

DESCOBERTA E CATALOGAÇÃO RASTREÁVEL DE FONTES PARA AVALIAÇÃO ECONÔMICA DE INVESTIMENTOS PÚBLICOS: O CASO DA ILUMINAÇÃO PÚBLICA DO DISTRITO FEDERAL

Bruno Fernandes Tavora¹, Wender Santos Guimaraes²

¹Faculdade Anhanguera, Brasília, Brasil (bruno.tavora@gmail.com)

²Faculdade de Informática e Administração Paulista, Brasília, Brasil

Resumo: Este trabalho desenvolveu e testou um catálogo rastreável de fontes públicas para subsidiar futuras avaliações econômicas da modernização da iluminação pública do Distrito Federal. Com Python e MongoDB, foram organizadas sete fontes, oito conjuntos de dados e 90 recursos. A execução em força bruta recuperou 85 arquivos e extraiu 1.373 páginas; a segunda processou cinco recursos e 148 páginas. O modelo preserva a cadeia completa de proveniência (fonte, arquivo, página e trecho), embora exija validação humana das evidências. Os resultados demonstram a viabilidade da abordagem para superar a dispersão informacional e melhorar a avaliabilidade de investimentos públicos.

Palavras-chave: Dados abertos governamentais; Catálogo de dados; Rastreabilidade da informação; Avaliação de investimentos públicos; Proveniência de dados.

INTRODUÇÃO

A ampliação das iniciativas de Dados Abertos Governamentais, dos portais de transparência e dos repositórios institucionais tem expandido significativamente as possibilidades de acompanhamento, monitoramento e controle da ação pública.

Nesse contexto, o movimento Linked Open Data (LOD) contribuiu para ampliar as possibilidades de compartilhamento e reutilização de dados. Conforme proposto por Berners-Lee (2006), os dados devem ser disponibilizados livremente e de forma conectada, favorecendo sua integração e reutilização em diferentes contextos.

No Brasil, a agenda de dados abertos ganhou relevância a partir da Lei de Acesso à Informação (Lei nº 12.527/2011) e foi fortalecida pelo Decreto nº 8.777/2016, que instituiu a Política de Dados Abertos do Poder Executivo Federal (Brasil, 2016). Entretanto, Santarém Segundo (2015, p. 2) observava que a cultura de publicação de dados abertos ainda era limitada e carente de infraestrutura adequada para armazenamento e acesso efetivo pelos usuários finais. Como resposta a parte desses desafios, o Plano de Ação da Infraestrutura Nacional de Dados Abertos (INDA) 2023–2025 passou a enfatizar aspectos relacionados à qualidade, governança, interoperabilidade, catalogação e reutilização da informação pública (Brasil, 2023).

Apesar desses avanços, persistem desafios relacionados à organização, descoberta e integração das informações públicas. Dados sobre investimentos públicos, documentos contratuais, registros financeiros, informações operacionais e dados territoriais frequentemente permanecem distribuídos entre instituições, modelos e plataformas distintas. Como resultado, embora a informação esteja disponível, ela permanece fragmentada em múltiplos sistemas e repositórios, dificultando sua recuperação, integração e utilização para análise e tomada de decisão.

Essa limitação torna-se particularmente relevante no contexto da avaliação de investimentos públicos. O Guia Geral de Análise Socioeconômica de Custo-Benefício de Projetos de Investimento em Infraestrutura estabelece que a análise dos investimentos deve considerar custos, benefícios, riscos, externalidades e impactos associados aos projetos (Brasil, 2021). Dessa forma, a produção de evidências para avaliação econômica depende da disponibilidade e da integração de informações provenientes de diferentes domínios informacionais.

No caso da modernização da iluminação pública do Distrito Federal por meio da substituição de luminárias convencionais por tecnologia LED, uma eventual avaliação econômica poderia demandar a integração de informações contratuais, financeiras, operacionais, territoriais e indicadores relacionados aos possíveis resultados do investimento. Assim, a obtenção de evidências depende da capacidade de



localizar, compreender e integrar informações produzidas por diferentes instituições.

