Metodologia de desenvolvimento do indicador

“Projetos de Pesquisa Fiocruz na Plataforma Lattes”

Abril

2025

Sumário

1) Introdução

As interações científicas ocorrem em variados contextos e múltiplos níveis de colaboração, representando um dos pilares do avanço do conhecimento. Embora as colaborações documentadas em coautorias de publicações científicas sejam amplamente abordadas na literatura e respaldadas por bases de dados bibliográficas padronizadas, as colaborações estabelecidas em projetos (sejam eles de pesquisa, ensino ou desenvolvimento) permanecem pouco documentadas, portanto, menos acessíveis para análise sistemática. Isso ocorre principalmente pela ausência de uma base de dados consolidada e específica para registros de projetos em nível nacional.

Nesse cenário, a Plataforma Lattes, apresenta-se como uma fonte alternativa e potencial para a coleta de dados sobre projetos. Este documento descreve o procedimento desenvolvido pelo Observatório C,T&I em Saúde da Fiocruz, para coletar, tratar e organizar informações sobre registros de projetos cadastrados por pesquisadores associados à Fiocruz na Plataforma Lattes. Todos os projetos cadastrados por servidores da Fiocruz (ativos e inativos) foram coletados e apresentados neste painel, cuja metodologia encontra-se descrita nas próximas páginas.

2) Representação esquemática da metodologia adotada

Para ilustrar as etapas e a transformação feita nos dados, na Figura 1, é apresentado de forma esquemática o processo realizado. Todos os dados dos projetos dos servidores da Fiocruz foram coletados a partir da Plataforma Lattes. Uma mudança na orientação dos dados foi necessária. O objetivo desta reorientação foi retirar a centralidade dos dados das pessoas (servidores) e atribuir uma centralização nos projetos feitos por servidores da Fiocruz.

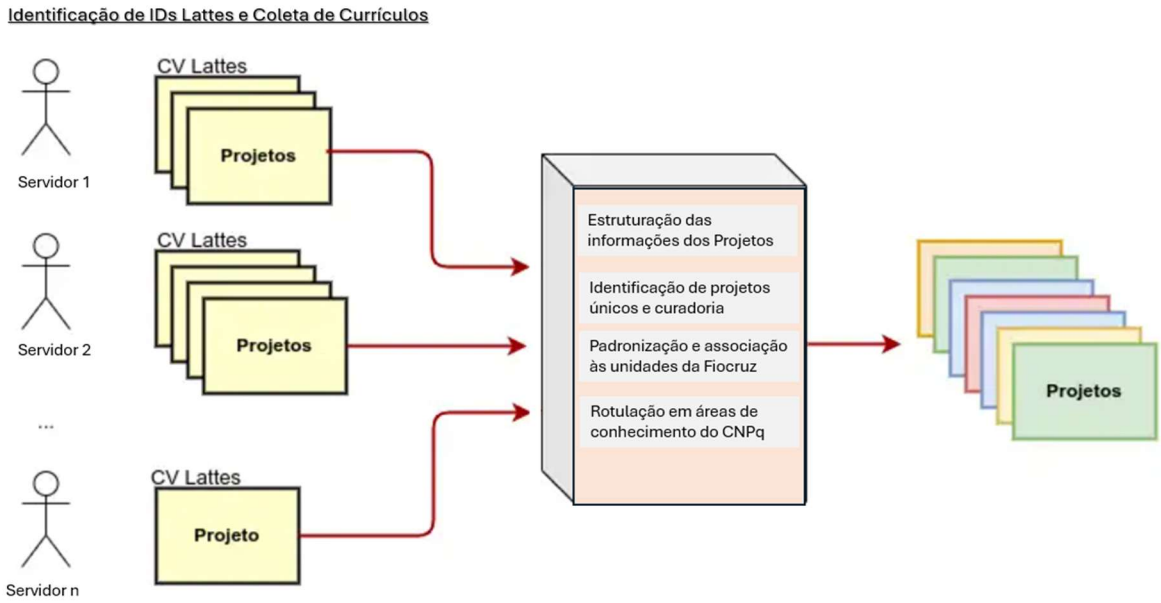

Figura 1. *Esquema macro da coleta de dados a partir dos CVs Lattes.*

3) Metodologia

3.1 Identificação dos IDs Lattes e coleta dos currículos

Utilizando o Base Lattes Fiocruz (BLF), sistema desenvolvido pela Coordenação-Geral de Gestão de Tecnologia de Informação (Cogetic), que importa e disponibiliza para consulta os currículos Lattes da instituição. A partir da relação de servidores fornecida pela Coordenação Geral de Gestão de Pessoas (Cogepe), extraiu-se do BLF os dados necessários para análises e construção do indicador. Foi necessário utilizar informações como a data de entrada e saída dos servidores da instituição, para adequar e contabilizar apenas os projetos pertencentes a servidores com matrícula ativa e vínculo institucional efetivo.

De posso dos IDs Lattes, os currículos correspondentes foram coletados no formato XML para cada servidor (ativo e inativo) da Fiocruz. Durante esta etapa, foi extraído especificamente informações sobre projetos cadastrados pelos pesquisadores, considerando todos os tipos existentes na plataforma: projetos de pesquisa, ensino, desenvolvimento, extensão e outros.

