Estudo de Viabilidade e Elicitação de Requisitos para uma Plataforma de Geração de Dados Sintéticos Configurável: Uma Abordagem Centrada na Flexibilidade Paramétrica

Cristhian Eduardo Kapelinski de Avilla[1], **Rafael da Silva Moral**[1], **Lucas Correa Rodrigues**[1]

[1] [1]Curso de Ciência da Computação
Universidade Federal do Pampa (UNIPAMPA) – Campus Alegrete
Alegrete, RS – Brasil

{cristhianavila.aluno, rafaelmoral.aluno, lucascr2.aluno}@unipampa.edu.br

Abstract. *This research presents a comprehensive feasibility study and systematic requirements elicitation methodology for developing a next-generation synthetic data generation platform. Current market solutions exhibit significant limitations in parametric flexibility, constraining practitioners in scenarios requiring fine-grained control over data characteristics, structural formatting, and quality attributes. Our investigation addresses three critical dimensions: content generation mechanisms (including pattern-based textual synthesis via regular expressions), structural configuration capabilities (encompassing delimiter specification and decimal notation), and data quality simulation (incorporating controlled outlier injection and missing value patterns). Through systematic analysis of existing tools and structured stakeholder consultation, this study establishes a validated requirements framework that prioritizes user-configurable flexibility as the primary design principle. The research contributes to software engineering methodology by providing a structured approach to requirements engineering in domain-specific tool development, with particular emphasis on balancing functional completeness with usability constraints.*

Resumo. *Esta pesquisa apresenta um estudo abrangente de viabilidade e uma metodologia sistemática de elicitação de requisitos para o desenvolvimento de uma plataforma de geração de dados sintéticos de nova geração. As soluções de mercado atuais exibem limitações significativas na flexibilidade paramétrica, restringindo profissionais em cenários que requerem controle granular sobre características dos dados, formatação estrutural e atributos de qualidade. Nossa investigação aborda três dimensões críticas: mecanismos de geração de conteúdo (incluindo síntese textual baseada em padrões via expressões regulares), capacidades de configuração estrutural (englobando especificação de delimitadores e notação decimal) e simulação de qualidade de dados (incorporando injeção controlada de outliers e padrões de valores ausentes). Através da análise sistemática de ferramentas existentes e consulta estruturada às partes interessadas, este estudo estabelece uma estrutura de requisitos validada que prioriza a flexibilidade configurável pelo usuário como princípio primário de design. A pesquisa contribui para a metodologia de engenharia de software ao fornecer uma abordagem estruturada para engenharia de requisitos no desenvolvimento de ferramentas específicas de domínio, com ênfase particular no equilíbrio entre completude funcional e restrições de usabilidade.*

1. Introdução

A crescente demanda por dados sintéticos na era da transformação digital representa um desafio multifacetado que transcende a simples geração de registros aleatórios [5]. A complexidade dos sistemas contemporâneos de software e a necessidade de conformidade com regulamentações de privacidade de dados, como a Lei Geral de Proteção de Dados (LGPD) no Brasil e o General Data Protection Regulation (GDPR) na União Europeia, estabeleceram novos paradigmas para o uso de dados em ambientes de desenvolvimento, teste e pesquisa [7].

O domínio de geração de dados sintéticos situa-se na intersecção crítica entre engenharia de software, ciência de dados e estatística computacional. Enquanto ferramentas tradicionais como Faker [4] e Mockaroo [3] atendem necessidades básicas de geração de dados, elas frequentemente carecem da granularidade de controle necessária para cenários avançados de engenharia de sistemas e modelagem estatística.

Esta investigação fundamenta-se na premissa teórica de que a **flexibilidade paramétrica** constitui o fator determinante na eficácia de ferramentas de geração de dados sintéticos. Definimos flexibilidade paramétrica como a capacidade do sistema de permitir configuração granular em múltiplas dimensões: conteúdo semântico, estrutura sintática e características de qualidade dos dados gerados.

O presente estudo representa a fase inicial de um projeto de desenvolvimento de software, aplicando metodologias consolidadas de engenharia de requisitos [1, 2] para estabelecer uma base empírica sólida que orientará as fases subsequentes de design e implementação.