Essas informações normalmente encontram-se distribuídas entre portais de transparência, sistemas administrativos, bases estatísticas, plataformas de geoinformação e repositórios documentais distintos, tornando necessária a identificação, localização, descrição e rastreabilidade prévias das fontes de informação. A simples disponibilização de recursos na internet não garante sua descoberta, recuperação ou reutilização efetiva como evidência para futuras análises.

Sob essa perspectiva, o problema investigado converge com os objetivos da INDA relacionados à catalogação, interoperabilidade e reutilização da informação pública (Brasil, 2023). Antes da realização de análises econômicas, avaliações de impacto ou estudos de custo-benefício, existe uma etapa informacional frequentemente negligenciada: descobrir, organizar e tornar rastreáveis as fontes necessárias à produção de evidências. Diante desse desafio, o presente estudo propõe um modelo de catalogação rastreável como etapa preliminar essencial para a produção de evidências.

O propósito deste artigo não é calcular o retorno socioeconômico da modernização da iluminação pública nem afirmar relações causais entre a adoção da tecnologia LED e eventuais mudanças nos indicadores observados. O problema delimitado é anterior. Busca-se compreender a avaliabilidade informacional do investimento, entendida como a existência, acessibilidade, rastreabilidade e potencial de integração das fontes públicas de informação capazes de sustentar avaliações futuras. Essa questão torna-se particularmente relevante diante da dispersão institucional das informações, da instabilidade de endereços eletrônicos, da predominância de documentos disponibilizados em formato PDF e da ausência de metadados padronizados que favoreçam a descoberta, recuperação e integração dos recursos informacionais.

Como estudo de caso, foi selecionada a modernização da iluminação pública do Distrito Federal, cuja documentação encontra-se distribuída entre portais institucionais, sistemas de transparência, bases de dados abertas, relatórios administrativos e demais documentos públicos produzidos por diferentes instituições.

Diante desse cenário, formula-se a seguinte questão de pesquisa: como descobrir, organizar e tornar rastreáveis fontes públicas dispersas, de maneira que

possam compor uma base de evidências para a futura avaliação econômica de um investimento público?

Assim, o objetivo deste artigo é desenvolver e analisar um catálogo de fontes, datasets, recursos informacionais e evidências relacionados à modernização da iluminação pública do Distrito Federal, complementado por uma prova de conceito implementada em Python e MongoDB, visando demonstrar mecanismos de descoberta, organização, rastreabilidade e integração de informações potencialmente úteis para futuras avaliações de investimentos públicos.

MATERIAL E MÉTODOS

2.1 DELINEAMENTO E ESTUDO DE CASO

O trabalho caracteriza-se como uma pesquisa aplicada, de natureza exploratória e descritiva, conduzida por meio de estudo de caso. Essa estratégia metodológica permite observar, em contexto real, os desafios associados à descoberta, recuperação e organização de informações necessárias à avaliação de um investimento público, preservando as características institucionais e tecnológicas dos ambientes em que os dados são produzidos e disponibilizados.

O estudo de caso compreende a modernização da iluminação pública do Distrito Federal. A escolha decorre da disponibilidade de documentação pública produzida por diferentes organizações e da potencial necessidade de integração de múltiplas fontes informacionais para subsidiar futuras avaliações econômicas do investimento. Além disso, o Guia Geral de Análise Socioeconômica de Custo-Benefício de Projetos de Investimento em Infraestrutura estabelece que a análise de investimentos públicos deve considerar custos, benefícios, riscos, externalidades e efeitos distributivos (Brasil, 2021). Essas dimensões orientaram a identificação e a busca por fontes jurídico-regulatórias, operacionais, econômico-financeiras, territoriais, de resultado e de impacto.

Diferentemente de um conjunto de dados concentrado em um único sistema, as informações relacionadas à modernização da iluminação pública encontram-se distribuídas entre diferentes instituições e ambientes digitais. Exemplos incluem portais de dados abertos dos governos federal e distrital, sistemas de transparência, plataformas de compras públicas, sítios institucionais do Governo do Distrito Federal, da Companhia Energética de Brasília (CEB), órgãos de controle, bases cartográficas, documentos regulatórios, contratos, relatórios técnicos e registros administrativos.



