PROTOCOLO DE COLETA DE DADOS BIBLIOGRÁFICOS DE PATENTES

Raulivan Rodrigo da Silva
https://orcid.org/0000-0002-2740-1045.
raulivan@cefetmg.br.
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) |
ROR https://ror.org/04ch49185 | Belo Horizonte, Brasil.

Thiago Magela Rodrigues Dias
https://orcid.org/0000-0001-5057-9936.
thiagomagela@cefetmg.br.
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) |
ROR https://ror.org/04ch49185 | Belo Horizonte, Brasil.

Washington Luís Ribeiro de Carvalho Segundo
https://orcid.org/0000-0003-3635-9384.
washingtonsegundo@ibict.br.
Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) |
ROR https://ror.org/006c42y96 | Brasília, Brasil.

Eixo temático: Técnicas, Ferramentas e Infraestruturas
Modalidade: Resumo expandido
DOI: 10.22477/ix.ebbc.361

Resumo: Este estudo visa contribuir com o avanço e disseminação de pesquisas envolvendo documentos de patentes, no contexto da Ciência da Informação. O principal objetivo é a proposição de um processo sistemático para a coleta e tratamento de grandes volumes de dados bibliográficos de patentes. Uma vez que diversos repositórios apresentam restrições relacionadas à quantidade de dados no processo de *download*. Os resultados obtidos mediante a aplicação das estratégias delineadas nesta pesquisa em um estudo de caso, são enriquecedores e contribuem de forma significativa, oferecendo assim uma alternativa eficaz aos desafios enfrentados na coleta de grandes conjuntos de dados de patentes.

Palavras-Chave: Patente. Protocolo. *Espacenet*.

1 INTRODUÇÃO

Neste ambiente dinâmico e de constante mudanças, acompanhar o desenvolvimento tecnológico, de um setor, ou um grupo de empresas tem se tornado um pré-requisito para empresas que desejam se destacar no meio tecnológico e de inovação. Dentre as diversas formas de acompanhar esse desenvolvimento, segundo Nascimento e Speziali (2020), o uso de informações tecnológicas contidas em documentos de patentes é uma excelente alternativa. A compreensão do estado da técnica da arte por meio de documentos de patentes, consequentemente apresenta um cenário mais assertivo a respeito de tendências tecnológicas, setores promissores, bem como, a possibilidade de novas tecnologias. Sanz-Casado (2006) aponta que estudos e análise de documentos de patentes permitem identificar o conhecimento científico e convertê-lo em conhecimento tecnológico.

Neste contexto, a coleta eficiente de dados de documentos de patentes é fundamental para o sucesso das pesquisas. A definição de bases de dados especializadas, técnicas de mineração de dados, análise textual entre outros elementos caracterizam um processo multidisciplinar envolvendo diversas áreas do conhecimento (Raturi *et al.*, 2010 apud Nascimento; Speziali, 2020).

Apesar de existirem diversos repositórios que disponibilizam o acesso pago ou gratuito aos dados de patente, como o BuscaWeb e Espacenet, ter acesso aos dados de documentos de patentes não é considerada uma tarefa trivial. Além de limitações como: limite no volume de dados a ser manipulado em um único processo; limite de horários de acesso aos dados; limite de filtros e critérios de seleção dentre outros. O que dificulta análises mais amplas, que abrangem um alto volume de dados, como por exemplo, avaliar o cenário de um país ou conjunto de instituições. Pires, Ribeiro e Quintella (2020) realizaram uma análise comparativa entre seis ferramentas de consulta de documentos de patentes, a saber, *Espacenet, Patentscope, Google Patents, Lens, Derwent Innovation Index* e *Orbit Intelligence*. Analisando 13 características distintas tais como o tipo de acesso, número de autoridades de patentes cobertas, tamanho da base e limite de exportação de dados. Logo, concluíram que todas as ferramentas analisadas possuem algum tipo de limitação, porém a ferramenta “Orbit Intelligence” se destaca, apresentando um melhor resultado em todos os critérios considerados na análise, no entanto, a ferramenta não é gratuita.

