XXIV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – XXIV ENANCIB

ISSN 2177-3688

GT 8 – Informação e Tecnologia

COLETA E INTEGRAÇÃO DE FONTES DE DADOS HETEROGÊNEAS SOBRE PATENTES

COLLECTION AND INTEGRATION OF HETEROGENEOUS PATENT DATA SOURCES

Raulivan Rodrigo da Silva – Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Thiago Magela Rodrigues Dias – Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Washington Luís Ribeiro de Carvalho Segundo – Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)

Modalidade: Trabalho Completo

Resumo: No campo da ciência da informação e ciência de dados, a integração e padronização de múltiplas fontes de dados são cruciais para garantir a consistência e comparabilidade dos resultados. Este trabalho aborda a necessidade de integrar dados heterogêneos relacionados a patentes coletadas de diferentes repositórios, destacando os desafios e benefícios dessa prática. Assim, este trabalho tem como objetivo principal estabelecer por meio do processo metodológico fundamento em estudo de caso, um processo sistemático de integração de dados relacionado a patentes provenientes de três fontes distintas: Espacenet, INPI e currículos da Plataforma Lattes. Como resultado, mediante as estratégias delineadas neste estudo foi possível estabelecer um esquema composto por oito entidades que visam normalizar os dados e estabelecer relacionamentos entre as diferentes fontes, de tal forma a viabilizar análises de diversas magnitudes. Embora tenha-se focado em dados oriundos da Espacenet, INPI e Plataforma Lattes, o modelo proposto pode ser adaptado para outras fontes de dados.

Palavras-chave: patentes; base de dados; heterogêneos; Espacenet; Plataforma Lattes.

Abstract: In the field of information science and data science, the integration and standardization of multiple data sources are crucial to ensure the consistency and comparability of results. This paper addresses the need to integrate heterogeneous patent-related data collected from different repositories, highlighting the challenges and benefits of this practice. Thus, the main objective of this paper is to establish, through the methodological process based on a case study, a systematic process of integrating patent-related data from three different sources: Espacenet, INPI, and curricula from the Lattes Platform. As a result, through the strategies outlined in this study, it was possible to establish a scheme composed of eight entities that aim to normalize the data and establish relationships between the different sources, in such a way as to enable analyses of different magnitudes. Although the focus was on data from Espacenet, INPI, and the Lattes Platform, the proposed model can be adapted to other data sources.

Keywords: patents; data base; heterogeneous; Espacenet; Plataforma Lattes.

1 INTRODUÇÃO

No contexto da ciência da informação e ciência de dados, a integração e padronização de múltiplas fontes de dados são essenciais para garantir a consistência e comparabilidade dos resultados. Neste sentido, este trabalho aborda a necessidade de integrar conjuntos de dados heterogêneos sobre patentes coletadas de diferentes repositórios, destacando os desafios e benefícios dessa prática. De acordo com Bernstein e Haas (2008) e Seligman *et al*. (2010) a integração de dados provenientes de diversas fontes é um dos maiores e mais custosos desafios na área da computação. Entre as principais dificuldades está a consolidação de fontes e repositórios heterogêneos em um formato único e uniforme, de maneira automática e segura.

Atualmente, em 2024, há uma variedade de repositórios de patentes disponíveis na internet, que podem ser de acesso aberto ou comercial. Esses repositórios podem ser classificados em dois grupos principais: internacionais e nacionais. Os repositórios nacionais, também denominados repositórios locais, estão vinculados aos respectivos escritórios de propriedade industrial, presentes em quase todas as nações que aplicam proteção à propriedade intelectual. Geralmente, esses repositórios disponibilizam apenas documentos de patentes depositados em seus próprios escritórios. Em contraste, os repositórios internacionais oferecem acesso a documentos de patentes provenientes de múltiplos escritórios, abrangendo diversas nacionalidades (Brandão, 2016).

Dada a diversidade de repositórios existentes, observa-se uma consequente variedade de estruturas de organização e formatos para a disponibilização dos dados relacionados ao depósito de patentes. Cada repositório adota sua própria política de armazenamento e disponibilização de informações, apresentando os dados em diferentes formatos, tais como XML, CSV, TXT, XLS, XLSX, PDF, RIS, BibTeX ou JSON (Pires; Ribeiro; Quintella, 2020).

Nesta circunstância, é uma prática comum no processo de coleta de dados sobre depósito de patentes ou em pesquisas na área de Propriedade Intelectual, surgir a necessidade de coletar dados em fontes distintas com intuito de enriquecer o conjunto de dados a ser usado para análises, uma vez que a informação disponibilizada pode variar de acordo com o repositório usado para a coleta (Pires; Ribeiro; Quintella, 2020). Entretanto, essa

prática pode gerar alguns desafios devido à falta de integração entre as bases de dados, como a duplicidade e inconsistência dos dados, que implica diretamente na qualidade dos resultados de análises.