Como consequência, a construção de uma base de evidências exige não apenas localizar documentos, mas também compreender suas origens, relações, limitações e potencial de integração.

Sob essa perspectiva, o estudo não busca analisar diretamente os impactos da política pública, mas investigar a disponibilidade e a rastreabilidade das informações que potencialmente poderão sustentar avaliações futuras. O foco recai, portanto, sobre a identificação das fontes, dos recursos informacionais e dos mecanismos necessários para reduzir os efeitos da dispersão informacional observada no ecossistema de dados públicos relacionado ao investimento.

Diante desse contexto, o estudo adota um modelo de catalogação rastreável para identificar, organizar e relacionar fontes, recursos e evidências produzidas em diferentes ambientes informacionais.

2.2 DESCOBERTA E ORGANIZAÇÃO DO CATÁLOGO

A descoberta das fontes foi iniciada por meio de navegação exploratória em portais institucionais, páginas de transparência e catálogos de dados abertos. Essa etapa manual foi mantida por permitir o reconhecimento prévio de vocabulários, estruturas de navegação, periodicidades de atualização e padrões de publicação, subsidiando posteriormente os processos de automação e coleta.

As fontes identificadas foram classificadas em oito categorias: institucional, governança, jurídico-regulatória, operacional, econômico-financeira, territorial, resultado e impacto. Essa classificação buscou refletir diferentes dimensões informacionais potencialmente relevantes para a construção de evidências relacionadas ao investimento analisado.

O modelo de catalogação foi estruturado em três entidades principais: fonte, dataset e recurso. A separação foi inspirada no Data Catalog Vocabulary (DCAT), que distingue os conceitos de catálogo, conjunto de dados e distribuição (W3C, 2024), bem como na organização adotada pela plataforma CKAN. No modelo proposto, a fonte representa a origem institucional da informação; o dataset corresponde a uma coleção lógica de dados ou documentos; e o recurso representa o arquivo, página ou serviço efetivamente recuperável.

Além dessas entidades, foram incorporadas as entidades evidência e evento de linhagem. A entidade evidência registra informações extraídas dos recursos catalogados consideradas potencialmente úteis para futuras análises. Já a entidade evento de linhagem

documenta transformações, relacionamentos e processos de obtenção dos dados. Dessa forma, o catálogo não se limita a registrar a localização dos documentos, mas também busca preservar a rastreabilidade das evidências identificadas e dos procedimentos que levaram à sua produção.

Tabela 1. Entidades do modelo e função no protótipo.

Entidade	Descrição	Função
Fonte	Origem institucional da informação	Identificar o produtor da informação
Dataset	Coleção lógica de dados	Agrupar recursos relacionados
Recurso	Arquivo, página ou serviço	Permitir acesso ao conteúdo
Evidência	Informação extraída	Registrar achados relevantes
Evento de Linhagem	Registro de transformação	Garantir rastreabilidade

2.3 IMPLEMENTAÇÃO EM PYTHON E MONGODB

A prova de conceito foi implementada em Python, responsável pela leitura do catálogo, descoberta de links, download, cálculo de hash, extração de texto e geração de métricas.

O MongoDB foi selecionado para armazenar metadados heterogêneos e documentos com estruturas diferentes sem exigir que todos os registros possuam os mesmos campos. O banco foi organizado nas coleções sources, datasets, resources, evidences e lineage_events.

O fluxo de ingestão preserva a cópia original e gera um resumo criptográfico SHA-256. A ISO 15489-1 (2016) destaca autenticidade, confiabilidade, integridade e usabilidade como características necessárias aos registros.

O hash não representa toda a proveniência, mas permite verificar se o conteúdo utilizado permanece idêntico.

Os eventos de processamento foram modelados com base na PROV-O, que relaciona entidades, atividades e agentes (W3C, 2013).

2.4 EXTRAÇÃO E REGISTRO DAS EVIDÊNCIAS

Após o download, o texto dos PDFs é extraído por página. Essa escolha permite que uma informação estruturada mantenha referência direta à localização em que foi encontrada.

