

# Inteligência Artificial Generativa Aplicada na Análise da Produção Científica

Programa de Pós-Graduação em Informática e Gestão do Conhecimento (PPGI) - Uninove  
Programa de Pós-Graduação em Gestão de Projetos (PPGP) - Uninove

Dr. Edson Melo de Souza

AULA 04

# Limpeza e Normalização de Metadados em Bibliometria

# Objetivos

- ▶ Capacitar alunos a identificar inconsistências em metadados bibliográficos e aplicar técnicas de normalização para garantir indicadores bibliométricos confiáveis.
- ▶ Explorar ferramentas e metodologias para limpeza de dados, abordando desafios comuns na análise bibliométrica.
- ▶ Dominar técnicas de tratamento e padronização de dados bibliográficos para garantir qualidade e replicabilidade.

# O Impacto do “Lixo” nos Dados

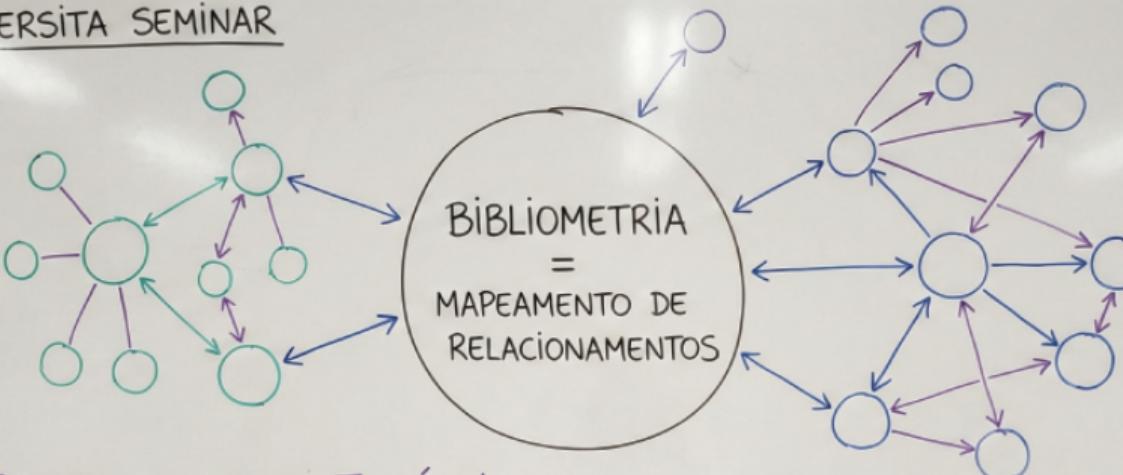
## Contextualização

A bibliometria não conta apenas documentos; ela mapeia relacionamentos. Se o “nó” do relacionamento (autor ou instituição) estiver duplicado com nomes diferentes, a rede se quebra ([PRITCHARD, 1969](#)).

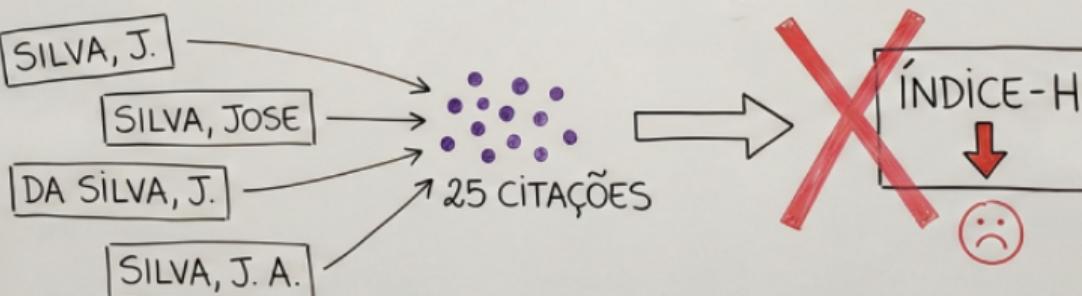
- ▶ Estudo de Caso: Imagine calcular o Índice-H do pesquisador “José Silva”, mas seus artigos estão listados sob “J. Silva”, “Jose Silva” e “J. S. da Silva”. O resultado? Um Índice-H subestimado.
- ▶ Outro Exemplo: Uma instituição chamada “Universidade Federal do Rio de Janeiro”, mas também aparece como “UFRJ” e “Federal University of Rio de Janeiro”. Isso pode levar a uma análise incorreta da produção científica da instituição.