Dessa maneira, devido ao grande volume de dados coletados, a tarefa de integração de diferentes conjuntos de dados pode se tornar exaustiva e suscetível a erros. Ademais, o processo de integração em alguns casos não pode ser inteiramente automatizado por meio de softwares, em virtude da complexidade inerente à representação das estruturas semânticas dos dados coletados, o que impõe obstáculos significativos à identificação e resolução de conflitos.

Mediante o exposto, este trabalho tem como objetivo principal, estabelecer um processo sistemático de integração de dados relacionados a patentes, provenientes de três fontes de dados distintas, a saber: Espacenet, INPI e currículos da Plataforma Lattes.

2 FUNDAMENTAÇÃO TEÓRICA

Pesquisadores na área de gerência de dados, tem adotado esforço em pesquisas que viabilizem um acesso comum para fonte de dados heterogêneos, objetivando fornecer acesso integrado às informações provenientes de distintas fontes de dados, preservando a integridade da informação armazenada (Almeida, 2021). Uma forma de promover a uniformidade de fontes de dados heterogêneos é definir um ou mais esquemas que representem uma visão que contemple os bancos de dados envolvidos. Com esse intuito, Batini, Lenzerini e Navathe (1986) estabelecem que esse processo é conhecido como integração de esquemas, caracterizado pela resolução de conflitos semânticos, descritivos e estruturais entre os esquemas das fontes de dados a serem integrados. Shelt e Larson (1990) concordam que a integração de esquemas se apresenta como excelente solução no processo de integração de diferentes fontes de dados.

Neste contexto, um esquema define como os dados são estruturados em um banco de dados de modelo relacional, incluindo nomes de tabelas ou entidades, campos, tipos de dados bem como seus relacionamentos. Desta forma a proposição de um esquema que contemple todas as bases de dados que se almeja integrar emerge como uma solução para o desafio de integrar fontes de dados heterogêneas (Lima,2016; Ram; Ramesh, 1999). O processo de proposição de um esquema unificado pode ser caracterizado como um processo que,

mediante a entrada de um conjunto de diferentes esquemas de banco de dados, obtém-se como saída, uma descrição única dos esquemas iniciais bem como a informação de mapeamento entre os esquemas (Batini, Lenzerini, Navathe, 1986). Este esquema unificado, pode viabilizar acesso flexível e eficiente a várias fontes de dados heterogêneas.

Existem três principais etapas, de acordo com Batini, Lenzerini e Navathe (1986), existentes nas diversas metodologias de integração de bases de dados heterogêneas. Etapa de pré-integração, em que todas as bases de dados iniciais são modeladas usando um modelo comum; Etapa de identificação de correspondência entre os esquemas, em que são identificados os objetos relacionados e os que geram conflitos entre as bases de dados; e por fim, a Etapa de geração do esquema unificado e da informação do mapeamento, que consiste basicamente te na construção de guia para integrar as bases de dados.

3 METODOLOGIA

Este trabalho trata-se de um estudo de caso, ou seja, um estudo de natureza empírica que investiga um determinado fenômeno, dentro de um contexto em que ainda há lacunas na literatura (Serrano; Gobbo Junior, 2014). Estudo de caso, de acordo com Yin (2005), pode ser caracterizado pelo estudo que "investiga um fenômeno contemporâneo dentro de seu contexto da vida real, especialmente quando os limites entre o fenômeno e o contexto não estão claramente definidos". Desta forma, a adoção do estudo de caso como estratégia de pesquisa, é lato, pois apesar de focar em casos específicos, um estudo de caso bem conduzidos permite contribuir para a generalização teórica ao identificar padrões ou princípios que podem se aplicar a contextos mais amplos.

Esta seção consiste na especificação de um esquema para integração de conjunto de dados heterogêneos sobre o depósito de patentes. Viabilizando assim a construção de uma base de dados local composta por dados provenientes da Espacenet, INPI e currículos da Plataforma Lattes. Partindo da conclusão prévia da coleta dos dados em seus respectivos repositórios, ou seja, este estudo não aborda os processos e métodos de coleta de dados e sim a integração dos dados em uma única base de dados.

3.1 Caracterização das entidades e atributos

Dito isto, a coleta eficiente de dados de documentos de patentes é fundamental para pesquisas envolvendo Propriedade Intelectual. 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 (Nascimento; Speziali, 2020). Entretanto, com a coleta dos dados concluída, se faz necessário integrar os dados coletados e estabelecer relacionamentos entre eles. Estabelecer uma padronização dos dados é fundamental para garantir a consistência e a comparabilidade entre os diferentes conjuntos de dados.