O programa identifica trechos candidatos por termos relacionados ao investimento, como luminárias,



consumo, investimento, manutenção, economia e prazo. Cada candidato é registrado com recurso, página, trecho, valor, unidade, período e método de extração.

A extração automática não foi tratada como validação definitiva. Um mesmo número pode representar meta, valor acumulado, previsão ou resultado realizado.

Por esse motivo, o fluxo distingue evidência candidata e evidência validada. A conferência humana preserva o diálogo entre o valor estruturado e seu contexto documental, reduzindo o risco de transformar uma ocorrência textual em evidência sem interpretação.

Versões futuras do protótipo poderão incorporar modelos de inteligência artificial (LLMs) e técnicas de processamento de linguagem natural para melhorar a precisão na identificação e classificação semântica das ocorrências, mantendo sempre a rastreabilidade até a página e o trecho original do documento.

2.5 CATÁLOGO NAVEGÁVEL E CRITÉRIOS DE AVALIAÇÃO

Os registros do MongoDB foram organizados para navegação entre fonte, dataset, recurso e evidência. A consulta permite identificar a instituição de origem, o arquivo utilizado, a página em que a informação foi localizada e o trecho que sustenta o registro candidato.

Os critérios de avaliação consideraram a cobertura do catálogo, a recuperabilidade dos recursos, a integridade dos arquivos, a rastreabilidade das ocorrências e o funcionamento do fluxo de processamento. Foram observadas as quantidades de fontes, datasets e recursos registrados; URLs localizadas; arquivos recuperados; resumos SHA-256 gerados; páginas extraídas; ocorrências candidatas; e evidências validadas.

2.6 SÍNTESE DAS DECISÕES METODOLÓGICAS

As decisões metodológicas foram definidas em função do problema de pesquisa e do objetivo de construir um catálogo capaz de apoiar a descoberta, organização e rastreabilidade de evidências relacionadas a um investimento público. O estudo de caso permitiu delimitar o contexto investigado e concentrar a análise em um conjunto específico de fontes informacionais. A modelagem baseada em fonte, dataset e recurso forneceu uma estrutura para organização dos dados identificados, enquanto os

princípios de metadados, gestão documental e proveniência contribuíram para o registro da origem, integridade e rastreabilidade dos recursos catalogados.

Complementarmente, a implementação em Python e MongoDB possibilitou operacionalizar essas definições conceituais por meio de um fluxo de coleta, armazenamento e documentação das evidências identificadas. Dessa forma, aspectos relacionados à descoberta de fontes, catalogação, preservação da integridade dos recursos e registro de eventos de processamento foram incorporados em uma arquitetura única, permitindo analisar de maneira sistemática a disponibilidade e a organização das informações relacionadas ao investimento estudado.

A relação entre os objetivos da pesquisa, as decisões metodológicas adotadas e sua implementação no protótipo é apresentada na Tabela 2.

Tabela 2. Decisões metodológicas e aplicação no protótipo.

Decisão adotada	Aplicação no protótipo	Fundamentação
Estudo de caso	Modernização da iluminação pública	Pesquisa aplicada
Fonte, dataset e recurso	Estrutura do catálogo	DCAT e CKAN
Oito categorias	Organização temática	Guia ACB e INDA
MongoDB	Coleções documentais	Flexibilidade do modelo
Python	Descoberta e ingestão	Processo computacional
SHA-256	Verificação dos recursos	ISO 15489
Eventos de lineage	lineage events	PROV-O
Página e trecho	evidences	Rastreabilidade

3 RESULTADOS E DISCUSSÃO

3.1 CONSTRUÇÃO E ORGANIZAÇÃO DO CATÁLOGO

A etapa de catalogação reuniu sete fontes institucionais, oito datasets e noventa registros de recursos relacionados à modernização da iluminação pública do Distrito Federal. Desses registros, 77 apresentaram identificadores originais únicos, indicando repetições no mapeamento inicial. A concentração ocorreu nas categorias operacional, com 49 registros; econômico-financeira, com 18; jurídico-regulatória, com 15; e resultado, com oito registros.

