CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DE MINAS GERAIS
CAMPUS TIMÓTEO

João Marcos Martins da Costa Cota

RECUPERAÇÃO DE INFORMAÇÃO QUÍMICA EM PATENTES

Timóteo

2016

João Marcos Martins da Costa Cota

RECUPERAÇÃO DE INFORMAÇÃO QUÍMICA EM PATENTES

Monografia apresentada à Coordenação de Engenharia de Computação do Campus Timóteo do Centro Federal de Educação Tecnológica de Minas Gerais para obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Leonardo Lacerda Alves

Timóteo

2016

João Marcos Martins da Costa Cota

RECUPERAÇÃO DE INFORMAÇÃO QUÍMICA EM PATENTES

Monografia apresentada à Coordenação de Engenharia de Computação do Campus Timóteo do Centro Federal de Educação Tecnológica de Minas Gerais para obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Leonardo Lacerda Alves

Trabalho aprovado. Timóteo, 14 de março de 2017:

Leonardo Lacerda Alves(Orientador)

Marcelo de Sousa Balbino

Odilon Corrêa da Silva

Timóteo, 14 de março de 2017

.

.

.

.

.

Dedico a
algumas pessoas.

Agradecimentos

A realização dessa pesquisa e a participação desse curso, contei com ajuda de muitos, deixo aqui meus agradecimentos:

- ao meu orientador, Leonardo, por ter me dado a oportunidade de realizar este trabalho.
- ao CEFETMG pela estrutura, ensino e tempo que estive estudando aqui.
- ao Maurílio pelo apoio e ajuda.
- aos meus Pais, João Anastácio e Maria da Conceição, por me darem oportunidade de chegar até aqui. Aos meus irmãos Mateus e Paulo por terem me apoiado.
- aos servidores e técnicos do CEFETMG.
- aos outros que não foram citados aqui, agradeço pessoalmente.

Resumo

Tendo em vista o grande numero de patentes existes, observando também um crescimento do registro de patentes. Existe uma demanda por encontrar informações relevantes nessas patentes. Há também uma descentralização dessas patentes, devido a cada pais ou bloco econômico possuir seu próprio órgão regulador. Devido a esses pontos apresentados, observa uma necessidade de centralizar esses bancos de patentes a fim de obter informações relevantes principalmente a área da química. Dessa forma essa pesquisa buscou um meio de desenvolver uma prova de conceito para poder buscar, centralizar e analisar informações químicas em patentes. Essa prova de conceito consistiu no desenvolvimento de um protótipo que realizasse buscas e tratamento de informações químicas em bancos de patentes públicos existentes.

Palavras-chave: recuperação de informação, centralização de informação, recuperação de informações químicas, integração de serviços.

Abstract

Considering the growing number of registered patents, also the large number of existent patents. Has observed a demand to retrieve relevant informations in patents. According to these points,

Keywords: information retrieval, information federation, chemical information retrieval, integration of services.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

São úteis formas padronizadas para a descrição e recuperação de informações diversas, como de fórmulas químicas (SUN et al., 2007), de citações bibliográficas (GILES; BOLLACKER; LAWRENCE, 1998), de informação corporativa (HU; SVENSSON, 2010) e de repositórios de patentes (CHEN; CHIU, 2011). Porém, repositórios de patentes são distribuídos entre vários países e há pequena interface entre eles. Para o usuário, existe então a necessidade de implementar uma forma que facilite que as informações sejam tratadas.

De acordo com Pasche et al. (2014), existem cerca de 50 milhões de patentes espalhadas no mundo. Então, surge a necessidade de se criar meios que possibilitem usuários realizarem buscas nesses bancos de patentes. Esses bancos muitas vezes variam suas formas de acesso e localização, já que cada pais ou federação comercial tem o seu próprio. Quando uma patente possui detalhes muito específicos (e.g.: informação química), a recuperação se torna ainda mais difícil pela inexistência da padronização desses detalhes.

Atualmente os usuários da área de química fazem sua busca de forma manual, entrando em contato com os diversos bancos de patentes existentes, muitas vezes sem ter um conhecimento de como otimizar suas buscas. Esse processo pode apresentar certa imprecisão ao obter os resultados, visto que pode se perder em buscas passadas, além de esbarrar em problemas de acesso a cada banco existente. A proposta deste trabalho então é realizar toda essa busca através de uma ferramenta que proponha integração continua entre essas bases, fazendo as comparações e trazendo os resultados de acordo com o que usuário deseja, garantindo uma gestão eficaz dos recursos.

Além disso, recuperar informações químicas sobre patentes é possível pela recuperação de patentes, sem fazer qualquer reconhecimento especial de características químicas presentes nos documentos, o que limita a capacidade dos usuários. Portanto, a proposta deste trabalho também é adaptar os serviços existentes para federar esses diversos bancos de patentes para que atendam melhor as necessidades de recuperar informação química por usuários desse serviço.

1.1 Justificativa

A existência dessa pesquisa se faz pela necessidade de implementação de uma infraestrutura de dados químicos para auxiliar a recuperação de informações químicas, e que, além disso, possa dar suporte adequado às tarefas de usuários da área de química. Para tal objetivo, deve-se buscar um entendimento mais amplo das informações químicas que são relacionadas a patentes, para então possibilitar uma evolução que não limite a um cenário reduzido.

Um facilitador para esse processo todo é a existência de uma padronização universal do formato dos metadados das patentes, tanto para área de química como para as demais áreas.

Entretanto, ainda assim é necessário a criação de uma infraestrutura de dados químicos. Essa infraestrutura é um arcabouço tecnológico e metodológico para lidar com os serviços que a infraestrutura fornece. Cada usuário pode consumir os serviços a partir dessa infraestrutura para se obter os dados necessários para a tomada de decisão sobre patentes.

As consultas para obtenção desses dados devem ser feitas com base em vários metadados de patentes, como códigos internacionais, titulo, co-citações e também por termos relacionados, que podem estar presentes nas descrições dessas patentes.

Para isso temos a necessidade de uma infraestrutura de dados que organize essas informações que estão dispersas em vários bancos de patentes. Essa desintegração abre espaço para esse trabalho, pois com a dispersão desses dados faz-se necessária uma federação de serviços que ajude o usuário a pesquisar em diversas bases de patentes e obtenha um único resultado que melhor lhe satisfaça.

Porém, a federação de serviços dos vários repositórios não é simples. É necessário estudar a melhor forma de se criar uma ferramenta que organize, federe, e reduza os obstáculos de uso dessas informações, tais como repetir buscas em diversas bases de patentes, ordenação imprecisa e inconsistente de resultados dos vários serviços, dentre outros.

1.2 Problema

As seguintes questões constituem o problema desta pesquisa: Quais são as classes de serviços que suportam o intercâmbio de informações químicas em patentes? Quais as métricas melhor representam o desempenho desses serviços na recuperação de informações químicas em patentes?

As características da informação química são baseadas na forma como esses dados são armazenados pelos diversos serviços, bem como a estrutura do composto químico e sua representação de forma mais legível para um computador. As classes de serviços que promovem isso estão relacionadas a plataformas de federação dos dados e referem-se aos diferentes grupos de serviços que executam trabalhos parecidos, tais como conversão, recuperação, visualização, etc.

Serviços para recuperação dessas informações são programas ou sistemas que podem ser acessados por meio de computador conectado a *internet* (SOUZA et al., 2006), portanto são aplicações *web*.

1.3 Objetivos

Para responder ao problema proposto, o objetivo geral deste trabalho é propor um conjunto de serviços que permitem melhorar o intercâmbio de informações químicas de patentes.

Ainda, o trabalho possui objetivos mais específicos como:

1. Propor um conjunto de serviços para federar diversos bancos de patentes;
2. Especificar serviços para agilizar e melhorar a busca, a recuperação e a visualização de informações químicas em patentes;
3. Desenvolver um protótipo de ferramenta que sirva como prova de conceito e possibilite ao usuário realizar busca em bancos de patentes;
4. Analisar as implicações desses serviços no desempenho dos sistemas automáticos de recuperação de informação químicas em patentes.