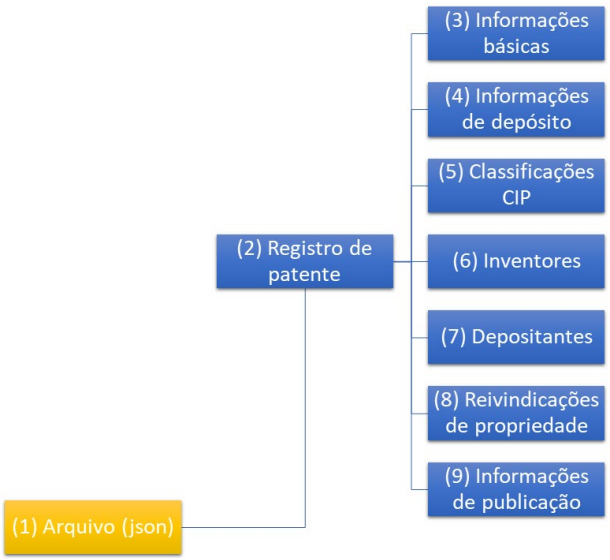
Assim, para integrar os conjuntos de dados coletados é especificado nesta seção um esquema de banco de dados fundamentado no Modelo Entidade Relacionamento introduzido por Peter Pin-Shan Chen, em 1976 (Date, 2003).

O esquema proposto neste estudo, estabelece uma estrutura lógica que define a organização e a forma de armazenamento dos dados coletados. Ele compreende a descrição formal das tabelas (entidades), suas colunas (atributos), tipos de dados, restrições, relacionamentos entre as entidades e outras estruturas que compõem um banco de dados relacional. Desta forma, o esquema serve como um conjunto de regras que especifica como os dados são organizados e como as relações entre diferentes conjuntos de dados são estabelecidas, garantindo assim a integridade e a consistência dos dados armazenados.

Nesse sentido, uma entidade é considerada uma estrutura de dados que representa a essência de um tipo específico de informação. Por exemplo, a entidade "Patente" é uma estrutura de dados que agrupa as informações pertinentes a uma patente, constituindo assim um registro de patente no banco de dados. Em síntese, uma entidade é formada por atributos que a caracterizam, e esses atributos podem ser utilizados como critérios de seleção, bem como na composição das informações a serem extraídas do banco de dados.

Para a construção das entidades, os conjuntos de dados coletados foram analisados a fim de identificar o esquema atual dos mesmos. Iniciando pelo conjunto coletado na Espacenet, as patentes coletadas são disponibilizadas em arquivos no formato .json e cada arquivo possui informações de uma ou mais patentes, a Figura 1 apresenta como é organizado o conteúdo de cada arquivo.

Figura 1 – Representação do arquivo de patentes coletadas na Espacenet



Fonte: Elaboração dos autores

O item (1) ilustra o arquivo gerado com os dados obtidos ao consultar uma patente na Espacenet; já o item (2) representa um registro de patente; no item (3) são as informações básicas de identificação da patente, tais como título, resumo e código de identificação da família a qual a patente pertence; o item (4) apresenta as informações de depósito da patente, como a data e número de depósito no formato original, tal como informado pelo INPI, e nos padrões EPODOC e DOCDB, estes dois últimos são regras de formatação aplicado no número de identificação da patente, DOCDB é usado na versão atual da Espacenet, e o EPODOC é usado na versão Classic da Espacenet; no item (5) são as classificações recebidas pela a patente; no item (6), a listagem dos nomes dos inventores; no item (7) contém a listagem dos nomes dos depositantes; no item (8) informações das reivindicações de propriedade; e por fim, no item (9) as informações de publicação da patente, como data de publicação e número de identificação da patente, assim como nos dados de depósito, os dados de publicação seguem os formatos original, EPODOC e DOCDB.

Já os dados coletados no INPI é um conjunto de dados denominado “BADEPI v8.3 Patentes” que reúne as informações bibliográficas sobre as Patentes (patente de invenção (PI), modelo de utilidade (MU) e certificado de adição (C)) que foram depositadas no INPI no período de 1997 a 2020. Conjunto este disponibilizado pelo INPI com o apoio da Organização Mundial da Propriedade Intelectual (OMPI).