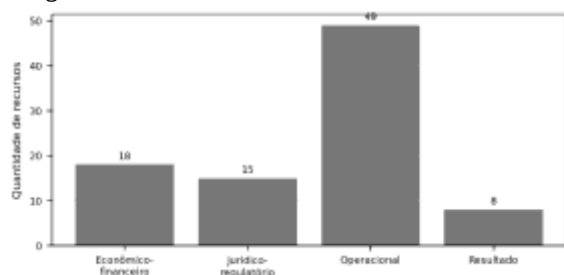
A predominância da categoria operacional decorre da periodicidade dos relatórios de atividades e dos termos de aceite produzidos durante a execução da

concessão. Em contraste, documentos contratuais, legislativos e regulatórios possuem menor frequência de publicação, embora sejam necessários para compreender obrigações, riscos e condições econômico-financeiras do investimento.

Tabela 3. Síntese quantitativa e distribuição do catálogo.

Entidade	Quantidade
Fontes	7
Datasets	8
Registros de recursos	90
Identificadores originais únicos	77
Recursos operacionais	49
Recursos econômico-financeiros	18
Recursos jurídico-regulatórios	15
Recursos de resultado	8

Figura 1. Distribuição dos recursos catalogados por categoria.



3.2 PRIMEIRA EXECUÇÃO: TESTE EXPLORATÓRIO EM FORÇA BRUTA

A primeira execução foi conduzida como teste exploratório em força bruta, entendido neste trabalho como o processamento em lote de todo o catálogo sem validação prévia individual das correspondências encontradas. O objetivo foi verificar a capacidade operacional do protótipo para localizar endereços, recuperar documentos, preservar arquivos, calcular resumos criptográficos e extrair conteúdo textual em escala superior à amostra mínima necessária à prova de conceito.

O processo atribuiu URLs aos noventa registros e concluiu 85 downloads, correspondendo a uma taxa de sucesso de 94,4%. Foram gerados resumos SHA-256 para 86 conteúdos recuperados, incluindo um retorno cujo formato diferia do esperado. A extração textual foi concluída para 85 recursos e totalizou 1.373 páginas. Quatro registros apresentaram erro de download e um documento exigiria OCR.

Tabela 4. Resultados da primeira execução do protótipo.

Indicador	Resultado
Recursos catalogados	90
Identificadores originais únicos	77
URLs localizadas	90
Downloads concluídos	85
Taxa de sucesso dos downloads	94,4%
Erros de download	4
Conteúdo diferente do esperado	1
Recursos com SHA-256	86
Extrações concluídas	85
Páginas extraídas	1.373
Documento que exigiria OCR	1
Ocorrências candidatas	4.645
Evidências validadas	0

Os resultados demonstraram que a arquitetura era capaz de percorrer um conjunto documental amplo e executar as principais etapas de ingestão. Entretanto, o teste em força bruta também revelou limitações que não seriam observadas apenas pela análise conceitual do modelo.

3.3 PROBLEMAS IDENTIFICADOS NA PRIMEIRA EXECUÇÃO

O primeiro problema esteve relacionado à correspondência entre os registros catalogados e os endereços encontrados. O mecanismo inicial utilizava principalmente similaridade textual entre o título do recurso e o texto dos links. Em alguns casos, títulos semelhantes receberam associação inadequada quanto ao período ou ao tipo documental. O relatório de fevereiro de 2026, por exemplo, foi relacionado ao documento de março de 2026, enquanto o registro de janeiro de 2026 recebeu o arquivo referente a fevereiro.

A comparação dos resumos criptográficos confirmou que o mesmo arquivo havia sido associado a registros de períodos diferentes. Esse resultado demonstrou que a localização de uma URL e o sucesso do download não asseguram, isoladamente, a identidade documental. A validação deve comparar título, texto



de âncora, período de referência, nome do arquivo e, quando possível, o conteúdo recuperado.

Também foram observados links dependentes de JavaScript, páginas intermediárias e conteúdo diferente do formato esperado. Esses casos mostram que a recuperabilidade dos documentos depende da forma como cada instituição implementa sua publicação, e não apenas da existência formal de um endereço eletrônico.

