



**SECRETARIA DE SEGURANÇA PÚBLICA
UNIVERSIDADE ESTADUAL DE GOIÁS – UEG
COORDENADORIA DE ENSINO – COE
COORDENAÇÃO DE ENSINO PRESENCIAL E DE PÓS-GRADUAÇÃO
ESPECIALIZAÇÃO EM GERENCIAMENTO DE SEGURANÇA PÚBLICA**

WILLIAM GOMES FERNANDES JÚNIOR

**EXTRAÇÃO DE DADOS ESTRUTURADOS EM LAUDOS PERICIAIS: UMA
METODOLOGIA DE SUPORTE À GESTÃO DA INFORMAÇÃO**

GOIÂNIA – GO

2025



WILLIAM GOMES FERNANDES JÚNIOR

**EXTRAÇÃO DE DADOS ESTRUTURADOS EM LAUDOS PERICIAIS: UMA
METODOLOGIA DE SUPORTE À GESTÃO DA INFORMAÇÃO**

Artigo apresentado como exigência parcial para conclusão da disciplina Metodologia Científica do Curso de Especialização em Gerenciamento de Segurança Pública - CEGESP, pela Secretaria de Segurança Pública do Estado de Goiás - SSP e pela Universidade Estadual de Goiás - UEG, sob a orientação do(a) Prof. Dr. Fernando Fortes Picoli.

GOIÂNIA – GO

2025



EXTRAÇÃO DE DADOS ESTRUTURADOS EM LAUDOS PERICIAIS: UMA METODOLOGIA DE SUPORTE À GESTÃO DA INFORMAÇÃO

STRUCTURED DATA EXTRACTION FROM FORENSIC REPORTS: A METHODOLOGY TO SUPPORT INFORMATION MANAGEMENT

William Gomes Fernandes Júnior*
Prof. Dr. Fernando Fortes Picoli**

Resumo: O estudo propõe uma metodologia automatizada para extrair dados estruturados de laudos periciais do tipo “Caracterização e Eficiência de Armas de Fogo e Munições”, produzidos pelo Laboratório de Balística Forense do Instituto de Criminalística Leonardo Rodrigues, entre setembro de 2024 e fevereiro de 2025. O problema investigado é a ausência de informações consolidadas que subsidiem a gestão do laboratório, ocasionada pela dependência de leitura manual caso uma análise seja desejada. O objetivo é implementar e validar um processo reproduzível, na Linguagem de Programação Python (utilizando-se das bibliotecas pdfminer.six e expressões regulares), capaz de identificar tópicos padronizados, classificar objetos (arma, munição, estojo e desconhecidos) e persistir os registros em formato JSON. Na fase de validação, 50 laudos foram avaliados; dos 44 compatíveis, a comparação com coleta manual revelou erro absoluto de 0,51 %, demonstrando alta fidelidade. Aplicada ao conjunto integral de 345 laudos, a técnica considerou 285 documentos válidos, gerando 2 394 registros de conjuntos de objetos extraídos. Os dados extraídos permitiram elaborar indicadores de produtividade por perito, distribuição de calibres e médias de dias por laudo, evidenciando potencial para subsidiar decisões e atender ao princípio da eficiência administrativa. Conclui-se que a metodologia é viável, requer apenas a padronização de software para mitigar incompatibilidades de documentos do tipo PDF e amplia a governança da informação no âmbito pericial.

Palavras-chave: Coleta de dados; Processamento textual; Data Mining; Gestão da Informação; Laudos periciais.

The study proposes an automated methodology for extracting structured data from forensic reports titled "Characterization and Efficiency of Firearms and Ammunition", issued by the Forensic Ballistics Laboratory of the Leonardo Rodrigues Criminalistics Institute between September 2024 and February 2025. The research problem addressed is the lack of consolidated information to support laboratory management, resulting from the need for manual review when analysis is required. The objective is to implement and validate a reproducible process in the Python programming language (using the pdfminer.six library and regular expressions), capable of identifying standardized report sections, classifying objects (firearms, ammunition, cartridge cases, and unidentified items), and persisting the extracted records in JSON format. During the validation phase, 50 reports were evaluated; of the 44 compatible ones, comparison with manual data

* Bacharel em Engenharia de Computação pela Universidade Federal de Goiás. Perito Criminal na Polícia Científica do Estado de Goiás. Especializando em Gerenciamento de Segurança Pública (SSP-GO/UEG).

** Doutor em Odontologia pela Universidade Federal de Goiás. Perito Criminal na Polícia Científica do Estado de Goiás.



collection revealed an absolute error rate of 0.51%, indicating high accuracy. When applied to the full set of 345 reports, the technique deemed 285 documents valid and generated 2,394 object set records. The extracted data enabled the development of performance indicators by examiner, caliber distribution charts, and average processing times per report, demonstrating the method's potential to inform decision-making and uphold the principle of administrative efficiency. The study concludes that the proposed methodology is feasible and requires only software standardization to mitigate compatibility issues with PDF documents. It enhances information governance within the forensic context.

Keywords: Data Collection; Text Processing; Data Mining; Information Management; Forensic Reports.

1. INTRODUÇÃO

A sociedade humana está inserida num contexto que decisões diversas precisam ser tomadas constantemente. Tais decisões variam desde as mais simples e individuais, como a satisfação de necessidades básicas, até aquelas cujo escopo influencia toda a sociedade, ou seja, decisões com consequências globais. Assim, seja um caçador que busca pistas no ambiente que realiza sua atividade de caça, ou seja um país que busca dados sobre atividades militares de uma outra nação inimiga, em ambos os casos, esses atores buscam um dos pilares da tomada de decisão: a informação.

Segundo Kaye (1995), instituições antigas como o Exército Romano, bem como a própria Igreja da era medieval já utilizavam técnicas de inteligência e informações em seu benefício. Ainda, o autor defende que nas instituições modernas a informação adquiriu um novo patamar, sendo caracterizada como um Recurso Organizacional. Diante dessa crescente importância, evoluímos na maneira que usamos os dados disponíveis, que inicialmente serviam apenas como suporte de um processo decisório baseado na mera intuição, para formas mais modernas, como o armazenamento de dados em bancos de dados estruturados ou, num contexto mais recente, na organização de dados para treinamentos de modelos de inteligência artificial.

