Pecha Kucha

DOI:10.21680/2447-7842.2023v9n2ID33661

Uso do ORCID como indicador persistente na construção do banco de dados de produção acadêmica da Fundação Oswaldo Cruz

Use of ORCID as a permanent indicator in the construction of the Oswaldo Cruz Foundation's academic production database

Marcos Wesley Soares Alves

Waldeyr Mendes Cordeiro da Silva [1]

Rafaela Lora Grando [2]

Vanessa de Arruda Jorge [3]

Submetido em: 17/04/2023	Aprovado na ConfOA: 14/06/2023	Publicado em: 04/12/2023

Resumo: Indicadores digitais persistentes são marcadores únicos e permanentes atribuídos a recursos digitais, como publicações científicas, pesquisadores, revistas científicas, ou outros tipos de dados e informações. O uso desses indicadores permite reconhecer um recurso digital em um contexto, eliminando duplicações e ambiguidades, possibilitando assim, o desenvolvimento e manutenção de bancos de dados mais precisos e completos. A proposta apresenta a aplicação do Open Researcher and Contributor ID (ORCID), um identificador digital persistente para

[1] Professor no Instituto Federal de Goiás em exercício na Presidência da República como Coordenador-Geral de Desenvolvimento de Sistemas.Atuo ainda como pesquisador colaborador e co-orientador voluntário na Universidade de Brasília e como pesquisador bolsista do Observatório Fiocruz.

[2] Graduada em Ciências Biológicas: Microbiologia pela Universidade Federal do Rio e Janeiro (UFRJ). Mestrado em Engenharia de Processos Químicos e Bioquímicos (UFRJ). Doutorado sanduíche UFRJ e Universidade Autônoma de Barcelona. MBA Gestão Inovação em Marketing na Pontifica Universidade Católica (PUC-RS).

[3] Doutora e mestre em Ciência da Informação pelo convênio IBICT/UFRJ e graduada em Arquivologia pela Universidade Federal do Estado do Rio de Janeiro (UNIRIO). Atualmente é Coordenadora de Informação e Comunicação da Vice-presidência de Educação, Informação e Comunicação (VPEIC) da Fiocruz, atuando na coordenação do Fórum de Editores Científicos, do Fórum de Ciência Aberta e Fórum de Preservação Digital da Fiocruz e Comitê Gestor do Arca Dados.

autores, como ferramenta para aprimorar as atividades de tratamento de dados de produção científica do Observatório em Ciência, Tecnologia e Inovação em Saúde da Fiocruz. Seu uso pode trazer benefícios para o gerenciamento e análise de dados, especialmente nas tarefas de extração, tratamento e carga de dados, aprimorando a detecção de relações entre autores e instituições, inclusive aquelas relativas à Ciência Aberta.

Palavras-chave: indicadores digitais persistentes; ORCID; desambiguação de nomes; banco de dados; padronização de dados.

Abstract: Persistent digital identifiers are unique and permanent markers assigned to digital resources, such as scientific publications, researchers, scientific journals, or other types of data and information. Using these indicators allows for recognizing a digital resource in a context, eliminating duplications and ambiguities, thus enabling the development and maintenance of more accurate and complete databases. The proposal presents the application of the Open Researcher and Contributor ID (ORCID), a persistent digital identifier for authors, as a tool to improve the activities of data processing of scientific production of the Observatório em Ciência, Tecnologia e Inovação em Saúde da Fiocruz. Its use enables earnings on data management and analysis, especially in data extraction, treatment, and loading tasks. Such benefits can improve the detection of relationships between authors and institutions, including those related to Open Science.

Keywords: digital persistent indicators; ORCID; name disambiguation; database; data standardization.

1 INTRODUÇÃO

Fruto dos esforços de universidades, centros de pesquisa e outras organizações do gênero, grande parte das informações científicas por elas

produzidas podem ser localizadas via bases de dados bibliográficas, catálogos de serviços de informação ou acervos físicos. Tais informações estão voltadas, sobretudo para a comunidade científica, embora também existam iniciativas de divulgação científica que visam torná-las mais acessíveis à população. Acessar e conhecer informações científicas confiáveis e certificadas pela academia e delas se apropriar, é de interesse de toda a sociedade (Santos, 2021).