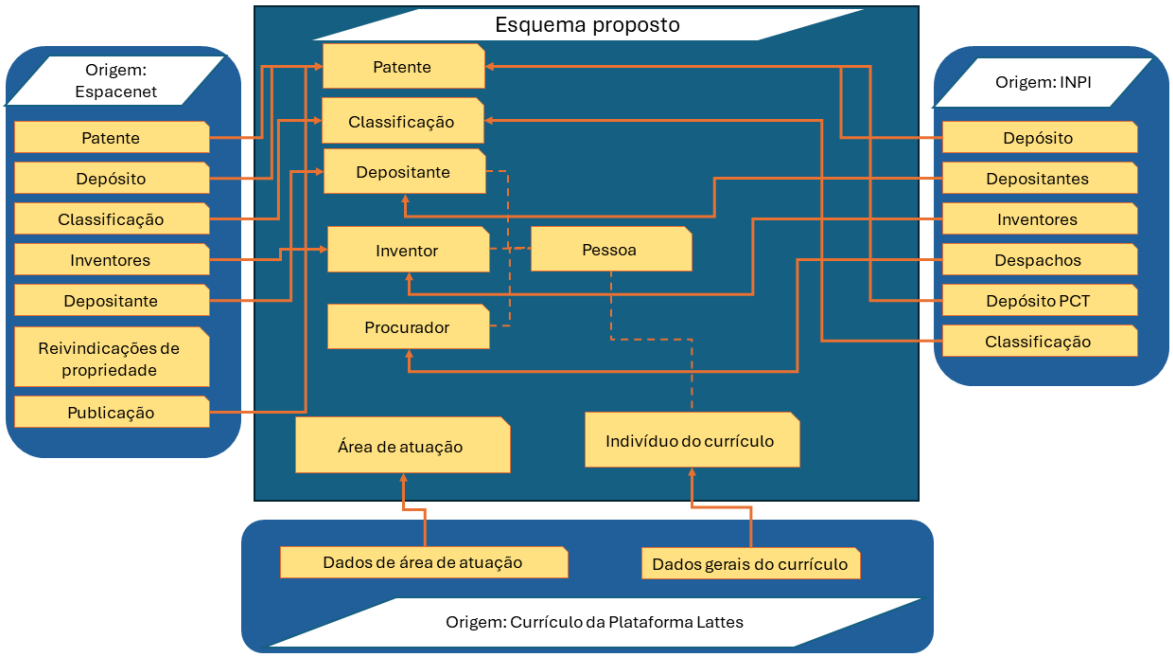
A base de dados BADEPI na versão 8.3, é composta por sete arquivos de extensão .CSV. O arquivo “PTN_DEPOSITOS”, com pedidos de patentes depositados no INPI; os arquivos “PTN_DEPOSITANTES” e “PTN_DEPOSITANTES_V2” com dados dos depositantes das patentes; o arquivo “PTN_INVENTORES” com os dados dos inventores; o arquivo “PTN_DESPACHOS” com informações dos despachos realizados no processo de depósito das patentes; o arquivo “PTN_PCTS” com as informações pedido internacional depositado através do sistema PCT; e por fim, o arquivo “PTN_CLASSIFICACOES” com as classificações atribuídas por cada patente durante o processo de depósito.

E por fim, os dados dos currículos da Plataforma Lattes, são em conformidade com a estrutura obtida quando recupera o currículo no formato XML. Considerando as informações contidas na chave “DADOS-GERAIS” e da chave “AREAS-DO-CONHECIMENTO”.

3.2 Mapeamento entre as fontes de dados

Concluída a compressão da estrutura dos conjuntos de dados a serem integrados, é necessário a identificação e mapeamento das entidades que irão compor o esquema proposto neste estudo. A Figura 2 apresenta a visão geral do mapeamento de entidade de acordo com sua respectiva fonte de dados.

Figura 2 – Identificação de correspondência entre os esquemas



Fonte: Elaboração dos autores

A figura apresenta quatro grupos de entidades, o grupo da esquerda são as entidades identificadas no conjunto de dados coletado na Espacenet, composto por sete entidades (dados da patente, dados referente ao depósito da patente, classificações recebidas pela patente, dados dos inventores, dados dos depositantes, dados dos processos de reivindicação de propriedade e por fim, os dados de publicação das patente. O conjunto alinha na parte inferior da imagem, estão as entidades dados gerais do currículo e dados de áreas de atuação extraídos do conjunto de currículos coletados na Plataforma Lattes. No lado direito da figura, estão as entidades identificadas no conjunto de dados oriundos do INPI (dados sobre o depósito de patentes, dados sobre os depositantes, dados dos inventores, dados dos despachos envolvidos no processo de depósito da patente, dados sobre os depósitos via PCT e os dados das classificações atribuída às patentes). E por fim no centro da figura, estão as entidades do esquema proposto neste estudo.

A entidade “Patente” do esquema proposto, agrupa os dados “Patente”, “Depósito” e “Publicação” do conjunto de dados coletados na Espacenet. De igual modo, agrupa os dados “Depósito” e “Depósito PCT” do conjunto coletado no INPI. A entidade “Classificação” agrupa os dados das entidades Classificação dos conjuntos de dados coletados na Espacenet e no INPI. Todas as pessoas envolvidas foram agrupadas na entidade “Pessoa” e por meio das entidades “Depositante”, “Inventor” e “Procurador” estabelecem o tipo de relacionamento da pessoa com a patente, considerando os dados da Espacenet e do INPI. E por fim, estão as entidades mais específicas, a entidade “Área de atuação” agrupa as áreas de atuação informada nos currículos e a entidade “Indivíduo Lattes” agrupa os dados de identificação do proprietário do currículo, estabelecendo um relacionamento com a entidade “Pessoa”.

