientific, technical, social and internationalization development, as well as follows the historical evolution of these contents. For this reason, there has been investment in projects of this nature, whether for extraction and even for its applicability in new services, research, and products in which it is worth investing.

Considering this process, the actions performed are aimed at generating skills and identifying specialists. The resource presented herein works as a basis for finding talents in the institution through resumes registered on Lattes Platform and on other open platforms.

For the Deans of Graduate Studies and Research, this directory of data helps to identify the specialists of the institution and even the collaborators of certain study themes. As example of practical application of this resource, the case of COVID-19 pandemic is mentioned, making it possible to verify the internal researchers and their main collaborators in the development of public health, health safety and even data science. There are other levels that can also be explored, such as issues of guidance in undergraduate, master's and doctoral levels.

There are four levels of content that can be explored to condense in a context of identifying talents through resumes in the Lattes Platform, such as (i) specialists in scientific productivity; (ii) specialists in guidance and participation in thesis/dissertations in themes/subjects examining commission; (iii) leaders of research groups, and; (iv) specialists in technological production.

The identification of this scenario can be seen from a general framework, combining all possible identifications, aforementioned in the previous paragraph, giving a percentage margin for each one. For instance, 25% for each item or studying which ones are more important and dividing the percentages according to the relevance of each item studied.

However, regardless of the order or more relevant item, it is possible to identify particularities in each one of the items in this possible talent pool. This is because they are traceable and accessible data, which have a standard use interoperability, which can be reused to determine department standards and research centers, as in a system for generating ranking and per person indicators.

When it comes to verification from orientation and participation in theses and dissertation examining commission, it can be based on absolute numbers of orientation, participation in examining commissions or on the relation between both.

The process to achieve this particularity of analysis has the advantage of being able to identify scientific families, such as identifying the professors who manage to guide their advisees since undergraduate course, going through the Master's program until doctorate. This can also be applied to the specialists who manage to guide their advisees in higher academic levels, verifying if the advisees were granted scholarships during this period (Costa; Pinto, 2016).

Following this line of reasoning, we have the open content of the Directory of Research Groups of CNPq, which holds a large data collection about the development of research groups, their participants, and collaborators and finally, the most valuable item, the leaders of the research groups. That information can be useful for identifying specialists in fields, themes, and subjects.

This data can be used separately, specifically to identify the researchers' profiles, as it can be blended in the content of the researchers' resumes.

This fusion results in other items in analysis, which deal exclusively with the productive dynamic of researchers in scientific issues (journal articles, works presented in records of events, published/edited books and chapters of books), as well as the technological issue (patents, technological models, graphic designer, among others).

The scientific capital can also be seen by identifying the most cited authors within their respective fields, providing more guidance in the context of expertise for the field, theme, or subject. The citation index can be generated through databases available in Capes Journal Portal (free for federal systems) or through Academic Google.

Finally, for the identification of these specialists and for building this talent pool, it is possible to verify the researchers' performance in open access publications, as well as in commercial contents such as magazines indexed in databases. The focus of this means of data verification is also to identify if the studies of a group of researchers have adherence at the internationalization level.

5.5 FINAL CONSIDERATIONS

This type of service is essential for UFSC, as well as for any other education institution, and its importance can be summarized in six fundamental points:

(1) It is used as support for Graduate Programs to follow, in real time, the evolution of scientific, technical, and artistic data, and the orientation of its professors. It is worth noting that the Information Observatory also provides services aimed at monitoring the development of a certain Graduate Program in the perspective of collaborators with the other programs in the same field in Brazil, as explained in the article "A bibliometric analysis of the scientific production and collaboration between graduate programs in manufacturing engineering in Brazil" that monitored the main Graduate Programs in Production Engineering from 2008 to 2017 (Dutra *et al.*, 2019);

(2) Explores and interoperates with FAIR applications within the institution, accessing valuable data for planning graduate programs for the Dean's office, as an alternative to add information and indicators of C&T in the Search portal;

(3) Similar to otter Lattes Platform *Scripts* extraction, it is accessible, easy to handle and interoperable. From resumes in SML format, it can reuse data, as seen in UFSC Extractor and Observatory;

(4) It works as informational basis to the Research Dean, and to the project of developing a monitoring laboratory in C&T at UFSC, inside the Information Observatory,

(5) Data generated by both UFSC Extractor and Observatory can work as reference to other public institutions, concerning monitoring per person of researchers' productivity and visibility, since UFSC is the best ranked in citation per capita in the country, when considering the elimination of auto citations, and;

(6) Finally, it is reinforced that the organization of data collected from Lattes Platform in compliance with FAIR principles works as a model for the construction of other systems and services that allow an easy information retrieval, as well as the projection of new indicators on C&T of an institution, its levels of national and international collaboration, and the provision for automatic collection of raw data openly gathered; whenever there is no legal restriction.

