MÉTODOS E FERRAMENTAS PARA COLETA DE DADOS INDÍGENAS NOS PORTAIS DE DADOS ABERTOS DO BRASIL E ESTADOS UNIDOS

RICARDO COSTA ROSSI*
MARIÂNGELA SPOTTI LOPES FUJITA**
ISIDORO GIL-LEIVA***

Resumo: *A heterogeneidade dos dados indígenas representa um obstáculo na sua coleta. Essa dificuldade decorre da insuficiência de informações sobre essas comunidades nos portais governamentais. Neste artigo investigamos os métodos e ferramentas que permitam coletar dados nos portais de dados abertos do Brasil e Estados Unidos. As ferramentas recuperadas foram obtidas através de repositórios bibliográficos e nos próprios portais utilizando «operadores booleanos» e «Termos referentes ao tema desta pesquisa», sendo processados e selecionados oito estudos e 36 API que abordam a web semântica ou métricas de avaliação de portais de dados. Foram destacadas a limitação de dados indígenas disponíveis nos portais e a necessidade de chave de acesso no portal brasileiro, o que viola um dos princípios de dados abertos. Concluiu-se que o portal brasileiro possui uma interface gráfica mais amigável. O portal dos EUA exige maior conhecimento técnico, mas oferece API para coleta de metadados.*

Palavras-chave: *Dados abertos; Ferramentas para coleta de dados; Web semântica.*

Abstract: *The heterogeneity of indigenous data represents an obstacle to its collection. This difficulty arises from the lack of information about these communities on government portals. In this article, we investigate the methods and tools that allow data collection on open data portals in Brazil and the United States. The tools retrieved were obtained through bibliographic repositories and on the portals themselves using «Boolean operators» and «Terms related to the theme of this research», and eight studies and 36 APIs that address the semantic web or data portal evaluation metrics were processed and selected. The limitation of indigenous data available on the portals and the need for an access key on the Brazilian portal were highlighted, which violates one of the principles of open data. It was concluded that the Brazilian portal has a more user-friendly graphical interface. The US portal requires greater technical knowledge, but offers APIs for collecting metadata.*

Keywords: *Open data; Data collection tools; Semantic web.*

* Universidade Estadual Paulista «Júlio de Mesquita Filho» (UNESP). Email: ricardo.rossi@unesp.br. ORCID: https://orcid.org/0000-0002-9910-0065.
** Universidade Estadual Paulista «Júlio de Mesquita Filho» (UNESP). Email: ricardo.rossi@unesp.br. ORCID: https://orcid.org/0000-0002-8239-7114.
*** Universidad de Murcia. Email: isgil@um.es. ORCID: https://orcid.org/0000-0002-7175-3099.

INTRODUÇÃO

A heterogeneidade dos dados abertos indígenas representa um dos maiores obstáculos na sua coleta. Essa dificuldade decorre em grande parte da insuficiência de informações específicas sobre comunidades indígenas disponíveis nos portais de dados governamentais. Segundo Bandeira et al. (2014), dados governamentais abertos frequentemente são difíceis de usar, pois vêm em formatos que máquinas não conseguem ler facilmente, são desorganizados ou proprietários.

A insuficiência dessas informações culmina na falta de políticas públicas para defesa desses povos. O Instituto Brasileiro de Geografia e Estatística (IBGE) ([s.d.]) enfatiza que os dados do Censo Demográfico são a principal fonte de dados para entender as condições de vida da população em todos os municípios, auxiliando a administração pública e o planejamento social e econômico.

Dessa forma, a disponibilização de dados abertos sobre os povos indígenas pode ter impacto direto na equidade e na inclusão, pois não só permite uma maior compreensão de suas realidades, mas também oferece aos pesquisadores a possibilidade de realizar estudos mais aprofundados e contextualizados. Os dados relacionados à situação dessas comunidades são cruciais para estudos em áreas como antropologia, medicina ou educação. Os estudos gerados a partir desses dados não só contribuem para o avanço do conhecimento em diferentes disciplinas, mas também podem influenciar na formulação de políticas públicas atreladas a problemas reais.

Lógico que quando nos referimos a dados indígenas não podemos deixar de lado a governança desses dados, pois referem-se a vasta e distinta quantidade de materiais culturais tangíveis e intangíveis e dessa forma devemos tratá-los com ética e respeito. Segundo Souza et al. (2023), os princípios CARE *(Collective Benefit, Authority to Control, Responsibility, Ethics*) promovem diretrizes para promover a governança e autodeterminação de dados indígenas que abrangem autoridade para controlar a responsabilidade e ética sobre esses dados.

Uma combinação que propulsiona a motivação desta pesquisa seria entre tecnologia e dados abertos que têm o potencial de transformar a forma como a pesquisa é conduzida e compartilhada. Borgman (2012) apresenta quatro amplas justificativas para o compartilhamento de dados de pesquisa: reprodutibilidade, atendimento ao interesse público, fazer novas perguntas e avançar a pesquisa. O compartilhamento de dados abertos pode contribuir para a resolução de problemas sociais e ambientais, permitindo que pesquisadores e formuladores de políticas tomem decisões informadas com base em evidências.