Diante disso, para fomentar a produção destes dados, diversas técnicas foram desenvolvidas com o passar do tempo. No campo do conhecimento científico, Marconi e Lakatos (2003) exemplificam que os dados podem ser provenientes de técnicas de coleta documental, observações, questionários, entrevistas, análise de conteúdo, medições em campo, além de outras técnicas. Já para Turban (2009), no contexto de análise de negócios, informações e conhecimentos podem ser provenientes de técnicas como, por exemplo, *text mining*, *data mining*, *web mining* e relatórios *ad-hoc*. Ainda, no ramo das Operações de Inteligência, Krizan (1999) esclarece que dados podem ser oriundos de fontes humanas (*experts* em um tópico, agentes infiltrados), de objetos (posição, tamanho, características físicas), irradiações (fenômenos detectáveis) e registros (banco de dados, pesquisas). Através destes casos, mencionados num rol exemplificativo, observa-se a complexidade e diversidade inerente à coleta de dados, permitindo-nos levantar questionamentos sobre a origem e verossimilhança dos dados com a realidade fática.

Dessa forma, no contexto competitivo e desafiador vivenciado pela gestão contemporânea, ter à disposição informações oriundas de fontes confiáveis é essencial para análise e modelagem de processos institucionais. Decisões baseadas apenas na intuição do gestor são cada

vez mais criticadas, sendo substituídas por decisões alicerçadas em dados estatísticos, informações compiladas, dados agrupados e classificados, além de outras operações de tratamento de dados brutos. Percebe-se, em consequência, que a Gestão da Informação no contexto institucional não é apenas uma tendência, mas uma necessidade de sobrevivência para as corporações que buscam otimizar e controlar suas operações de forma eficiente e eficaz.

Dentro deste contexto, temos, no Estado de Goiás por meio da Polícia Científica, uma fonte inestimável de informações confiáveis, analisadas e verificadas por Peritos Criminais. Estes processam diversos tipos de vestígios, sejam eles coletados de cenas de crimes ou requisitados diretamente pelas autoridades competentes, e, após suas análises, os Peritos materializam suas conclusões em Laudos Periciais, nos quais descrevem informações sobre os vestígios analisados de modo quantitativo e qualitativo. Dessa forma, o Laudo Pericial promove informações diretas ao requisitante da perícia, bem como, de maneira indireta, pode fornecer dados subsidiários para gestão de Segurança Pública.

A decisão baseada em evidências tem sido difundida dentre as práticas de gestão, a exemplo do que já acontece nos outros campos do conhecimento. Diante dessa necessidade, é de suma importância que os processos de coleta forneçam dados de qualidade, assegurando que as projeções e interpretações derivadas de suas análises retratem de maneira fidedigna a realidade factual, considerando a aplicação coerente e correta das referidas análises. Como exemplo dessa importância, podemos mencionar, no contexto de Segurança Pública do Estado de Goiás, as ocorrências que eventualmente são tipificadas de maneira incorreta no Registro de Atendimento Integrado, e que levam a análises e estatísticas incorretas, refletindo nos resultados gerais observados nos relatórios de Segurança Pública, e que, ocasionalmente, podem levar os gestores a desenvolver Políticas de Segurança Pública destoantes da realidade. Sob essa ótica, a ausência de informação talvez seja menos prejudicial que a informação incorreta, que pode levar a uma tomada de decisão imprecisa.

Sob outro aspecto, para entender a magnitude da importância da coleta de informações de fontes confiáveis, podemos observar o rigor das citações nos diversos trabalhos científicos publicados, sendo primordial a apresentação da origem de informações importantes na fundamentação de estudos realizados. Desta maneira, revisores desses trabalhos podem consultar tais fontes, julgar sua credibilidade e seu respectivo impacto na pesquisa analisada. De modo similar, laudos periciais fornecem informações embasadas em métodos e técnicas científicas

consolidadas, com seus procedimentos descritos e formalizados no laudo, juntamente com todo contexto do exame pericial. Dessa forma, coletar o dado que foi lançado no laudo pericial é a garantia de que a informação mais precisa está sendo obtida, ou seja, que é a informação mais acurada possível.

Diante desta situação, no intuito de suprir uma lacuna de coleta de informações vigente no aspecto atual da Polícia Científica do Estado de Goiás, especificamente no escopo de gestão do Laboratório de Balística Forense do Instituto de Criminalística Leonardo Rodrigues, propõe-se este trabalho, no intuito de estudar uma metodologia para fomentar a produção de dados explorando a forma indireta do Laudo Pericial fornecer informações relevantes ao gestor. Para tanto, serão estudados dados de laudos periciais dos exames de Caracterização e Eficiência de Arma de Fogo e Munições, produzidos nos últimos 6 meses pelo Laboratório de Balística Forense do Instituto de Criminalística Leonardo Rodrigues, doravante referenciado como LABAL-ICLR, este vinculado à Superintendência de Polícia Técnico-Científica (SPTC-GO), subordinada à Secretaria de Estado da Segurança Pública do Estado de Goiás (SSP-GO). Estes laudos fornecerão a base para análise de sua estrutura de padronização básica, a fim de evidenciar padrões e estruturas que permitam a extração automatizada de dados.

Além disso, a pesquisa precisa verificar a viabilidade da leitura automatizada dos laudos na forma em que são armazenados, bem como entender quais dados podem ser levantados através de uma metodologia automatizada de extração de estruturas padronizadas.

Por fim, é importante discutir como a manipulação desses documentos produzidos em um período e sua respectiva extração de dados podem apresentar contextos significativos, os quais formarão a base de análises de dados mais complexas e que os auxiliarão na tomada de decisões no cotidiano de suas atividades laborais.

Conforme mencionado por Turban (2009), setores públicos e privados são pressionados a responder rapidamente às mudanças que ocorrem nos ambientes que operam. Dessa forma, o autor afirma também que essa pressão leva à necessidade da rápida e eficiente tomada de decisões que, em alguns casos, precisam de dados consideráveis e relevantes, a serem compilados em informações e conhecimento, para embasar uma melhor análise dentro da estrutura do processo decisório. Desse modo, fica evidenciado que a ausência desses dados relevantes traz prejuízo aos processos de tomada de decisão. Portanto, essa ausência de assertividade, de modo geral, impacta todos os níveis de gestão.

Especificamente na esfera pública, o aumento da eficiência e a adequada gestão de recursos disponíveis está alicerçada no caput do art. 37 da Constituição da República Federativa do Brasil (Brasil, 1988), norma que vincula a Administração Pública ao princípio da eficiência. Desse modo, a Administração Pública tem intrinsecamente o dever de buscar a eficiência de seus processos, conforme mencionado por Di Pietro (2020 p. 249 apud Meirelles 2003 p. 102). Para Di Pietro (2020), um dos aspectos do princípio da eficiência está vinculado ao modo estrutural da Administração Pública, sua organização e disciplina, bem como na apresentação de resultados melhores nas suas prestações de serviços.

