
Documentos Presidenciais da Biblioteca da Presidência da República:

Base de Metadados (versão 1)

Nota Técnica

Esta nota técnica descreve a construção, organização e finalidade da base de metadados de documentos presidenciais da Biblioteca da Presidência da República. O conjunto de dados reúne informações estruturadas sobre 8.606 documentos oficiais produzidos por diferentes Presidentes do Brasil, disponibilizados publicamente no portal da Biblioteca da Presidência da República.

O objetivo principal desta base é oferecer uma infraestrutura de pesquisa confiável, rastreável e cientificamente válida para estudos em políticas públicas, agenda governamental, comunicação presidencial, análise de discursos, prioridades políticas e comportamento institucional do Executivo federal.

Os documentos foram coletados exclusivamente a partir do portal oficial da Biblioteca da Presidência da República¹. Foram incluídos os seguintes tipos de documentos, em suas categorias originais conforme disponibilizadas pela fonte: discursos; entrevistas; pronunciamentos; mensagens oficiais; comunicações presidenciais; notas e outros documentos associados.

Para o processo de coleta de dados, foram utilizadas técnicas de extração automatizada a partir das páginas HTML, mantendo sempre a integridade dos metadados originais. Cada documento passou pelas seguintes etapas: a) identificação da URL principal b) extração de título, data e presidente responsável c) registro do tipo documental conforme a categoria original d) conversão de datas para formato ISO quando possível e) armazenamento do nome de arquivo e informações adicionais. Não foram aplicadas classificações temáticas, recodificações ou filtros analíticos. O objetivo desta versão é preservar a estrutura mais próxima possível da fonte primária.

Os campos presentes na planilha incluem:

- document_id: identificador único interno
- presidente: nome completo do presidente autor do documento
- tipo_documento: categoria documental conforme a fonte

¹ <https://www.biblioteca.presidencia.gov.br/>

- data: data ISO (YYYY-MM-DD), quando disponível
- ano e mes: extraídos ou derivados da data - título: título ou heading do documento conforme publicado
- url_original: link direto para o documento na Biblioteca da PR - document_filename: nome de arquivo da versão consultada
- language: idioma (pt-BR) - fonte: Biblioteca da Presidência da República
- notas: campo livre para particularidades

A padronização se limitou a limpeza de espaços em branco nos nomes de presidentes, conversão de datas incompletas para formatos consistentes, normalização de colunas, reorganização da estrutura para facilitar depósito e reuso. Nenhuma transformação de conteúdo foi aplicada aos textos originais, que não foram incluídos nesta versão. Importante destacar que há documentos apresentam datas incompletas, assim como categorias documentais heterogêneas e que dependem da taxonomia da fonte.

Algumas aplicações potenciais da base de dados incluem: estudos sobre a agenda presidencial (agenda retórica, agenda simbólica, agenda comunicacional); análises de discursos políticos; pesquisas sobre prioridades e mudanças governamentais; investigação de mudanças na agenda governamental ao longo do tempo; comparações entre diferentes presidentes e em diferentes momentos do tempo; treinamento de modelos de PLN com documentos estruturados.

Em versões futuras buscamos ter bases de dados adicionais contendo classificações temáticas segundo o Comparative Agendas Project codebook; textos completos; atualizações periódicas do corpus, incluindo documentos de governos mais recentes; código-fonte do sistema de raspagem e mecanismos de busca.

Formato de citação recomendada:

Capella, A. C. N. (2025). Base de Metadados de Documentos Presidenciais da Biblioteca da Presidência da República (versão 1). Repositório Institucional da UNESP.