Logo, é necessário estabelecer um processo estruturado para recuperação dos dados. Os autores Moehrle *et al.* (2010) estabeleceram um processo de recuperação de dados patentários que pode ser dividido em três etapas, etapa de pré-processamento, caracterizada pela coleta dos dados, etapa de análise dos dados, em que são aplicados métodos e ferramentas de análise e por fim, a etapa de descoberta do conhecimento envolvendo análise dos resultados de forma visual e mensurável. Neste mesmo contexto, os autores Abbas, Zhang e Khan (2014) propuseram um processo dividido em três camadas classificadas em pré-processamento, processamento e pós-processamento. O processo inicia com a definição do repositório de patentes que será utilizado como fonte de dados, em sequência na etapa de pré-processamento é realizada a coleta das patentes que correspondem aos critérios pré-estabelecidos, para posteriormente filtrar e agrupar os dados de acordo com campos discriminados. Já na etapa de processamento, conjuntos menores são selecionados para serem utilizados na análise. E por fim, na etapa de pós-processamento, consiste em examinar e interpretar os conjuntos de dados para identificar padrões, análise de tendências, previsão tecnológica entre outros tipos de análise.

De acordo com o apresentado, o presente estudo tem como objetivo proposição de um protocolo sistemático de coleta e tratamento de dados patentários (Protocolo CTDP), fundamentado no processo genérico proposto pelos autores Abbas, Zhang e Khan (2014)..

2 METODOLOGIA

O processo metodológico descreve o funcionamento do "Protocolo de Coleta e Tratamento de dados Patentários (Protocolo CTDP)", composto por três etapas: definição do repositório de patentes; coleta dos dados bibliográficos; e por fim, construção de uma base de dados local.

2.1 ANTELAÇÃO DE REPOSITÓRIO DE DADOS DE PATENTES

Essa é a primeira etapa estabelecida pelo protocolo, caracterizada pela na definição de qual repositório de patentes deverá ser adotado como fonte de dados. Portanto, são estabelecidos 11 aspectos genéricos, os quais são utilizados para balizar a escolha do repositório de patente que mais se adequa às necessidades de pesquisa. A saber: Cobertura Geográfica (avaliar se o repositório abrange as patentes de todas as regiões geográficas de interesse da pesquisa); Amplitudes de Documentos (avaliar se entre a variedade de tipos de documentos disponibilizado pelo o repositório, como patentes, invenção, modelos de utilidade, *design*, famílias, citações, entre outros, são condizentes com o escopo da pesquisa); Áreas especializadas (avaliar se o repositório cobre os setores industriais específicos de interesse de pesquisa. Alguns repositórios de dados de patentes são especializadas em áreas específicas como química, biotecnologia, farmacêutica, entre outras); Atualização e Manutenção (avaliar a frequência que os dados do repositório são atualizados. Informações desatualizadas podem prejudicar a precisão das análises); Interface de Busca (avaliar a usabilidade da interface de busca dos dados. Uma interface intuitiva com muitos recursos filtros e seleção podem facilitar o processo de extração de informações); Recursos de Análise (avaliar se o repositório disponibiliza recursos para análise sem a necessidade de realizar download dos dados. Ferramentas de análise avançadas podem otimizar o processo eliminando o tempo de extração das informações); Idioma (avaliar se o repositório disponibiliza as patentes nos idiomas relevantes para pesquisa. O idioma pode interferir diretamente em análises textuais, como identificação de termos frequentes, como remoção de *stopwords*); Custo (avaliar os custos de acesso ao repositório de dados de patentes. Alguns repositórios são de acesso gratuito, enquanto outros exigem assinatura ou pagamento por acesso); Relevância do repositório (avalie a relevância e reputação do repositório de dados, bem como avaliações de outros pesquisadores); Restrições de Uso (avaliar se as restrições no uso dos dados, como limitações de *download*, compartilhamento ou uso comercial são compatíveis com as necessidades do estudo); API (avaliar se o repositório disponibiliza de recursos tecnológicos, como API e *Webservices* para viabilizar a construção de processos computacionais de coleta de dados).

