

## **PROGRAMA UNIFICADO DE BOLSAS DE ESTUDO PARA APOIO À FORMAÇÃO DE ESTUDANTES DE GRADUAÇÃO (PUB-USP)**

### **Extração Automatizada de Dados do Campus da USP São Carlos a partir do Anuário Estatístico da USP**

#### **Resumo**

Este projeto tem como objetivo desenvolver um sistema automatizado para extrair e organizar informações referentes ao campus da USP São Carlos a partir do Anuário Estatístico da Universidade de São Paulo (USP). A proposta inclui a documentação do sistema e o treinamento de um membro da equipe para garantir sua reutilização em anos futuros. O bolsista também deverá possuir conhecimentos em programação e manipulação de dados. O projeto terá duração de 12 meses, com dedicação semanal de 10 horas.

#### **Justificativa**

O Anuário Estatístico da USP é uma ferramenta valiosa para análise e planejamento institucional, mas apresenta os dados de forma agregada, o que dificulta a obtenção de informações específicas por campus. Atualmente, os dados dos "Números do Campus" são coletados de forma manual, por meio de solicitação direta às Unidades do campus, com posterior tabulação interna. Essa abordagem, embora funcional, demanda tempo e apresenta limitações em termos de corretude, atualização e padronização.

A automatização da extração de dados diretamente do Anuário da USP - fonte oficial, pública e regularmente atualizada - permitirá acelerar esse processo, aumentar a precisão dos levantamentos e facilitar a atualização anual dos indicadores locais. A coleta sistematizada de dados do campus é essencial para subsidiar a formulação de políticas de permanência, melhorias estruturais, gestão de recursos e análise da evolução do campus ao longo dos anos.

## **Apresentação de resultados anteriores**

Embora esta seja a primeira edição formalizada do projeto, a equipe já realizou, de forma manual, levantamentos parciais dos dados do campus da USP São Carlos a partir de dados coletados diretamente com Unidades e Órgãos do Campus e de anuários anteriores, o que demonstrou a viabilidade e relevância do desenvolvimento de uma solução automatizada.

## **Objetivos**

- Desenvolver um sistema automatizado para extração de dados das edições anuais do Anuário da USP relacionados ao campus da USP São Carlos;
- Organizar os dados extraídos de forma reutilizável e acessível;
- Documentar o funcionamento do sistema;
- Capacitar um membro da equipe para reutilização do sistema em edições futuras do Anuário.

## **Métodos**

Serão utilizados scripts em Python para realização de web scraping e extração de informações de arquivos PDF, Excel e CSV. Os dados serão filtrados por campus e organizados em tabelas ou bases reutilizáveis. Toda a lógica do sistema será documentada e estruturada para permitir sua reconfiguração anual. Utilitários como Jupyter Notebook poderão ser utilizados para garantir reprodutibilidade.

## **Detalhamento das atividades a serem desenvolvidas pelo(a)(s) bolsista(s)**

- Levantamento da estrutura e formato dos dados do Anuário Estatístico da USP;
- Desenvolvimento de scripts de extração e filtragem de dados por campus;
- Organização dos dados extraídos em formatos reutilizáveis;
- Testes com anuários de anos anteriores;
- Elaboração de documentação técnica;

- Treinamento de um membro da equipe para operação futura do sistema.

### **Inclusão de pré-requisitos para seleção de estudantes**

- Conhecimento básico em linguagem de programação (preferencialmente Python);
- Familiaridade com manipulação de dados em formatos como PDF, Excel e CSV;
- Noções de automação de tarefas e extração de dados (web scraping, scripts);
- Capacidade de filtrar e organizar dados com base em critérios específicos;
- Desejável: noções básicas de SQL e reprodutibilidade de código (Jupyter Notebook ou RMarkdown).

### **Resultados previstos e seus respectivos indicadores de avaliação**

- Sistema funcional para extração de dados (indicador: script executável com entrada e saída definidas);
- Base de dados organizada com informações filtradas por campus (indicador: conjunto de arquivos estruturados entregues);
- Documentação do sistema (indicador: manual de uso e reconfiguração);
- Capacitação de membro da equipe (indicador: treinamento realizado e testado com sucesso).

### **Conclusão e outras informações que sejam relevantes para o processo de avaliação**

O projeto contribuirá com a sistematização de dados institucionais, gerando eficiência em processos internos de gestão da informação e planejamento. Além disso, promove o desenvolvimento de competências técnicas por parte do bolsista, fortalecendo sua formação em áreas como programação, automação e gestão de dados.

**Cronograma de execução das atividades previstas (12 meses de vigência)**

<b>Mês</b>	<b>Atividades principais</b>
1º	Levantamento de requisitos e estrutura dos dados no Anuário; planejamento inicial.
2º - 3º	Desenvolvimento dos módulos iniciais de extração e leitura de dados.
4º - 5º	Filtros por campus e organização dos dados.
6º	Testes com anuários anteriores e ajustes.
7º	Documentação técnica preliminar e versão funcional.
8º - 9º	Otimizações e adaptações para novos formatos.
10º	Documentação final e início do treinamento.
11º	Consolidação do uso pela equipe.
12º	Entrega final e avaliação do projeto.