Outro problema ocorreu na extração das ocorrências candidatas. O mecanismo baseado em palavras-chave e expressões regulares simples produziu 4.645 registros, mas números presentes em datas, títulos, códigos e numerações de seções foram interpretados como valores. Números escritos no padrão brasileiro também podiam ser fragmentados. Assim, o total não corresponde a 4.645 evidências distintas, mas a ocorrências brutas que exigem interpretação.

3.4 SEGUNDA EXECUÇÃO: AMOSTRA CONTROLADA

Após a análise da primeira execução, foi realizada uma segunda rodada com uma amostra controlada de cinco registros de recursos. O propósito não foi reproduzir o processamento integral, mas observar com maior precisão o funcionamento da cadeia de ingestão e rastreabilidade.

Os cinco registros foram baixados com sucesso, receberam resumos SHA-256 e tiveram seu conteúdo textual extraído. Ao todo, foram processadas 148 páginas, sem erros de download e sem necessidade de OCR. A amostra possui finalidade técnica, não estatística, e busca demonstrar o percurso entre o registro catalogado, o arquivo recuperado e o conteúdo extraído por página.

Tabela 5. Resultados da segunda execução do protótipo.

Indicador	Resultado
Registros selecionados	5
Downloads concluídos	5
Taxa de sucesso	100%
Recursos com SHA-256	5
Extrações concluídas	5
Páginas extraídas	148
Recursos que exigiram OCR	0
Ocorrências candidatas	1.063
Evidências validadas	0

A execução controlada manteve visíveis os problemas de identidade documental observados

anteriormente. Dessa forma, cinco registros processados não significam necessariamente cinco documentos distintos e corretamente associados. A ocorrência reforça a necessidade de validar o documento antes de transformar seus trechos em evidência.

3.5 IDENTIFICAÇÃO DE CONTEXTOS CANDIDATOS

A extração baseada em palavras-chave produziu 1.063 ocorrências candidatas nos cinco registros processados. Os resultados foram distribuídos entre desempenho, luminárias, manutenção, prazo, receitas e despesas e consumo de energia.

Tabela 6. Distribuição das ocorrências candidatas na segunda execução.

Indicador candidato	Ocorrências
Desempenho	520
Luminárias	339
Manutenção	164
Prazo	25
Receitas e despesas	8
Consumo de energia	7
Total	1.063

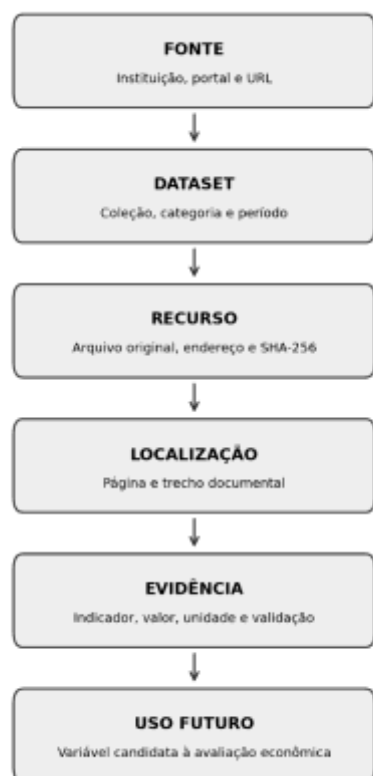
Nenhuma ocorrência foi classificada automaticamente como evidência validada. Esse resultado decorre da simplicidade do motor de busca, que localizou termos e contextos relacionados ao investimento, mas não conseguiu distinguir com segurança se determinado número representava meta contratual, previsão, resultado realizado, acumulado, data ou identificação de seção.

A ausência de valores validados não significa que o contexto de busca ou o documento não tenham sido encontrados. O protótipo recuperou o arquivo, identificou a página e preservou o trecho relacionado à categoria pesquisada. Assim, o sistema reduz o espaço de busca e permite que a validação seja realizada manualmente.

3.6 RASTREABILIDADE E POSSIBILIDADES DE EVOLUÇÃO

O principal resultado técnico não foi a extração automática de valores, mas a preservação da cadeia entre a origem institucional e a ocorrência localizada. A arquitetura mantém as relações entre fonte, dataset, recurso, arquivo armazenado, página, trecho e evidência candidata.