A visibilidade da coleta de dados depende de ser realizada com eficiência e para isso esses dados devem estar estruturados em conjuntos de dados disponibilizados publicamente, permitindo que qualquer pessoa possa acessá-los, utilizá-los e redistribuí-los sem restrições.

Para que os conjuntos de dados sejam criados de forma organizada surgiu a criação do plano de dados abertos que se refere a um conjunto de diretrizes e estratégias para disponibilizar informações de forma transparente, acessível e reutilizável para o público em geral. Para medir o desempenho de instituições governamentais na abertura de dados foi estabelecida a Open Government Partnership (OGP) ou Parceria para Governo Aberto para que os «Planos de Ação Nacionais sejam constituídos de compromissos de Estado alinhados aos princípios do Governo Aberto, quais sejam: Transparência, Accountability, Participação Cidadã e Tecnologia e Inovação» (Bertin et al. 2019, p. 2).

Isso envolve a publicação de dados em formatos abertos e interoperáveis, garantindo a privacidade e segurança das informações, promovendo a transparência e a participação cívica, incentivando a integração de dados públicos com outras bases de dados. Entretanto o Governo deve atuar como provedor de infraestrutura de dados, assegurando acesso simples e confiável às informações brutas, de modo que permita sua coleta e reutilização (Robinson et al. 2009).

Diante do exposto, esta pesquisa visa identificar métodos e ferramentas em repositórios bibliográficos e nos portais de dados abertos do Brasil e Estados Unidos que permitam coletar automaticamente dados indígenas, bem como explorar e analisar os recursos disponibilizados nesses portais, imprescindíveis para a utilização dessas ferramentas de forma eficaz.

1. DADOS ABERTOS: CONCEITOS E APLICAÇÕES

Os dados abertos fomentam a colaboração entre diferentes setores, como governo, academia e sociedade civil, para resolver problemas complexos e promover o bem comum.

Segundo os autores Isotani e Bittencourt (2015), dados abertos são dados que podem ser livremente utilizados, reutilizados e redistribuídos por qualquer pessoa, que é fundamental para promover a transparência, a inovação e a participação cidadã, especialmente em contextos governamentais e sociais.

Ainda nesse sentindo Neves (2013) afirma que dados abertos significam que qualquer pessoa ou organização pode usar informações públicas de forma gratuita para criar aplicativos, análises ou até produtos. Para serem considerados abertos, esses dados devem ser fáceis de acessar, usar e redistribuir, sem restrições. Além disso, precisam ser facilmente encontrados em locais indexados, ou seja, estruturados, lidos por máquinas e sem impedimentos legais.

Para que haja eficiência no uso das informações públicas esses dados devem estar estruturados, e podem ser definidos como aqueles que se encontram minimamente organizados em colunas, sendo o formato *CSV* um padrão comum para o seu armazenamento. Amaral (2016) descreve o inverso dizendo que dados

não estruturados são como aqueles que não possuem qualquer tipo de estrutura definida, muitas vezes não possuem sequer qualquer tipo de metadado e mesmo quando os possuem esses metadados não são de muita ajuda para os processos de análise.

Quando nos referimos a dados abertos, não podemos deixar de lado os princípios e pilares dos dados abertos que são amplamente discutidos em literatura especializada. Desse modo é importante verificar como é feita a sua disponibilização dos dados abertos, se seguem os 8 princípios de dados abertos estabelecidos em um encontro realizado em dezembro de 2007 em Sebastopol, na Califórnia, que reuniu pesquisadores, representantes de organizações da sociedade civil e ativistas norte-americanos.

Os princípios dos dados abertos estipulam que dados públicos devem ser completos, primários (apresentados como coletados da fonte), atuais (publicados rapidamente para preservar seu valor), e acessíveis à maior quantidade de pessoas. Além disso, devem ser compreensíveis por máquina (estruturados para processamento automático, como em *CSV* ou XML), não discriminatórios (disponíveis sem necessidade de cadastro), não proprietários (nenhuma entidade deve ter controle exclusivo), e livres de licenças (não sujeitos a *copyrights*, patentes ou segredos industriais) (W3C [s.d.]a).

Por fim, a combinação de dados abertos e tecnologia pode gerar insumos para localizar informações indígenas de forma mais eficiente beneficiando diretamente essas comunidades.