Dessarte, é necessário elencar os repositórios de patentes de interesse e classificá-los de acordo com os aspectos estabelecidos. Pontuando com 1(um) quando o repositório atende e 0 (zero) quando não atende ao aspecto de acordo com as necessidades da pesquisa.

2.2 COLETA DE DADOS

Após a definição do repositório, é necessário a seleção e coleta dos dados. Entretanto, cada repositório pode apresentar suas particularidades no processo de disponibilização dos dados. Assim, considerando os repositórios *Derwent Innovation Index, Espacenet Google Patents, INPI, Latipat, Lens, Orbit Intelligence* e *Patentscope*, nesta primeira versão, foi definido a coleta no repositório *Espacenet*. A escolha de iniciar pela a *Espacenet*, se deu ao fato de estar em conformidade com todos os 11 axiomas estabelecidos na seção anterior e pela sua relevância internacional. Além de disponibilizar um serviço *web* denominado OPS (*Open Patent Services*) que fornece acesso gratuito ao banco de dados do EPO (*European Patent Office*) por meio de API (*Application Programming Interface*). O que torna viável o desenvolvimento de algoritmos computacionais para automatizar a consulta e coleta dos dados. Ademais, a *Espacenet* disponibiliza as patentes no idioma original, ou seja, todas as patentes depositadas no Brasil são disponibilizadas no idioma portugues do Brasil.

Por conseguinte, fazendo uso da OPS, é possível definir filtros para a seleção das patentes de interesse. Seguindo a documentação (OPS, 2023), é possível pesquisar por mais de 20 campos seguindo a sintaxe CQL (*Contextual Query Language*), dispensando assim a necessidade conhecer previamente os números de depósito das patentes. Por exemplo, para consultar patentes brasileiras publicadas no período de janeiro a dezembro de 2021, é possível construir a seguinte consulta: "pn=BR and pd="20210101 20211231", em que no parâmetro "pn" é informado o código do país e utilizando o conector "and" é informado o parâmetro "pd" em que é informado o período de publicação da patente, destacando a data inicial e final que devem ser informadas no formato AAAAMMDD (ano, mês, dia).

Em vista disso, para a segunda etapa do protocolo, foi implementado um algoritmo utilizando uma linguagem de programação que viabilize executar requisições com protocolo *Hypertext Transfer Protocol* (HTTP) e manipule arquivos no formato JSON. Desta forma é possível automatizar todo o processo de coleta. Para evitar possíveis bloqueios de acesso é realizado apenas cinco requisições por minuto. E para não extrapolar o limite de download estabelecido pelo repositório, é realizada uma requisição para cada dia dentro do intervalo de publicação informado como critério de consulta, logo, é necessário sempre informar um intervalo de datas de publicação como critério de seleção das patentes além dos demais filtros de interesse. As patentes coletadas na Espacenet, são armazenadas em arquivos no formato ".json", formato este, já adotado pela OPS para retorno dos resultados, evitando assim a conversão para outros formatos.

2.3 BASE DE DADOS LOCAL

E por fim, com a coleta dos dados concluída. A terceira etapa do protocolo consiste unificar os dados coletados e estabelecer relacionamentos entre eles. A padronização dos dados é fundamental para garantir a consistência e a comparabilidade entre diferentes conjuntos de dados coletados. Assim, ao final desse processo, os dados estarão prontos para serem explorados e analisados.

Portanto, foi adotado o uso de um banco de dados relacional, baseado no Modelo Entidade Relacionamento introduzido por Peter Pin-Shan Chen, em 1976 (Date, 2003). O banco de dados definido

foi o SQLite, por se tratar de um banco de dados que não tem a necessidade de um servidor dedicado, além de não demandar nenhum tipo de configuração, possibilitando o acesso local. Assim, seu uso como base de dados se torna, além de fácil, mais fluído, dinâmico e leve (The Sqlite Consortium, 2023).