2. Fundamentação Teórica e Trabalhos Relacionados

2.1. Taxonomia de Ferramentas de Geração de Dados Sintéticos

A literatura especializada identifica três categorias principais de ferramentas de geração de dados sintéticos, cada uma com características e limitações específicas:

1. **Geradores Baseados em Templates**: Ferramentas como Faker utilizam bibliotecas pré-definidas de padrões para gerar dados categóricos comuns (nomes, endereços, números de telefone). Embora eficientes para casos de uso padronizados, apresentam limitações significativas quando aplicados a domínios especializados ou formatos proprietários.
2. **Plataformas de Configuração Visual**: Soluções como Mockaroo oferecem interfaces gráficas para configuração de esquemas de dados. Apesar da usabilidade superior, frequentemente restringem a expressividade do usuário através de menus limitados e opções pré-estabelecidas.
3. **Bibliotecas Programáticas**: Frameworks como NumPy e SciPy permitem geração de dados através de código personalizado. Embora ofereçam flexibilidade máxima, demandam expertise técnica significativa e investimento temporal considerável para implementação.

2.2. Lacunas Identificadas na Literatura

A análise sistemática da literatura revela três lacunas principais nas soluções existentes:

- **Ausência de Mecanismos Unificados**: Não existe uma solução que integre efetivamente geração baseada em padrões textuais (expressões regulares) com modelagem estatística avançada em uma interface coesa.
- **Controle Limitado sobre Formatação de Saída**: Ferramentas comerciais frequentemente impõem formatos de exportação rígidos, negligenciando requisitos específicos de integração de sistemas.
- **Simulação Inadequada de Imperfeições de Dados Reais**: A maioria das soluções gera dados "perfeitos", falhando em reproduzir características estatísticas de datasets do mundo real, incluindo outliers, valores ausentes e distribuições não uniformes.

3. Metodologia de Pesquisa

3.1. Abordagem Metodológica

Esta investigação adota uma abordagem metodológica híbrida, combinando revisão sistemática de literatura com pesquisa empírica junto ao público-alvo. A metodologia fundamenta-se nos princípios da Design Science Research [1], estruturada em quatro fases distintas:

1. **Fase de Diagnóstico**: Análise comparativa de ferramentas existentes utilizando framework de avaliação multidimensional
2. **Fase de Elicitação**: Aplicação de questionários estruturados e entrevistas semiestruturadas com profissionais target
3. **Fase de Análise**: Categorização e priorização de requisitos utilizando técnicas de análise de conteúdo
4. **Fase de Validação**: Verificação cruzada dos requisitos elicitados através de grupos focais

3.2. Critérios de Seleção da Amostra

O público-alvo da pesquisa compreende três categorias principais de profissionais:

- **Engenheiros de Software**: Profissionais envolvidos em desenvolvimento de sistemas e engenharia de qualidade
- **Cientistas de Dados**: Especialistas em análise de dados, machine learning e estatística aplicada
- **Pesquisadores Acadêmicos**: Docentes e discentes de pós-graduação em áreas correlatas

4. Análise do Problema de Pesquisa

4.1. Dimensões da Flexibilidade Paramétrica

A análise preliminar identifica três dimensões críticas onde a flexibilidade paramétrica se manifesta como requisito fundamental:

4.1.1. Dimensão de Geração de Conteúdo

A capacidade de especificar semanticamente o conteúdo dos dados constitui o núcleo funcional de qualquer ferramenta de geração de dados sintéticos. Nossa análise identifica duas abordagens complementares:

- **Geração Baseada em Padrões Textuais**: Utilização de expressões regulares (regex) para definição precisa de formatos de string, permitindo geração de identificadores únicos, códigos de produto, e formatos proprietários específicos de domínio.
- **Modelagem Estatística Paramétrica**: Implementação de distribuições probabilísticas configuráveis (gaussiana, linear, exponencial, uniforme) com controle granular sobre parâmetros estatísticos.