2. PROCEDIMENTOS METODOLÓGICOS

A tendência global é de que cada vez mais países adotem políticas e legislações que promovam a transparência e a disponibilização de dados governamentais em formatos abertos gerando insumos para pesquisadores desenvolverem estudos em diversas áreas. Dessa forma para execução dessa pesquisa será utilizada uma abordagem exploratória e descritiva nos portais de dados dividida em duas fases: a primeira fase consiste em identificar ferramentas de Tecnologia da Informação (TI) em repositórios científicos e nos portais de dados governamentais de âmbito federal do Brasil e dos Estados Unidos da América (EUA) e a segunda fase consiste na análise dos recursos tecnológicos disponibilizados nos portais investigados nesta pesquisa. Abaixo segue o detalhamento de casa fase:

Primeira fase — *Identificação de ferramentas de tecnologia da informação*: é dividida em duas etapas:

a) Busca e seleção nas bases de dados Scopus e Google Acadêmico utilizando os operadores booleanos «AND», «OR» e os termos «Open Data», «Native People», «Web Semântica» e «Web Scraping» que abordem: (I) Tecnologias

que envolvem a *web* semântica; (II) Técnicas de *web scraping*[1] — para extração automatizada de conjuntos de dados nos portais; (III) Processos de ETL[2] (*Extract, Transform, Load*) — para coleta e tratamento dos dados.

b) Busca e seleção de *API* nos portais de dados utilizando o termo *API*, observando: a quantidade de *API*[3] disponibilizadas e usabilidade[4] do portal e das ferramentas/*API*.

Segunda fase — *Análise dos recursos tecnológicos disponibilizados nos portais*: análise das amostras de dados indígenas nos portais de dados brasileiro e dos EUA. Para coleta destas amostras foram utilizadas as palavras-chave «tribes» no portal data.gov e «Povos Indígenas» no portal dados.gov.br que englobam os formatos de arquivos disponibilizados (CSV, XLS, XML, JSON) e disponibilizam metadados sobre os conjuntos de dados.

3. RESULTADOS

Os portais de dados abertos governamentais explorados apresentam método de interoperabilidade de dados com uso de interface gráfica como é o caso do Brasil e outro que utiliza comandos em linguagem de programação para coletar metadados e dados que é o portal dos Estados Unidos. Os resultados foram organizados de acordo com os métodos e ferramentas identificados a partir da pesquisa bibliográfica e das *API* localizadas nos portais de dados.

3.1. Métodos e ferramentas identificados pela pesquisa bibliográfica

Foram identificados oito estudos que abordam: a *web* semântica para coleta de metadados, ferramentas para publicação de metadados envolvendo informações sobre conjunto de dados e ferramentas para descoberta de conjunto de dados. Nesses estudos foram identificadas as ferramentas para coleta de dados ou metadados com seu respectivo método para utilização das ferramentas. A tabela abaixo demonstra os métodos acompanhados das respectivas ferramentas:

[1] *Web scraping* — é uma técnica para converter dados da *web* não estruturados em dados estruturados (Thomas e Mathur 2019).
[2] ETL (*Extract, Transform, Load*) — processamento de dados, garantindo que as informações extraídas de diferentes fontes sejam transformadas e carregadas de forma eficiente para sistemas de informação (Vida et al. 2021).
[3] *API* são mecanismos que possibilitam a comunicação entre componentes de *software* por meio de definições e protocolos (Amazon Web Services [s.d.]).
[4] Usabilidade — a medida na qual um produto pode ser usado por usuários específicos para alcançar objetivos específicos com eficácia, eficiência e satisfação em um contexto específico de uso (International Organization for Standardization 1998).

Tabela 1. Estudos, métodos e ferramentas identificadas na pesquisa bibliográfica

Id	Título	Método	Ferramenta	Referência
1	*Applications of Semantic Web in integrating open data and bibliographic records: a development example of an infomediary of Taiwanese indigenous people*	Integração de diversas fontes de dados através de consultas federadas *SPARQL*	*SPARQL* — é uma linguagem de consulta em arquivos que utilizam *RDF*.	Sung e Chi 2021
2	*Proposta de Arquitetura de Publicação Automatizada de Dados Abertos Conectados Utilizando Meta-Dados e Ontologias*	Criação de arquitetura de *software* para publicação de dados abertos. Possui 3 camadas: Extração de dados, indexação semântica e busca semântica. Arquitetura de *software* — refere-se à organização geral do *software* e aos modos pelos quais proporciona integridade ao sistema (Shaw e Garland, 1995)	*UnB-LOD* e *DBgoldbr* Integração dos dados *UnBGOLD* Indexação e publicação de metdados	Martins 2019
3	*Towards the Ethnic Understanding of Taiwanese Indigenous Peoples: A Mashup Based on Semantic Web and Open Data*	Método dividido em três procedimentos: 1) Coleta de dados e formato RDF e Não *RDF*. 2) Coversão dos dados Não *RDF* em *RDF*. 3) Consultas federadas *SPARQL* para implementar Marshup de dados.	*SPARQL* — Linguagem de consulta para dados representados em *RDF*	Chi, Sung e Lien 2020
4	*Metadados para descrição de datasets e recursos informacionais do «Portal Brasileiro de Dados Abertos»*	Identificação dos principais elementos dos metadados recuperados, seguindo o padrão de metadados do governo eletrônico (E-PMG1.1). Utilizando a técnica *Crosswalk*.	*Crosswalking*. estabelece relações entre esquemas ou vocabulários diferentes.	Moreira et al. 2017
5	*Arquitetura de publicação de dados abertos conectados governamentais da Universidade de Brasília*	Método divido em duas etapas: 1) Buscar elementos em metadados, ontologias e arquivos RDF. 2) Definir quais elementos podem ser usados. Formular uma arquitetura de software através da ferramenta *UnBGOLD*.	A ferramenta *UnBGOLD*	Victorino et al. 2020
6	*Dados governamentais abertos: métricas e indicadores de reuso*	Modelo de avaliação de dados abertos governamentais com base em métricas e indicadores internacionais.	Avaliação de portais com base na métrica *DGAbr*.	Silva 2018
7	*Google Dataset Search: visão geral e perspectivas para indexação e disponibilização de conjuntos de dados científicos abertos*	Identificar, indexar e disponibilizar pela internet conjunto de dados	Ferramenta de pesquisa de conjuntos de dados. *Google Dataset Search*	Pinto e Amaral 2020
8	*JavaScript Web Scraping Tool for Extraction Information from Agriculture Websites*	Categorizar diversas técnicas, ferramentas e bibliotecas para extração de informações de conteúdo *web* não estruturado.	*Ferramentas de Web Scraping (Octorparse e Parse Hub.*	Zhekova e Yumer 2024