1.4 Estrutura do trabalho

Essa pesquisa possui uma estrutura em cinco capítulos, onde eles podem ser resumidos em:

- Capítulo atual de introdução;
- O capítulo 2 apresenta os procedimentos metodológicos necessários para desenvolvimento desse trabalho, incluindo etapas para o projeto do protótipo e estratégias de validação.
- As bases teóricas são apresentadas no capítulo 3.
- No capítulo 4 é detalhada uma avaliação do protótipo da ferramenta de federação. São então avaliadas a utilidade e a eficiência de um modelo de recuperação baseado na recuperação de patentes em dois serviços já existentes. Dois serviços de recuperação de patentes são escolhidos e analisados para identificar campos comuns entre eles para propor uma ferramenta que agrupe as consultas que serão executadas nos dois.
- Finalmente, no capítulo 5 são analisados os resultados e apresentadas as conclusões e considerações finais, principais contribuições, limitações e indicadas algumas direções para trabalhos futuros.

2 Procedimentos metodológicos

Os procedimentos metodológicos adotados neste trabalhos são organizados em etapas:

1. reunir e estudar trabalhos de recuperação de informação, especialmente aqueles voltados a patentes – parte da pesquisa é bibliográfica e serve para obter conhecimento a respeito de recuperação de informações químicas em patentes;
2. escolher bancos de patentes do domínio publico e replicar um conjunto de patentes para agilizar o processo de busca das informações – sendo que parte dessa pesquisa pode ser classificada como pesquisa documental;
3. projetar e implementar um *software*, como prova de conceito, que federe e indexe a massa de informações de patentes usando estruturas de dados específicas – parte da pesquisa que pode ser classificada como estudo de caso;
4. analisar as implicações que uma federação de serviços produz no desempenho de uma infraestrutura de dados química e avaliar como novos trabalhos podem contribuir para a criação dessa infraestrutura.

2.1 Revisão da literatura

A revisão de literatura consiste num levantamento bibliográfico a respeito de trabalhos relacionados à área de recuperação de informação. Portanto, em âmbito mais geral, espera-se que os trabalhos pesquisados incluam meios de recuperação de informações mais gerais e também trabalhos mais específicos sobre recuperação de informações químicas em patentes.

Tais técnicas de recuperação serão melhor discutidas no capítulo 3, sobre fundamentação teórica e metodológica. As abordagens de interesse relacionam recuperação de informação dentro da área de engenharia de computação. As abordagens de interesse ainda devem ser úteis para a recuperação e formas de classificação da informação química presente em patentes.

2.2 Coleções de patentes

Para desenvolver a prova de conceito deste trabalho serão usados documentos aleatórios da coleção *TREC-Chem* [1], e um pequeno conjunto de dez patentes obtidos do repositório USPTO.

Os documentos e as patentes servem apenas para experimentação do protótipo para a prova de conceito.

[1] *Text REtrieval Conference* para química, que tem por objetivo desenvolver e avaliar tecnologia para pesquisa e recuperação de texto de grande escala em documentos relacionados com química, incluindo trabalhos acadêmicos e patentes.

2.3 Prototipação de um sistema de recuperação de informação químicas em patentes

Na prototipação é proposto um modelo de *software*. Esse então é um protótipo de um sistema de recuperação de informação, implementado em linguagem C# e usando a bibliotecas fornecidas por vários centros de patentes e escritas em varias linguagens. Pretende-se executar as funções de indexar e recuperar informações químicas nas patentes disponíveis. O protótipo é documentado no capítulo 5.

A prototipação se baseia no trabalho de Anastácio (2009), pela adoção de mecanismos comuns de classificação, busca, indexação, coleta e recuperação de documentos e sua adaptação para necessidades especiais. Desta forma para essa monografia os requisitos do protótipo são: federação e indexação da informação, tratamento das buscas e reconhecimento de entidades químicas e relacionadas a patentes. Pela implementação desses requisitos, um sistema de recuperação de informação comum torna-se um sistema de recuperação de informação químicas em patentes.

Os requisitos funcionais do protótipo incluem o tratamento de desenhos químicos, como de moléculas sendo tratada através da incorporação de bibliotecas para tal finalidade; e a federação dos repositórios de patentes através da extração desses documentos e a geração de um índice federado. Os requisitos não-funcionais desse protótipo inclui a recuperação de informação, a indexação, a interface de busca com o usuário e a busca em lote para avaliação de várias expressões de busca em conjunto.

2.4 Avaliação e validação

O objeto de informação deste trabalho é uma prova de conceito, sobre a qual existe pequena capacidade de avaliação e validação.

A experimentação sobre a prova de conceito não é conclusiva neste trabalho e exige trabalho futuro que adote avaliação com usuário, testes de desempenho ou outros procedimentos metodológicos.

Desse modo alguns resultados de avaliação e validação são detalhados no capítulo 5, juntamente com os resultados da avaliação, sua análise e discussões.

3 Fundamentos teóricos e metodológicos

O objetivo deste capítulo é o apresentar a fundamentação teórica necessária para que os principais conceitos relacionados à recuperação de informações químicas em patentes. Além de definir os conceitos, o capítulo também apresenta alguns fundamentos históricos, teóricos e metodológicos sobre o qual esta pesquisa se baseia.

Este capítulo se divide nas seguintes seções:

- A primeira seção faz uma breve apresentação sobre algumas ferramentas existentes e as implicações delas para esta pesquisa.
- A segunda seção faz uma breve apresentação sobre a recuperação e o tratamento das informações federadas.
- A terceira seção trata de estudos sobre como a informação química em patentes é recuperada e analisada.
- A quarta seção discute como os repositórios de patentes são organizados.
- A quinta seção apresenta os fundamentos sobre as operações de serviços web (WS) e os códigos de operação que podem ocorrer durante a utilização de uma possível infraestrutura de dados químicos.

3.1 Ferramentas computacionais

De acordo com WIPO (2016a), a *World Intellectual Property Organization* (WIPO) desenvolveu um sistema de busca por termos químicos em patentes, que pode ser acessado pela ferramenta online PATENTSCOPE e permite fazer buscas usando as notações INCHI, Smiles, termos IUPAC, dentre outras notações. A ferramenta também permite analisar imagens com fórmulas e até mesmo desenhar as fórmulas químicas.

Até o momento dessa pesquisa, a ferramenta PATENTSCOPE era gratuita, possivelmente por estar em estágio de testes ainda. Porém, no próprio site PATENTSCOPE, existem seções que são pagas e essa pesquisa por termos químicos usa de ferramentas proprietárias para desenhos das moléculas. É o caso do ICedit que foi abordado por Schmid (2010). É necessário um registro para que se possa fazer uso do serviço.

Gratuitas ou pagas, não parece haver uma interface de acesso remoto (como uma API) nos serviços de busca por patente (WIPO, 2016a). A ferramenta estava disponível durante a realização desta pesquisa e portanto ela serviu de referência para conhecer os requisitos de uma ferramenta desse tipo.

Durante o desenvolvimento desta pesquisa, a WIPO remodelou o seu próprio site, e a pesquisa relacionada a quimicos e patentes passou a estar disponível mediante uma taxa pelo uso do serviço. Dessa forma, a infraestrutura proposta neste trabalho também passou a se justificar pela possibilidade de se criar serviços gratuitos na Internet.

Muitos dos serviços e das ferramentas disponíveis na Internet fazem uso da Chemistry Development Kit (CDK) (STEINBECK et al., 2003). Essa biblioteca possui diversas funções como tradução de termos entre diferentes notações (inchi, iupac, smile, etc), possibilidade de desenhar moléculas quimicas e conversão de imagens para compostos químicos. A biblioteca é bem completa, com uma grande comunidade de desenvolvimento e implementada na linguagem JAVA (GOSLING et al., 2014).

3.2 Recuperação e tratamento das informações

Com a evolução da Internet, as pessoas passaram a obter informações em fontes baseadas na Internet. Isso ocorreu devido a eficácia do meio, além da maior difusão da Internet, tornando-a então um meio padrão e preferido para busca de informações.

Assim, o campo de recuperação de informação também passou a investigar os documentos e serviços web como fonte de informação (MANNING et al., 2008). A Recuperação e tratamento de informações têm como objetivo fornecer acesso e tratamento a dados estruturados e não-estruturados e não começou com a web. O campo evoluiu de várias fontes e métodos de recuperação. Na Web, inicialmente, começou com publicações científicos, publicações em bibliotecas e páginas web simples, evoluindo e se espalhando por outras formas de conteúdo, tais como notícias, documentos jurídicos, documentos médicos e outros (MANNING et al., 2008).