Por isso, sete tabelas foram modeladas para compor o banco de dados local. A tabela *PATENTS* para armazenar dados básicos de uma patente, como título, resumo, data de publicação e depósito; tabela *PERSONS* para armazenar dados de identificação de toda as pessoas envolvidas, como depositantes e inventores; tabela *APPLICANTS* que estabelece um relacionamento do tipo depositante entre as tabelas "*PATENTS*" e "*PERSONS*; a tabela *INVENTORS* que estabelece um relacionamento do tipo inventor entre as tabelas "*PATENTS*" e "*PERSONS*"; a tabela *CLASSIFICATIONS* que armazena dados das classificações recebidas pelas as patentes; a tabela *IDENTIFICATIONS* que armazena todas as variações do número de depósito e de publicação de uma patente; e por fim, a tabela *PERSON_CITATIONS* que armazenado as diferente formas em que uma pessoa é referenciada na patente.

Posto isto, antes de armazenar os dados nas respectivas tabelas, alguns tratamentos foram realizados objetivando a limpeza dos dados. O primeiro foi estabelecer critérios para definir que uma entrada de dados é válida, ou seja, definir o mínimo informação que um determinado tipo de dado deve possuir. Para uma patente ser considerada válida, deve possuir o número de depósito e o título informados.

Em sequência, no conjunto de patentes consideradas válidas foi realizada a deduplicação dos dados, considerando como critério o número de depósito, o título e o ano de depósito, devido ao fato da Espacenet disponibilizar todas as versões de publicação de uma patente, ou seja, a versão de publicação gerada após as alterações. E por fim, os nomes próprios foram transformados em caixa alta e removido o excesso de espaços.

Desta forma, o processo de transformação, deduplicação e padronização foram implementados utilizando a linguagem de programação *Python*, permitindo assim a automação de todo o processo.

Mediante as estratégias de coleta e tratamento de dados apresentadas, foi possível estabelecer um protocolo consistente e sistemático de coleta e armazenamento de dados bibliográficos de patentes.

3 RESULTADOS

Mediante as estratégias definidas neste estudo, foi possível obter como resultado um conjunto ferramental escrito na linguagem de programação Python que viabiliza a aplicação do protocolo proposto neste estudo, o Protocolo CTDP.

Com o objetivo de realizar uma prova de conceito, com o objetivo de verificar a eficácia do protocolo proposto, o mesmo foi aplicado para coletar todas as patentes depositadas no Brasil e disponibilizadas no repositório da Espacenet entre os anos de 1900 a 2022. O processo durou quatro meses e 20 dias, recuperando o total de 903.979 patentes. O elevado tempo de execução, devido às estratégias adotadas para evitar as restrições do repositório, torna o repositório com baixo desempenho em relação ao tempo de processamento. Contudo, foi possível recuperar 20,5 gigabytes de dados, sem intercorrências. A Tabela 1 apresenta o quantitativo de dados coletados.