3.3 Estratégia de coleta de dados

Esta subseção é destinada à exposição de uma estratégia de coleta de dados sobre depósito de patentes, fazendo uso da linguagem de programação Python. A linguagem de programação Python é amplamente reconhecido pela sua simplicidade e poder no desenvolvimento de ferramentas de mineração de dados e análise, o que faz dele a escolha ideal para o desenvolvimento de ferramentas de coleta de dados, além do fato de atualmente (2024), estar entre as mais utilizadas no setor tecnológico.

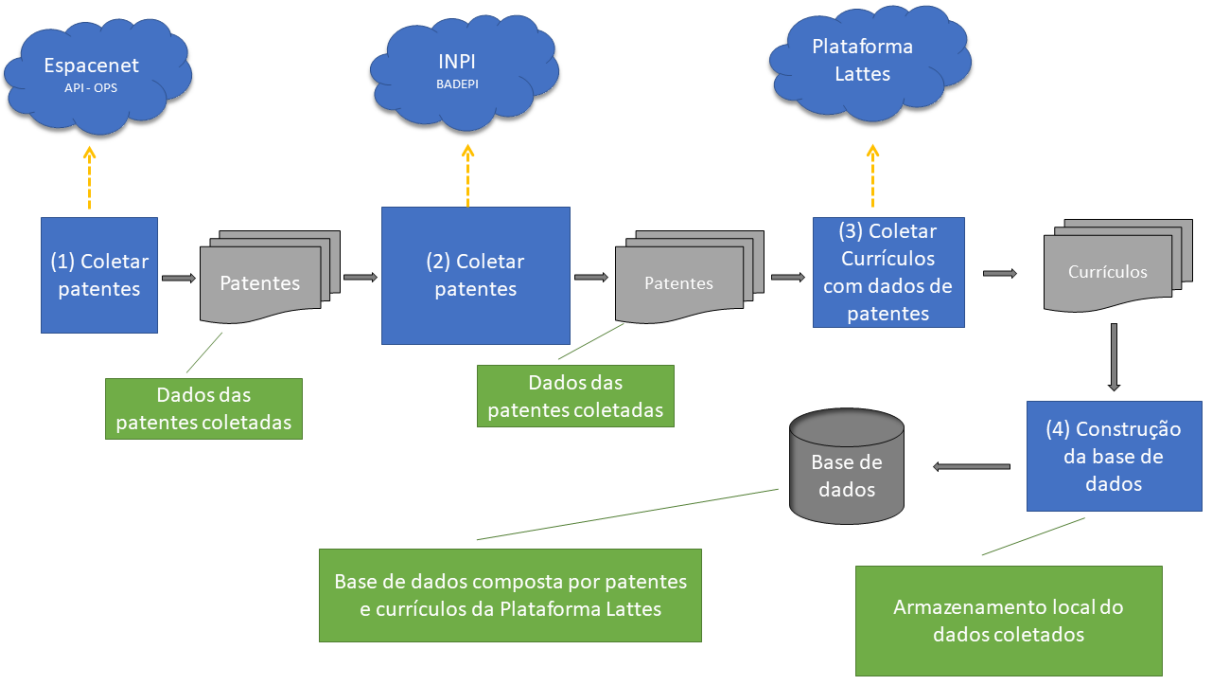
Dito isto, na primeira etapa coleta de dados são definidas as fontes de dados. Conforme apresentado anteriormente foram agregados dados de três diferentes fontes: Espacenet, INPI e Plataforma Lattes, esta última é relevante pois disponibiliza dados sobre os indivíduos que normalmente não são disponibilizados em documentos de patentes, como a titulação acadêmica, áreas de atuação entre outras.

Para a coleta de dados de patentes no repositório Espacenet é realizada através do serviço OPS (*Open Patent Services*) que fornece acesso aberto ao banco de dados do EPO *(European Patent Office*). A fim de complementar a coleta realizada na Espacenet, em sequência, foi coletado a Base de Dados Estatísticos sobre Propriedade Industrial (BADEPI), um conjunto de dados disponibilizado pelo INPI contendo as principais estatísticas relativas aos serviços prestados. Este conjunto de dados se torna relevante pois contém informações que não foram obtidas no processo de coleta na Espacenet, como informações da nacionalidade e cidade de residência dos depositantes e inventores da patente.