Fonte: Elaborado pelos autores

Desses oito estudos selecionados nesta pesquisa, a maioria utiliza ferramentas para captura de dados, incluindo tecnologias de *web* semântica, *web scraping*. Entre os exemplos práticos, destacam-se: a utilização do UnBGOLD para publicação e indexação de metadados e o desenvolvimento de um infomediário voltado a povos indígenas de Taiwan, demonstrando o potencial dessas tecnologias.

Para melhor entendermos onde seria melhor o funcionamento dessas ferramentas dividimos em 3 grupos: Grupo 1 (artigos 1 e 2) refere-se a coleta de metadados; Grupo 2 (artigos 3, 4 e 5) refere-se a ferramentas para publicação de metadados e por fim o Grupo 3 (artigos 6, 7 e 8) trata da descoberta de dados ou avaliação de dados governamentais.

O grupo 1 — Coleta de dados utilizando a *web* semântica: Sung e Chi (2021) propõem o uso de consultas federadas SPARQL[5] baseadas em ontologia para integrar dados e realizar buscas semânticas. Complementarmente, Chi, Sung e Lien (2020) desenvolvem consultas federadas SPARQL para integrar dados em formato RDF[6] facilitando a compreensão étnica. A conexão entre eles é evidente no uso de SPARQL e RDF para a *web* semântica.

O grupo 2 — Publicação de metadados: Martins (2019) propõe uma arquitetura para publicação de metadados automatizada utilizando a ferramenta UnBGOLD. Moreira et al. (2017) discutem metadados para descrever conjuntos de dados, associando os elementos de metadados identificados nos *datasets* com os elementos do e-PMG 1.1 utilizando a técnica *Crosswalking*. Por fim, Victorino et al. (2020) apresentam a ferramenta UnBGOLD para auxiliar na publicação de metadados.

O grupo 3 — Trata da coleta de dados ou avaliação de dados governamentais: para que seja realizada uma coleta de dados eficiente é necessário avaliar como esses dados são disponibilizados, neste caso Silva (2018) avalia dados governamentais abertos com base na métrica DGABr. Pinto e Amaral (2020) exploram o Google Dataset Search como ferramenta com o propósito de identificar, indexar e disponibilizar pela internet *datasets* e Zhekova e Yumer (2024) apontam duas ferramentas eficazes para *web scraping*: Octorparse: esta ferramenta extrai textos, vídeos e imagens de *websites* e devido a sua interface interativa possibilita que o usuário crie fluxos de trabalho personalizados para a extração de dados. ParseHub: se destaca por coletar dados de *sites* dinâmicos que utilizam AJAX e JavaScript, usando *machine learning* para transformar conteúdo *web* em dados estruturados.

Outra ferramenta para *web scraping* é a Apify, que também merece destaque. Sua integração com outras *API* a torna particularmente útil para pesquisas que necessitam de um alto nível de automação e conectividade com outras plataformas.

[5] SPARQL — facilita a consulta e a manipulação de conteúdo de arquivos RDF na *web* ou em um repositório RDF (W3C [s.d.]c).
[6] O Resource Description Framework (RDF) é um *framework* para representar informações na *web*. Este documento define uma sintaxe abstrata que serve para conectar todas as linguagens e especificações baseadas em RDF (W3C [s.d.]b).

3.2. *API (Application Programming Interface)*

Além dos métodos e ferramentas demonstrados acima, dando início ao item b) da fase 1 da metodologia foram encontradas 36 soluções para coleta de dados disponibilizadas pelos portais de dados governamentais do Brasil e dos Estados Unidos da América (EUA), porém sua utilização demanda conhecimentos intermediários em tecnologia mesmo com o auxílio de inteligência artificial, devido à necessidade de parametrização dos sistemas. Comparativamente, as *API* do portal brasileiro mostraram-se mais intuitivas e acessíveis que as do portal dos EUA, que apresentaram complexidade excessiva e deficiências em usabilidade.

