



10° EBBC

Encontro Brasileiro de
Bibliometria e Cientometria



DANDARADATA: programa para identificação de Comunidades Remanescentes de Quilombos na produção científica indexada na SciELO

Josafha Pereira de Carvalho

<https://orcid.org/0009-0006-8446-1489>

josafha.carvalho@itec.ufpa.br

Universidade Federal do Pará (UFPA)

Ramiro Arthur Sobrinho Freire Silva

<https://orcid.org/0009-0003-4240-1435>

ramiro.silva@itec.ufpa.br

Universidade Federal do Pará (UFPA)

Ediane Maria Gheno

<https://orcid.org/0000-0003-2743-4557>

edianeigheno@ufpa.br

Universidade Federal do Pará (UFPA)

Resumo:

O estudo apresenta o desenvolvimento de um programa (DandaraData) para automação da extração de dados da produção científica sobre Comunidades Remanescentes de Quilombos indexada da SciELO (1996-2024). O objetivo foi solucionar o desafio metodológico da identificação de entidades geográficas em textos não estruturados, especificamente no mapeamento de Comunidades Remanescentes de Quilombos certificadas pela Fundação Cultural Palmares em artigos científicos. O DandaraData utiliza expressões regulares, validação de contexto semântico e priorização geográfica para cruzar metadados da produção científica com a lista das comunidades certificadas. Os resultados demonstraram alta precisão na recuperação de informações, eliminando falsos positivos e viabilizando a análise exata em larga escala das comunidades quilombolas que foram objetos de pesquisa na produção científica.

Palavras-Chave: Mineração de Texto. Expressões Regulares. DandaraData. Comunidades Remanescentes de Quilombos. Bibliometria.

EIXO TEMÁTICO:

Técnicas, Ferramentas e
Infraestruturas

MODALIDADE:

Iniciação Científica

DOI: 10.22477/x.ebbc.999

COMO CITAR ESTE ARTIGO:

CARVALHO, Josafha Pereira de; SILVA, Ramiro Arthur Sobrinho Freire; GHENO, Ediane Maria. DandaraData: programa para identificação de Comunidades Remanescentes de Quilombos na produção científica indexada na SciELO. In: EBBC - Encontro Brasileiro de Bibliometria e Cientometria. 10. **Anais [...]**. Curitiba, 2026. DOI: 10.22477/x.ebbc.999. Disponível em: <https://doi.org/10.22477/x.ebbc.999>.



1 INTRODUÇÃO

Os estudos métricos da informação dependem da qualidade dos metadados extraídos nas bases científicas. A identificação de entidades geográficas em textos não estruturados constitui desafio recorrente, especialmente quando há homônimas e ambiguidades. No mapeamento da produção científica sobre Comunidades Remanescentes de Quilombos (CRQs), se observou que muitos nomes coincidem com termos variados, dificultando a recuperação automática da informação. O cruzamento manual entre metadados da produção científica e a lista oficial da Fundação Cultural Palmares (FCP) torna-se inviável em larga escala, enquanto abordagens probabilísticas podem gerar muitos falsos positivos indesejáveis.

Nesse contexto, questionou-se: quais CRQs certificadas pela FCP foram objeto de estudo na produção científica indexada na *SciELO* sobre comunidades quilombolas (1966-2024)? Para responder a essa questão, o presente trabalho apresenta o desenvolvimento do DandaraData, programa computacional concebido para realizar cruzamento determinístico entre metadados da produção científica com a planilha oficial de CRQs certificadas pela FCP. Portanto, o objetivo foi solucionar o desafio metodológico da identificação de entidades geográficas em textos não estruturados, especificamente no mapeamento de CRQs certificadas pela FCP em artigos científicos. Como objetivo específico, buscou-se: identificar o número de CRQs certificadas pela FCP que foram pesquisadas e avaliar a precisão do programa Dandaradata.