Dessa forma, reconhecendo que o LABAL-ICLR é parte da Administração Pública Direta do Estado de Goiás, é imperativo que este laboratório evolua seus processos de tomada de decisão, acompanhando as tendências mais eficazes que fomentam a consciência situacional dos gestores, e, de modo mais específico, constantemente reavalie seus processos de análise e produção de dados de maneira crítica, buscando entender quais dados são produzidos nos seus processos cotidianos, como identificá-los, classificá-los, armazená-los, e especialmente como utilizá-los de maneira adequada como suporte ao processo decisório.

Ainda, a gestão eficaz da informação no âmbito do laboratório não apenas beneficia as autoridades requisitantes de laudos periciais, como também a própria Administração Pública no atendimento de um de seus princípios, ou seja, ao servir com eficiência o cidadão. Ainda, informações produzidas no laboratório podem ser bases de fomento a níveis superiores de gestão, visto que dados relacionados às armas de fogo tem destaque nacional nas políticas de segurança governamentais. No contexto social, as consequências estão diretamente relacionadas à celeridade da justiça, percebida através do célere atendimento aos procedimentos processuais vinculados, que em suma visam a proteção dos bens jurídicos mais fundamentais e essenciais para a convivência social, além da proteção da dignidade humana. Além do aspecto social direto, uma gestão mais eficiente evita desperdícios e melhora processos, reduzindo custos operacionais e permitindo a alocação oportuna em outras áreas que lhe forem convenientes, promovendo também um benefício social indireto.

Portanto, com objetivo de fomentar a produção de dados como suporte a gestão da informação do Laboratório de Balística do Instituto Leonardo Rodrigues, a metodologia proposta por esse estudo foi desenvolvida, especificamente no intuito de extrair dados relevantes provenientes dos laudos periciais do tipo “Caracterização e Eficiência de Armas de Fogo e

Munições” finalizados no período compreendido entre o início de setembro de 2024 até o final de fevereiro de 2025. Fornecendo ao gestor uma base auxiliar para que seja processada e convertida em conhecimento útil aplicável em sua gestão.

2. REVISÃO DA LITERATURA

2.1. Aspectos sobre a confiabilidade de dados em exames periciais

Conforme mencionado por Saad Neto (2017 p. 9 apud Tucci p. 71), exame de corpo de delito é compreendido como o conjunto de procedimentos e técnicas científicas que tem por intuito evidenciar, analisar, constatar, demonstrar e documentar a materialidade delitiva que deixou vestígios. Diante desse entendimento, verifica-se estrita relação dos exames periciais com o método científico, de modo que os dados produzidos nos exames periciais são embasados em técnicas verificadas e aceitas na comunidade científica, afastando a opinião pessoal de quem aplica o método ou a técnica. Por outro lado, além da robustez do próprio método, existe a responsabilização jurídica do Perito Oficial no desempenho de suas funções, com fulcro nos dispositivos do Art. 342 do Código Penal (Brasil, 1940), que trata sobre as consequências penais da realização de falsa perícia, ou seja, são previstas penas por afirmações falsas ou por omissões perante a veracidade das análises na atuação dos peritos. Estes aspectos, de metodologia e responsabilização contribuem para a qualidade dos dados produzidos no âmbito das perícias em geral.

2.2. Considerações sobre o formato de armazenamento

Os laudos do LABAL-ICLR são produzidos por editores de texto convencionais e posteriormente são convertidos em formato digital tipo PDF para aplicação de assinatura digital. Em seguida, são armazenados no Sistema de Informações Criminalísticas ODIN (Sistema ODIN), que está implementado na infraestrutura de *datacenter* estatal utilizada pela Secretaria de Estado da Segurança Pública. A compreensão desse formato de armazenamento é primordial para entender a viabilidade e as limitações do desenvolvimento de uma metodologia automatizada.

2.2.1. Particularidades do formato de arquivo PDF

Conforme discutido por Berg, Oepen e Read (2012), o formato de armazenamento de arquivos no formato PDF foi desenvolvido pela Adobe Systems no início dos anos 1990, para poder gerar documentos digitais que pudessem ser vistos em telas e impressoras de características e recursos diferentes, ao passo que o formato visual do conteúdo fosse preservado. Por esse motivo, em essência, este formato de arquivo não foi originalmente desenvolvido buscando a preservação textual em formato digital, e sim sua disposição geométrica num espaço predefinido. Em suas etapas iniciais, o documento digital possuía informações de objetos (símbolos, caracteres de fontes, imagens) e suas respectivas posições num plano cartesiano pré-definido, como por exemplo as dimensões de uma folha de papel do tipo A4.

Este formato de arquivo evoluiu com o passar do tempo e com sua respectiva popularização, melhorando aspectos como usabilidade e tecnologias assistivas. Essas evoluções foram essenciais para que os arquivos gerados neste formato fossem interpretados de maneira adequada por ferramentas automatizadas. Representações visuais de fontes e caracteres foram substituídas por coleções de frases e grupos textuais completos, e em alguns casos com metadados com a respectiva ordem de apresentação de conteúdo para auxílio em tecnologias assistivas a pessoas com deficiência.

Embora o formato tenha evoluído, ele sempre apresentou retrocompatibilidade com suas versões mais primitivas. Deste modo, é esperado que existam documentos, bem como *softwares* de produção de documentos, que apresentarão características primitivas do formato PDF e que apresentam baixa compatibilidade com a extração textual automatizada.

2.3. Considerações sobre as estruturas padronizadas e técnicas de extração textuais

Por meio da Portaria 128/2019 da SSP-GO, foram estabelecidos o formato padrão dos laudos periciais, de observância obrigatória e com abrangência vinculada a todos laudos periciais da SPTC-GO. Esta portaria estabelece marcadores textuais básicos, como o tipo do laudo pericial, a identificação dos Peritos Oficiais envolvidos, datas de início e fim da elaboração textual, marcadores de título padronizados e diversos outros elementos visuais e textuais. Esses elementos são essenciais para extração de autoria, datas e procedimentos envolvidos, bem como a identificação textual de cada tópico do laudo.

Embora a portaria supracitada tenha estabelecido uma padronização estrutural básica, ela não menciona a padronização de apresentação dos dados da análise pericial propriamente dita. Neste ínterim, o Laboratório de Balística Forense, internamente já utiliza formatos padronizados de apresentação de dados, graças ao uso intensivo de ferramentas de automação textual por seus integrantes, fornecendo um ambiente favorável a extração de dados estruturados.