Tabela 1 – Quantitativos dos dados coletados

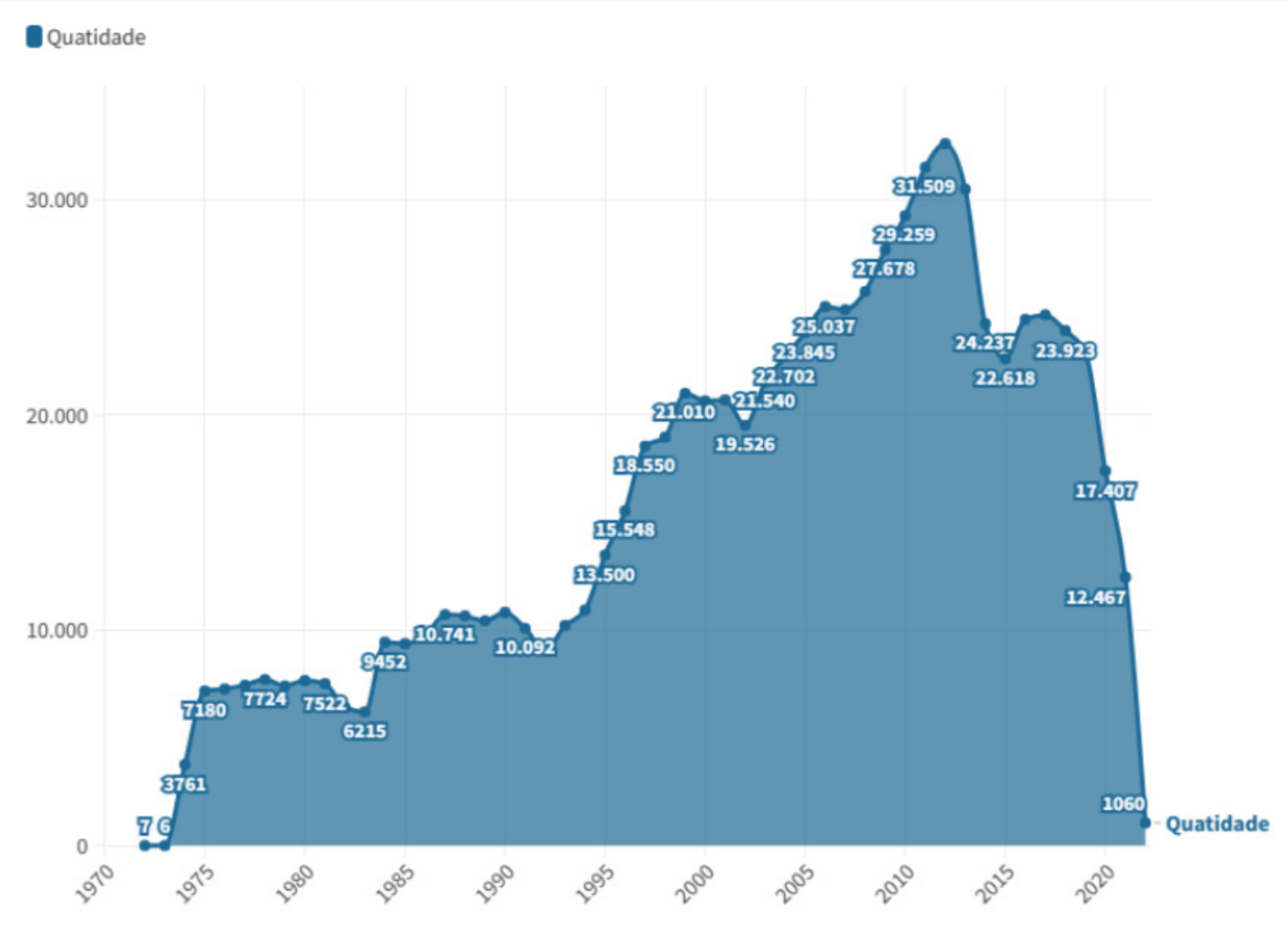
Tipo de registro	Quantidade
Patentes coletadas na Espacenet	903.979
Patentes distintas coletadas na Espacenet	797.747

Fonte: Elaboração do autor (2023)

Conforme apresentado, mediante a aplicação do protocolo proposto neste estudo, foi possível construir uma base de dados composta por 903.979 registros de patentes oriundos da Espacenet, em que considerando apenas a última publicação de cada patente, obtém-se o total de 797.747 patentes únicas.

Objetivando explorar a base de dados local estabelecida, considerando todo o conjunto de patentes coletadas, a Figura 1 apresenta a evolução temporal do número de depósito de patentes ao longo dos anos.

Figura 1 – Evolução temporal do depósito de patentes por ano



Fonte: Elaboração do autor (2023).

É possível observar que no período analisado, de 1970 a 2022 no Brasil foram realizados em média aproximadamente 16.000 depósitos de patentes por ano, sendo 2012 o ano com maior número de

patentes depositadas, contendo um total de 32.608 patentes. Destaca-se ainda que houve um crescimento contínuo no número de depósitos entre os anos de 1992 e 2012, podendo ser considerado um período de grande desenvolvimento tecnológico. É possível observar que posterior a 2012 há uma queda considerável, em que uma hipótese para tal fato pode estar relacionado a possibilidade de existirem pedidos de patentes que foram depositados no INPI, porém ainda não foram disponibilizados para a Espacenet, ficando assim, fora do conjunto de dados considerados pela análise. Destacando ainda, a possibilidade de existirem patentes ainda no período de sigilo, corroborando para queda no número de patentes mais recentes.