As fontes desse tipo de informação, em geral, são conhecidas e acessadas pela comunidade científica. Entretanto, há barreiras para a recuperação eficaz e assertiva da informação, o que pode trazer insegurança aos pesquisadores quanto à qualidade das buscas (Santos, 2021). Muitas vezes é necessário combinar resultados de várias bases de informação científica, ou mesmo validar a informação de uma base em outra. No contexto de alto volume de informações a serem combinadas, validadas e deduplicadas a curadoria manual, além de ineficiente se torna inviável.

Observatório em Ciência, Tecnologia e Inovação em Saúde (Observatório CT&I em Saúde) é uma iniciativa da Vice-presidência de Educação, Informação e Comunicação da Fundação Oswaldo Cruz (Fiocruz) que, entre outros objetivos, visa contribuir para a gestão da informação científica produzida na Instituição.

Diante da complexidade e do tamanho da instituição, que conta com uma unidade técnica de apoio, para produção de animais de laboratório e derivados de animais, 4 escritórios regionais no Ceará, Mato Grosso do Sul, Piauí e Rondônia, além de 16 unidades técnico-científicas (Fiocruz [202?]), o Observatório CT&I em Saúde está organizando estas informações dispersas em 10 estados na forma de indicadores de produção acadêmica. O objetivo é permitir que a Instituição fomente um ecossistema de pesquisa científica, considerando bases de dados internas e externas, ainda o intuito central da organização deste conhecimento é a construção de uma rede de inteligência científica, na qual o gerenciamento e análise integrada de informação envolvendo CT&I da Fiocruz sirva para aprimorar o planejamento dos recursos e a tomada de decisão.

Sendo assim, foram realizadas buscas nas bases *Web of Science, Scopus, Pubmed, Lilacs* a partir do nome institucional Fiocruz e suas variações, na base

Lattes a partir do CPF dos servidores da Fiocruz e no repositório institucional Arca. Os dados buscados sobre a produção acadêmica institucional restringiram-se ao período de janeiro de 2008 a setembro de 2022. Os dados coletados destas seis bases de dados foram inseridos em um banco de dados relacional especialmente modelado e implementado em um Sistema Gerenciador de Bancos de Dados (SGBD) PostgreSQL.

Uma vez inseridas as informações no banco de dados, diversos processos foram desenhados e implementados para tratar os campos e estabelecer vínculos entre eles para posterior criação de indicadores. Tais processos de tratamento dos dados foram implementados em linguagem de programação Python e concentraram-se, principalmente, em tratar unidades, instituições parceiras, veículos de publicação, dentre outros. Muitos campos, por sua natureza, foram tratados como entidades, e passaram a necessitar de um identificador digital único.

Apesar do SGBD criar identificadores únicos de forma automática, à medida que os dados são inseridos, há características importantes e únicas quanto aos nomes dos autores. Neste contexto, a padronização dos nomes de autores torna-se uma tarefa complexa e laboriosa, por tais características únicas, quais sejam: variações de um mesmo nome, homônimos, abreviações, nomes de citação e acentuação e uma extensa lista de quesitos a serem tratados para identificar corretamente os autores a partir do nome declarado nas publicações em que participam. Por exemplo, sobrenomes comuns no território brasileiro como Silva, Oliveira e Alves, podem facilmente ser confundidos quando o primeiro nome estiver abreviado (Super Interessante, 2017). Ainda há situações em que há mudança de nome ao longo da vida por motivo de casamento ou divórcio, por exemplo.

A identificação autoral é ainda mais complexa quando se faz uso de bases de informação como o currículo Lattes, cujo preenchimento é livre. Muitas abordagens têm sido propostas para resolver este problema, tais como soluções baseadas em grafos (Zhang & Al Hasan, 2017) ou redes de colaboração (Zhang, Saha & Al Hasan, 2014), agrupamentos por similaridades usando abordagens de aprendizado de máquina supervisionadas e não-supervisionadas (Ferreira, Gonçalves & Laender, 2012; Smalheiser & Torvik, 2009; Tekles & Bornmann, 2020). O tratamento deste

problema é, portanto, uma tarefa não-trivial, especialmente para grandes volumes de dados, como é o caso da Fiocruz.

Uma solução recorrente na literatura é a adoção de um identificador digital persistente que permita reconhecer ou agrupar autores homônimos a partir das várias formas de escrita de seus nomes (Santos, 2021). Entre as soluções existentes desse tipo, tais como ID Lattes, ID Scopus, e outros, o *Open Researcher and Contributor ID* (ORCID) (Cress, 2019), este último tem ganhado espaço e relevância pelo seu caráter multinacional.