Após a identificação dos marcadores vinculados pela portaria e outros elementos padronizados observados, os dados de seus subitens podem ser estudados, e conforme mencionado por Chen (2023), dentro de estruturas textuais podem ser realizadas análises de sintaxe, por meio da análise de padrões rígidos, ou mesmo semântica, através de outras técnicas baseadas em *Machine Learning*.

De acordo com Goyvaerts e Levithan (2012), vários padrões textuais podem ser explorados pelo uso de expressões regulares, de modo direto ou mesclados com regras de expressões regulares. Essa flexibilidade permite utilizar desde padrões rígidos a padrões mesclados de caracteres alfanuméricos que atendam a padrões textuais específicos. Deste modo, é possível explorar tais padrões para a extração textual de dados diversos, que atendam a uma estrutura previsível dentro de um grupo de dados de interesse, explorando eventuais padronizações previamente estabelecidas.

3. METODOLOGIA

3.1. Conjunto de dados

Foram coletados 345 laudos periciais de exames do tipo “Caracterização e Eficiência de Armas de Fogo e Munições”, utilizando como critério de inclusão aqueles documentos que tinham sido finalizados no período compreendido entre o início do mês de setembro de 2024 até o final do mês fevereiro de 2025. Os nomes dos peritos criminais envolvidos foram suprimidos e nenhum elemento identificador pessoal será utilizado na apresentação dos resultados, visando preservar a privacidade pessoal dos servidores envolvidos indiretamente na pesquisa. Informa-se que os participantes indiretos serão referenciados apenas como “Perito X”, onde o caractere “X” representa um ou mais dígitos.

3.2. Ambiente de implementação e ferramentas utilizadas

A metodologia foi desenvolvida utilizando a linguagem de programação Python, na versão 3.12.1, bem como as seguintes bibliotecas de funções:

- Pdfminer.six – versão 20240706, para análise e extração do conteúdo de documentos no formato PDF sem manipulação (*raw data extraction*);
- Biblioteca “re”, nativa do Python, para manipulação de expressões regulares;
- Bibliotecas auxiliares de menor relevância – bibliotecas como “pathlib”, “os”, “json”, “datetime” e “numpy” foram utilizadas para manipulação de arquivos, datas e operações vetoriais.

3.3. Extração

Antes da explicação das técnicas de extração propriamente ditas, deve-se entender os pré-requisitos básicos para a eficácia da metodologia de extração de dados estruturados. Inicialmente, é imperativo que o formato de armazenamento digital utilizado, PDF, seja compatível com a extração automatizada do conteúdo digital sem distorções significativas, de modo a não se enquadrar nas extrações com baixa compatibilidade de extração automatizada enunciada no tópico 2.2.1. Complementarmente, é essencial que os dados apresentem uma estrutura padronizada, conforme enunciado nos primeiros parágrafos do tópico 2.3.

Feitas essas considerações, passamos à seleção de dados-alvos a serem extraídos por esta metodologia. Os dados selecionados não são um rol exaustivo de dados passíveis de extração, porém um rol exemplificativo do que se pode ser obtido ao aplicar esta metodologia automatizada de extração. Ressalta-se ainda que, quanto mais dados e mais específicos são os dados alvos, mais complexo é o estudo sobre os dados e mais custosa é a implementação da metodologia automatizada. Neste caso específico, os seguintes registros a seguir foram selecionados para extração automatizada.

1. Número de RG (Registro Geral do Laudo Pericial);
2. Ano de RG (Ano do Registro Geral);
3. Lista de Peritos envolvidos;
4. Data de início;
5. Data de finalização;

6. Lista de armas periciadas identificadas, bem como seus atributos de tipo, calibre nominal, marca, de modificação e indicador de calibre nominal ou compatível;
7. Lista de munições periciadas, bem como seus atributos de quantidade e calibre nominal;
8. Lista de estojos periciados, bem como de seus atributos de quantidade e calibre nominal;
9. Lista de itens desconhecidos periciados, identificados como objetos a serem periciados, porém que não foram categorizados pela metodologia;
10. Lista de envelopes de segurança utilizados para acondicionar o material após o exame pericial.

Com os dados de interesse definidos, os passos descritos em seguida foram implementados para extraí-los dos documentos analisados.

3.3.1. Extração textual do documento PDF e manipulação inicial

A extração do texto base do arquivo PDF, sem nenhum tratamento, foi realizada utilizando a função “*extract_pages*” da biblioteca *pdfminer.six*. Ao aplicar essa função nos dados foram obtidos dois resultados-base: um deles apresentando a extração coerente e esperada do conteúdo textual do documento PDF, enquanto o outro apresentou uma extração divergente do texto visualizado no documento original, com diversas repetições textuais e alteração de ordem do conteúdo apresentado. Ao inspecionar atenciosamente os resultados falhos, foi identificado que eles são resultados de versões primitivas do formato PDF, conforme detalhado no tópico 2.2.1 deste artigo. Foi necessária a implementação de algoritmos com heurísticas de detecção de repetições textuais, em paralelo com detecção esperada do cabeçalho de laudos do LABAL-ICLR. Deste modo, foi possível agrupar extrações textuais como válidas e inválidas. Extrações inválidas possuem pouco valor para extração automatizada, pois a ausência de ordem coerente no documento, bem como demasiadas repetições desnecessárias influenciariam na interpretação correta dos dados. Portanto, a metodologia deve ser aplicada somente nas extrações classificadas como válidas.

3.3.2. Limpeza de dados de cabeçalho e rodapé

O resultado do processo de extração é um conjunto de dados textuais sobre cada página do documento original, em que cada conjunto contém uma lista que contém os textos extraídos das linhas dessas páginas. Ainda, os laudos analisados possuem dados de cabeçalho e rodapé, que geram ruído desnecessário em todas as suas páginas. Essa parte textual é de pouca relevância e deve ser removida para obter-se apenas o texto limpo do documento, sem distorções. Portanto, foram desenvolvidas funções automatizadas de detecção de cabeçalho e rodapé, baseadas na detecção de repetições recorrentes no início de cada página do documento.

Os dados do cabeçalho, apesar de pouca relevância no conteúdo geral do documento, possuem dois dados de interesse: a informação do número do Registro Geral do laudo (RG) e a informação do ano deste RG. Então, após o algoritmo estabelecer o texto de cabeçalho e rodapé, a seguinte expressão regular é aplicada neste conteúdo:

- `r"RG\s+\d+\s+\d+\s+\s(LABAL|SEBAL)/ICLR"`

Esta expressão regular é responsável por identificar padrões textuais como “RG 1234/2024 - LABAL/ICLR” bem como variações numéricas e textuais, obedecendo as regras intrínsecas desta expressão. Após a obtenção da informação de interesse, os dados de cabeçalho e rodapé são excluídos dos conjuntos de dados resultantes. Desta forma, os agrupamentos de páginas contêm apenas informações mais relevantes ao contexto da análise pericial.