**Consequência:** Em vez de um autor com 100 citações, temos 4 autores com 25 citações cada. O índice H cai drasticamente, prejudicando financiamentos e rankings.

## UNIVERSITA SEMINAR



### ESTUDO DE CASO: JOSÉ SILVA



CONSEQUÊNCIA: ÍNDICE-H CAI DRASTICAMENTE. PREJUDICA FINANCIAMENTOS E RANKINGS.

- Pró  
1. In  
2. IN  
3. TR  
ch  
4. EN  
(pr  
er

# Discussão

## Por que a Limpeza de Dados é Crucial em Bibliometria?

- ▶ **Precisão dos Indicadores:** Dados inconsistentes podem levar a métricas imprecisas, afetando avaliações de impacto e produtividade.
- ▶ **Credibilidade da Pesquisa:** Análises baseadas em dados sujos podem comprometer a confiança nos resultados.
- ▶ **Tomada de Decisão Informada:** Instituições e pesquisadores dependem de dados limpos para decisões estratégicas.
- ▶ **Eficiência na Análise:** Dados bem estruturados facilitam análises mais rápidas e eficazes.

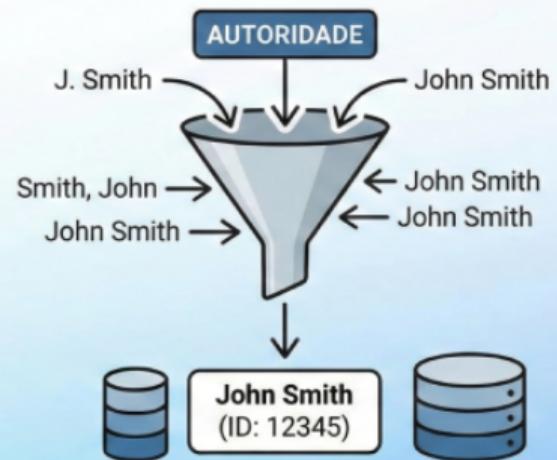
# LIMPEZA

(Remover sujeira)



# NORMALIZAÇÃO

(Padronizar variantes)



# Diagnósticos de Problemas Comuns

- Afiliações Institucionais Divergentes:** Diferentes formas de nomear a mesma instituição. **Exemplo:** “Univ Sao Paulo”, “USP”, “Universidade de São Paulo”, “University of Sao Paulo”, “Luiz de Queiroz College of Agriculture”. **Desafio:** Separar “Departamento” de “Instituição” em registros bibliográficos.
- Ambiguidade e Dispersão de Autores:** Ausência de informações cruciais como ano de publicação ou título. **Homônimos:** Wang, Y. (Existem milhares na China). Como saber se é o mesmo? **Variação de Nome:** García-Márquez, G. vs Márquez, G. G.
- Títulos de Periódicos e Conferências:** Registros repetidos que distorcem análises quantitativas. **Abreviações vs. Títulos completos:** " (J. Am. Chem. Soc. vs Journal of the American Chemical Society).
- Dados Duplicados:** Mesmos registros inseridos várias vezes. **Exemplo:** Um artigo listado múltiplas vezes em bases diferentes.