REFERENCES

COSTA, Airton; PINTO, Adilson Luiz. **De bolsista a cientista:** a experiência da UFSC com o Programa de Iniciação Científica no processo de formação de pesquisadores (1990 a 2012). Florianópolis: EdUFSC, 2016.

DIAS, Thiago Magela Rodrigues. **Um Estudo Sobre a Produção Científica Brasileira a partir de dados da Plataforma Lattes**. 2016. 181 p. Thesis (Doctorate in Mathematical and Computational Modeling) - Graduate Program Course in Mathematical and Computational Modeling, Federal Center for Technological Education of Minas Gerais, Belo Horizonte, 2016.

DIGIAMPIETRI, Luciano Antônio *et al*. Minerando e caracterizando dados de currículos lattes. *In:* BRAZILIAN WORKSHOP ON SOCIAL NETWORK ANALYSIS AND MINING, 2., 2012, Curitiba. **Annals [...]** Curitiba: Brasnam, 2012.

DUTRA, Silvana Toriani *et al*. A bibliometric analysis of the scientific production and collaboration between graduate programs in manufacturing engineering in Brazil. **Informação & Sociedade**, v. 29, n. 1, p. 117-136, 2019. Available from: https://periodicos.ufpb.br/ojs2/index.php/ies/article/view/44852. Access on: Sept. 27, 2020.

FLADUNG, Rainer. ***Scientific communication:*** economic analysis of the eletronic journal market. Stuttgart: Ibidem-Verlag, 2007.

FREIRE, José Donizetti. **CNPq e o acesso aberto à informação científica**. 2011. 275 p. Thesis (Doctorate) - Graduate Program Course, Faculty of Information Science, University of Brasília, Brasília, 2011.

HENNING, Patrícia Corrêa et al. Desmistificando os Princípios FAIR: conceitos, métricas, tecnologias e aplicações inseridas no ecossistema dos Dados FAIR. *In:* XIX Encontro Nacional de Pesquisa em Ciência da Informação, 19, 2018, Londrina. **Annals [...]** Londrina: UEL, 2018.

HENNING, Patrícia Corrêa et al. GO FAIR e os princípios FAIR: o que representam para a expansão dos dados de pesquisa no âmbito da Ciência Aberta. **Em Questão**, v. 25, n. 2, p. 389-412, 2019. DOI: https://doi.org/10.19132/1808-5245252.389-412. Available from: https://seer.ufrgs.br/index.php/EmQuestao/article/view/84753. Access on: Sept. 27, 2020.

KURAMOTO, Hélio. Informação científica: proposta de um novo modelo para o Brasil. **Ciência da Informação**, v. 35, n. 2, p. 91-102, 2006. Available from: https://www.scielo.br/j/ci/a/RcPCvVSyQ6dx7RcmJFLnbxL/abstract/?lang=pt. Access on: Sept. 27, 2020.

LEITE, Fernando César Lima. **Gestão do conhecimento científico no contexto acadêmico: proposta de um modelo conceitual**. 2016. 240 p. Dissertation (Master in Information Science) - Department of Information and Documentation Science, University of Brasilia, Brasília, 2016.

MENA-CHALCO, Jesús Pascoal; CESAR JUNIOR, Roberto Marcondes. ScriptLattes: An open-source knowledge extraction system from the Lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31-39, 2009.

SHAVELL, Steven. Should copyright of academic works be abolished?. **Journal of Legal Analysis**, n. 1, v. 2, p. 301-358, 2010. Available from: http://jla.oxfordjournals.org/content/2/1/301.short. Access on: Oct. 01, 2020.

SUBER, Peter. **Open Access Overview: focusing on open access to peer-reviewed research articles and their preprints**. Creative Commons, 2007. Available from: http://legacy.earlham.edu/~peters/fos/overview.htm. Access on: Aug. 9, 2020.

VEIGA, Viviane Santos de Oliveira *et al*. Plano de gestão de dados fair: uma proposta para a Fiocruz. **Liinc em Revista**, v. 15, n. 2, p. 275-286, 2019. DOI: https://doi.org/10.18617/liinc.v15i2.5030. Available from: https://revista.ibict.br/liinc/article/view/5030. Access on: Aug. 9, 2020.

How to cite this chapter: PINTO, Adilson Luiz; DIAS, Thiago Magela Rodrigues; CANTO, Fábio Lorensi do; CARVALHO SEGUNDO, Washington Luís Ribeiro de. Open data of the lattes platform according to FAIR principles: examples of extractor and UFSC information observatory. *In*: SALES, Luana Faria; VEIGA, Viviane Santos de Oliveira; VIDOTTI, Silvana Aparecida Borsetti Gregório; HENNING, Patrícia; SAYÃO, Luís Fernando (org.). **FAIR Principles Applied To Research Data Management**: Brazilian Experiences. Brasília, DF: Editora Ibict, 2024. cap. 5, p. 71-79. DOI: 10.22477/9788570131959.cap5.