Os dados coletados na Espacenet e no INPI, são limitados em relação a informações sobre os indivíduos relacionados a produção técnica. Assim, objetivando contornar este limitador e de se propor uma estratégia de validação dos dados coletados com outras fontes, utilizando o *framework LattesDataXplorer* desenvolvido por Dias (2016), realizou-se a coleta de currículos cadastrados na Plataforma Lattes do CNPq, que possuem informações sobre depósito de patentes. Agregar dados de currículos da Plataforma Lattes além de enriquecer a base de dados com informações tais como formação acadêmica, áreas de atuação, instituições de vínculo, endereço profissional, permite a proposição de uma estratégia para verificar a consistência dos dados de patentes informados pelos proponentes em seus currículos. Outra alternativa para a obtenção de dados dos currículos da Plataforma Lattes é através da Plataforma BrCris, um sistema agregador que possibilita a recuperação, certificação e visualização de dados e informações pertinentes aos diversos atores envolvidos na pesquisa científica no contexto brasileiro mantida pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT) (IBICT, 2024). O BrCris possui em sua interface recursos que viabilizam o download dos resultados de busca no formato CSV.

A Figura 3, apresenta a visão geral do fluxo de coleta de dados estabelecido.

Figura 3 – Visão geral da coleta de dados



Fonte: Elaboração do autor.

Mediante ao exposto, todo o processo de coleta dos dados foi dividido em quatro etapas distintas, (1) coleta dos dados de patentes no repositório da Espacenet, (2) coleta de patentes no INPI, (3) a coleta de dados dos currículos da Plataforma Lattes que possuem referências a depósito de patentes, e por fim, (4) a construção da base de dados local composta pelos dados coletados.

4 RESULTADOS

Mediante ao apresentando, como resultado do processo metodológico obtém-se um esquema composto por um conjunto de oito entidades. Essas entidades foram modeladas conforme os dados recuperados nas bases Espacenet, INPI e Plataforma Lattes. Entretanto, caracteriza um modelo que pode ser facilmente adaptado para outras fontes de dados. Assim, a primeira entidade do esquema é a entidade “Patente” detalhada pelo Quadro 1.

Quadro 1 – Entidade Patente

Atributos	**Tipo**	**Descrição**
Número do documento	Texto	Número de depósito da patente
País	Texto	Código do país de depósito da patente
Código de publicação	Texto	Código da versão de publicação da patente
Última publicação	Sim/Não	É a última versão de publicação da patente?
Data de depósito	Data	Data de depósito da patente
Data de publicação	Data	Data de publicação da patente
Data de concessão	Data	Data de concessão da patente
Primeira Classificação CIP	Texto	Código da primeira classificação da patente
Título	Texto	Título da patente
Resumo	Texto	Resumo da patente
Identificação da família	Texto	Código de identificação da família da patente (Espacenet)

Fonte: Elaboração dos autores.

A "Entidade Patente" é composta por 12 atributos que a caracterizam. Os atributos da entidade Patente foram definidos conforme os dados bibliográficos da patente, fundamentados na norma ST.9. Assim, para a construção da "Entidade Patente" se faz necessário percorrer todo o conjunto de dados coletados e extrair e agrupar tais informações pelo número de depósito, a fim de construir uma entidade para cada informação de depósito de patente recuperada. Em sequência, apresentado pelo Quadro 2 a "Entidade Classificação".

Quadro 2 – Entidade Classificação

Atributos	**Tipo**	**Descrição**
Classificação	Texto	Classificação recebida pela patente
Seção	Texto	Seção da classificação de acordo com a CIP
Classe	Texto	Classe da classificação de acordo com a CIP
Subclasse	Texto	Subclasse da classificação de acordo com a CIP
Grupo	Texto	Grupo da classificação de acordo com a CIP
Subgrupo	Texto	Subgrupo da classificação de acordo com a CIP
Ordem	Numérico	Ordem de atribuição das classificações da patente
Patente	Patente	Patente que recebeu a classificação

Fonte: Elaboração dos autores.

Devido ao fato de uma patente poder receber mais de uma classificação, se fez necessário ter uma entidade para representar tais classificações. A "Entidade Classificação" é composta por 8 atributos, o valor da classificação e os valores dos elementos que compõem a

classificação. O atributo ordem estabelece a ordem de prioridade das classificações atribuídas a patente e o atributo Patente, é utilizado para estabelecer uma relação entre a patente e suas classificações.

Prosseguindo, uma patente dentre as suas informações, possui informações referente aos indivíduos parte do processo de depósito, sejam eles depositante, inventor ou procurador. Assim, para representar tais indivíduos foi definida a "Entidade Pessoa", detalhada pelo Quadro 3.