Ambos os portais, o Portal Brasileiro de Dados Abertos e o data.gov dos Estados Unidos, utilizam o CKAN como plataforma subjacente para seus catálogos de dados abertos. Isso significa que a funcionalidade central de suas *API CKAN* é inerentemente semelhante, a diferença é que o portal de dados brasileiro disponibiliza uma interface gráfica para a utilização deste recurso.

As *API* do portal brasileiro podem ser acessadas através do *link*[7] que lista as *API* disponíveis para acesso a conjuntos de dados públicos. No caso do portal de dados abertos brasileiro, a *API* funciona mediante *login* e senha da plataforma gov.br. Após logado na plataforma é necessário gerar um *token* (chave de acesso) para acessar os dados. Após gerar a chave de acesso é só escolher a *API*, preencher os campos para gerar o código, executá-lo e obter os dados. Esses procedimentos contrariam um dos 8 princípios de dados abertos, ou seja, o acesso identificado do usuário através do *token* fere o princípio do número 6 que é não ser discriminatório, os dados devem estar disponíveis para qualquer pessoa, sem necessidade de cadastro ou qualquer outro procedimento que impeça o acesso.

Com relação ao portal de dados dos Estados Unidos, quase no rodapé do portal encontra-se o repositório de dados contendo *links* para acesso às *API* e estudos de caso, que tendem a ser mais difíceis exigindo um conhecimento mais técnico em informática comparando com as do portal brasileiro, dessa forma exigindo maior familiaridade com linguagens de programação para sua utilização. No entanto, o portal norte-americano oferece manuais, orientações detalhadas e estudos de caso que auxiliam no desenvolvimento de *API* personalizadas para fins específicos como coleta de metadados.

A seção *Data Tools* (disponível em resources.data.gov) faz parte do portal de dados norte-americano, uma iniciativa do governo dos EUA que consolida ferramentas, recursos e diretrizes para apoiar agências federais, desenvolvedores e o público no gerenciamento, publicação e uso eficiente de dados abertos. Esta seção funciona como um catálogo de ferramentas de dados disponibilizadas pelo portal, promovendo

[7] Disponível em: https://dados.gov.br/swagger-ui/index.html.

acessibilidade e reutilização de dados governamentais, em conformidade com políticas como o *OPEN Government Data Act*.

Dentre as 14 *API* de dados disponibilizadas nesta seção, selecionamos a *API CKAN* para que possamos fazer uma comparação entre os dois portais. No caso do portal de dados dos Estados Unidos este nos direciona para uma página (ckan.org) onde podemos escolher entre «CKAN para governo» ou «CKAN para empresas». Ao selecionar a opção «CKAN para governo», o *website* nos apresenta uma lista de portais de dados disponíveis, incluindo opções como Singapura, Canadá e Estados Unidos. Com o *link* para acesso ao portal de dados dos EUA, o *website* redireciona para a página inicial do data.gov (portal americano de dados abertos), sem demonstrar como é utilizada a *API CKAN* para captura de dados nos conjuntos de dados disponibilizados pelo portal, caracterizando falta de usabilidade do portal.

As *API DKAN*[8] e Assistente de Visualização de Dados são soluções desenvolvidas sobre a plataforma Drupal[9]. No entanto, as interfaces geradas por essas ferramentas apresentam significativos desafios de usabilidade. Ao acessá-las, os usuários são direcionados para *websites* com estruturas complexas e fluxos de navegação pouco intuitivos, resultando em uma experiência fragmentada. Essa complexidade acaba restringindo o uso efetivo principalmente a especialistas familiarizados com a ferramenta Drupal.

Essas tecnologias seriam uma ferramenta poderosa para acessar e integrar os conjuntos de dados disponibilizados nos portais investigados nesta pesquisa. Tal abordagem é particularmente relevante para temas específicos como os dados indígenas, que constituem a problemática central deste estudo, focando na heterogeneidade de dados publicados. Abaixo a Tabela 2 retrata um panorama das ferramentas ou *API* identificadas nesta pesquisa.