3.2 Estruturação das Informações dos Projetos

Nesta etapa, as informações coletadas foram organizadas em uma estrutura tabular (CSV). Entre os dados estruturados estão:

- ID-Lattes do CV Lattes.
- Nome completo do autor principal do projeto;
- Período de execução do projeto (Início e fim);
- Título do projeto;
- Natureza do projeto;
- Situação do projeto;
- Quantitativos de formação de alunos (graduação, especialização, mestrado e doutorado)
- Descrição completa do projeto (se houver);
- Informação sobre financiamento (se existe ou não);

3.3 Identificação dos projetos únicos e curadoria dos dados

Na medida em que os dados estruturados foram analisados, verificou-se que muitos projetos foram registrados por mais de um integrante do grupo1, ou seja, o mesmo projeto foi coletado de dois ou mais currículos, gerando duplicidade. Para remover os duplicados um script foi construído. Algumas decisões para determinar qual informação seria mantida como referência nos casos duplicados, foram tomadas. Como critério optou-se por manter o registro com maior completude das descrições, títulos e listas de participantes. Ressalta-se que por ser tratar de uma base auto preenchível os servidores digitam as informações, é praticamente impossível todas as repetições serem removidas. Para assegurar uma curadoria mais efetiva, uma segunda rodada de limpeza para remoção de duplicidades foi conduzida. Após os scripts

[1] A participação de um servidor em um projeto de pesquisa pode ser de diferentes formas, como coordenador, coordenador adjunto, como aluno de doutorado etc. No currículo não há essa diferenciação quanto a participação, de tal modo que optamos por chamar de **integrante do projeto** qualquer servidor que declarou, em seu CV Lattes, ter participado do respectivo projeto.

computacionais, os títulos dos projetos foram imputados no software Vantage Point®, onde uma limpeza e junção de repetições utilizando algoritmos de logica fuzzy foram aplicados.

Outro campo onde aplicou-se curadoria foi o período de duração dos projetos. Quando detectados projetos semelhantes ou idênticos, mas com datas diferentes, considerou-se o período mais abrangente possível. Por exemplo, se um integrante do projeto colocou iniciou em 2010 e outro integrante considerou o ano de início 2008, a referência mantida para o início foi 2008. Se um integrante colocou final em 2015 e o outro integrante colocou 2017, o ano de 2017 foi mantido, por ser o mais recente. Vale destacar, que por ser auto preenchível muitos pesquisadores esquecem de retornar aos seus currículos e atualizar a data de fim do projeto. Sendo assim, uma avaliação quanto ao período de duração dos projetos foi feita. Neste caso, considerou-se que projetos com duração superior a 15 anos que não tivessem ano de fim com *Duração incompatível*.

Por fim, para contabilizar apenas os projetos com servidores ativos vinculados a Fiocruz, adicionou-se uma verificação quanto as datas de entrada e eventual saída (inativos) dos servidores. Esta checagem se fez necessária para compatibilizar a duração do projeto com o período delimitado ao vínculo institucional, eventuais períodos de sobreposição foram mantidos.

3.4 Padronização e associação às unidades da Fiocruz

Após a deduplicação, e curadoria dos períodos de cada projeto foi realizada a padronização das informações, destacando a correta associação dos pesquisadores às suas respectivas unidades dentro da Fiocruz. Esse procedimento, foi realizado com auxílio da planilha de servidores, com sua respectiva localização Institucional, fornecida pela Cogepe. A unidade de localização atual do servidor foi escolhida como referência.

3.5 Atribuição do rótulo de financiamento

Por fim, atribuiu-se um indicador quantitativo que mostra se cada projeto recebeu algum tipo de financiamento, seja interno ou externo. Atualmente, este indicador apenas diferencia projetos financiados dos não financiados, permitindo uma rápida avaliação do impacto do financiamento sobre as atividades científicas da instituição. No currículo não há informações qualitativas sobre valores financiados.

3.6 Rotulação dos projetos em áreas de conhecimento

Para facilitar a identificação das áreas de pesquisa às quais os projetos estão relacionados, utilizou-se um processo de rotulação automática baseado em um modelo de Inteligência Artificial (IA) desenvolvido pelo Observatório C,T&I em saúde. Esse modelo tem como objetivo classificar cada projeto em uma Grande Área, Área e Subárea do conhecimento (os três principais níveis na taxonomia de classificação oferecido pelo CNPq), de acordo com a sua temática predominante.

O modelo utiliza a combinação de técnicas de TFIDF-LCPN-SVM elaborado especificamente para rotulação de publicações científicas (uma atividade também realizada pelo Observatório). Uma das vantagens desta abordagem é que ela não exige hardware especializado como GPUs para ser executada. Para mais detalhes do modelo e sua implementação, consultar a metodologia de produção científica.