4.1.2. Dimensão de Configuração Estrutural

A formatação de saída representa um aspecto frequentemente negligenciado, mas crítico para integração de sistemas:

- **Especificação de Delimitadores**: Configuração de separadores de campo (`;`, `,`, `|`), delimitadores de texto (`"`, `'`), e separadores decimais (`.`, `,`).
- **Codificação e Formatação**: Controle sobre encoding de caracteres, formatação de datas e representação numérica.
- **Formatos de Exportação Múltiplos**: A plataforma deverá suportar a exportação dos dados gerados em múltiplos formatos tabulares padrão de mercado, como CSV, XLSX e ODT, além de um formato JSON estruturado que preserve a organização tabular, otimizado para testes de integração de sistemas e APIs.

4.1.3. Dimensão de Qualidade e Realismo

A simulação de imperfeições de dados reais emerge como requisito diferenciador:

- **Injeção Controlada de Outliers**: Capacidade de introduzir valores atípicos em colunas numéricas com distribuição e frequência configuráveis.
- **Simulação de Valores Ausentes**: Implementação de padrões de missing data, incluindo missing completely at random (MCAR), missing at random (MAR) e missing not at random (MNAR).
- **Distribuições Categóricas Ponderadas**: Especificação de frequências relativas para dados categóricos, permitindo simulação de distribuições não uniformes.

4.2. Formulação do Problema de Pesquisa

Com base na análise multidimensional apresentada, formulamos o seguinte problema de pesquisa:

Como operacionalizar a flexibilidade paramétrica em uma ferramenta de geração de dados sintéticos, de modo a maximizar a capacidade de configuração do usuário nas dimensões de conteúdo, estrutura e qualidade, mantendo simultaneamente critérios de usabilidade e eficiência computacional?

5. Objetivos da Investigação

5.1. Objetivo Geral

Estabelecer uma especificação de requisitos empiricamente validada para o desenvolvimento de uma plataforma de geração de dados sintéticos que implemente flexibilidade

paramétrica como princípio arquitetural fundamental, capacitando usuários a exercer controle granular sobre múltiplas dimensões dos dados gerados.

5.2. Objetivos Específicos

1. **Caracterizar o estado da arte** em ferramentas de geração de dados sintéticos através de análise comparativa sistemática, identificando limitações e oportunidades de inovação.
2. **Investigar a demanda empírica** por mecanismos avançados de geração de conteúdo, com ênfase na utilização de expressões regulares para síntese textual e distribuições estatísticas para modelagem numérica.
3. **Quantificar a relevância** de controles de configuração estrutural, incluindo personalização de delimitadores, separadores e formatos de exportação.
4. **Avaliar a criticidade** da simulação de imperfeições de dados, abrangendo outliers estatísticos, padrões de valores ausentes e distribuições categóricas não uniformes.
5. **Estabelecer uma taxonomia** de requisitos funcionais e não funcionais priorizados segundo critérios de impacto e viabilidade técnica.
6. **Desenvolver um framework conceitual** que oriente a arquitetura de software da solução proposta, integrando princípios de engenharia de software e usabilidade.
7. **Validar empiricamente** a especificação de requisitos através de consulta estruturada às partes interessadas, garantindo alinhamento com necessidades reais do domínio.

6. Justificativa e Contribuições Esperadas

6.1. Relevância Científica

Esta investigação contribui para o avanço do conhecimento em múltiplas dimensões:

- **Engenharia de Requisitos**: Apresentação de uma metodologia estruturada para elicitação de requisitos em domínios técnicos especializados, com ênfase na operacionalização de conceitos abstratos como ”flexibilidade”.
- **Engenharia de Software**: Estabelecimento de princípios arquiteturais para sistemas de alta configurabilidade, explorando trade-offs entre flexibilidade e complexidade.
- **Ciência de Dados**: Contribuição para a compreensão de requisitos de qualidade em dados sintéticos, incluindo métricas de realismo e fidelidade estatística.