A visão do ORCID alinha-se à identificação em nível global de todos os participantes de pesquisas, bolsas de estudos e inovação. Esta identificação única permitiria conectar o autor às suas contribuições através de disciplinas, fronteiras e tempo (ORCID, [202?a]). O ORCID é um identificador único de 16 dígitos que pode ser atribuído a cada pesquisador para manter um registro das atividades de pesquisa associando-as a cada identificador (Enago Academy. ([202?]). De acordo com Gasparyan e colaboradores (Gasparyan *et al.*, 2014), o ORCID oferece uma nova solução para o problema persistente de transcrição variável e ordem de nomes complexos, omissão de nomes do meio e iniciais, mudanças em nomes de mulheres casadas e divorciadas e existência de nomes comuns na maioria dos países e continentes ainda os autores sugerem que seu uso pode melhorar o rastreamento de artigos com registros bibliográficos variáveis das mesmas fontes em vários bancos de dados (Gasparyan *et al.*, 2014).

Este trabalho é um estudo de caso de como o Observatório CT&I em Saúde da Fiocruz vem utilizando o ORCID como indicador persistente para identificação, desambiguação e deduplicação dos autores de produções acadêmicas institucionais.

2 DESENVOLVIMENTO

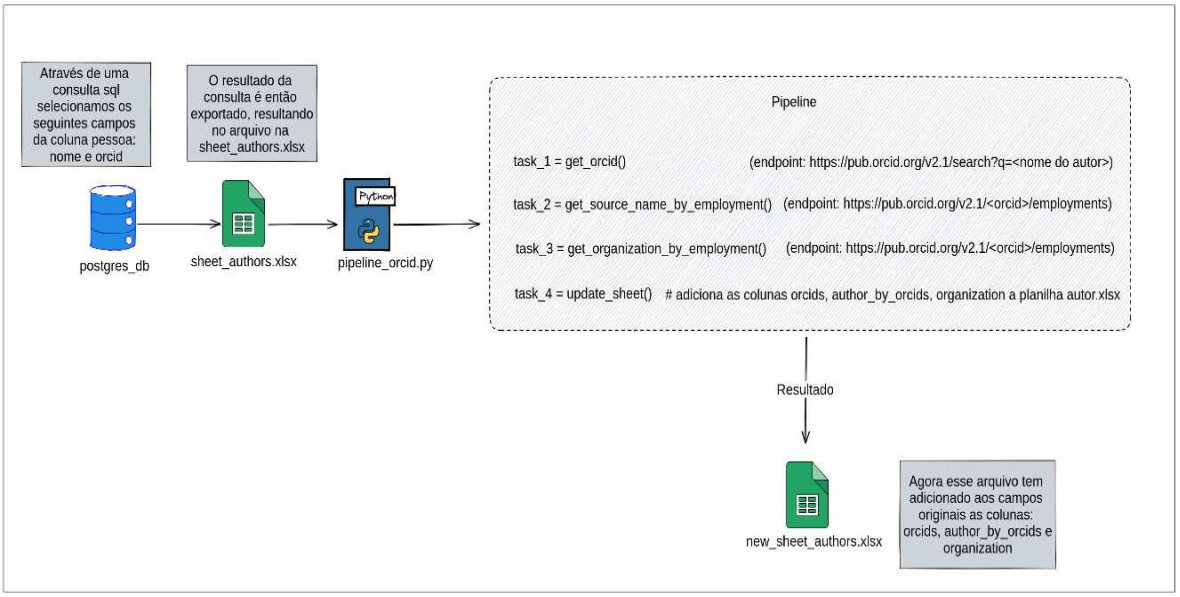
Entre os indicadores produzidos pelo Observatório CT&I em Saúde da Fiocruz, está o indicador de produção acadêmica, para o qual um *pipeline* de coleta, tratamento, combinação e validação dos dados foi elaborado. Este tem sido refinado

ao longo do tempo a cada *release* do indicador. Entre os tratamentos e combinação de dados coletados, está a desambiguação de nomes de autores.