# Intervalo de 15 Minutos

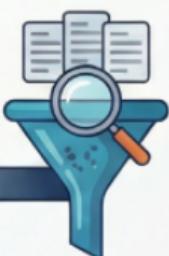
- ▶ **Excel:** Planilha eletrônica amplamente utilizada para manipulação básica e limpeza de dados.
- ▶ **OpenRefine:** Ferramenta poderosa para limpeza e transformação de dados. Permite identificar e corrigir inconsistências em grandes conjuntos de dados bibliográficos ([OpenRefine Development Team, 2023](#)).
- ▶ **Python com Pandas:** Biblioteca versátil para manipulação de dados. Scripts personalizados podem ser criados para automatizar a limpeza e normalização de metadados.
- ▶ **R com Bibliometrix:** Pacote específico para análise bibliométrica, que inclui funções para limpeza e padronização de dados ([ARIA; CUCCURULLO, 2017](#)).

# Guia Visual: Higienização de Dados Bibliométricos



## Elimine Registros Duplicados

Use o DOI ou a combinação de Título, Ano e Autor para identificação.



## Padronize Nomes e Instituições

Agrupe variações como "E. M. Souza" e "Souza, Edson M." em uma identidade única.



## Corrija Erros de Digitação

Assegure a consistência de termos e títulos, corrigindo a formatação.



## Prepare para a Análise

Exporte os dados limpos para um arquivo CSV compatível com o software de análise.



## Princípios Essenciais



### Mantenha Sempre o Original

Antes de iniciar, faça uma cópia de segurança dos dados brutos.



### Documente Todas as Alterações

Garanta transparência e a possibilidade de replicar o processo de limpeza.



### Automatize Tarefas Repetitivas

Considere o uso de scripts para otimizar a limpeza e reduzir erros.



### Verifique os Resultados Finais

Após a limpeza, confira se os dados estão corretos e consistentes.

# Exercício Prático de Limpeza de Dados com Excel

Utilizando a planilha fornecida com metadados bibliográficos, vamos aplicar técnicas de limpeza e normalização para corrigir inconsistências e preparar os dados para análise bibliométrica. O que vamos fazer:

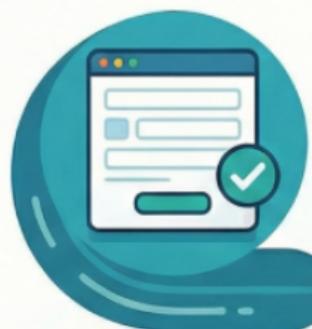
- ▶ Identificar e remover duplicatas (usando DOI ou combinação Title + Year + Authors).
- ▶ Padronizar nomes de autores e instituições (“E. M. Souza”, “Souza, Edson M.”, “Edson Melo de Souza” → uma única identidade). Converter para “Universidade Nove de Julho (UNINOVE)”
- ▶ Corrigir erros de digitação e formatação (“Bibliometric Analisys” → “Bibliometric Analysis”).
- ▶ Criar um CSV para usar no VOSviewer.

- ▶ **OpenRefine:** Importar os dados limpos do Excel para o OpenRefine para uma análise mais aprofundada. Utilizar suas funcionalidades para identificar padrões e inconsistências adicionais.
- ▶ **VOSviewer:** Exportar os dados finalizados para o VOSviewer para visualização e análise bibliométrica. Criar mapas de coautoria, coocorrência de termos e outras redes relevantes ([ECK; WALTMAN, 2010](#)).

- ▶ Adote padrões consistentes ao inserir novos dados bibliográficos.
- ▶ Utilize identificadores únicos, como ORCID para autores e ROR para instituições.
- ▶ Realize verificações regulares de qualidade dos dados para identificar e corrigir problemas rapidamente.
- ▶ Mantenha-se atualizado com as melhores práticas em gestão de dados bibliográficos.

# Garantindo a Qualidade dos Dados Bibliográficos

Para assegurar a integridade e a usabilidade de informações acadêmicas, é fundamental seguir um conjunto de boas práticas. Estas diretrizes formam um ciclo contínuo de gestão para manter a excelência dos dados bibliográficos.



## 1. Padronize a Inserção de Dados

Adote padrões consistentes ao inserir novas informações bibliográficas.



## 2. Use Identificadores Únicos

Utilize ORCID para autores e ROR para instituições para evitar ambiguidades.



## 4. Mantenha-se Atualizado

Acompanhe as melhores e mais recentes práticas em gestão de dados.