4 CONSIDERAÇÕES FINAIS

O protocolo CTDP foi concebido fundamentado no modelo genérico delineado por Abbas, Zhang e Khan em 2014. Este protocolo oferece diretrizes para orientar a seleção do repositório de patentes mais adequado, levando em consideração as exigências específicas da pesquisa. Além disso, propõe estratégias para a recuperação eficiente de grandes conjuntos de patentes e a conversão desses dados em uma estrutura acessível que facilite a consulta e a extração de conjuntos de informações específicas.

Destacando que a base de dados gerada por meio da aplicação do Protocolo CTDP pode ser facilmente integrada a diversas ferramentas de análise e mineração de dados, tais como Gephi, Carrot2 e Microsoft Power BI. Essa integração é viabilizada pelo fato de a base de dados ser estruturada sob o modelo relacional, o que possibilita a conexão com qualquer ferramenta que suporte a interação com bancos de dados SQLite.

Contudo, algumas limitações foram identificadas neste estudo, a ausência de aplicação do Protocolo CTDP em diversas abordagens analíticas para validar sua eficácia, o que resulta em uma falta de testes mais abrangentes do protocolo. Além disso, outra limitação é a restrição da integração apenas com a base de dados da Espacenet, o que inviabiliza, nesta versão inicial, a utilização de outras bases de dados, como a Patentscope.

Ademais, espera-se que este estudo possa fornecer uma contribuição substancial para a ampliação do escopo dos estudos relacionados a documentos de patentes.

REFERÊNCIAS

ABBAS, A.; ZHANG, L.; KHAN, S. U. A literature review on the state-of-the-art in patent analysis. **World Patent Information,** [*S. l.*], v. 37, p. 3, 2014. DOI: https://doi.org/10.1016/j wpi.2013.12.006. Disponível em: https://www.sciencedirect.com/science/article/abs/piiS0172219013001634?via%3Dihub. Acesso em: 6 mar. 2024.

DATE, C. J. **Introdução a Sistemas de Banco de Dados.** 8. ed. [*S. l.*]: Elsevier, 2003.

MOEHRLE, M. G. *et al.* Patinformatics as a business process: a guideline through patent research tasks and tools. **World Patent Information,** [*S. l.*], v. 32, n. 4, p. 291, 2010. DOI: https://doi.org/10.1016/j.

wpi.2009.11.003. Disponível em: https://ideas.repec.org/a/eee/worpat/v32y2010i4p291-299.html. Acesso em: 6 mar. 2024.

NASCIMENTO, Raphael da Silva; SPEZIALI, Marcelo Gomes. Patentometria: a utilização de dados contidos em patentes como mecanismo de análise da predominância tecnológica dos nits. *In*: ENCONTRO INTERNACIONAL DE GESTÃO, DESENVOLVIMENTO E INOVAÇÃO, 4. ed., 2020, **Anais [...].** [*S. l.*]: Eigedin, 2020.

OPS. **Open Patent Services RESTful Web Services.** 1.3.16. ed. [*S. l.*], 2023. Web service. Disponível em: https://www.epo.org/searching-for-patents/data/web-services/ops.html. Acesso em: 13 jan. 2024.

PIRES, E. A.; RIBEIRO, N. M.; QUINTELLA, C. M. Sistemas de busca de patentes: análise comparativa entre espacenet, patentscope, google patents, lens, derwent innovation index e orbit intelligence. **Cadernos de Prospecção,** [*S. l.*], v. 13, n. 1, p. 13, mar. 2020. DOI: https://doi.org/10.9771/cp.v13i1.35147. Disponível em: https://periodicos.ufba.br/index.php/nit/article/view/35147. Acesso em: 24 abr. 2024.

SANZ-CASADO, Elías. **Los estudios métricos de la información y la evaluación del a actividad científica: conceptos básicos.** São Paulo: ECA/USP, 2006. Material didático do curso “Os estudos métricos da informação”, ministrado no Programa de Pós-Graduação em Ciência da Informação da ECA/USP, nov. 2006.

THE SQLITE CONSORTIUM. **The Sqlite Consortium.** versão 3.45.3. [*S. l.*], 2024. Website. Disponível em: https://sqlite.org/index.html. Acesso em: 13 jan. 2024.