Na primeira etapa do *pipeline* para geração de um indicador de produção acadêmica, os dados são coletados de diversas bases, em seus diversos formatos. Em seguida os dados coletados são convertidos para um formato comum (JSON) e armazenados em um banco de dados. Todas essas atividades são executadas de forma semiautomática por *scripts in-house*, desenvolvidos na linguagem de programação Python 3.

Dentre as informações coletadas nas bases, quando disponível, é coletado de forma associada ao nome do autor o seu ORCID. Entretanto, a coleta de publicações nas bases de dados estende-se para datas anteriores à existência do ORCID, e por isso muitos autores que hoje têm seu ORCID registrado, não o tem citado nas publicações anteriores ao registro. É a estes casos que esta solução se aplica. O Observatório CT&I em Saúde da Fiocruz criou *scripts Python* que consultam a API pública do ORCID (ORCID, [202?b]) para, através do nome de um determinado autor no banco de dados, obter o seu ORCID. O processo se dá em 4 partes, conforme Figura 1.

Figura 1- Detalhamento do processo de obtenção e aplicação do ORCID como identificador persistente

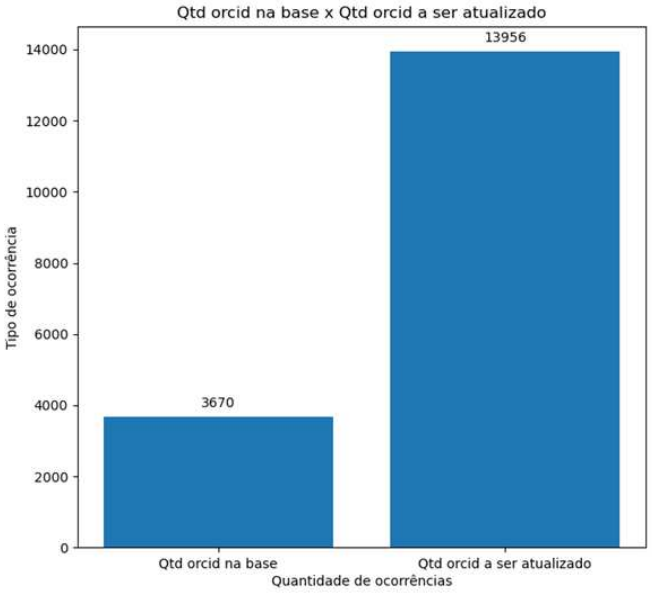


Fonte: Elaboração própria.

1 – Caso os valores do campo ORCID proveniente do banco de dados estejam preenchidos, este será utilizado, caso contrário, é realizada uma busca pelo nome do autor na API, pelo *endpoint*: */v2.1/search?q=<nome do autor>* para recuperar o ORCID do autor;
2 – Uma vez com o ORCID recuperado, é realizada a obtenção do nome do autor e das instituições a ele associadas. A obtenção do nome a partir do *endpoint* é realizada para possibilitar a validação do ORCID, fazendo uma comparação do nome existente na base com o nome retornado pela API. O *endpoint* utilizado é o: */v2.1/<orcid>/employments;*
3 – Obtenção da organização de trabalho do autor. A obtenção dessa informação também é proveniente do *endpoint*: */v2.1/<orcid>/employments;*
4 – Atualização da planilha que foi utilizada como entrada de dados. Além dos campos nome e orcid, foram adicionados os seguintes campos provenientes da API: orcids, author_by_orcids e organization.

O banco de dados utilizado nesta análise e padronização contém 66 mil produções científicas com pelo menos um autor vinculado à Fiocruz. Os autores presentes no banco totalizaram 179.789 registros de autores afiliados à Fiocruz ou a instituições parceiras. Neste conjunto de dados, a quantidade de autores com ORCID era de 3.670 registros. Após a aplicação do método descrito neste trabalho, esse número passou a ser de 13.956. Com isso, houve um ganho de 10.286 novos ORCID identificados. Após a aplicação do método descrito neste trabalho, houve um incremento de 23% na vinculação de ORCID a autores das produções (Figura 2). Esse processo de busca e validação de ORCID oportunizou uma maior precisão e qualidade nas informações armazenadas em nosso sistema, proporcionando a desambiguação e agrupamento de uma elevada quantidade de nomes de autores.

Figura 2 - Resultados obtidos após a consulta do ORCID



Fonte: Elaboração própria.