6.2. Relevância Prática

O impacto prático desta pesquisa manifesta-se em múltiplos contextos:

- **Otimização de Processos de Desenvolvimento**: A capacidade de gerar dados sintéticos altamente customizados reduz significativamente o tempo de preparação de ambientes de teste e desenvolvimento.
- **Robustez de Sistemas**: A simulação controlada de imperfeições de dados permite teste mais eficaz de pipelines de ETL e algoritmos de machine learning, resultando em sistemas mais resilientes.
- **Conformidade Regulatória**: A geração de dados sintéticos estruturalmente idênticos, mas semanticamente distintos dos dados de produção, facilita a conformidade com regulamentações de privacidade.

6.3. Inovação Tecnológica

A principal inovação desta proposta reside na integração sinérgica de múltiplos paradigmas de geração de dados em uma plataforma unificada. Enquanto soluções existentes tratam diferentes tipos de dados como domínios isolados, nossa abordagem propõe uma arquitetura que permite composição arbitrária de mecanismos de geração, maximizando a expressividade do usuário sem comprometer a usabilidade.

A utilização de expressões regulares como mecanismo primário para geração textual representa uma inovação significativa no domínio, permitindo que usuários especifiquem formatos complexos através de uma linguagem formal bem estabelecida, eliminando a dependência de templates pré-definidos.

7. Cronograma e Metodologia de Execução

A execução desta investigação está estruturada em quatro fases principais, distribuídas ao longo de um período de 3 meses:

1. **Revisão Sistemática de Literatura** (Semanas 1-3): Análise abrangente do estado da arte em ferramentas de geração de dados sintéticos
2. **Desenvolvimento e Aplicação de Instrumentos** (Semanas 4-6): Elaboração e aplicação de questionários estruturados ao público-alvo
3. **Análise e Categorização** (Semanas 7-9): Processamento dos dados coletados e categorização de requisitos
4. **Síntese e Validação** (Semanas 10-12): Elaboração da especificação de requisitos e validação com especialistas

8. Considerações Éticas

Esta pesquisa será conduzida em estrita conformidade com os princípios éticos estabelecidos pelo Conselho Nacional de Ética em Pesquisa (CONEP). Todos os participantes serão informados sobre os objetivos da pesquisa e fornecerão consentimento informado antes da coleta de dados. A confidencialidade dos dados será garantida através de técnicas de anonimização, e os resultados serão apresentados de forma agregada, preservando a identidade individual dos participantes.

9. Resultados Esperados

Ao término desta investigação, esperamos produzir:

1. Uma **especificação de requisitos validada** que servirá como base para o desenvolvimento da plataforma de geração de dados sintéticos
2. Um **framework conceitual** para operacionalização da flexibilidade paramétrica em ferramentas de software
3. **Contribuições metodológicas** para a área de engenharia de requisitos em domínios técnicos especializados
4. **Publicações científicas** em conferências e periódicos de engenharia de software e ciência de dados

Referências

[1] Pressman, R. S.; Maxim, B. R. (2016). *Engenharia de Software: Uma Abordagem Profissional*. 8ª ed. Porto Alegre: AMGH Editora.

[2] Sommerville, I. (2019). *Engenharia de Software*. 10ª ed. São Paulo: Pearson Prentice Hall.

[3] Mockaroo Inc. (2025). Mockaroo - Random Data Generator and API Mocking Tool. Disponível em: `https://www.mockaroo.com/`. Acesso em: 23 set. 2025.

[4] Daniele Faraglia et al. (2025). Faker: A Python Package that Generates Fake Data. *Faker Documentation*. Disponível em: `https://faker.readthedocs.io/`. Acesso em: 23 set. 2025.

[5] Chen, R. J.; Lu, M. Y.; Chen, T. Y.; Williamson, D. F.; Mahmood, F. (2021). Synthetic Data in Machine Learning for Medicine and Healthcare. *Nature Biomedical Engineering*, 5(6), 493-497.

[6] Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. (2019). Modeling Tabular Data using Conditional GAN. *Advances in Neural Information Processing Systems*, 32, 7335-7345.

[7] Jordon, J.; Yoon, J.; van der Schaar, M. (2019). PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. *International Conference on Learning Representations*.

[8] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

[9] Little, R. J.; Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. 3ª ed. Hoboken: John Wiley & Sons.

[10] Dankar, F. K.; Ibrahim, M. K. (2021). Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. *Applied Sciences*, 11(5), 2158.