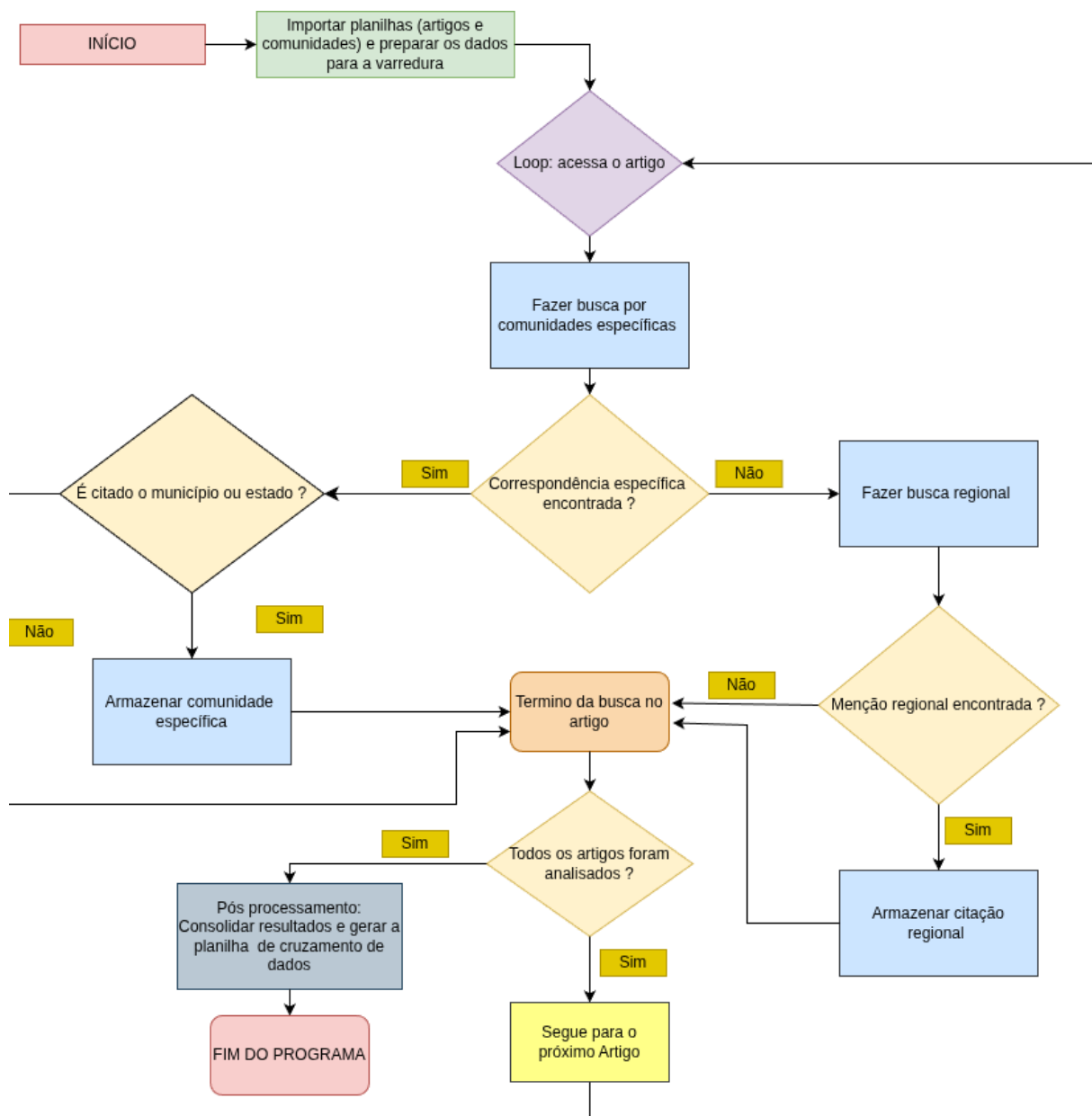
O nome do programa é uma homenagem a Dandara, do Quilombo dos Palmares, símbolo de resistência ao sistema escravista.

2 METODOLOGIA

O DandaraData foi desenvolvido em Python, com a biblioteca Pandas, com o paradigma de Casamento de Padrões utilizando Expressões Regulares (Regex). A lógica do programa ocorre em módulos sequenciais, ilustrados na Figura 1.

Segundo Friedl (2006), as expressões regulares permitem ao pesquisador dominar seus dados, oferecendo poder de varredura de grandes volumes de informações e verificar entradas com precisão extrema. No DandaraData, esse domínio sobre os dados se materializa alinhado às regras de validação para uma máxima exatidão na recuperação das entidades nomeadas. A base do DandaraData utiliza Regex associada a verificações lógicas estruturadas para identificar se CRQs (Fundação Cultural Palmares, 2025), bem como seu estado e/ou município, estão presentes nos metadados (título, palavras-chave e resumo) do corpus bibliográfico: artigos em periódicos indexados na *SciELO* sobre comunidades quilombolas, período de 1996-2024. Para recuperar os registros, foi utilizado a seguinte expressão de busca: `TS=(Quilombo* OR Kilombo* OR Quilombis* OR Quilombagem OR Aquilomba*)`. Não havendo casos de duplicidade, o total de artigos foi de 542. As demais tipologias de documento não foram consideradas. Tal procedimento de coleta, inclusão e exclusão foram ancorados em Pantoja, Borralho e Gheno (2025).

Figura 1 - Fluxo lógico e arquitetura de decisão do algoritmo DandaraData



Fonte: elaborada pelos autores (2026).