Na organização e recuperação automatizada de informação, os modelos de busca clássicos que foram estudados por Pasche et al. (2014), tais como, como máquinas de pesquisa booleanas e probabilísticas, são comumente usadas em sistemas automatizados de recuperação. Porém, no contexto de patentes não são modelos muito eficientes, pois exigem muito esforço computacional e são imprecisos, não sendo viáveis para uma organizaçào de patentes.

Experimentalmente, Pasche et al. (2014) propuseram um re-ranqueamento baseado nas citações presentes nas patentes e obtiveram melhores resultados. Ao normalizar dados comuns, tais como termos e tipo de arquivos, notou-se uma melhora nos resultados das buscas. Uma análise feita com cerca de 13 milhões de patentes armazenadas em Oracle DB, pouco mais de 1 milhão foram extraídas usando *query* em SQL e armazenadas em arquivos XML, consumindo cerca de 33 dias. Então os dados foram armazenado em sistema de arquivos Hadoop.

Primeiramente, os dados foram normalizados, sendo então tratados usando recuperação de informações de Terrier, que é um ranqueamento baseado nos métodos do framework Lucene. Os resultados então foram dispostos em uma matriz, ranqueados e exibidos ao usuário. Quando Pasche et al. (2014) resolveram avaliar os impactos em performance nos resultados, notou-se que a descrição não melhora a performance. Portanto, eles removeram esse campo das patentes analisadas.

Três tipos de classificação (PASCHE et al., 2012) são mais conhecidas. A primeira é usada para avaliar buscas usando a mesma metodologia do *TREC-Chem*, constituído por tópicos como titulo, resumo e autoria da patente. A segunda é uma forma empírica de pesquisa, onde as buscas são feitas com base em palavras-chaves. E a terceira, que é uma forma de busca também empírica, baseia-se em uma única patente-alvo, por meio da qual outras parecidas e compatíveis são procuradas.

Pasche et al. (2014) também utilizaram também o sistema de *Top-Precision* (P0) que obteve o melhor resultado dentre outros métodos de ordenação. Os autores também reali-

zaram uma normalização baseada em ontologias do conteúdo de patente, termos e identificadores que são armazenados em metadados.

Também foram realizados testes em dois métodos de busca, utilizando algoritmos BM25 e PL2, sendo o segundo baseado em estimativa de Poisson de aleatoriedade. Após essa avaliação, os autores utilizaram um re-ranqueamento usando redes de citação para melhorar os resultados, com as patentes mais citadas. Os autores também avaliaram o uso de classes IPC, a classificação internacional para patentes, para verificar se elas trazem melhorias nos resultados.

Pasche et al. (2014) concluíram que o campo de descrição não melhora os resultados, mostrando que em alguns casos o campo chegou a diminuir a precisão de alguns resultados. Ao retirar esse campo, a indexação ficou mais rápida e mais eficiente.

Também foi observado que o uso de metadados ajudou a melhorar a precisão dos resultados. Ao utilizar o algoritmo BM25 se conseguiu obter um desempenho maior, significando tempo menor para retorno das buscas. O ranqueamento pelas citações melhorou significativamente os resultados. E também códigos IPC permitiram melhorar a busca por patentes.

3.3 Análise de informação química em patentes

Para formular buscas de patentes, devem ser usados detalhes das patentes que podem estar ou não presentes nos metadados dos documentos. Um detalhe que pode ser importante em algumas patentes é a composição química e o processo químico adotados na invenção.

Além disso, o objetivo dessa pesquisa é caracterizar a informação química em patentes e então permitir a recuperação dessas informações, explorando as características das informações relacionadas a patentes e relacionando-as com as informações químicas, permitindo que patentes sejam recuperadas com maior precisão.

A informação química existe em diversas fontes. Um exemplo é dado por Pence e Williams (2010) com o Chemispider. Esse é um grande tesauro de informações quimicas diversas, porém ele não faz relação com bancos de patentes, de forma que ao buscar determinado composto químico na base do Chemispider, só obtém como resposta dados relativos ao composto escolhido.

Outro exemplo já citado é o PATENTSCOPE. Mais sofisticado, esse possui uma ferramenta de busca por termos químicos em patentes, porém ainda de forma experimental e disponível somente mediante pagamento.

Como mencionado por Alberts et al. (2011), que citam sobre os desafios encontrados na recuperação de informações de patentes no capitulo 5, a recuperação de patentes é mais complexa que uma recuperação tradicional pois trata-se de uma especialização da recuperação de informações. As informações de patentes precisam de uma análise adicional, uma ordenação específica e, quando está relacionada a estruturas químicas presentes em patentes, é ainda mais especializada.

Com relação aos meios de armazenar, tratar as informações dessas patentes e retornar análises importantes para o campo da química, há estudos sobre a evolução dos bancos de dados químicos. Ao longo dos anos, os bancos de dados especializados deixaram de ser algo utilizado apenas para organizar informação, passando a ser uma poderosa ferramenta para descobrir novos compostos, como foi observado por Miller (2002). Os bancos de dados

e bancos de patentes são públicos e servem para encontrar compostos patenteados e de domínio público, mas também são úteis para indicar novos meios para pesquisa e descoberta de novos compostos e então criar novas patentes.

Muitos desses bancos comumente usam a notação SMILES (WEININGER, 1988) como notação padrão das informações químicas, já que estão presentes muitos atributos relacionados a rotação, posição, quiral, além de outros atributos relacionados à formação do composto químico. A informação armazenada de acordo com a estrutura é de interesse também, pois a forma e ordem do composto podem ser armazenadas de duas formas.

A primeira forma, onde se guarda o tipo de conectividade entre os átomos do elemento, com ou sem modelos bidimensionais. E a segunda forma que também guardam dados tridimensional sobre a molécula.

Três entidades químicas devem ser armazenadas para uma adequada indexação química:

> One consideration is which chemical entities are to be tracked. The three basic entities are: compound (a generalization of the chemical structure, which generally corresponds to the synthetic chemist's view that was discussed above); form (the compound plus counterion(s), solvent(s) and isotopically labelled atoms, as in the curator's view); and lot (a sample of the compound, which is often termed 'batch') (MILLER, 2002, p. 46).

Uma maneira de buscar compostos químicos é por meio das estruturas químicas. Elas podem ser desenhadas por meio de um programa computacional, tal como *ISIS/Draw*, *ChemDraw* ou *ChemSketch*. Então toda vez que é sintetizado um novo composto, esse é buscado no banco de dados. A estrutura molecular também pode ser confundível sendo que compostos diferentes podem apresentar semelhanças entre si, caso venham a ser alinhados seguindo uma certa orientação, para tal existe uma orientação comum (MANNING et al., 2008). Para melhorar as buscas são gravados apenas compostos discretos, que são compostos completos de forma a serem amostras puras, o que evita ambiguidades. E desse modo cria-se bibliotecas que conseguem ler e gerar compostos com base nesses compostos já existentes (KRALLINGER et al., 2013).

Uma forma comum de realizar buscas por compostos é pelo termo exato do composto. Essa é a forma mais simples, rápida e computacionalmente barata de se buscar. Outro tipo de busca é por subestrutura bidimensional, onde o usuário pode desenhar a estrutura. Há algumas questões com respeito a estequiometria, aromaticidade e as limitações de representação atômica em um pacote de software especial, que poderia levar a resultados anômalos para o usuário que executa uma pesquisa por subestrutura, mas essas geralmente podem ser resolvidas por um estudo mais minucioso do pacote de software usado (MILLER, 2002).

Quando são feitas buscas por estruturas tridimensionais, o usuário pode inserir mais dados na hora da busca. Assim como é feito no modelo bidimensional, é feita uma pretriagem do dado que será buscado para que seja normalizado. Mesmo assim, podem ocorrer ambiguidades com outras estruturas, embora o resultado ainda seja mais refinado. Existem também outros tipo de busca, como busca por similaridade onde deseja-se buscar elementos parecidos e busca por propriedades similares, onde se busca por propriedades similares entre compostos químicos (MANNING et al., 2008).

Na recuperação de informação, também são de interesse os métodos para refino das buscas. Para casos onde existe muito a ser analisado, pode-se usar ambiente computacionais em *cluster* (BUYYA, 1999). Outro método também usado é análise intelectual. Esse último é

um método demorado, mas necessário para casos de que os resultados apresentam muitos erros (MILLER, 2002).

Existem fornecedores que oferecem softwares básicos de registro, busca, visualização e análise de estruturas químicas. A diferença entre eles se dá pela utilidade das buscas de banco de dados químicos que contribuem para a descoberta de elementos. Outras vezes as diferenças dependem de fatores externos à capacidade de executar as buscas (INFOCHEM, 2017).