Este resultado, para além dos benefícios óbvios, promove meios de identificar através de novas combinações de informações, como buscas na base DOAJ (https://doaj.org), relações de Ciência Aberta na produção científica constante no banco de dados do Observatório CT&I em Saúde da Fiocruz. Adicionalmente, ao promover o uso do ORCID, o Observatório CT&I em Saúde da Fiocruz contribui para a padronização e a transparência na atribuição de créditos e reconhecimento aos autores de trabalhos científicos, promovendo a integridade da pesquisa e a valorização do trabalho acadêmico (Silva et al., 2022).

3 CONSIDERAÇÕES FINAIS

A utilização do ORCID como identificador persistente para autores de produção científica existente no banco de dados do Observatório CT&I em Saúde da Fiocruz tornou-o mais confiável, assegurando que um número maior de autores estará livre de ambiguidades. A atualização do banco de dados também permite melhorias na identificação de metadados das produções científicas, tais como

indicadores de Ciência Aberta. Além disso, o ORCID é amplamente reconhecido e utilizado pela comunidade científica, o que aumenta a visibilidade e a reputação dos pesquisadores vinculados aos nossos registros.

Como trabalho futuro, o Observatório CT&I em Saúde da Fiocruz pretende usar os demais *endpoints* da API para melhorar continuamente a qualidade dos dados de seu banco de dados, incluindo outros indicadores, como grupos de pesquisa e patentes.

REFERÊNCIAS

Cress, P. E. (2019). Why Do Academic Authors Need an ORCID ID? *Aesthetic Surgery Journal*, 39(6), 696–697. https://doi.org/10.1093/asj/sjz042

Enago Academy. ([202?]). *How ORCID is Changing the Publishing Landscape - Enago Academy*. Enago Academy. Recuperado de: https://www.enago.com/academy/orcid-changing-publishing-landscape/

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. *ACM SIGMOD Record*, 41(2), 15–26. https://doi.org/10.1145/2350036.2350040

FIOCRUZ. ([202?]). *Unidades e escritórios*. Fiocruz. Recuperado de: https://portal.fiocruz.br/unidades-e-escritorios

Gasparyan, A. Y., Akazhanov, N. A., Voronov, A. A., & Kitas, G. D. (2014). Systematic and Open Identification of Researchers and Authors: Focus on Open

Researcher and Contributor ID. *Journal of Korean Medical Science*, 29(11), 1453. https://doi.org/10.3346/jkms.2014.29.11.1453

Santos, T. V. d. (2021). *Identificadores persistentes: aplicabilidade na organização e acesso à informação científica* [Dissertação de mestrado, Universidade Federal de São Paulo]. https://www.teses.usp.br/teses/disponiveis/27/27151/tde-02052022-115537/pt-br.php

ORCID. ([202?a]). *About ORCID*. Recuperado de: https://info.orcid.org/what-is-orcid/

ORCID. ([202?b]). *Public API*. Recuperado de: https://info.orcid.org/documentation/features/public-api/

Smalheiser, N. R., & Torvik, V. I. (2009). Author name disambiguation. *Annual Review of Information Science and Technology*, *43* (1), 1–43. https://doi.org/10.1002/aris.2009.1440430113

Silva, M. V. P., Jorge, V. A., Silva, W. M. C., Grando, R. L., Fonseca, F. L. (2022). *Impacto da taxa de processamento de artigos em uma instituição de pesquisa em saúde: um estudo de caso da Fundação Oswaldo Cruz (Fiocruz)*. Universidade Federal de Alagoas.

Super Interessante. (2017). *A origem dos 50 sobrenomes mais comuns do Brasil*. Super Interessante. Recuperado de: https://super.abril.com.br/especiais/a-origem-dos-50-sobrenomes-mais-comuns-do-brasil/#silva

Tekles, A., & Bornmann, L. (2020). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *Quantitative Science Studies*, 1(4), 1510–1528. https://doi.org/10.1162/qss_a_00081

Zhang, B., & Al Hasan, M. (2017). Name Disambiguation in Anonymized Graphs using Network Embedding. In CIKM '17: *ACM Conference on Information and Knowledge Management.* ACM. https://doi.org/10.1145/3132847.3132873

Zhang, B., Saha, T. K., & Al Hasan, M. (2014). Name disambiguation from link data in a collaboration graph. In 2014 *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE. https://doi.org/10.1109/asonam.2014.6921563