Quadro 3 – Entidade Pessoa

Atributos	**Tipo**	**Descrição**
Nome	Texto	Nome do indivíduo
Pessoa Jurídica	SIM/NÃO	Seção da classificação de acordo com a CIP
Nacionalidade	Texto	Nacionalidade do indivíduo

Fonte: Elaboração dos autores.

O objetivo da "Entidade Pessoa" é concentrar todos os indivíduos participantes do processo de depósito, sejam eles pessoas físicas ou jurídicas, objetivando preparar um ambiente que viabilize a identificação e caracterize única de cada indivíduo. E para estabelecer os vínculos entre os indivíduos e as patentes, são estabelecidas as entidades "Inventor", "Depositante" e "Procurador" (Quadro 4).

Quadro 4 – Entidade Depositante, Entidade Inventor e Entidade procurador

Entidade Depositante		
Atributos	**Tipo**	**Descrição**
Sequência	Numérico sequencial	Ordem de participação informada no processo
Pessoa	Pessoa	Entidade Pessoa
Patente	Patente	Entidade Patente
Entidade Inventor		
Atributos	**Atributos**	**Atributos**
Sequência	Numérico sequencial	Ordem de participação informada no processo
Pessoa	Pessoa	Entidade Pessoa
Patente	Patente	Entidade Patente
Entidade Procurador		
Atributos	**Atributos**	**Atributos**
Sequência	Numérico sequencial	Ordem de participação informada no processo
Pessoa	Pessoa	Entidade Pessoa
Patente	Patente	Entidade Patente

Fonte: Elaboração dos autores.

A “Entidade Depositante”, estabelece o vínculo dos depositantes das patentes, já a “Entidade Inventor” estabelece vínculo de inventores de uma patente e por fim, a “Entidade Procurador” que estabelece o vínculo entre os procuradores envolvidos no processo de depósito da patente.

Em sequência, tem-se a “Entidade Indivíduo Lattes” (Quadro 5). Uma entidade formada como fonte de dados complementar formada a partir de currículos registrados na Plataforma Lattes que possuem informações de depósito de patentes em seus currículos. Essa fonte de dados se torna relevante pois trata-se de uma promissora fonte de dados sobre os indivíduos envolvidos no depósito de patentes.

Quadro 5 – Entidade Indivíduo Lattes

Atributos	**Tipo**	**Descrição**
ID Lattes	Texto	Número de identificação do currículo na plataforma
Nome	Texto	Nome do proprietário do currículo
Data de Atualização	Data	Data da última atualização do currículo
ORCID	Texto	ORCID do individuo
Pais	Texto	Pais de residência informado no currículo
Estado	Texto	Estado de residência informado no currículo
Cidade	Texto	Cidade de residência informado no currículo
Area de Atuação	Texto	Primeira área de atuação informada
Maior Título	Texto	Maior titulação informada
Pessoa	Pessoa	Entidade Pessoa

Fonte: Elaboração dos autores.

É possível observar que a entidade possui as informações provenientes dos currículos coletados na Plataforma Lattes acrescido do atributo Pessoa” útil para estabelecer o vínculo entre os proprietários dos currículos com os indivíduos parte do processo de depósito da patente. E por fim, a “Entidade Área de Atuação” (Quadro 6) responsável por representar as áreas de atuação informadas nos currículos da Plataforma Lattes.

Quadro 6 – Entidade Área de Atuação

Atributos	**Tipo**	**Descrição**
Área	Texto	Área de atuação informada
Especialidade	Texto	Especialidade informada
Subárea	Texto	Subárea informada
Grande área	Texto	Grande área informada
Sequência	Texto	Ordem de atribuição das áreas de atuação
Currículo Lattes	Individuo Lattes	Currículo de vínculo com a área de atuação

Fonte: Elaboração dos autores.

A “Entidade Área de Atuação”, armazena as informações de grande área, área e subárea de atuação informadas nos currículos. Além de possuir o atributo “Currículo Lattes” que estabelece um vínculo com a “Entidade Indivíduo Lattes”, o proprietário do currículo mapeado.

Mediante o apresentando, as oito entidades modeladas têm como objetivos normalizar os dados e estabelecer relacionamento entre as diferentes fontes de dados. Ressalta-se que, embora este estudo tenha se dedicado à integração de dados oriundos da Espacenet, INPI e Plataforma Lattes, há possibilidade de adaptação para outras fontes de dados. Ademais, o objetivo deste trabalho não é propor um modelo físico de entidade e relacionamento, mas sim apresentar uma estratégia para a integração de fontes de dados distintas relativas a documentos de patentes.

5 CONSIDERAÇÕES FINAIS

A necessidade de fazer uso de uma ou mais fontes de dados relacionados ao depósito de patentes em estudos na área da Propriedade Intelectual é de fato uma prática comum, seja para enriquecer a base de dados ou atender alguma necessidade específica do projeto de pesquisa. Contudo, relacionar conjuntos de formatos distintos é um processo que exige muito esforço e que em muitos casos não é possível automatizar, devida a semântica dos dados.