Primeiramente os usuários devem ter a capacidade de realizar buscas sobre estruturas químicas no contexto dos dados relacionados, incluindo propriedades químicas, atividade biológica e dados estruturais de proteína, por exemplo. Os usuários também devem ter a capacidade de procurar os resultados no contexto de informações de buscas relacionadas, a fim de obter respostas adequadas a partir dos dados. As especificidades destes requisitos variam de organização para organização, mas a adaptabilidade e a possibilidade de integrá-lo a outras ferramentas analíticas e de visualização são cruciais (MILLER, 2002).

3.4 Repositórios de patentes

A necessidade de federação das fontes de dados de patentes se justifica pela dispersão das patentes no mundo. Há cerca de 300 escritórios de patentes espalhados pelo mundo, e outros 10 escritórios regionais que organizam patentes de um bloco de países (WIPO, 2016b).

Algum desses escritórios fornecem meios de acessarem suas patentes, seja por uma busca no próprio sistema de informação ou fazendo uso a algum tipo de *Application Programming Interface* (API). Porém, não existe uma forma centralizada de acesso a essas informações, o que corresponderia a uma federação de serviços e repositórios de patentes (MASIAKOWSKI; WANG, 2013).

Por outro lado, mesmo federados, os dados ainda precisam ser adequadamente organizados para favorecer a recuperação por usuários de patentes. Para essa pesquisa foram selecionados dois repositórios de patentes a fim de criar um protótipo de uma ferramenta com funcionalidades mínimas para que em trabalhos futuros possa ser expandida para mais repositórios.

Esses repositórios de patentes iniciais são os repositórios da União Europeia e do Estados Unidos. Podendo futuramente ser adicionado outros tantos como o repositório chines, o brasileiro Instituto Nacional de Propriedade Industrial (INPI), dentre outros.

3.4.1 *United States Patent and Trademark Office - USPTO*

O repositório de patente do Estados Unidos é aberto, possuindo diversas formas de acessar seus serviços web, com diversos tipos de agrupamentos, seja por patentes, por datas ou até mesmo reivindicações. O nome do serviço de busca de patentes é *PatentsView* (USPTO, 2016).

O serviço de exposição dos dados é baseado em *rest API*, sendo compatível com as tecnologias atuais. Os dados são passados do cliente para o serviço e tem suas respostas no formato *Json*, sem ambiguidades além de requerer pequena largura de banda.

A API possui uma linguagem de busca que proporciona pouca ambiguidade, permitindo assim efetuar comparações de vários termos, bem como datas, pesquisar palavras em frases, e pesquisar por todos os campos presentes em uma patente. Essa consulta é feita

somente por *json*, onde a estrutura deve obedecer o padrão existente no *json* e a hierarquia criada pelo serviço.

3.4.2 *European Patent Office - EPO*

O escritório europeu de patentes possui quatro serviços disponíveis para consulta, sendo que dois deles são para consultas não automatizadas: *Espacenet* e *Open Patent Services (OPS)*. Os serviços oferecem acesso gratuito a cerca de 90 milhões de patentes (OFFICE, 2015). O serviço que foi adotado nesta pesquisa foi o *Open Patent Services (OPS)*.

3.5 Operações de um Serviço Web

Um serviço web (WS, de *web service*) é um conjunto de métodos que podem ser consumidos por programas independentemente da linguagem de programação e da plataforma que permite a execução dos métodos. Aos serviços web importam apenas os protocolos de comunicação e de troca de dados (SOUZA et al., 2006).

As codificações são independentes de operações, completas e acessíveis para a maioria das linguagens atuais. Um WS suporta uma combinação das solicitações, tal como, passar dados por XML e solicitar JSON de retorno. Desde que essas informações sejam solicitadas corretamente (LEE, 2011).

Para que alguém consiga acesso à API, é necessário que esteja em conformidade com o tipo de codificação e dados suportados pelo servidor de acordo com a padronização já existente. No caso do protótipo deste trabalho, a padronização está descrita na seção 3.5.1. Também é necessário respeitar a sintaxe aceita por XML e JSON (LEE, 2011).

3.5.1 Códigos de operação

Nos serviços web projetados neste trabalho, os códigos de operação e erros são baseados nos erros do padrão HTTP (FIELDING et al., 1999). Eles são explicados e descritos na tabela 1.

Tabela 1: Descrição códigos *HTTP*.

Código	**Nome**	**Descrição**
0	Unreachable	Algo impede a interface de conectar com api
200	Success	Sucesso na operação
400	Bad Request	Requisição efetuada de forma não compatível
404	Not Found	Recurso solicitado não foi encontrado
500	Internal Server Error	Quando ocorre algum erro interno no servidor ou conversão incompatível
503	Service Unavailable	Serviço indisponível no momento

Como exibido na tabela 1, o código de erro 0 indica quando não ha conexão entre a aplicação cliente e a API, seja por falta de internet, seja por algum fator no meio que impede o cliente de encontrar a api. O código 400 indica uma requisição de forma incompatível, por

exemplo, indicar para o servidor que os dados que serão passados estão no formato XML, porém enviar um JSON. O código 404 indica que o recurso solicitado não foi encontrado, um exemplo disso seria efetuar uma chamada de uma api que não existe, ou um caminho com o nome errado dentro do escopo do serviço. Para os códigos de erro 500 e 503, esses indicam que ocorreu um erro no processamento de alguma requisição, como falha na conversão e até mesmo erro interno no serviço de hospedagem (BERNERS-LEE; FIELDING; FRYSTYK, 1996).

Para o código de resposta 200, esse indica sucesso na requisição. Mesmo que a requisição tenha retornado nenhuma resposta, ele ainda indica sucesso na comunicação HTTP.

Existem outros códigos de resposta HTTP não apresentados na tabela, mas que não são objeto de estudo e avaliação neste trabalho.

4 Desenvolvimento e Implementação

Esse capítulo apresenta os detalhes do projeto e da implementação da infraestrutura de dados químicos para recuperação de informação em patentes, proposta nesta pesquisa. O capítulo explica detalhes técnicos, usos de linguagens, exemplos de chamadas à API e definição do Serviço Web.

Relacionado aos aspectos de funcionamento da infraestrutura de federação, são propostos métodos que compõem os serviços web.

4.1 Elementos básicos do serviço web

De início, são descritos aspectos do comportamento do Serviço Web (WS).

Dessa forma o serviço web possui independência das operações, ou seja, os métodos expostos para um programa cliente conseguem por si só obter uma resposta.

4.1.1 Especificação do serviços e seus metadados

Um serviço web está de acordo com as definições e seus metadados de serviço. Um servidor pode suportar várias versões.

Caso o servidor receba uma solicitação com um número de versão que ele não suporta, o servidor deve gerar uma exceção *BadRequest*. Para tal, as solicitações devem estar de acordo com as definições a seguir.

Estas definições definem dois métodos de codificação de pedidos ao WS. O primeiro usa XML como linguagem de codificação. A segunda codificação usa pares de chave-valor (JSON) para codificar os vários parâmetros de um pedido.

Um exemplo desse par de chave-valor é

```
1 "glossary": {"title": "example glossary" }
```

onde “glossary” é uma chave “{"title": "example glossary"}” é um valor para a chave.

A tabela a seguir relaciona os metodos do WS com seus respectivos tipos de requisição e retorno.

Tabela 2: Métodos padrão da API.

Operação	Codificação de Solicitação	Codificação de Retorno
InchiToSmile	XML ou JSON	XML ou JSON
InchiToMol	XML ou JSON	XML ou JSON
InchiToInchiKey	XML ou JSON	XML ou JSON
SmileToInchi	XML ou JSON	XML ou JSON
SmileToMol	XML ou JSON	XML ou JSON
SmileToInchiKey	XML ou JSON	XML ou JSON
MolToSmile	XML ou JSON	XML ou JSON
MolToInchi	XML ou JSON	XML ou JSON

Continua na próxima página

Tabela 2 – Continuação da página anterior

Operação	Codificação de Solicitação	Codificação de Retorno
MolToInchiKey	XML ou JSON	XML ou JSON
MolToImage	XML ou JSON	XML ou JSON
InchiToImage	XML ou JSON	XML ou JSON
SmileToImage	XML ou JSON	XML ou JSON
InchiKeyToImage	XML ou JSON	XML ou JSON
ConvertFromTo	XML ou JSON	XML ou JSON
FromValues	-	XML ou JSON
ToValues	-	XML ou JSON
GetFullPatent	XML ou JSON	XML ou JSON
FetchByFields	XML ou JSON	XML ou JSON

A Tabela 2 mostra as assinaturas de um conjunto de métodos preliminares desta pesquisa, outros poderão ser adicionados futuramente. Na tabela 2 são descritos os tipos de codificação aceitos pelas entradas e a codificação retornada pelos métodos. Note que os métodos que aceitam um parâmetro de entrada, esses parâmetros devem ser passados no formato XML ou JSON. Os métodos que não precisam de parâmetros não têm essa informação. Na seção 4.2.1 são mostrados detalhes desses métodos, como as chaves e os tipos de dados suportados por cada chave.

