15ª Conferência Lusófona de Ciência Aberta (ConfOA)
Acesso Aberto e Dados de Investigação Abertos: sistemas, políticas e práticas
Modalidade: Pecha Kucha

Desambiguação de nomes de autores: um desafio para os repositórios

Disambiguation of authors' names: A challenge for repositories

Maria Eduarda Pereira Rodrigues (MEPR)
Instituto Politécnico de Castelo Branco (IPCB); Centro de Estudos de Recursos Naturais, Ambiente e Sociedade (CERNAS-IPCB)
Castelo Branco, Portugal
Orcid: 0000-0001-9842-3412
erodrigues@ipcb.pt

António Moitinho Rodrigues (AMR)
Escola Superior Agrária – Instituto Politécnico de Castelo Branco (ESA/IPCB); Centro de Estudos de Recursos Naturais, Ambiente e Sociedade (CERNAS-IPCB)
Castelo Branco, Portugal
Orcid: 0000-0002-5862-3898
amrodrig@ipcb.pt

RESUMO:
Nos repositórios institucionais o controlo de autoridade dos nomes dos autores e a desambiguação permanecem um desafio, sendo essenciais à gestão da informação. Este estudo pretendeu verificar o nível de desambiguação dos nomes dos autores nos repositórios de 8 institutos politécnicos portugueses. Analisaram-se 800 entradas de autor em 100 **registos** por repositório e codificaram-se as dúvidas/ocorrências. Registaram-se 124 dúvidas/ocorrências. Verificou-se escassa utilização dos identificadores ORCID e CIÊNCIA-ID. Identificou-se a necessidade de desambiguar nomes de autores, o que permitirá agregar a produção individual dos autores num único ponto de acesso, melhorando os resultados da pesquisa.

Palavras-chave: controlo de autoridade; gestão da informação; recuperação de informação

INTRODUÇÃO

Nas bibliotecas das instituições de ensino superior e nos repositórios institucionais (RI) o controlo de autoridade dos nomes dos autores e a sua desambiguação constituem um enorme desafio. A sua ausência produz impactos negativos na recuperação da informação (Morgan e Eichenlaub, 2018). Nos dois sistemas, este é um aspeto essencial à gestão da informação, uma vez que o nome dos autores é um dos indicadores mais utilizados na pesquisa de informação. Para garantir o controlo de qualidade, incluindo erros de ortografia, inconsistências nos dados, informação incompleta ou truncaturas erradas (Ortiz et al., 2017), alguns autores propõem a utilização de identificadores como o Open Researcher Contributor ID (ORCID), o Scopus ID (Elsevier) ou o Virtual International Authority Fila (VIAF) (Tarver et al., 2013; Myntti e Cothran, 2013; Walker e Armstrong, 2014). De acordo com Downey (2019) o controlo de autoridade nos RI tem sido negligenciado, em parte, devido à diminuição do detalhe dos metadados nos RI, o que, segundo Sanyal, Bhowmick e Das (2021), pode inclusivamente gerar desconfiança nos resultados das pesquisas.

Assim, diversos autores (Pazzini, 2022; Cho, 2022; Kim e Owen-Smith, 2021; Mandal, 2023) convergem na necessidade de automatizar o processo utilizando identificadores persistentes, maioritariamente o ORCID, para: a) criação de pontos de acesso únicos aos autores; b) normalização da forma do nome; e c) desambiguação dos nomes dos autores nos RI o que vai ao encontro do defendido por Truta, et al. (2020) no contexto do Projeto RCAAP. De salientar que a interoperabilidade entre os sistemas é, também, um fator a considerar no controlo de autoridade dos RI (Gaitanou et al., 2024) diminuindo erros de digitação manual.

O presente estudo teve como objetivo analisar o índice de autores dos RI das instituições portuguesas de ensino superior politécnico, alojadas no Serviço de Alojamento de Repositórios Institucionais (SARI), para verificar a situação relativa ao controlo de autoridade de nomes de autor e respetiva desambiguação e propor sugestões para a sua melhoria.