Contudo, apesar da complexidade envolvida, a integração de bases de dados heterogêneas é uma tarefa essencial que impacta diretamente na qualidade dos resultados de uma pesquisa. O objetivo do desenvolvimento deste estudo é contribuir para o processo de integração de fontes de dados com diferentes estruturas, servindo como um guia para orientar os pesquisadores no processo de integração de diferentes conjuntos de dados sobre o depósito de patentes. Ademais, a integração eficiente de diferentes conjuntos de dados melhora a precisão e a abrangência das análises realizadas, também permite identificar e remover dados duplicados. A metodologia proposta visa minimizar os erros de cruzamento de informações e redundâncias, contribuindo assim para a confiabilidade dos resultados.

REFERÊNCIAS

ALMEIDA, João Gabriel Quaresma de. **Aqüeducte:** um serviço para integração de dados heterogêneos em cidades inteligentes. 96 f. Dissertação (Pós-Graduação em Sistemas e Computação) — Centro de Ciências Exatas e da Terra da Universidade Federal do Rio Grande do Norte, Natal, 2021.

BATINI, C.; LENZERINI, M.; NAVATHE, S B. A comparative analysis of methodologies for database schema integration. **ACM computing surveys,** New York, NY, v. 18, n. 4, p. 323–364 Dec. 1986.

BERNSTEIN, P. A., HAAS, L. M. Information integration in the enterprese: a guide to the tools and core technologies for merging information from disparete sources. **Communications of the ACM**, New York, NY, v. 51, n. 9, p. 72–79, Sept., 2008.

BRANDÃO, F. G. **Democratização da informação a partir do uso de repositórios digitais institucionais : da comunicação científica às informações tecnológicas de patentes**. Dissertação (Mestrado) — Universidade Regional Integrada do Alto Uruguai e das Missões, set. 2016. Disponível em: https://lume.ufrgs.br/handle/10183/179853. Acesso em: 24 maio 2024.

DATE, C. J**. Introdução a sistemas de banco de dados**. 8. ed. Rio de Janeiro : Campus, 2003.

DIAS, Thiago Magela Rodrigues. **Um estudo da produção científica brasileira a partir de dados da Plataforma Lattes**. 181 f. Tese (Doutorado em Modelagem Matemática e Computacional) — Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte, 2016.

IBICT. **BrCris**. 2024. Disponível em: https://brcris.ibict.br/. Acesso em: 16 abr. 2024.

LIMA, Cláudio de. **Projeto lógico de bancos de dados noSQL documento a partir de esquemas conceituais entidade-relacionamento estendido (EER)**. 140 f. Dissertação (Ciência da Computação) — Universidade Federal de Santa Catarina, Florianópolis, 2016.

NASCIMENTO, M. G.; SPEZIALI, Raphael da S. Patentometria: a utilização de dados contidos em patentes como mecanismo de análise da predominância tecnológica dos nits. **Anais do IV Encontro Internacional de Gestão, Desenvolvimento e Inovação (EIGEDIN)**, v. 4, n. 1, 30 out. 2020. Edição online realizada de 3 a 6 nov. 2020.

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**, Salvador, v. 13, n. 1, p. 13, mar 2020. Disponível em: https://periodicos.ufba.br/index.php/nit/article/view/35147. Acesso em: 1 maio 2024.

RAM, S.; RAMESH, S. Schema integration: past, present and future. Management of

heterogeneous and autonomous databases systems. In: Elmagarmid, Ahmed; Rusinkiewicz, Marek; Sheth, Amit (ed.). **Management of heterogeneous and autonomous database systems**. San Francisco, USA: Morgan Kaufmann Publishers, 1999. p. 119–155.

SELIGMAN. *et al*. **Openll**: AN open source information integration toolkit. **Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data**, New York, USA, p. 1057–1060, 2010. Trabalho apresentado no evento realizado de 6 a 10 jun. 2010, em Indianapolis, Indiana, USA.

SERRANO, B.; GOBBO JUNIOR, J. A. Redes de inovação: mapeamento de inventores de patentes em uma empresa do setor de cosméticos. **Revista GEPROS: gestão da produção, operações e sistemas**, v. 9, n. 1, p. 101–113, jan. 2014.

SHELT, A.P.; LARSON J. Federated database systems for managing heterogeneous, distributed and autonomous Databases. **ACM Computing Surveys**, New York, USA, v. 22, n. 3, p. 183–236, Sept. 1990.

YIN, R. K. **Estudo de caso**: planejamento e métodos. 2. ed. Porto Alegre : Bookman, 2005.