4.2 Elementos comuns

Para cada instância do WS, deve ser atribuído um identificador único que é atribuído pelo servidor quando o recurso é criado. Esse identificador atua sobre todas operações que serão executadas pelos servidor e ao final o identificador é apagado, o que significa que um identificador de recurso não pode ser reutilizado depois de ter sido atribuído. Identificadores de recursos não se destinam a associar recursos WS com objetos do mundo real e os valores não têm que ser significativos fora do âmbito de uma instância do WS.

O parâmetro para a função deve ser o nome de uma propriedade e o seu valor. O valor pode ser simplesmente um texto ou uma lista de elementos que são os valores da propriedade nomeada. A função deve resolver todos os recursos referenciados localmente e, caso necessário, o serviço pode resolver referências remotas. No caso em que o servidor só suporta recursos referenciados localmente, e ele encontra um recurso remotamente referenciado, o servidor deve gerar uma exceção Unreacheable.

4.2.1 Métodos de acesso e seus parâmetros

A tabela 3 mostra os métodos, as suas chaves e os valores suportados. Os métodos têm todos os seus acessos feitos por *POST* para maior segurança dos dados.

Tabela 3: Descrição dos Métodos.

Tipo Retorno	Operação	Chaves de Parâmetros	Tipos de Parâmetros
String	InchiToSmile	term	String
Continua na próxima página			

Tabela 3 – Continuação da página anterior

Retorno	Operação	Chaves de Parâmetros	Chaves de Parâmetros
String	InchiToMol	term	String
String	InchiToInchiKey	term	String
String	SmileToInchi	term	String
String	SmileToMol	term	String
String	SmileToInchiKey	term	String
String	MolToSmile	term	String
String	MolToInchi	term	String
String	MolToInchiKey	term	String
String	MolToImage	term	String
String	InchiToImage	term	String
String	SmileToImage	term	String
String	InchiKeyToImage	term	String
String	ConvertFromTo	from to term	String String String
Map<String, Long>	FromValues	-	-
Map<String, Long>	ToValues	-	-
Patent	GetFullPatent	Patente	Patent
List<Patent>	FetchByFields	fields ordernation limit page search	String String Long Long String

O capítulo descreve o funcionamento das duas classes principais que constituem o funcionamento da infraestrutura desta pesquisa.

A classe de conversão é demonstrada pelo diagrama de classes da Fig. 1. Essa classe é responsável pelas conversões das diversas notações químicas existentes e previstas. Os métodos de conversão são métodos simples que recebem uma *string* com o termo, que é uma notação química a ser convertida, e retorna uma outra *string* com o termo convertido. Para métodos que convertem notação química em imagem, o retorno também será uma string no formato *BASE64* (JOSEFSSON, 2006) [1].

Para os métodos de conversão avançada, *ConvertFromTo* precisa da passagem de dois parâmetros, *from* que é o tipo do termo que será convertido e *to* que será o retorno da conversão, além do *term* que é o termo propriamente dito para conversão. Os valores que devem ser passados para esse método devem ser obtidos usando as funções *FromValues* para os valores de entrada e *ToValues* para os valores de retorno, que retornam uma listagem composta por chave e valor de cada tipo de conversão, e deverá ser fornecido o valor para a função *ConvertFromTo*. Esses valores podem ser combinados de formas variadas.

[1] *BASE64* de acordo com RFC 3548 tem como objetivo estabelecer um alfabeto comum em codificação de forma a reduzir a ambiguidade, levando a uma melhor interoperabilidade.

Figura 1: Diagrama de classes para descrever a API de conversão.

Fonte: Criação do Autor

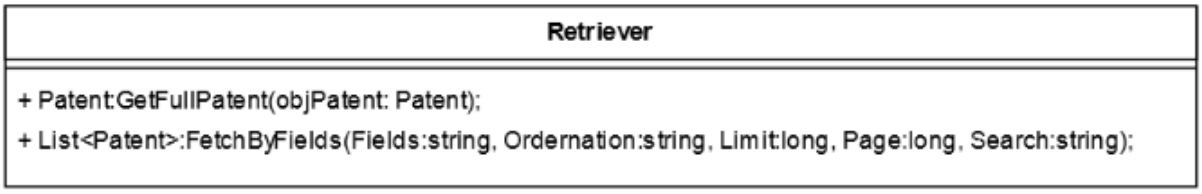
Para ambas as classes de conversão, caso sejam fornecidos valores incompatíveis, é fornecido um erro de *BadRequest*. Caso os valores fornecidos não possam ser convertidos é fornecido um erro de *InternalServerError*.

A classe de recuperação é demonstrada pelo diagrama de classes da Fig. 2. Ela possui método que recupera a patente completa *GetFullPatent*, inclusive com sua descrição, e o método *FetchByFields* que retorna lista de patentes.

Essa classe possui uma peculiaridade, visto que o tempo de resposta à chamada dos métodos dela pode ser alto dependendo dos tamanhos de retorno. O método *GetFullPatent* retornará todo o conteúdo de uma patente, inclusive o texto completo de sua descrição, além da URL dessa patente.

Os métodos que retornam lista de dados deverão ser invocados cautelosamente para evitar que estoure o tempo limite do cliente. O método *FetchByFields* tem um retorno massivo de dados, podendo o valor de *Limit* ser um número grande, esse é responsável pela quantidade de registros de retorno. Esse valor pode ser reduzido quando solicitado em paginas, onde *Page* é o campo responsável por essa paginação, onde para cada página será retornada a quantidade de registros informada pelo limite. Esses método não retorna o texto

Figura 2: Diagrama de classes para descrever a API de recuperação de patentes.



Fonte: Criação do Autor

completo da patente, só os dados básicos, resumo, título e a URL da patente.

Para o método *FetchByFields*, existe a necessidade de informar uma *string* contendo os campos de retorno das patentes e o campo de ordenação. Esses campos são preenchidos com os números relacionados aos campos das patentes de acordo com WIPO (2013), sendo separados por virgulas. Caso o campo de *fields* não seja preenchido, ocasiona em um erro de *BadRequest*. Caso o campo de *ordering* não seja preenchido, é considerado a ordenação padrão por data de publicação da patente.

Para o campo *search* é definida uma linguagem de busca, que é tratada mais adiante na seção 4.3. Se esse campo não for preenchido é informado um erro de *Unreacheable*, indicando que o recurso não pode ser alcançado com a busca solicitada.

Em relação a classe *patent*, ela possui uma abstração de uma patente real, onde os seus campos foram separados na ordem que aparecem nas numerações de descrição disponíveis em WIPO (2013). Os campos podem ser divididos em:

- (10) Relatório a identificação da patente;
- (20) dados relativos ao pedido de uma patente;
- (30) dados relativos a prioridade no âmbito da Convenção de Paris ou do Acordo sobre os Aspectos dos Direitos de Propriedade Intelectual Relacionados ao Comércio (Acordo TRIPS);
- (40) datas de disponibilização do público;
- (50) informações técnicas;
- (60) referências a outros documentos de patentes nacionais ou domésticos, legalmente ou processualmente conexas, incluindo aplicações inéditas;
- (70) identificação das partes interessadas;
- (80)(90) identificação de dados relacionados a outras convenções internacionais;

Para cada item descrito anteriormente, existem ainda outros itens relacionados. Esses estão todos presentes na classe *patent*, que também inclui o texto completo da patente bem como a sua *URL*.

4.3 A linguagem suportada pelo campo *search*

Para o campo de busca, é utilizada a linguagem de consulta já existente em uma das APIs, presente no repositório USPTO, sem um nome conhecido. Tal linguagem permite a associação de operadores lógicos para efetuar operações de comparação e associação de valores com atributos, de tal forma que é possível efetuar comparações.