3.3.3. Identificação das partes informativas e estrutura de tópicos do documento

Ao analisar o conjunto de dados, bem como a portaria de padronização dos laudos, foi identificado que o laudo apresenta a seguinte estrutura básica:

1. Elementos informativos, contendo os dados do tipo do exame pericial, de procedimentos envolvidos, requisitante, destinatários do laudo, pessoas envolvidas, unidade pericial, peritos envolvidos no exame e data de início do exame.
2. Estrutura de tópicos do laudo, contendo itens como Histórico, Objetivos, Objetos examinados, Resultados dos exames e Considerações finais. Neste caso a estrutura é flexível e adaptável conforme a necessidade do exame.

3.3.4. Extração de dados das partes informativas

Dados apresentados nessa seção apresentavam um padrão bem definido, mantendo núcleos textuais facilmente identificados. Os dados foram extraídos dessa seção apenas localizando o núcleo textual de interesse por busca simples, como por exemplo a localização da expressão regular `r".erit.+?rimina.+?:"`, responsável por detectar padrões como “Perito criminal”, “Peritos criminais”, “Perita Criminal” e variações de acordo com as regras dessa expressão. De modo similar, porém com busca textual simples foi encontrado o texto referente à data de início do exame, pela busca textual “Início do exame”. Após localizado o núcleo de interesse, o dado apresentado na linha subsequente era relativo ao dado de interesse associado àquele núcleo textual, obtendo, portanto, o registro dos peritos envolvidos no laudo analisado.

3.3.5. Identificação e validação da estrutura de tópicos

A estrutura de tópicos nos laudos analisados foi identificada pela análise do padrão de expressão regular dado pelo padrão `r"\d\s+[A-Z]{2}.+[A-Z]{2,10}"`, que identifica elementos textuais que iniciam com um dígito, seguido de um ou mais espaços, subseguidos de dois caracteres maiúsculos e terminados entre 2 a 10 caracteres maiúsculos. Essa análise foi feita no início de cada linha do documento, e os resultados positivos são marcados. Posteriormente, no intuito de eliminar eventuais falsos positivos, os resultados obtidos foram confrontados com o resultado esperado possível para o número do tópico, como por exemplo, o HISTÓRICO do laudo somente poderia ser associado ao número 1, como observado no elemento “1. HISTÓRICO”, bem como a observância da sequência incremental natural dos tópicos principais nos laudos analisados.

3.3.6. Separação dos tópicos de interesse

Após a identificação das localizações de tópico principal do documento, o texto de cada um deles foi isolado, e tópicos importantes para obter os dados-alvo dessa metodologia se encontravam majoritariamente no histórico, porém também havia dados importantes presentes no último tópico principal do laudo, denominado majoritariamente como “CONSIDERAÇÕES FINAIS” e por vezes como “ENCERRAMENTO”.

3.3.7. Análise de dados no grupo de histórico

Neste grupo, era possível localizar os seguintes dados de interesse relacionados às armas periciadas, munições, estojos e outros objetos relativos ao exame pericial, bem como suas respectivas quantidades. Nos laudos analisados, foi observado um padrão de apresentação dos objetos periciados, fato que a metodologia toma proveito para localizar e analisar cada objeto. Foi observado que existe um marcador de descrição de objeto, dado pelo símbolo “•”. Dessa maneira, para determinar quais objetos foram periciados, foi necessário analisar o texto entre os marcadores de objetos, e considerar o último objeto, que não está descrito entre marcadores, porém que está delimitado pela existência de um novo parágrafo ou do encerramento do texto do tópico do histórico. Deste modo, eram obtidos diversos registros de diferentes objetos, que necessitam de avaliação individual para avaliação de seu tipo correspondente.

3.3.8. Identificação do tipo de objeto e suas características

Nesta etapa, foram utilizadas as seguintes expressões regulares, com o intuito de identificar o tipo de objeto referente a cada registro de objetos gerado.

- `r"\d+\s+arma\s+de\sfogo"`: responsável pela identificação de registros de armas de fogo com elemento textual precedente numérico de um ou mais dígitos;
- `r"\d+\s+(muniç(ão|ões)|cartucho)"`: responsável pela identificação de registros de munições, cartuchos e variações similares, com elemento textual precedente numérico de um ou mais dígitos;
- `r"\d+\s+estoujo"`: responsável pela identificação de registros do tipo estojo e estojos, com elemento textual precedente numérico de um ou mais dígitos;
- `r"\d+\s+arma\s+de+\s+pressão"`: responsável pela identificação de registros do tipo armas de pressão, com elemento textual precedente numérico de um ou mais dígitos;

Desse modo, os registros são classificados como tipo “Arma” (arma de fogo ou pressão), “Munição”, “Estojo” e “Desconhecido”. Os três primeiros tipos descritos necessitam de avaliação adicional para a determinação de quantidade, calibre nominal e outras características relevantes.

Para determinação desses atributos, foram desenvolvidas funções de análise textual específicas para cada tipo de dados-alvo. Para cada objeto, eram feitas as análises a seguir.

1. Na análise do atributo relacionado à quantidade de objetos, os primeiros dígitos no texto de registro de objeto foram extraídos, e transformados em dado do tipo numérico, representando a quantidade de objetos daquele registro. A título de exemplo, ao aplicar essa técnica no segmento “02 munições de calibre nominal 9mm Luger.”, os dígitos “02” são extraídos e transformados em informação numérica representando a quantidade de objetos referentes ao objeto (munição de calibre nominal 9mm Luger). Essa análise é aplicada nos registros tipo “Arma”, “Munição” e “Estojo”.
2. Na análise do atributo relacionado ao calibre nominal, efetua-se a busca do termo “calibre nominal”, e, após localizado, o texto subsequente é comparado com uma lista pré-definida de calibres nominais cadastrada. Ao aplicar essa técnica no segmento “02 munições de calibre nominal 9mm Luger.”, o texto “9mm Luger.” é localizado após o segmento “calibre nominal”. Ao verificar a existência de algum calibre pré-cadastrado no segmento de texto restante, é encontrada a equivalência com o calibre “9mm Luger”. Caso não haja equivalência, o calibre é cadastrado como “desconhecido”. Essa análise é aplicada nos registros tipo “Arma”, “Munição” e “Estojo”.
3. Nos registros do tipo “Arma”, verificações adicionais são feitas, para determinar se há informação de fabricação semiartesanal, de compatibilidade entre calibre (quando não há informação explícita do calibre nominal), se foi feita alguma adaptação, o tipo de arma e sua respectiva marca.