Tabela 2. Ferramentas Identificadas nesta pesquisa

<table>
<tr><th>API</th><th>Webscraping</th><th>ETL</th></tr>
<tr><td>Portal de dados - Brasil</td><td rowspan="4">Foram identificadas duas ferramentas:

(OctoParse e ParseHub),

que são soluções visuais para coleta automatizada de dados da web.</td><td rowspan="4">Foram listadas quatro ferramentas:
• Apache NiFi;
• Talend;
• Pentaho Data Integration;
• Power Query.
Cada uma com características distintas, desde soluções open-source até integrações com plataformas como Microsoft Excel e Power BI</td></tr>
<tr><td>O portal brasileiro oferece 22 API's, com um processo de autenticação via GOV.BR e geração de token. A interface é interativa, permitindo a parametrização e geração de códigos diretamente no portal.</td></tr>
<tr><td>Portal de dados - EUA</td></tr>
<tr><td>O portal norte-americano disponibiliza 14 API's, sendo mais técnicas e menos interativas.</td></tr>
</table>

Fonte: Elaborado pelos autores

[8] DKAN — É responsável por criar catálogos de dados abertos modernos utilizando recursos do módulo drupal, focados em padrões e priorizando API para uma variedade de casos de uso específicos. (CivicActions [s.d.])
[9] O Drupal é um sistema de gerenciamento de conteúdo (CMS – Content Management System) *open source* escrito em PHP, usado para criar e gerenciar *sites* complexos e aplicações *web*.

Encerramos a primeira fase e na segunda fase realizamos a análise das amostras de dados indígenas coletadas nos portais de dados brasileiro e dos EUA em conjunto com as ferramentas de dados coletadas.

3.3. Análise dos recursos tecnológicos disponibilizados nos portais: dados indígenas X Metadados

Esta seção se dedicou à análise das amostras de conjuntos de dados indígenas coletados nos portais de dados investigados nesta pesquisa. A princípio foram analisados quais arquivos em formatos estruturados (JSON, XML e CSV) são disponibilizados nesses conjuntos e quais podem ser efetivamente utilizados com as ferramentas de coletas de dados selecionadas. Para embasar essa análise, realizamos uma coleta inicial de conjuntos de dados centrados na temática indígena, utilizando as palavras-chave «Tribes» e «Povos Indígenas».

No portal de dados dos Estados Unidos, a busca pela palavra-chave «Tribes» resultou em 433 conjuntos de dados. Desses, selecionamos cinco conjuntos que continham a palavra-chave no título, conforme detalhado na tabela abaixo:

Tabela 3. Conjuntos de dados capturados no portal de dados dos EUA

Nome do conjunto de dados	Tipo de ação para coleta de dados
TIGER/Line Shapefile, 2020, Nation, U.S., American Indian Tribal Subdivisions	Disponibiliza um arquivo no formato *XML* com metadados sobre os conjuntos de dados, como título, descrição e outras informações descritivas
EPA Tribes (3 of 6): American Indian Reservations	Não disponibiliza arquivo para *download*, mas um *link* para uma página com um arquivo em *XML* contendo os metadados sobre os conjuntos de dados, como título, descrição e outras informações descritivas.
EPA Tribes (2 of 6): Alaska Native Villages	Não disponibiliza arquivo para *download*, mas um *link* para uma página com um arquivo em *XML* contendo os metadados sobre os conjuntos de dados, como título, descrição e outras informações descritivas.
Federally Recognized Tribal Lands	Disponibiliza um arquivo para *download* no formato *CSV*, porém contém o nome das reservas e sua geolocalização
TIGER/Line Shapefile, 2022, Nation, U.S., American Indian/ Alaska Native/Native Hawaiian Areas (AIANNH)	Disponibiliza para *download* um arquivo para ser utilizado com a a *API ERIS2OPEN*

Fonte: data.gov [s.d.]

Observamos que a maioria dos conjuntos de dados no portal dos Estados Unidos fornece metadados, mas não os dados em si. Isso posiciona o portal como um catálogo centralizado de metadados, que disponibiliza essas informações principalmente em arquivos XML. Consequentemente, a experiência do usuário muitas vezes se limita à descoberta de descrições e não o acesso aos dados.

Um problema recorrente é a disponibilização inadequada dos arquivos *CSV*. Por exemplo, o conjunto de dados «Federally Recognized Tribal Lands» contém dados geoespaciais sobre reservas, sem informações adicionais relevantes, não oferecendo dados em comum uns com outros e dificultando a criação, por exemplo, de um repositório. Essa disponibilização inadequada de dados nos portais é o que chamamos de «arquivos de dados heterogêneos».

O portal de dados brasileiros com a palavra-chave «Povos Indígenas» recuperou 33 conjuntos de dados. Desses, selecionamos cinco conjuntos, conforme detalhado na Tabela 4 abaixo:

Tabela 4. Conjuntos de dados capturados no portal de dados do Brasil

Nome do conjunto de dados	Tipo de ação para coleta de dados
Tabela Povos Indígenas do Brasil	Todos os 5 conjuntos de dados selecionados disponibilizam arquivos no formato *CSV*, e também disponibilizam informações sobre metadados contendo título, descrição do conjunto de daos dentre outros.
Gráficos sobre demandas de Ingresso em Terra Indígena	
Informações sobre indígenas respondendo a processos criminais	
Tabela de aldeias indígenas	
Tabelas de terras indígenas	

Fonte: gov.br. [s.d.]

No portal de dados brasileiro, através da seleção desses 5 conjuntos de dados, observamos que é possível a utilização de *API* que capturem metadados dos conjuntos de dados e seu conteúdo. Lembramos que para a coleta de dados o portal de dados brasileiro possui uma interface gráfica facilitando a coleta. O ponto fraco seria a utilização de *token* para acesso aos dados.