Essa Linguagem é baseada em JSON, de tal forma que a sua conversão para tipo de objetos é facilitada. Esse é um dos motivos para escolha dessa linguagem para ser a linguagem da API deste trabalho. Além disso, ela é compartilhada com uma das API, o que reduz o trabalho de conversão.

Essa linguagem apresenta grande vantagens, como a existência de vários operadores lógicos e relacionais de tal forma que evita ambiguidade, além de ser de fácil escrita e compreensão. Mais detalhes da linguagem são explicados a seguir. Esses detalhes incluem configuração de chave e valor e critérios de busca.

Os critérios para busca são definidos da seguinte forma:

```
criterios
    pares
    "_eq" : {par simples}
    "_neq" : {par simples}
    "_gt" : {par simples}
    "_gte" : {par simples}
    "_lt" : {par simples}
    "_lte" : {par simples}
    "_begins" : {par simples}
    "_contains" : {par simples}
    "_text_all" : {par simples}
    "_text_any" : {par simples}
    "_text_phrase" : {par simples}
    "_not" : {crit rio}
    "_and" : [{crit rio}, ...]
    "_or" : [{crit rio}, ...]
par
    par simples
    "campo" : [valor, ...]
par simples
    "campo" : valor
```

De forma que as buscas simples, como um exemplo que pesquisa pela patente de número 7861317, podem ser feitas seguindo o padrão:

```
{
    "patent_number":"7861317"
}
```

Buscas com união de campos, onde busca patentes com sobrenome do autor, como Withney, e possua e frase *coton gin* ou o sobrenome do autor seja Hopper e o contenha no texto a palabra COBOL:

```
{
```

```
    "_or": [
        {
            "_and": [
                {
                    "inventor_last_name": "Whitney"
                },
                {
                    "_text_phrase": {
                        "patent_title": "cotton gin"
                    }
                }
            ]
        },
        {
            "_and": [
                {
                    "inventor_last_name": "Hopper"
                },
                {
                    "_text_all": {
                        "patent_title": "COBOL"
                    }
                }
            ]
        }
    ]
}
```

Os exemplos mostram o emprego de operadores de comparação, que comparam inteiros, decimais, datas e campos de texto. Os campos de data seguem o padrão ISO 8601 (WOLF; WICKSTEED, 1998) *YYYY-MM-DD*. Para esses tipos, pode-se usar os operadores:

- _eq – igual a
- _neq – diferente de
- _gt – maior que
- _gte – maior ou igual a
- _lt – menor que
- _lte – menor ou igual a

Para campos de texto e texto completo, é possível fazer consultas usando:

- _begins – a frase começa com o valor solicitado
- _contains – a frase contem o valor solicitado
- _text_all – o texto contem todas as palavras solicitadas

- _text_any – o texto contem alguma das palavras solicitadas
- _text_phrase – o texto contem o exato valor solicitado

Devido a limitação das API, não foram disponibilizados todos os campos para realização da busca, de forma que os campos disponibilizados são os campos mais comuns entre elas. Esses campos são listados a seguir, com a descrição curta de cada um deles e o tipo de dados que eles recebem.

- appcit_app_number

  ID do aplicativo (emitido pelo USPTO) para aplicação citada pela patente selecionada.

  Tipo de dado: *string*.

- appcit_category

  Entidade que citou um pedido na patente selecionada.

  Tipo de dado: *string*.

- appcit_date

  Data de apresentação citado no pedido da patente.

  Tipo de dado: *date*.

- appcit_kind

  Tipo de aplicação citada na patente.

  Tipo de dado: *string*.

- app_country

  País em que foi apresentado o pedido de patente. Padrão U.S.

  Tipo de dado: *string*.

- app_date

  Data de pedido de patente (data de depósito).

  Tipo de dado: *date*.

- app_number

  Código de aplicação atribuído pela USPTO.

  Tipo de dado: *string*.

- app_type

  Tipo de Aplicação de Patente. 02 a 28 = Aplicativo utilitário, 29, D = Aplicação de projeto, 60 = Aplicação provisória, 90 = Pedido de reexame.

  Tipo de dado: *string*.

- assignee_city

  Cidade do cessionário da patente.

  Tipo de dado: *string*.

- assignee_country

  País do cessionário conforme listado na patente.

  Tipo de dado: *string*.

- assignee_first_name

  Primeiro nome, se for único cessionário.

  Tipo de dado: *string*.

- assignee_first_seen_date

  Primeira data de concessão de patentes para todas as patentes de um destinatário no banco de dados.

  Tipo de dado: *date*.

- assignee_id

  Código único de identificação do cessionário no banco de dados.

  Tipo de dado: *string*.

- assignee_last_name

  Ultimo nome, se for único cessionário.

  Tipo de dado: *string*.

- assignee_last_seen_date

  Data mais recente entre todas as patentes do concessionário presente no banco de dados.

  Tipo de dado: *date*.

- assignee_lastknown_city

  Cidade mais recente do cessionário desde a data de concessão da patente mais recente (relacionado com assignee_last_seen_date).

  Tipo de dado: *string*.

- assignee_lastknown_country

  País mais recente do cessionário desde a data de concessão da patente mais recente (relacionado com assignee_last_seen_date).

  Tipo de dado: *string*.

- assignee_lastknown_location_id

  Código de identificação único da localização do cessionário relacionado a data da ultima publicação (relacionado com assignee_last_seen_date).

  Tipo de dado: *string*.

- assignee_lastknown_state

  Estado mais recente do cessionário desde a data de concessão da patente mais recente (relacionado com assignee_last_seen_date).

  Tipo de dado: *string*.

- assignee_location_id

  Código de identificação único para a localização do cessionário, conforme listado na patente.

  Tipo de dado: *int*.

- assignee_organization

  Nome da organização, se o cessionário for uma organização.

  Tipo de dado: *string*.

- assignee_sequence

  Ordem na qual o cessionário aparece na patente.

  Tipo de dado: *int*.

- assignee_state

  Estado do cessionário na qual aparece na patente.

  Tipo de dado: *string*.

- assignee_total_num_inventors

  Número total de inventores na base de dados para um determinado cessionário (conforme indicado por assignee_id).

  Tipo de dado: *int*.

- assignee_total_num_patents

  Número total de patentes na base de dados para um determinado cessionário (conforme indicado por assignee_id).

  Tipo de dado: *int*.

- assignee_type

  Classificação do cessionário. 2 - Empresa ou Corporação dos EUA, 3 - Empresa ou Corporação Estrangeira, 4 - Pessoa dos EUA, 5 - Pessoa Estrangeira, 6 - Governo dos EUA, 7 - Governo Estrangeiro, 8 - Governo do País, 9 - Governo do Estado. "1"que aparece antes de qualquer um destes códigos significa parte de interesse

  Tipo de dado: *string*.

- cited_patent_category

  Categoria da patente citada.

  Tipo de dado: *string*.

- cited_patent_date

  Data de concessão da patente.

  Tipo de dado: *date*.

- cited_patent_kind

  Tipo de patente da patente citada (relacionado com patent_kind).

  Tipo de dado: *string*.

- cited_patent_number

  Numero de citação da patente.

  Tipo de dado: *string*.

- cited_patent_title

  Titulo da patente citada.

  Tipo de dado: *string*.

- citedby_patent_category

  Categoria de citação da patente.

  Tipo de dado: *string*.

- citedby_patent_date

  Data de concessão que cita a patente selecionada.

  Tipo de dado: *date*.

- citedby_patent_kind

  Tipo de citação da patente (relacionado com patent_kind).

  Tipo de dado: *string*.

- citedby_patent_number

  Numero da citação da patente.

  Tipo de dado: *string*.

- citedby_patent_title

  Titulo da citação da patente.

  Tipo de dado: *string*.

- cpc_category

  Categoria de nível superior da classificação de patentes corporativas CPC.

  Tipo de dado: *string*.

- cpc_first_seen_date

  Data de concessão da primeira patente no banco de dados dentro de uma subseção do CPC.

  Tipo de dado: *date*.

- cpc_group_id

  Código do grupo do CPC.

  Tipo de dado: *string*.

- cpc_group_title

  Titulo do grupo do CPC.

  Tipo de dado: *string*.

- cpc_last_seen_date

  Data de concessão da ultima patente no banco de dados dentro de uma subseção do CPC.

  Tipo de dado: *date*.