MATERIAL E MÉTODOS

Os dados foram recolhidos nos repositórios SARI dos 8 institutos politécnicos portugueses assim codificados: RePol1; RePol2; RePol3; RePol4 RePol5; RePol6; RePol7; e RePol8. Registou-se a data de recolha para cada RI (**Tabela 1**). Os elementos foram obtidos na *homepage* de cada RI registando-se a dimensão, o número total de entradas por autor, o número de autores com nomes não portugueses e o número de autores com identificador.

No menu principal de cada repositório institucional foi selecionada a opção Autor. Obteve-se um "*display*" de 20 registos por página/repositório. Seguidamente, identificaram-se todas as entradas de autor que constavam das primeiras cinco páginas de cada repositório institucional SARI, num total de 100 entradas de nome de autor/repositório. A amostra foi constituída por 800 entradas de autor, que foram todas analisadas, tendo-se recolhido, registado, codificado e descrito, em folha de recolha própria, as diversas ocorrências,

construindo-se uma matriz global. Para o efeito, foi criado um código de classificação do tipo de ocorrências registadas (DT) com o seguinte detalhe: DT1- Formas diferentes do nome para um mesmo autor; DT2- Autores diferentes com o mesmo nome; DT3- Problemas na truncatura; DT4- O mesmo autor com e sem identificadores (p. ex. ORCID); DT5 - Abreviaturas do nome sem ponto; DT6- Nomes hispânicos; DT7- Erro no nome do autor; DT8- Sem elementos suficientes para identificação do autor; DT9- Autor com entradas de nome iguais; e DT10- Nome abreviado e sem elementos para identificação do autor. As dúvidas/ocorrências e a sua localização na página de pesquisa do RI foram registadas. Para a construção da matriz de DT considerou-se a forma do nome, a informação no ORCID, os documentos arquivados e, em alguns casos, as outras fontes.

A análise qualitativa teve em consideração as DT encontradas na amostra. Na análise quantitativa, foram efetuadas contagens, calculadas percentagens (média e desvio padrão [±dp]) e determinados coeficientes de correlação de Pearson, utilizando o *software* IBM SPSS.

RESULTADOS E DISCUSSÃO

À data de realização do estudo, os 8 repositórios SARI analisados possuíam, na totalidade, 96 508 registos de documentos, variando entre 27 031 no RePol1 e 2 118 no RePol8. Verificou-se que quanto maior o número de documentos depositados no RI, maior o número de entradas de autor, situação confirmada pela correlação positiva elevada entre estes 2 parâmetros (r=0,883; p=0,004).

Analisando-se os resultados de cada um dos RI, individualmente, verificou-se que o número médio de documentos/autor/RI foi de 0,97 (±0,29), variando entre 1,55 no RePol1 e 0,75 documentos/autor/RI no RePol8 (**Tabela 1**).

Tabela 1 – Data de recolha de dados nos Repositórios Institucionais (RI) dos oito Institutos Politécnicos Portugueses (SARI) e mapa resumo de resultados obtidos.

Código RI	Data recolha dados	Registo total documentos	Número documentos por Entrada autor	% Autores sem ID	% Autores c/ ORCID	% Autores c/ CIENCIA ID	Entradas Autor com Nomes não Portugueses
RePol1	02/04/2024	27 031	1,55	95,0%	100,0%	60,0%	68
RePol2	02/04/2024	8 684	0,84	77,0%	95,7%	68,2%	30
RePol3	02/04/2024	8 533	1,29	92,0%	100,0%	37,5%	15
RePol4	03/04/2024	16 283	0,79	82,0%	100,0%	22,2%	33
RePol5	03/04/2024	4 257	0,78	94,0%	100,0%	16,7%	28
RePol6	03/04/2024	6 769	0,87	92,0%	87,5%	57,1%	29
RePol7	04/04/2024	22 833	0,91	92,0%	100,0%	37,5%	33
RePol8	05/04/2024	2 118	0,75	100,0%	0,0%	0,0%	19
	Valor médio (±dp) dos 8 RI	12 064	0,97 (±0,29)	90,5% (±0,07)	85,4% (±0,35)	37,4% (±0,24)	31,9 (±15,97)