## 3. Verifique Regularmente

Realize auditorias de qualidade para identificar e corrigir problemas rapidamente.



# Material de Apoio

## Checklist para Higienização de Dados Bibliográficos

- ▶ [ ] **Remoção de duplicatas:** O mesmo artigo aparece duas vezes (ex: pré-print e publicado)?
- ▶ [ ] **Padronização de Afiliações:** Todas as variantes da instituição alvo foram mapeadas para o nome oficial?
- ▶ [ ] **Disambiguação de Autores:** Verifiquei se “Silva, J.” é o mesmo autor em todos os registros usando a afiliação ou área temática como dica?
- ▶ [ ] **Ano de Publicação:** Existem datas impossíveis (ex: 2029) ou vazias?
- ▶ [ ] **Tipo de Documento:** Estou misturando Artigos com Editoriais ou Correções? (Geralmente devem ser analisados separadamente).

# Exercício de Fixação

**Cenário:** Você recebeu uma lista de publicações para avaliar a colaboração internacional da “Universidade Federal do Rio de Janeiro”.

**Problema:** Ao fazer a nuvem de palavras, aparecem:

- ▶ UFRJ
- ▶ Univ. Fed. Rio de Janeiro
- ▶ Federal University of Rio de Janeiro
- ▶ COPPE - UFRJ

**Tarefa:** Utilizando o OpenRefine ou Excel, crie uma tabela “De -> Para” unificando todas as variantes sob a sigla “UFRJ”.

- ▶ **Leitura 1:** bibliometrix: An R-tool for comprehensive science mapping analysis  
<https://doi.org/10.1016/j.joi.2017.08.007>
- ▶ **Leitura 2:** Software survey: VOSviewer, a computer program for bibliometric mapping  
<https://doi.org/10.1007/s11192-009-0146-3>
- ▶ **Leitura 3:** Statistical bibliography or bibliometrics?  
<https://doi.org/10.1108/eb026468>

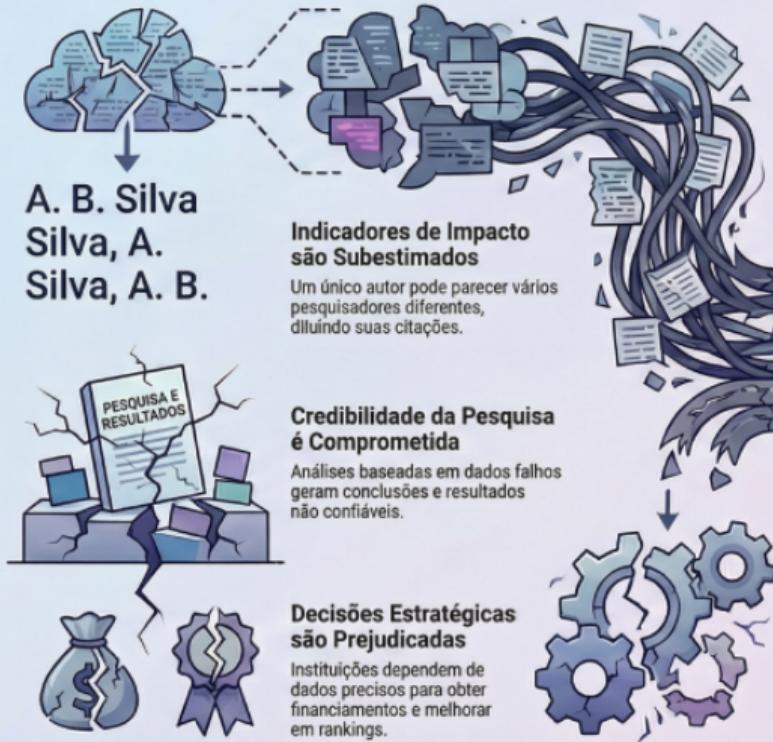
# Resumo

Nesta aula foram apresentadas informações sobre fontes de dados científicos e APIs, comparando bases proprietárias (Scopus e Web of Science) com a plataforma aberta OpenAlex ([PRIEM et al., 2022](#)). O texto detalha o funcionamento técnico de APIs RESTful e orienta sobre a extração automatizada de metadados acadêmicos, utilizando os endpoints e filtros da API do OpenAlex para fins de pesquisa e bibliometria.