Nos casos dos objetos do tipo “Desconhecido”, o trecho de descrição do objeto é salvo para posterior análise. Nessa análise, haverá objetos que não contemplados pela análise desta metodologia, situação perfeitamente normal, por não serem tratados dados desse tipo. No entanto, a análise individual desses registros poderá revelar não conformidades na adoção dos padrões praticados pelo LABAL-ICLR.

3.3.9. Persistência dos registros obtidos

Após extrair diversos registros de informações de interesse, procedeu-se com o armazenamento desses registros de maneira clara e estruturada, utilizando o formato JSON (*JavaScript Object Notation*). O uso desse formato permite a recuperação posterior desses registros e a sua efetiva manipulação de maneira rápida e confiável.

3.3.10. Validação do modelo

Para validar a metodologia, foram selecionados aleatoriamente 50 laudos periciais que não pertenciam à amostra daqueles incluídos no estudo e que compuseram a amostra de validação. Esses laudos foram submetidos ao processamento da metodologia implementada. Os dados coletados foram comparados com dados coletados manualmente (padrão de referência) dos mesmos laudos. Os resultados obtidos pela metodologia desenvolvida foram, então, comparados com o padrão de referência.

4. RESULTADOS E DISCUSSÕES

O estudo desenvolvido apresentou-se como uma alternativa para coleta de dados de interesse a partir de laudos periciais padronizados. Essa iniciativa consiste em um diferencial para coleta e armazenamento de dados que inicialmente não se esperava que fossem necessários serem registrados, além de uma boa forma de verificar outras fontes de dados.

A etapa de validação do estudo indicou que o modelo desenvolvido apresentou ótima acurácia em relação ao padrão de referência. Em 12% (6 laudos) da amostra de validação foi identificado um problema de leitura relacionado à versão do documento PDF utilizada. Dessa forma, a amostra de validação contemplou os 44 laudos restantes, que seus dados puderam ser lidos e o conteúdo pode ser avaliado pela metodologia automatizada.

A tabela a seguir demonstra a quantidade de dados lidos e contabilizados pela metodologia em cada registro alvo, em contraste com dados lidos manualmente.

Tabela 1 – Comparativo do número de registros obtidos através da metodologia em contraste com registros obtidos manualmente, referente aos 44 laudos utilizados na validação

Nome do registro	Extraídos pela metodologia	Extraídos pela leitura manual	Erro absoluto / erro percentual
Número RG	44	44	0 / 0%
Ano RG	44	44	0 / 0%
Lista de peritos	51	51	0 / 0%
Data de início	44	44	0 / 0%
Data de finalização	44	44	0 / 0%
Armas	54	55	1 / 1,81%
Munições	49	49	0 / 0%
Estojo	11	11	0 / 0%
Itens desconhecidos	10	9	1 / 11,11%
Envelopes de segurança	42	42	0 / 0%
Total	393	393	2 / 0,51%

Fonte: Elaborado pelo autor

Desta maneira, os resultados apresentados, oriundos dos laudos compatíveis com a metodologia, apresentaram ótima eficácia global, representando significativamente seus respectivos registros de objetos.

Considerando a ótima eficácia verificada na fase de validação, procedeu-se à análise total da amostra do trabalho, verificando-se que 285 laudos, de um total de 345 foram passíveis de leitura pela metodologia. Os demais 60 laudos apresentaram falha de leitura relacionada ao documento PDF de forma análoga ao observado na amostra de validação.

Tabela 2 – Quantidade de registros de cada dado-alvo obtidos pela aplicação da metodologia nos laudos aplicáveis

Nome do registro	Extraídos pela metodologia
Número RG	285
Ano RG	285
Lista de peritos	310
Data de início	285
Data de finalização	285
Armas	321
Munições	312
Estojos	80
Itens desconhecidos	41
Envelopes de segurança	270
Total	2.394

Fonte: elaborado pelo autor

Ainda, a partir da análise dos registros extraídos, envolvendo quantidades de itens, tipos específicos, podemos apresentar os dados a seguir.

Tabela 3 – Informações compiladas a partir da manipulação dos dados coletados

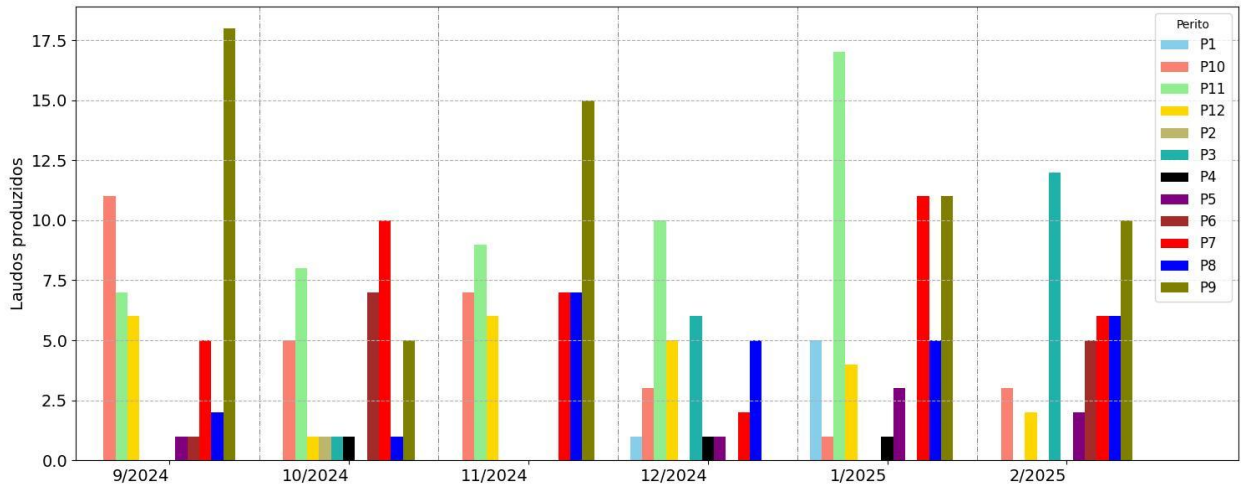
Registro obtido da análise de dados extraídos	Quantidade
Laudos examinados	285
Quantidade de armas totais	321 armas periciadas
Munições	6005 munições periciadas
Estojos	743 estojos periciados
Itens desconhecidos	38 itens