±dp – desvio padrão da amostra. Fonte: Autores

A análise da existência de identificadores de autor revelou que a percentagem de autores sem identificador é muito elevada, correspondendo a um valor médio de 90,5% (±0,07) dos registos de autor analisados por RI (**Tabela 1**). Aquela percentagem variou entre 100% de autores sem identificador no RePol8 e 77% no RePol2. Apenas 76 dos 800 autores amostrados possuía identificador no RI, número que variou entre 23 no RePol2 e 0 no RePol8. Em média, dos autores que possuíam identificador em cada RI, 85,4% (±0,35) tinham identificador ORCID e 37,4% (±0,24) tinham identificador CIENCIA ID (**Tabela 1**). Foi calculada uma correlação positiva elevada (r=0,835; p=0,01) entre os autores que possuíam os dois identificadores, ORCID e CIENCIA ID.

Todos os oito RI analisados registaram DT nos nomes dos autores, mas os dados obtidos mostram existir uma baixa correlação entre a dimensão do RI e o número de DT identificados nos autores. A correlação negativa não significativa encontrada (r=-0,543; p=0,164) parece sugerir que nos repositórios com maior número de documentos existem alguns mecanismos de controlo de autoridade relativamente aos nomes dos autores.

Nos 800 autores amostrados foram identificadas 255 entradas de autor com nomes não portugueses. O valor médio determinado foi de 31,9 (±15,97) entradas de autores com nomes não portugueses por RI, variando entre 68 no RePol1 e 15 no RePol3 (**Tabela 1**), tendo-se determinado uma correlação positiva não significativa entre o número de entradas de autor por RI e o número de entradas de autor com nomes não portugueses por RI (r=0,524: p=0,183). Ao mesmo tempo, a correlação negativa encontrada entre o número de entradas de autor com nomes não portugueses e as DT identificadas (r=-0,318; p=0,488), parece sugerir maior nível de normalização na identificação daqueles autores nos seus documentos.

Conforme se pode verificar na tabela 2, relativamente aos registos amostrados, identificaram-se um total de 124 DT, destacando-se DT1- Formas diferentes do nome para um mesmo autor RI (n=50), DT3 - Problemas na truncatura (n=18) e DT4 - O mesmo autor com e sem identificadores (n=17). O maior problema identificado está relacionado com as diferentes formas que o nome de um mesmo autor pode assumir, dificultando muito a tarefa de registo e identificação clara de autoria no RI, podendo ocorrer no arquivo mediado como no autoarquivo, enviesando os resultados de pesquisa por autor ao não refletir com exatidão toda a produção científica que um mesmo autor pode ter depositada no RI da sua instituição.

Verificou-se, também, que existem grandes diferenças entre os vários RI relativamente às diversas DT. Destaca-se pela positiva o RePol1 com 1 DT identificada (DT-1) e pela negativa o RePol7 com 27 DT, distribuídas por todas as DT, com exceção da DT-10 (**Tabela 2**).

Tabela 2 – Discriminação das 124 Dúvidas/Ocorrências (DT) identificadas nos 8 Repositórios Institucionais (RePol1 a RePol8), correspondendo a 15,5% da amostra analisada (n=800).