- cpc_section_id

  Código da seção do CPC, A = Necessidades Humanas, B = Operações de transporte operacional, C = Química Metalúrgica , D = Papel, Têxtil, E = Construção Civil, F = Engenharia Mecânica, Iluminação, Aquecimento, Armas, Motores, Bombas Hidráulicas, G = Física, H = Eletricidade, Y = Marcação geral de novos desenvolvimentos tecnológicos.

  Tipo de dado: *string*.

- cpc_sequence

  Ordem da classificação CPC na lista de classificações da patente selecionada,

  Tipo de dado: *int*.

- cpc_subgroup_id

  Código do subgrupo do CPC.

  Tipo de dado: *string*.

- cpc_subgroup_title

  Titulo do subgrupo do CPC.

  Tipo de dado: *string*.

- cpc_subsection_id

  Código da subseção do CPC.

  Tipo de dado: *string*.

- cpc_subsection_title

  Titulo da subseção do CPC.

  Tipo de dado: *string*.

- cpc_total_num_assignees

  Número total de cessionários únicos em patentes dentro de uma subseção CPC.

  Tipo de dado: *int*.

- cpc_total_num_inventors

  Número total de inventores únicos em patentes dentro de uma subseção CPC.

  Tipo de dado: *int*.

- cpc_total_num_patents

  Número total de patentes dentro de uma subseção CPC.

  Tipo de dado: *int*.

- govint_contract_award_number

  Número do contrato ou do prêmio conforme relatado na declaração de interesse do governo sobre patentes (se disponível).

  Tipo de dado: *string*.

- govint_org_id

  Código da organização da agência governamental dos EUA relatada na declaração de interesse do governo sobre patentes (se disponível).

  Tipo de dado: *int*.

- govint_org_name

  Nome da organização da agência governamental dos EUA relatada na declaração de interesse do governo sobre patentes (se disponível).

  Tipo de dado: *string*.

- govint_raw_statement

  A declaração completa de interesse do governo, conforme relatado em uma dada patente (se disponível).

  Tipo de dado: *string*.

- inventor_city

  Cidade do inventor conforme listado na patente selecionada.

  Tipo de dado: *string*.

- inventor_country

  País do inventor conforme listado na patente selecionada.

  Tipo de dado: *string*.

- inventor_first_name

  Primeiro nome do inventor;

  Tipo de dado: *string*.

- inventor_first_seen_date

  Primeira data de concessão de patentes entre as demais de um inventor no banco de dados.

  Tipo de dado: *date*.

- inventor_id

  Código de identificação único do inventor;

  Tipo de dado: *string*.

- inventor_last_name

  Sobrenome do inventor;

  Tipo de dado: *string*.

- inventor_last_seen_date

  Ultima data de concessão de patentes entre as demais de um inventor no banco de dados.

  Tipo de dado: *date*.

- inventor_lastknown_city

  Cidade do Inventor a partir de sua data de concessão de patente mais recente (relacionado com inventor_last_seen_date).

  Tipo de dado: *string*.

- inventor_lastknown_country

  País do Inventor a partir de sua data de concessão de patente mais recente (relacionado com inventor_last_seen_date).

  Tipo de dado: *string*.

- inventor_lastknown_location_id

  Código da localização do Inventor a partir de sua data de concessão de patente mais recente (relacionado com inventor_last_seen_date).

  Tipo de dado: *string*.

- inventor_lastknown_state

  Estado do Inventor a partir de sua data de concessão de patente mais recente (relacionado com inventor_last_seen_date).

  Tipo de dado: *string*.

- inventor_location_id

  Código exclusivo para a localização do inventor conforme listado na patente selecionada.

  Tipo de dado: *string*.

- inventor_sequence

  Ordem na qual o inventor está listado na patente.

  Tipo de dado: *int*.

- inventor_state

  Estado do inventor listado na patente selecionada.

  Tipo de dado: *string*.

- inventor_total_num_patents

  Quantidade total de patentes para um determinado inventor (conforme indicado pelo inventor_id).

  Tipo de dado: *int*.

- ipc_class

  Segundo nível hierárquico do sistema IPC, as seções são subdivididas em classes.

  Tipo de dado: *string*.

- ipc_classification_value

  "I"definindo "informação de invenção"ou "N"definindo "informação de não invenção".

  Tipo de dado: *string*.

- ipc_first_seen_date

  Data da concessão da primeira patente no banco de dados dentro de um grupo do IPC.

  Tipo de dado: *string*.

- ipc_last_seen_date

  Conceder a data da patente mais recente na base de dados dentro de um grupo IPC.

  Tipo de dado: *string*.

- ipc_main_group

  Subdivisões de subclasses IPC.

  Tipo de dado: *string*.

- ipc_section

  Os maiores níveis hierárquicos do IPC correspondente a campos técnicos: A = Necessidades Humanas, B = Operações de transporte operacional, C = Química Metalúrgica , D = Papel, Têxtil, E = Construção Civil, F = Engenharia Mecânica, Iluminação, Aquecimento, Armas, Motores, Bombas Hidráulicas, G = Física, H = Eletricidade.

  Tipo de dado: *string*.

- ipc_sequence

  Ordem da classificação IPC na lista de classificações da patente selecionada.

  Tipo de dado: *string*.

- ipc_subclass

  Subdivisões das classes IPC.

  Tipo de dado: *string*.

- ipc_subgroup

  Subdivisões do grupo principal do IPC.

  Tipo de dado: *string*.

- ipc_symbol_position

  Posição inicial ou final dos símbolos. A posição da primeira classificação de informação de invenção pode ser reconhecida por este campo. As letras "F"(*first*) e "L"(*last*) são usadas para a primeira posição e final, respectivamente.

  Tipo de dado: *string*.

- ipc_total_num_assignees

  Número total de cessionários em patentes dentro de uma classe IPC

  Tipo de dado: *int*.

- ipc_total_num_inventors

  Número total de inventores em patentes dentro de uma classe IPC

  Tipo de dado: *int*.

- nber_category_id

  Identificação da categoria de tecnologia do *National Bureau of Economic Research* (NBER) (consulte nber_category_title para obter detalhes)

  Tipo de dado: *string*.

- nber_category_title

  Titulo da categoria do NBER.

  Tipo de dado: *string*.

- nber_first_seen_date

  Data da concessão da primeira patente dentro de uma subcategoria NBER.

  Tipo de dado: *date*.

- nber_last_seen_date

  Data da concessão da patente mais recente dentro de uma subcategoria NBER.

  Tipo de dado: *date*.

- nber_subcategory_id

  Código da subcategoria NBER (relacionado com nber_subcategory_title, para obter detalhes).

  Tipo de dado: *string*.

- nber_subcategory_title

  Titulo da subcategoria NBER.

  Tipo de dado: *string*.

- nber_total_num_assignees

  Quantidade de cessionários em patentes dentro de uma subcategoria NBER.

  Tipo de dado: *int*.

- nber_total_num_inventors

  Quantidade de inventores em patentes dentro de uma subcategoria NBER.

  Tipo de dado: *int*.

- nber_total_num_patents

  Quantidade de patentes dentro de uma subcategoria NBER.

  Tipo de dado: *int*.

- patent_abstract

  Resumo da Patente.

  Tipo de dado: *string*.

- patent_average_processing_time

  Tempo médio de processamento para patentes na mesma classe da UPSC que a patente selecionada.

  Tipo de dado: *int*.

- patent_date

  Data de concessão da patente.

  Tipo de dado: *date*.

- patent_firstnamed_assignee_city

  Cidade do local do primeiro cessionário (isto é, primeiro na lista) na patente.

  Tipo de dado: *string*.

- patent_firstnamed_assignee_country

  País do local do primeiro cessionário (isto é, primeiro na lista) na patente.

  Tipo de dado: *string*.

- patent_firstnamed_assignee_id

  Código do cessionário (assignee_id) para o cessionário nomeado pela primeira vez (isto é, primeiro na lista) na patente.

  Tipo de dado: *string*.

- patent_firstnamed_assignee_location_id

  Código único para o local do primeiro cessionário nomeado (isto é, primeiro na lista) na patente.

  Tipo de dado: *string*.

- patent_firstnamed_assignee_state

  Estado do local do primeiro cessionário (isto é, primeiro na lista) na patente.

  Tipo de dado: *string*.

- patent_firstnamed_inventor_city

  Cidade do local do primeiro inventor (isto é, primeiro na lista) na patente.

  Tipo de dado: *string*.

- patent_firstnamed_inventor_country

  País do local do primeiro inventor (isto é, primeiro na lista) na patente.