Figura 2. Fluxo de transformação de fontes públicas em evidências rastreáveis.



Essa estrutura permite que o pesquisador abra o documento correspondente, verifique a página e interprete o trecho no contexto em que foi publicado.

Desse modo, o modelo não depende exclusivamente do mecanismo inicial de palavras-chave. A identificação pode ser manual, baseada em expressões regulares, apoiada por busca semântica ou executada por modelos de linguagem e agentes de inteligência artificial.

Em uma segunda versão, mecanismos mais sofisticados poderão normalizar números no padrão brasileiro, validar automaticamente o período do documento, comparar título e conteúdo, identificar unidades e classificar semanticamente os trechos.

Mesmo com essa evolução, cada resposta deverá permanecer vinculada ao documento, à página e ao trecho que a sustenta.

3.7 POTENCIAL PARA FUTURAS AVALIAÇÕES ECONÔMICAS

O catálogo desenvolvido não constitui uma avaliação econômica do investimento. Entretanto, organiza tipos de informação que podem assumir funções analíticas distintas em estudos posteriores. Valores contratuais e registros de receitas e despesas podem apoiar a análise dos custos; quantidades de luminárias e indicadores de desempenho podem representar produtos e resultados operacionais; dados territoriais podem subsidiar análises distributivas; e consumo energético e manutenção podem contribuir para a identificação de benefícios e custos evitados.

Tabela 7. Evidências potenciais e possíveis utilizações analíticas.

Evidência	Possível utilização
Valor contratual	Custos de implantação
Receitas e despesas	Análise financeira
Quantidade de luminárias	Produtos físicos do investimento
Indicadores de desempenho	Monitoramento operacional
Localização territorial	Análises distributivas
Consumo energético	Benefícios potenciais
Registros de manutenção	Custos operacionais

A existência desses registros não demonstra impactos nem estabelece relações causais. Sua contribuição está na avaliabilidade informacional do investimento: a capacidade de identificar quais informações existem, onde estão, a que período se referem e como podem ser verificadas antes da aplicação de métodos econômicos.

3.8 LIMITAÇÕES

As limitações identificadas incluem a dispersão institucional das informações, a inexistência de padrões uniformes de publicação, a predominância de documentos em PDF, a presença de links contextuais ou intermediários e a ocorrência de identificadores repetidos. A descoberta automática também apresentou erros de correspondência temporal e documental.

O motor de extração priorizou abrangência e apresentou baixa precisão. A identificação automática de evidências exigiria regras adicionais para números, unidades, períodos e contexto semântico. A ausência de validação humana impediu que as ocorrências fossem tratadas como valores consolidados para avaliação econômica.



Por fim, a segunda execução concentrou-se em cinco registros operacionais e não representa a totalidade do catálogo. Não obstante essas limitações, o protótipo validou a viabilidade técnica do fluxo de rastreabilidade e abre caminho para evoluções com o uso de inteligência artificial, constituindo uma base sólida para trabalhos futuros.

CONCLUSÃO

A pesquisa partiu do problema de descobrir, organizar e tornar rastreáveis fontes públicas dispersas capazes de compor uma base de evidências para a futura avaliação econômica de um investimento em infraestrutura pública. O estudo de caso demonstrou que a documentação relacionada à modernização da iluminação pública do Distrito Federal encontra-se distribuída entre diferentes grupos documentais, instituições e ambientes digitais, apresentando desafios associados à descoberta, recuperação, identificação e integração das informações.

Como resposta a esse problema, foi desenvolvido um catálogo estruturado nas entidades fonte, dataset, recurso, evidência e evento de linhagem, implementado por meio de uma prova de conceito em Python e MongoDB. O catálogo reuniu sete fontes, oito datasets e noventa registros de recursos. A primeira execução, realizada como teste exploratório em força bruta, recuperou 85 arquivos, gerou SHA-256 para 86 conteúdos e extraiu 1.373 páginas. A segunda execução processou uma amostra de cinco registros e 148 páginas, preservando a associação entre recurso, arquivo, página e trecho documental.