Esta pesquisa identificou 3 pilares para coletas de dados: Ferramentas de *web* semântica (coleta ou publicação de metadados), *API* (coleta de metadados ou dados) e Ferramentas de *web scraping* e ETL (coleta, fluxo ou integração de dados) e, por fim, analisou amostras de dados indígenas disponibilizadas por esses portais.

CONCLUSÃO

Essa pesquisa demonstrou o potencial significativo das ferramentas de tecnologia da informação na coleta de dados indígenas. A segunda fase da metodologia, dividida em duas etapas, nos levou a identificar três pilares principais:

a) Ferramentas para coleta de metadados (*web* semântica): Aplicáveis aos conjuntos de dados.
b) Ferramentas para coletas de dados: *web scraping*, ETL integração de dados e busca por conjuntos de dados.
c) *API* de dados: Disponibilizadas pelos próprios portais.

Essas ferramentas contribuem para o desenvolvimento de mecanismos eficazes de proteção para os povos indígenas podendo diminuir a heterogeneidade dos dados indígenas que atrapalha a coleta dificultando o andamento da pesquisa. Isso pode acontecer devido aos governos disponibilizarem dados envolvendo características diferentes, como foi observado nas amostras de dados indígenas coletadas nos portais dificultando a importação de dados. Os autores Chi, Sung e Lien (2020) vão ao encontro do combate à heterogeneidade desses dados pois o estudo propõe melhorar a compreensão étnica dos povos indígenas através de uma coleta de dados eficaz.

Além disso, a *API CKAN*, pela interface gráfica, exige identificação para acesso aos dados, isso pode ser um obstáculo, já que vai contra um dos oito princípios dos dados abertos.

Notamos diferenças claras de usabilidade entre os portais brasileiro e norte-americano. Embora o portal brasileiro ofereça ferramentas mais interativas, a capacidade técnica do usuário é decisiva para uma coleta de dados bem-sucedida. Reconhecer e lidar com essas limitações é essencial para futuras iniciativas que busquem usar a tecnologia da informação para apoiar os povos indígenas.

Para projetos futuros, um bom começo seria a realização de testes das ferramentas/*API* identificadas, começando pela *API* Apify, pois através desta é possível coletar dados ou integrar com *API*.

REFERÊNCIAS

AMARAL, F., 2016. *Introdução à ciência de dados: mineração de dados e big data.* Rio de Janeiro: Alta Books Editora.

AMAZON WEB SERVICES, [s.d.]. *O que é uma API (interface de programação de aplicações*)? [Em linha] [consult. 2025-03-26]. Disponível em: https://aws.amazon.com/pt/what-is/api.

BANDEIRA, J. M., et al., 2014. *Dados abertos conectados* [Em linha] [consult. 2025-05-18]. Disponível em: https://www.researchgate.net/publication/283569633_Dados_Abertos_Conectados.

BERTIN, P. R. B., et al., 2019. *A Parceria para Governo Aberto como plataforma para o avanço da Ciência Aberta no Brasil*. *Transinformação*. **31**, e190020. DOI: https://doi.org/10.1590/2318-0889201931e190020.

BORGMAN, C. L., 2012. The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*. **63**(6), 1059-1078. DOI: https://doi.org/10.1002/asi.22634.

CHI, Yu-Liang, Han-Yu SUNG, e Ying-Yuan LIEN, 2020. Towards the Ethnic Understanding of Taiwanese Indigenous Peoples: A Mashup Based on Semantic Web and Open Data. Em: P.-L. P. RAU, ed. *HCII 2020: HCI International 2020 – Late Breaking Papers: Cognition, Learning and Games* [Em linha]. Cham: Springer International Publishing, pp. 287-297 [consult. 2025-06-18]. Disponível em: https://link.springer.com/chapter/10.1007/978-3-030-49788-0_21.

CIVICACTIONS, [s.d.]. *DKAN Open Data Platform* [Em linha] [consult. 2025-06-18]. Disponível em: https://getdkan.org.

DATA.GOV, [s.d.]. T*he Home of the U.S. Government's Open Data* [Em linha] [consult. 2025-06-10]. Disponível em: https://data.gov.

GOV.BR, [s.d.]. *Bem-vindo ao Portal Brasileiro de Dados Abertos e Catálogo Nacional de Dados!* [Em linha] [consult. 2025-06-10]. Disponível em: https://dados.gov.br.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, [s.d.]. *Censo Populacional: Sobre Localidades indígenas* [Em linha] [consult. 2025-06-10]. Disponível em: https://www.ibge.gov.br/en/statistics/social/labor/22836-2022-census 3.html?edicao=.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 1998. ISO 9241-11. Requisitos ergonômicos para trabalho de escritório com terminais de exibição visual (VDTs). *Organização Internacional para Padronização*. **45**(9), 22.