Fonte: compilado pelo autor

Levando em consideração que, nos dados de interesse selecionados pela metodologia, há informações do autor, do perito subscritor do laudo, da quantidade de objetos e alguns insumos utilizados, podemos criar visualizações gráficas que apresentam informações relevantes ao gestor do laboratório. A título exemplificativo, nas figuras a seguir podemos demonstrar algumas

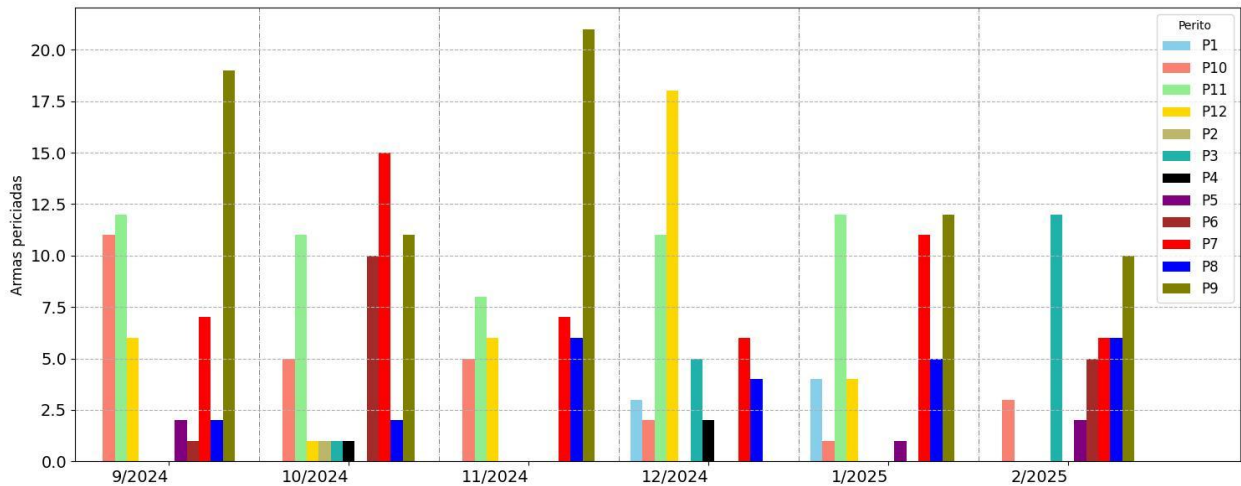
informações que podem ser compiladas a partir desses dados. Cabe ressaltar que os nomes dos peritos foram trocados pela sequência alfanumérica de P1 até P12.

Gráfico 1 - Laudos por perito por mês



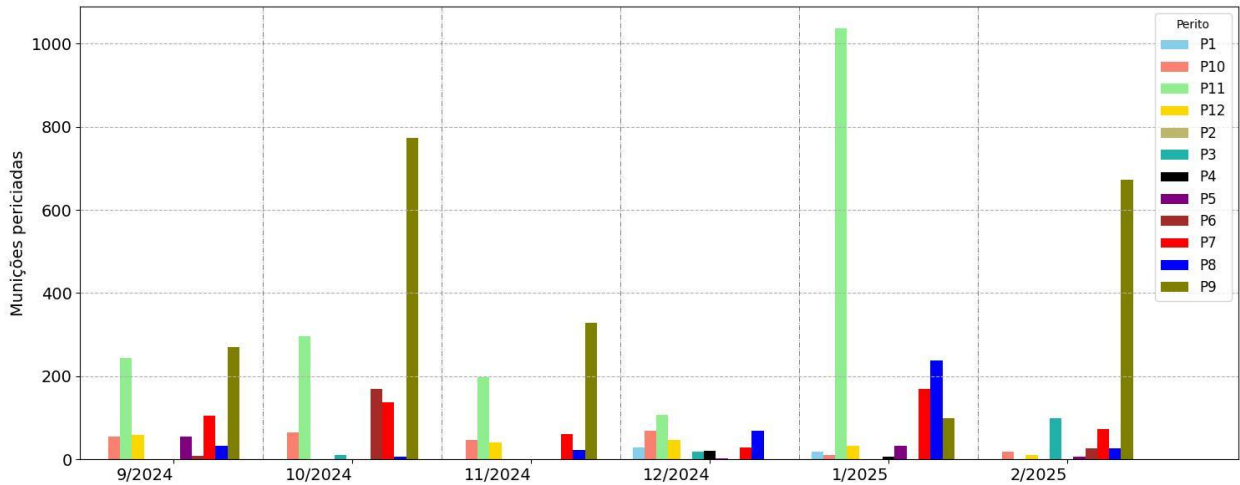
Fonte: compilado pelo autor

Gráfico 2 – Armas periciadas por perito por mês



Fonte: compilado pelo autor

Gráfico 3 – Quantidade munições periciadas por perito por mês



Fonte: compilado pelo autor

O Gráfico 1 ilustra a quantidade de laudos, abrangidos pela metodologia, cada perito produziu por mês. Isoladamente esse gráfico não possui significância acurada para mensuração de produtividade dos peritos, pois um laudo pode ser bastante heterogêneo, com uma infinidade de objetos distintos a serem periciados. É necessário um controle mais fino, que forneça ao gestor uma visão mais detalhada do que realmente foi trabalhado por cada servidor do laboratório. Logo, não se propõe que os dados sejam utilizados para a finalidade de mensuração de produtividade do trabalho pericial. Ao contrário, utilizar essa métrica isoladamente para verificar esse parâmetro seria um erro.

Para auxiliar na análise de produtividade, o gestor pode fazer uso dos Gráficos 2 e 3, que fornecem adicionalmente a quantidade de armas e munições periciadas por cada analista. Deste modo, a análise de produtividade passa a ser algo mais próximo da realidade, baseado em dados coletados diretamente dos documentos produzidos. Ainda assim os dados quantitativos não contemplam os aspectos qualitativos do exame pericial. Com base nos dados poder-se-ia afirmar quantas armas um determinado perito examinou naquele mês, mas não qual o nível de dificuldade de cada exame. Logo, apesar de consistir em uma ferramenta de gestão, a avaliação do gestor imediato quanto aos aspectos qualitativos dos exames é fundamental para o melhor gerenciamento dos recursos humanos e materiais.

Além dessas possibilidades, inúmeras outras análises podem ser realizadas. Analisar os dados, após sua coleta e devida estruturação, é uma tarefa relativamente simples. Bastaria o gestor

determinar como ele necessita uma visualização de dados, como quer dados agrupados que simples *scripts* ou *dashboards* podem ser gerados. A tabela a seguir demonstra algumas outras possibilidades de análise com os dados obtidos pela metodologia.