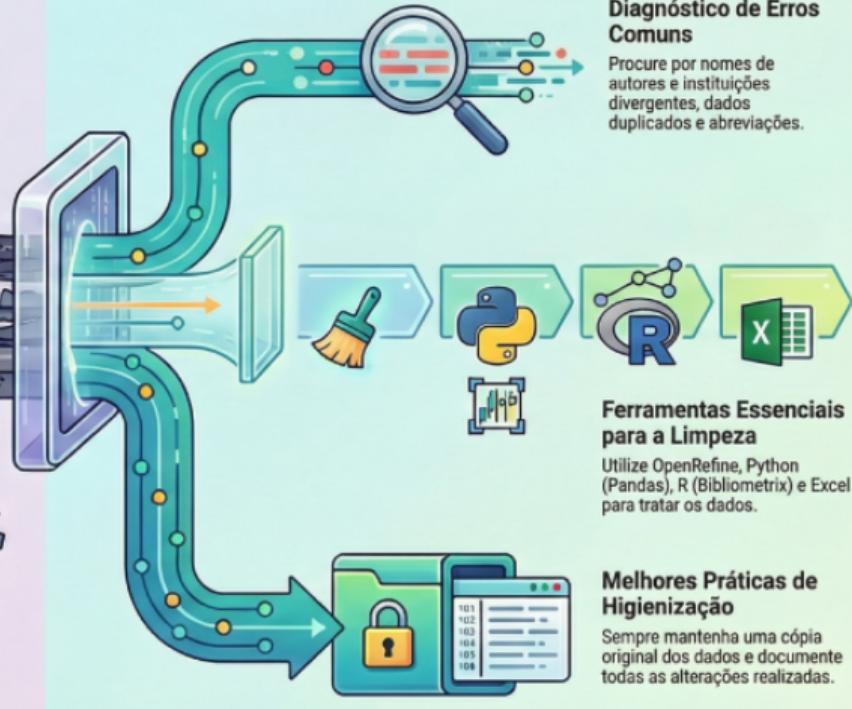
- ▶ **Bases de dados científicas** são essenciais para pesquisa e análise.
- ▶ **APIs RESTful** permitem extração automatizada de dados científicos.
- ▶ **OpenAlex** oferece acesso gratuito e fácil integração via API.
- ▶ **OpenRefine e Excel** são ferramentas úteis para limpeza e normalização de dados bibliográficos.
- ▶ **VOSviewer** é eficaz para visualização e análise bibliométrica.

# O Lixo nos Dados: A Importância da Limpeza de Metadados na Bibliometria

**O PROBLEMA:** Dados "Sujos" Distorcem a Realidade Científica



**A SOLUÇÃO:** Diagnóstico, Limpeza e Prevenção



## Referências

- ARIA, M.; CUCCURULLO, C. **Bibliometrix: An R-tool for comprehensive science mapping analysis.** *Journal of Informetrics*, v. 11, n. 4, p. 959–975, 2017.
- ECK, N. J. V.; WALTMAN, L. **Software survey: VOSviewer, a computer program for bibliometric mapping.** *Scientometrics*, v. 84, n. 2, p. 523–538, 2010.
- OpenRefine Development Team. **OpenRefine: A free, open-source tool for cleaning messy data.** 2023. Software. Disponível em: <https://openrefine.org>.
- PRIEM, J.; PIWOWAR, H.; ORR, R. **OpenAlex: A fully-open index of scholarly works, authors, venues, and institutions.** *arXiv preprint arXiv:2205.01833*, 2022. Disponível em: <https://arxiv.org/abs/2205.01833>.
- PRITCHARD, A. **Statistical bibliography or bibliometrics?** [S.l.]: Journal of Documentation, 1969. v. 25. 348–349 p.