3.7 Visualização dos dados

Uma view consolidada destes dados foi construída e um painel do tipo dashboard foi elaborado. Neste ambiente, foi desenvolvido um painel interativo, utilizando a ferramenta Power BI, que organiza as informações de forma estruturada e permite a rápida visualização de indicadores básicos sobre os diferentes projetos e sua evolução ao longo do tempo.

Além da visualização interativa, o painel permite a análise das redes de colaboração entre pesquisadores, a identificação das unidades mais atuantes e a definição dos pesquisadores centrais em termos de participação em projetos institucionais.

4) Aprendizado e próximos passos

Como ocorre em qualquer abordagem de coleta de dados, essa estratégia apresenta vantagens e limitações que devem ser cuidadosamente consideradas na definição do processo. A principal vantagem dessa abordagem é a organização estruturada das informações, uma vez que os dados são obtidos diretamente a partir da Plataforma Lattes dos pesquisadores vinculados a cada unidade da Fiocruz. Com isso, a responsabilidade pela atualização e completude das informações recai sobre o próprio pesquisador, o que torna o processo descentralizado e escalável.

Por outro lado, a principal limitação está justamente na dependência da qualidade e atualização dos currículos, o que pode comprometer a completude dos dados extraídos. Ainda assim, considerando a possibilidade de automatização da coleta e o alcance da base Lattes, entendemos que os benefícios superam as limitações, especialmente quando o objetivo é obter um panorama inicial ou complementar de projetos vinculados à instituição.

Em uma nova etapa deste projeto, pretende-se avaliar em conjunto com a Cogepe, de forma mais abrangente, a questão dos inativos. Caso um servidor tenha se aposentado em 2017, por exemplo, poderá ser verificado se este servidor ainda tem algum vínculo institucional, bolsista, celetista ou outro. Em casos positivos, a partir deste novo vínculo, incluir dados de projetos desenvolvidos pelo período que ainda está vinculado a instituição.

Com o objetivo de ilustrar o potencial analítico das redes de colaboração, elaboramos uma figura que representa graficamente as interações entre os projetos desenvolvidos pelos servidores e suas respectivas unidades institucionais.

Na Figura 3, cada aresta conecta duas unidades que realizaram projetos em conjunto. A espessura da aresta é proporcional ao número de projetos desenvolvidos em parceria, ou seja, quanto mais espessa a linha, maior o volume de colaboração entre as unidades correspondentes.

Essa abordagem evidencia o potencial das redes para retratar, de maneira objetiva, como ocorre a articulação entre diferentes grupos dentro da instituição. Trata-se de uma linha de análise, que será aprofundada nos próximos meses pelo Observatório.

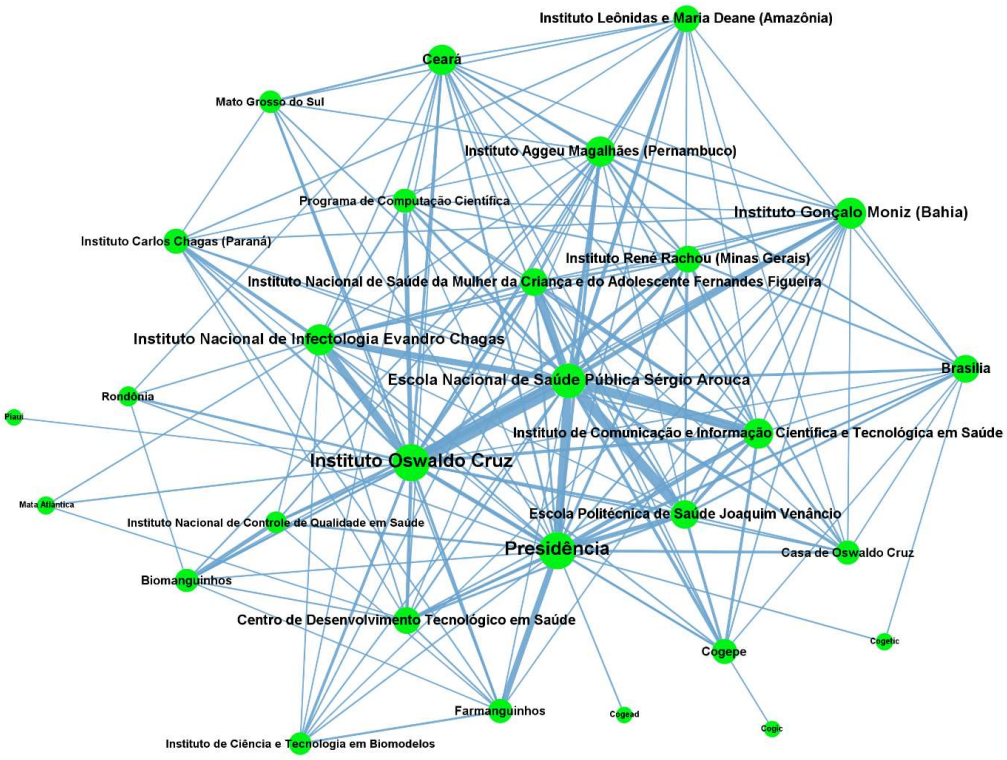

Figura 3. *Rede de coparticipação em projetos. Cada vértice representa uma unidade.*