Tabela 4 – Outras informações obtidas através da manipulação dos dados obtidos pela metodologia automatizada

Informação	Valor
Número médio de dias por laudo	2,24 dias
	.38 Special: 69 unidades
Calibre das armas mais frequentemente periciadas	9mm Luger: 48 unidades
	.22 L.R.: 38 unidades
	9mm Luger: 1543 unidades
Calibre das munições mais frequentemente periciadas	.22 L.R.: 1477 unidades
	.38 Special: 641 unidades
Relação de embalagens de segurança por laudo	0,95 embalagens / laudo

Fonte: compilado pelo autor

Por fim, resta claro que após a obtenção de dados, o gestor pode analisá-los para obter diversos tipos de informações de seu interesse. Os dados obtidos por meio dessa metodologia permitem filtros por nome de perito, intervalos de datas de interesse, tipos de armas envolvidas, bem como outras operações de filtragem e agrupamento de dados.

5. CONSIDERAÇÕES FINAIS

Este estudo permitiu confirmar que os laudos periciais do tipo “Caracterização e Eficiência de Armas de Fogo e Munições” possuem padronização suficiente para adotar a metodologia de extração automatizada proposta neste artigo, visto que a taxa de erro observada na etapa de validação foi de apenas 0,51%, no que tange a obtenção e classificação de registros.

Além disso, foi constatada uma falha significativa no procedimento de leitura dos documentos do tipo PDF, em que alguns desses documentos, especificamente naqueles produzidos numa versão antiga desse formato. Para mitigar esse problema, recomenda-se que a gestão atualize os parques tecnológicos de sua abrangência e instrua seus subordinados a operarem com versões mais atuais dos editores de texto responsáveis por gerar o documento tipo PDF para publicação na plataforma criminalística (sistema ODIN).

No âmbito da abrangência prática da metodologia desenvolvida, pode-se destacar a análise de previsibilidade no uso de recursos oriundos do laboratório de balística, como o uso de envelope de segurança para encaminhamento dos objetos após perícia, inferências sobre a quantidade de munições utilizadas no túnel de tiro presente no laboratório, correções de registros em desconformidades antes da liberação de laudos periciais (erros de descrição, tipografia), bem como da própria mensuração da produtividade dos peritos responsáveis pelos exames periciais do tipo “Caracterização e Eficiência de Armas de Fogo e Munições”.

Ademais, sugere-se, em estudos futuros, a ampliação da metodologia para adequar todos os procedimentos periciais envolvendo o Laboratório de Balística Forense, no intuito de se obter uma visão geral dos processos periciais que ocorrem no laboratório. Estender a abrangência permitiria ao gestor uma visão mais ampla do uso de insumos, bem como da produtividade geral das atividades desenvolvidas pelo laboratório. De forma oportuna, sugere-se ao gestor a determinação de padronização da forma de apresentação dos dados de todos os tipos de exames periciais ali produzidos, sem prejuízo à autonomia do perito na utilização de métodos e ferramentas.

Ainda no aspecto da abrangência, é importante desenvolver padrões estruturados de apresentação de dados relevantes, não somente no laboratório de balística, mas em todos os departamentos vinculados à produção de laudos. Uma padronização deste tipo em efetivo uso num laboratório químico, aliada a uma metodologia análoga, poderia dizer a frequência de apreensão de determinado entorpecente, sua quantidade e relevância ao longo do tempo. De um modo geral, o gestor passa a ter dados disponíveis e pode utilizar de técnicas de análise de dados para obtenção de informações relevantes em nível institucional.

Apresentar os dados de maneira estruturada permite sua posterior recuperação e análise, substituindo as análises semânticas, as quais muitas vezes aliadas a sistemas baseados em inteligência artificial, que demandam mais recursos computacionais e frequentemente envolvem o uso de infraestrutura de terceiros para seu efetivo uso, aumentando o custo monetário e colocando em risco a compartimentalização adequada dos dados custodiados pela Polícia Científica.

Dessa forma, o presente estudo não apenas oferece uma solução tecnológica viável para a otimização da gestão da informação no contexto pericial específico, mas também reforça a importância da padronização e da exploração de dados estruturados como alicerces para uma administração pública mais eficiente e baseada em evidências.

REFERÊNCIAS

BERG, Øyvind Raddum; OEPEN, Stephan; READ, Jonathon. **Towards High-Quality Text Stream Extraction from PDF**. Technical Background to the ACL 2012 Contributed Task. *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, 2012. Disponível em: <https://aclanthology.org/W12-3211/>. Acesso em: 09 mai. 2025.

BRASIL. **Código Penal**. Decreto-Lei nº 2.848, de 7 de dezembro de 1940. Art. 342.

BRASIL. **Constituição da República Federativa do Brasil**, art. 37 caput. Brasília, DF: Senado, 1988.

CHEN, Qiaochu et al. **Data extraction via semantic regular expression synthesis**. *Proc. ACM Program. Lang.*, v. 7, n. OOPSLA2, Art. 287, Out. 2023. 30 p. DOI: 10.1145/3622863. Disponível em: <https://doi.org/10.1145/3622863>. Acesso em: 22 mar. 2025.

DI PIETRO, Maria Sylvia Zanella. **Direito administrativo**. 33. ed. Rio de Janeiro: Forense, 2020. ISBN 978-85-309-8972-9.

GOYVAERTS, Jan; LEVITHAN, Steven. **Regular expressions cookbook: second edition**. Sebastopol: O'Reilly Media, Inc., 2012. ISBN 978-1-449-31943-4.

KAYE, David. The importance of information. **Management Decision**, v. 33, n. 5, p. 5-12, Reino Unido, 1995. ISSN 1758-6070.

KRIZAN, Lisa. **Intelligence Essentials for Everyone**: Occasional Paper Number Six. Washington, DC: Joint Military Intelligence College, Jun. 1999.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 5. ed. São Paulo: Atlas, 2003.

SAAD NETTO, Cláudio. **Exame de corpo de delito**. *Enciclopédia jurídica da PUC-SP*. Celso Fernandes Campilongo, Alvaro de Azevedo Gonzaga e André Luiz Freire (coords.). Tomo: Processo Penal. Marco Antonio Marques da Silva (coord. de tomo). 1. ed. São Paulo: Pontifícia Universidade Católica de São Paulo, 2017. Disponível em: <https://enciclopediajuridica.pucsp.br/verbete/441/edicao-1/exame-de-corpo-de-delito>.

TURBAN, Efraim et al. **Business intelligence: um enfoque gerencial para a inteligência de negócio**. Tradução: Fabiano Bruno Gonçalves. Porto Alegre: Bookman, 2009. Título original: *Business Intelligence: a managerial approach*. ISBN 978-0-13-234761-7.