Conforme a Figura 1, para que o casamento de padrões entre a CRQ e o artigo seja validado, o sistema exige, obrigatoriamente, a coocorrência do nome da CRQ com seu respectivo município ou estado nos metadados do artigo. Atendido esse critério, aplica-se uma priorização geográfica: município e estado recebem nível de prioridade alto; apenas o município recebe nível médio; e a apenas o estado, atribui-se o nível baixo. Isto é de grande importância para resolver homônimas

entre localidades distintas com o mesmo nome. Em seguida, para resolver ambiguidades semânticas de comunidades cujos nomes são palavras de uso comum, o algoritmo impõe uma validação de contexto (janela de quatro palavras precedentes), buscando descritores obrigatórios como “comunidade” ou “quilombo”. Se após o fim das buscas não houver menções a nenhuma comunidade, o DandaraData procura por menções generalistas, via busca regional (ex., comunidades do sudoeste de Bahia). No pós-processamento, os resultados são consolidados por identificadores únicos para evitar duplicidade de entidades transfronteiriças. Por fim, devido à imprevisibilidade da escrita nos artigos (erros de digitação, apelidos não oficiais), a metodologia exige uma etapa de curadoria manual (checagem humana) para corrigir inconsistências (falsos positivos e falsos negativos). O *Script* do programa DandaraData está disponível no Github: <https://github.com/Mensu-reLab/DandaraData>.

3 RESULTADOS

A eficiência do DandaraData foi testada no processamento de 542 artigos científicos indexados na *SciELO* de 1996-2024, cuja varredura gerou 142 menções a CRQs. A seguir, deu-se a etapa de validação manual que atestou as métricas reais do sistema: das 142 menções extraídas, 16 se tratavam de falsos positivos. Esses ruídos dividiram-se em nove (9) erros de identificação pura e sete (7) duplicações por ambiguidade geográfica (ocorrida na ausência do nome do município em estados com múltiplas comunidades homônimas), totalizando 126 menções corretas. Após esse processo de validação, identificou-se ainda 30 falsos negativos (menções reais não capturadas) decorrentes de erros de grafia dos autores (ex: “pimenteira” ao invés de “Pimenteiras”), apelidos ou ausência total de marcadores geográficos, os quais foram incluídos na análise. Desse modo, somadas às men-

ções corretas (26), totalizou um montante de 156 menções¹.

Ao avaliar o desempenho do DandaraData pelas métricas de precisão (proporção de correspondências positivas corretamente identificadas entre todas as correspondências positivas realizadas) e *recall* (proporção de correspondências positivas entre todos os casos positivos reais), foi registrada a precisão de 88,7% e *recall* de 80,7%.

A consolidação das 156 menções corretas de CRQs certificadas pela FCP, observou-se que elas estavam distribuídas em 127 artigos científicos (representando 23,4% do total). Após análise de ocorrências, foram identificadas 124 CRQs únicas, o que evidencia que algumas CRQs foram objeto de estudo em mais de um artigo.

4 CONSIDERAÇÕES FINAIS

O DandaraData demonstrou eficácia na extração de entidades em metadados. A precisão de 88,7% viabiliza estudos bibliométricos em larga escala, automatizando o cruzamento massivo de dados. Além disso, a taxa de *recall* (80,7%) corrobora que algoritmos determinísticos em textos de linguagem natural exigem validação manual para lidar com a imprevisibilidade da escrita humana. O programa atua como um suporte essencial para mapear a visibilidade de comunidades quilombolas na ciência, sendo sua lógica replicável para outros domínios. Pretende-se aprimorar o DandaraData e manter um banco de dados de acesso aberto de evidências científicas sobre CRQs, contribuindo para políticas públicas e enfrentamento à casos de justiça que os territórios enfrentam.

¹ Formalizando o total de menções, seja: P trata-se do número de menções recuperadas pelo programa; FP, número de falsos positivos identificados na validação manual; e FN, o número de falsos negativos encontrados durante essa validação. O número total de menções R é dado por: $R = P + FP - FN$.



REFERÊNCIAS

PANTOJA, Lidiane Vilhena; BORRALHO, Anderson do Rosario; GHENO, Ediane Maria. Comunidades quilombolas: produção científica, coautoria e temas emergentes indexados na *SciELO* (1996-2023). **Biblios Journal of Librarianship and Information Science**, Chile, (esp.), e008, 2025. DOI <https://doi.org/10.5195/biblios.2025.1389>. Acesso em: 15 jun. 2026

FRIEDL, Jeffrey. **Mastering Regular Expressions**. 3. ed. Sebastopol: O'Reilly Media, 2006.

FUNDAÇÃO CULTURAL PALMARES. **Certificação Quilombola**. 2025. Disponível em: <https://www.gov.br/palmares/pt-br>. Acesso em: 12 maio 2025.