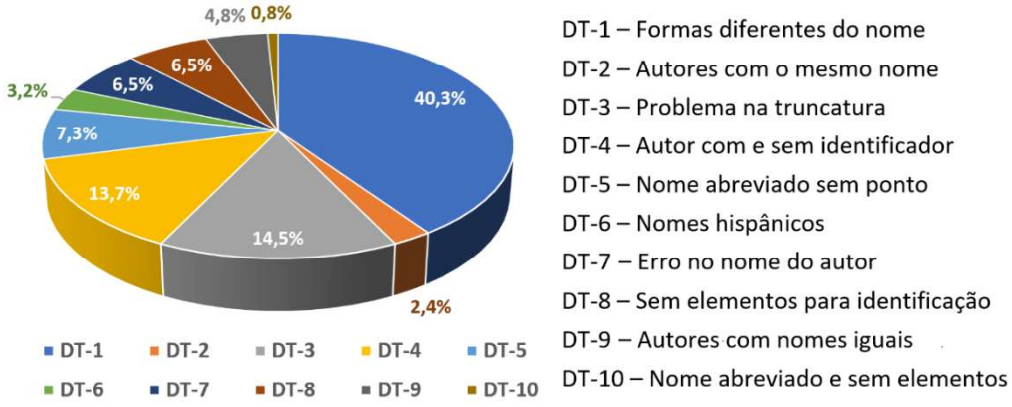
DT	RePol1	RePol2	RePol3	RePol4	RePol5	RePol6	RePol7	RePol8	Total
DT-1 – Formas diferentes do nome	1	3	2	3	11	9	6	15	50
DT-2 – Autores com o mesmo nome	0	1	1	0	0	0	1	0	3
DT-3 – Problema na truncatura	0	3	0	2	4	4	5	0	18
DT-4 – Autor com e sem identificador	0	5	0	2	3	3	4	0	17
DT-5 – Nome abreviado sem ponto	0	3	0	2	0	1	1	2	9
DT-6 – Nomes hispânicos	0	0	0	0	1	0	2	1	4
DT-7 – Erro no nome do autor	0	0	0	2	1	2	1	2	8
DT-8 – Sem elementos para identificação	0	1	0	1	1	1	2	2	8
DT-9 – Autores com nomes iguais	0	0	0	0	0	0	5	1	6
DT-10 – Nome abreviado e sem elementos	0	0	1	0	0	0	0	0	1

DT	RePol1	RePol2	RePol3	RePol4	RePol5	RePol6	RePol7	RePol8	Total
Total	1	16	4	12	21	20	27	23	124

Fonte: Autores

A representação gráfica das DT (**Gráfico 1**) ilustra os resultados da **Tabela 2**, evidenciando a necessidade de maior cuidado na alimentação dos RI, com enfoque na forma do nome do autor. Também é relevante controlar as truncaturas nos nomes dos autores.

Gráfico 1 – Representação gráfica da percentagem de Dúvidas/Ocorrências (DT) (n=124) identificadas por Repositórios Institucionais (RePol1 a RePol8)



Fonte: Autores

CONSIDERAÇÕES FINAIS

Os resultados obtidos permitem concluir que há muito trabalho ainda a realizar no âmbito da desambiguação de nomes de autor nos RI analisados.

A maioria dos autores não tem identificador associado. Quanto aos identificadores eletrónicos, constatou-se que o ORCID é o mais utilizado, seguido do CIENCIA ID, o que vai ao encontro do indicado pelo projeto RCAAP para os RI SARI. Existem autores de registo único e outros que não é possível desambiguar por falta de elementos. Muitos autores não pertencem à instituição dona do RI, o que dificulta o controlo e a desambiguação dos seus nomes.

Será muito importante definir instrumentos normalizadores que ajudem os gestores dos RI no processo de controlo de autoridade e desambiguação dos nomes dos autores.

Outras inconsistências, foram encontradas nos diversos RI, que não cabem dentro do âmbito do presente estudo, mas que poderão ser objeto de análise futura.

REFERÊNCIAS

CHO, James. H. Cataloging for a celebration: metadata for an Institutional Repository from the Ground Up. **Cataloging & Classification Quarterly**, v.60, n.2, p.141–163, 2013.

DOWNEY, Moira. Assessing author identifiers. **Journal of Library Metadata**, v.19, n.1–2, p.117–136, 2019. Disponível em:https://www.tandfonline.com/doi/full/10.1080/01639374.2021.2018633. Acesso em: 12 jan. 2024. DOI: https://doi.org/10.1080/01639374.2021.2018633.