As execuções também revelaram limitações importantes. A utilização de similaridade textual produziu associações incorretas entre títulos e períodos, enquanto o mecanismo de identificação baseado em palavras-chave gerou ocorrências numéricas sem contexto suficiente para validação automática. Como consequência, nenhuma ocorrência foi classificada automaticamente como evidência validada.

Entretanto, a ausência de evidências automaticamente validadas não comprometeu a prova de conceito. O protótipo foi capaz de localizar documentos, registrar seus contextos de origem e preservar sua rastreabilidade ao longo de todo o processo de tratamento da informação. Dessa forma, a arquitetura proposta permite validação manual e pode incorporar, em versões futuras, mecanismos de busca semântica, modelos de linguagem e agentes de inteligência artificial sem perder a referência à fonte

institucional, ao recurso documental, à página e ao trecho que deram origem à evidência.

O catálogo desenvolvido contribui para a avaliabilidade informacional do investimento ao transformar um processo de busca disperso em uma estrutura organizada, verificável, rastreável e potencialmente reutilizável. A principal contribuição do estudo consiste na proposição de um modelo de catalogação rastreável capaz de integrar fontes, datasets, recursos, evidências e eventos de linhagem, preservando o vínculo entre a informação estruturada e sua origem documental.

Conclui-se que a produção de evidências para avaliação de investimentos públicos depende não apenas da existência dos dados, mas também de sua descoberta, organização, preservação e rastreabilidade. Nesse sentido, o modelo proposto constitui uma etapa preliminar para a construção de bases de evidências, criando condições para futuras avaliações econômicas, estudos de impacto e análises de custo-benefício. A avaliação econômica permanece como etapa subsequente, condicionada à validação das evidências e à integração de informações financeiras, operacionais, territoriais, contratuais e de resultado. Por fim, a incorporação de técnicas de inteligência artificial para apoio à extração e validação das evidências configura uma direção promissora para trabalhos futuros, desde que preservados os mecanismos de proveniência, transparência e controle que sustentam a confiabilidade do processo.

AGRADECIMENTOS

Às minhas filhas, Laura e Helena, e à minha parceira, Tiara, pelo carinho, apoio e incentivo ao longo da realização deste trabalho.

REFERÊNCIAS

- BERNERS-LEE, T. Linked Data. W3C, 2006. Disponível em: <<https://www.w3.org/DesignIssues/LinkedData.html>>. Acesso em: 21 jul. 2026.
- BRASIL. Lei nº 12.527, de 18 de novembro de 2011. Regula o acesso a informações. Diário Oficial da União: Brasília, DF, 18 nov. 2011.
- BRASIL. Decreto nº 8.777, de 11 de maio de 2016. Institui a Política de Dados Abertos do Poder Executivo Federal. Diário Oficial da União: Brasília, DF, 12 maio 2016.
- BRASIL. Ministério da Economia. Secretaria de Desenvolvimento da Infraestrutura. Guia geral de análise socioeconômica de custo-benefício de projetos de investimento em infraestrutura. Versão 2. Brasília: SDI/ME, 2021.



BRASIL. Controladoria-Geral da União. Plano de Ação da Infraestrutura Nacional de Dados Abertos 2023–2025. Brasília: CGU, 2023.

CKAN PROJECT. CKAN documentation: datasets and resources. [S. l.], 2026. Disponível em: <<https://docs.ckan.org/en/latest/>>. Acesso em: 21 jul. 2026.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO 15489-1:2016: Information and documentation — Records management — Part 1: Concepts and principles. Geneva: ISO, 2016.

SANTARÉM SEGUNDO, J. E. Web Semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente às iniciativas internacionais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. Anais [...]. João Pessoa: ENANCIB, 2015.

W3C. PROV-O: The PROV Ontology. W3C Recommendation, 30 abr. 2013. Disponível em: <<https://www.w3.org/TR/prov-o/>>. Acesso em: 21 jul. 2026.

W3C. Data Catalog Vocabulary (DCAT) — Version 3. W3C Recommendation, 22 ago. 2024. Disponível em: <<https://www.w3.org/TR/vocab-dcat-3/>>. Acesso em: 21 jul. 2026.