ISOTANI, S., e I. I. BITTENCOURT, 2015. *Dados abertos conectados: em busca da web do conhecimento*. São Paulo: Novatec Editora.

MARTINS, L. C. B., 2019. *Proposta de Arquitetura de Publicação Automatizada de Dados Abertos Conectados Utilizando Meta-Dados e Ontologias* [Em linha]. Dissertação de mestrado, Universidade de Brasília – Instituto de Ciências Exatas, Departamento de Ciência da Computação [consult. 2025-06-10]. Disponível em: https://repositorio.unb.br/handle/10482/34816.

MOREIRA, F. M., et al., 2017. Metadados para descrição de datasets e recursos informacionais do «Portal Brasileiro de Dados Abertos». *Perspectivas em Ciência da Informação* [Em linha]. **22**(3), 158-185 [consult. 2025-06-10]. Disponível em: https://www.scielo.br/j/pci/a/nsNf68fmh3y4tNnh3XpjCZG/?lang=pt.

NEVES, O. M. C., 2013. Evolução das Políticas de Governo Aberto no Brasil. Em: *VI CONSAD de Gestão Pública, 16, 17 e 18 de abril de 2013, Brasília/DF* [Em linha] [consult. 2025-05-26]. Disponível em: http://consad.org.br/wp-content/uploads/2013/05/092-EVOLUÇÃO-DAS-POLÍTICAS-DE-GOVERNO-ABERTO-NO-BRASIL.pdf.

PINTO, A. L., e E. D. AMARAL, 2020. Google Dataset Search: Visão geral e perspectivas para indexação e disponibilização de conjuntos de dados científicos abertos. *Ciência da Informação*. **49**(3), 173-187.

ROBINSON, D., et al., 2009. Government data and the invisible hand. *Yale Journal of Law & Technology* [Em linha]. **11**, 160-175 [consult. 2025-06-18]. Disponível em: https://heinonline.org/HOL/LandingPage?handle=hein.journals/yjolt11&div=6&id=&page=.

SHAW, M., e D. GARLAN, 1995. Formulations and formalisms in software architecture. Em: Jan van LEEUWEN, ed. *Computer Science Today*. Berlin; Heidelberg: Springer-Verlag, pp. 307-323.

SILVA, P. N., 2018. *Dados governamentais abertos: métricas e indicadores de reuso* [Em linha]. Tese de doutorado, Universidade Federal de Minas Gerais [consult. 2025-06-18]. Disponível em: https://repositorio.ufmg.br/handle/1843/BUBD-AYNG4U.

SOUZA, L. P., et al., 2023. Promovendo a justiça epistêmica: uma análise dos princípios CARE na gestão de dados de pesquisa em relação aos povos indígenas. Em: *Anais do VI Workshop de Informação, Dados e Tecnologia – WIDaT* [Em linha] [consult. 2025-06-01]. Disponível em: https://labcotec.ibict.br/widat/index.php/widat2023/article/view/73.

SUNG, Han-Yu, e,Yu-Liang CHI, 2021. Applications of Semantic Web in integrating open data and bibliographic records: a development example of an infomediary of Taiwanese indigenous people. *The Electronic Library* [Em linha]. **39**(2), 337-353 [consult. 2025-06-01]. Disponível em: https://www.emerald.com/insight/content/doi/10.1108/el-09-2020-0258/full/html.

THOMAS, D. M., e S. MATHUR, 2019. Data analysis by web scraping using python. Em: *3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA), 12-14 June 2019, Coimbatore, India*. [S.l.]: IEEE, pp. 450-454.

VICTORINO, M. C., et al., 2020. Arquitetura de publicação de dados abertos conectados governamentais da Universidade de Brasília. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação* [Em linha]. **25**, 1-25 [consult. 2025-06-10]. Disponível em: https://www.redalyc.org/journal/147/14763386013/14763386013.pdf.

VIDA, E. S., et al., 2021. *Data Warehouse. Grupo A* [Em linha]. E-book. ISBN 9786556901916 [consult. 2025-06-10]. Disponível em: https://integrada.minhabiblioteca.com.br/#/books/9786556901916.

W3C, [s.d]a. *World Wide Web consortium* [Em linha] [consult. 2025-06-10]. Disponível em: https://www.w3.org.

W3C, [s.d]b. *RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation 25 February 2014* [Em linha] [consult. 2025-06-10]. Disponível em: https://www.w3.org/TR/rdf11-concepts.

W3C, [s.d.]c. *SPARQL 1.1 Overview. W3C Recommendation 21 March 2013* [Em linha] [consult. 2025-06-10]. Disponível em: https://www.w3.org/TR/sparql11-overview.

ZHEKOVA, M., e E. YUMER, 2024. JavaScript Web Scraping Tool for Extraction Information from Agriculture Websites. *BIO Web of Conferences* [Em linha]. **102**, 03008 [consult. 2025-06-10]. Disponível em: https://www.bio-conferences.org/articles/bioconf/abs/2024/21/bioconf_foset2023_03008/bioconf_foset2023_03008.html.