GAITANOU, P.; ANDREOU, I.; SICILIA, M.-A.I; GAROUFALLOU, E. Linked data for libraries: creating a global knowledge space, a systematic literature review. **Journal of Information Science**, v.50, n.1, p.204-244, 2024. DOI: https://doi.org/10.1177/01655515221084645. Disponível em: https://journals.sagepub.com/doi/full/10.1177/01655515221084645. Acesso em 19 jan. 2024.

KIM, J.; OWEN-SMITH, J. ORCID-linked labeled data for evaluating author name disambiguation at scale. **Scientometrics**, n.126, p.2057–2083, 2021.

MANDAL, Sukumar. World linking identifiers for authority and bibliographic records. **Library Philosophy and Practice (e-journal)**. 7875, 2023, 2023. Disponível em:. https://digitalcommons.unl.edu/libphilprac/7875. Acesso em 19 jan. 2024.

MORGAN, M.; EICHENLAUB, N. Author identifier analysis: name authority control in two institutional repositories. In Int'l Conf. on Dublin Core and Metadata Application. **Proceedings**. Toronto, 2018. DOI: https://doi.org/10.23106/dcmi.952139036. Disponível em: https://dl.acm.org/doi/10.5555/3308533.3308551. Acesso em 12 jan. 2024.

MYNTTI, Jeremy; COTHRAN, Nate. Authority control in a digital repository: preparing for linked data. **Journal of Library Metadata**, v.13, n.2–3, p.95–113, 2013.

ORTIZ, José; SEGARRA, José; SUMBA, Xavier; CULLCAY, José; ESPINOZA, Mauricio; SAQUICELA, Victor. Authors semantic disambiguation on heterogeneous bibliographic sources. *In* **XLIII Latin American Computer Conference** (CLEI), Cordoba, 2017. DOI: https://doi.org/10.1109/CLEI.2017.8226389. Disponível em: https://ieeexplore.ieee.org/document/8226389. Acesso em: 2 fev. 2024.

PIAZZINI, Tessa. Bibliographic control and institutional repositories. **JLIS.It**, v.13, n.1, p.132-42, 2022. DOI: https://doi.org/10.4403/jlis.it-12717. Disponível em: https://www.jlis.it/index.php/jlis/article/view/426. Acesso em 25 jan. 2024.

SANYAL, Debarshi. K.; BHOWMICK, P. Kumar; Das, Partha P. A review of author name disambiguation techniques for the PubMed bibliographic database. **Journal of Information Science**, n.47, p.227-254, 2019.

TARVER, Hanna; WAUGH, Laura; PHILLIPS, Mark, HICKS, Will. **Implementing name authority control into institutional repositories: a staged approach**. 2013. Disponível em: https://digital.library.unt.edu/ark:/67531/metadc172365/. Acesso em: 8 mai. 2024.

TRUTA, R.; CARVALHO, J.; LOPES, P.; RIBEIRO, F.; GRAÇA, P. Associação dos identificadores CIÊNCIA ID e ORCID a autores nos Repositórios Institucionais integrados no SARI do Projeto RCAAP. **Páginas a&b**, Porto, 3.ª série, p.281–282, 2020.

WALKER, Lizzy A.; ARMSTRONG, Michelle. "I cannot tell what the dickens his name is": name disambiguation in institutional repositories. **Journal of Librarianship and Scholarly Communication**, v. 2, n.2, eP1095, 2014. DOI: https://doi.org/10.7710/2162-3309.1095. Disponível em: doi.org/10.7710/2162-3309.1095 Acesso em 8 mai. 2024.

AGRADECIMENTOS

Este trabalho foi financiado por fundos nacionais através da FCT – Fundação para a Ciência e Tecnologia, I.P., no âmbito do projeto CERNAS UIDB/00681/2020.