  Tipo de dado: *string*.

- patent_firstnamed_inventor_id

  Código do inventor (inventor_id) para o inventor nomeado pela primeira vez (isto é, primeiro na lista) na patente

  Tipo de dado: *string*.

- patent_firstnamed_inventor_location_id

  Código único para o local do primeiro inventor nomeado (isto é, primeiro na lista) na patente.

  Tipo de dado: *string*.

- patent_firstnamed_inventor_state

  Estado do local do primeiro inventor (isto é, primeiro na lista) na patente.

  Tipo de dado: *string*.

- patent_kind

  Padrão ST.16 da Organização Mundial da Propriedade Intelectual (OMPI) Código

  Tipo de dado: *string*.

- patent_num_cited_by_us_patents

  Número de vezes que a patente foi citada por outras patentes.

  Tipo de dado: *int*.

- patent_num_combined_citations

  Número de patentes e pedidos citados pela patente seleccionada. Esta é a soma de citações de patentes dos EUA, patentes estrangeiras e aplicações dos EUA.

  Tipo de dado: *int*.

- patent_num_foreign_citations

  Número de patentes estrangeiras citadas pela patente selecionada.

  Tipo de dado: *int*.

- patent_num_us_application_citations

  Número de aplicações dos EUA citadas pela patente selecionada.

  Tipo de dado: *int*.

- patent_num_us_patent_citations

  Número de patentes dos EUA citadas pela patente selecionada.

  Tipo de dado: *int*.

- patent_number

  Número da Patente dos EUA, conforme designado pelo USPTO.

  Tipo de dado: *string*.

- patent_processing_time

  Prazo desde a data de apresentação do pedido até a data de concessão da patente.

  Tipo de dado: *int*.

- patent_title

  Titulo da patente.

  Tipo de dado: *string*.

- patent_type

  Categoria de patente. Existem 6 tipos possíveis: "Publicação Defensiva- 509, "Design- 474736, "Planta- 21052, "Reedição- 16416, "Registro de Invenção Estatutária- 2254, "Utilidade- 4910906.

  Tipo de dado: *string*.

- patent_year

  Ano de concessão da patente

  Tipo de dado: *int*.

- rawinventor_first_name

  Nome do inventor antes da desambiguação conforme listado na patente selecionada

  Tipo de dado: *string*.

- rawinventor_last_name

  Sobrenome do inventor antes da desambiguação conforme listado na patente selecionada

  Tipo de dado: *string*.

- uspc_first_seen_date

  Data de concessão da primeira patente no banco de dados dentro de uma classe principal USPC.

  Tipo de dado: *date*.

- uspc_last_seen_date

  Data de concessão da patente mais recente no banco de dados dentro de uma classe principal USPC.

  Tipo de dado: *date*.

- uspc_mainclass_id

  Código de identificação de classe principal da USPC.

  Tipo de dado: *string*.

- uspc_mainclass_title

  Descrição da classe principal da USPC.

  Tipo de dado: *string*.

- uspc_sequence

  Ordem da classificação USPC na lista de classificações para a patente selecionada.

  Tipo de dado: *int*.

- uspc_subclass_id

  Código de identificação de subclasse da USPC.

  Tipo de dado: *string*.

- uspc_subclass_title

  Descrição da subclasse da USPC.

  Tipo de dado: *string*.

- uspc_total_num_assignees

  Número total de cessionários de patentes dentro de uma classe principal USPC.

  Tipo de dado: *int*.

- uspc_total_num_inventors

  Número total de inventores de patentes dentro de uma classe principal USPC.

  Tipo de dado: *int*.

- uspc_total_num_patents

  Número total de patentes dentro de uma classe principal USPC.

  Tipo de dado: *int*.

- wipo_field_id

  Código de identificação da WIPO.

  Tipo de dado: *int*.

- wipo_field_title

  Descrição do campo da WIPO.

  Tipo de dado: *string*.

- wipo_sector_title

  Descrição do setor da WIPO.

  Tipo de dado: *string*.

- wipo_sequence

  Ordem do domínio tecnológico da WIPO na lista de domínios tecnológicos para a patente selecionada.

  Tipo de dado: *int*.

4.4 Codificação dos serviços

A codificação desse serviço deve obedecer a norma padrão para *HTTP* respeitando as requisições de *POST* da API, que é a forma de acesso a todos os métodos. Tanto as respostas quanto as requisições devem obedecer esse padrão, tanto para *XML* como para *JSON*. Caso não seja obedecido esse padrão de requisições, a mensagem de erro será dada pelo próprio roteamento da API, retornando o erro de *Bad Request*.

5 Testes, análises e considerações finais

Este trabalho objetiva propor e discutir uma infraestrutura de dados químicos como suporte à recuperação de informação de patentes. A infraestrutura de dados químicos atua como federação de serviços dos vários repositórios de patentes existentes no mundo, agrupados por meio de interfaces de programação de aplicações (API) existentes. Essas APIs correspondem a conversores de termos químicos e serviços de recuperação de patentes. Para efeito de testes, foram implementados clientes que consomem duas APIs de repositórios públicos de patentes e duas APIs de conversões, uma pública e outra usando a biblioteca CDK.

O objetivo desta pesquisa não foi gerar um protótipo de uma interface homem-computador, mas produzir uma interface de programação que facilitasse a criação de programas que utilizam de um serviço unificado, facilitando assim a implementação de novos programas e serviços de recuperação de patentes, especialmente patentes com conteúdo químico.

5.1 Testes e Análises

Dois tipos de testes foram empreendidos neste trabalho. O primeiro tipo, de conformidade. O segundo tipo, de viabilidade da prova de conceito.

Os testes de conformidade ficaram concentrados nos serviços de conversão. O objetivo era traduzir um termo químico de uma notação para outra, usando um ou vários serviços de conversão. Os serviços funcionaram adequadamente, inclusive aqueles baseados na biblioteca CDK.

Os testes de viabilidade da prova de conceito ficaram concentrados nos serviços de busca.

As buscas foram feitas primeiramente a partir de termos simples, como datas de publicação e número de patente. Essas buscas funcionaram corretamente, sendo possível a busca em cache local ou o acesso às API dos serviços remotos, diretamente.

As buscas mais complexas agregavam os termos, data de publicação e sobrenome de autores/inventores. Essas buscas testaram a velocidade da recuperação de uma mesma busca automaticamente submetida a vários repositórios, contra várias buscas manualmente submetidas a vários repositórios.

É trivial considerar que uma única busca, traduzida para outros formatos e submetida automaticamente para vários repositórios, teve potencial de reduzir o tempo de resposta percebido pelo usuário. Além disso, o usuário não precisa aprender a variadas notações de consulta, dos vários repositórios de patentes. Por outro lado, o custo de seleção dos vários resultados foi acumulado pelo sistema federado de recuperação de informação, algo que precisa ser calculado.

5.2 Considerações finais

O principal objetivo dessa pesquisa foi propor e especificar um conjunto de serviços federados para a recuperação de informação química em patentes.

Dessa forma, foi dada ênfase aos requisitos funcionais e não-funcionais do sistema automatizado de recuperação de informação de patentes e da infraestrutura de dados químicos, como uma forma de investigar a viabilidade de desenvolvimento e consequentemente, a contribuição dos serviços web para a criação de novos sistemas de informação.

A revisão de literatura apontou que pouca atenção tem sido dada a criação de serviços federados para o acesso a patentes, embora exista consenso que a diversidade de escritórios e países dificulte a recuperação de informação. A literatura aponta as infraestruturas de dados como solução em outros campos, como em dados geográficos e bibliográficos, sem qualquer discussão iniciada sobre o campo de propriedade industrial.

A padronização de interfaces de serviços foi a primeira atividade realizada, tendo em vista que e um prerrequisito para que outros serviços possam consumir os serviços primitivos de recuperação, conversão e outros. O que foi proposto, se aperfeiçoado, pode servir como um padrão global de recuperação de informação química em patentes, havendo uma única infraestrutura que permita que os bancos de patentes fossem federados, assim como ocorre com outros dados.

Um passo importante já havia sido dado pela indústria, com a padronização dos metadados de uma patente. Porém, cada bloco econômico ou país tem seu próprio meio de guardar e disponibilizar essas patentes. Alguns escritórios regionais já disponibilizam uma API de serviços. A variedade de serviços não é um problema para a federação. Ao contrário, a variedade é um problema apenas para a criação de programas-finais, para pesquisadores e universidades, que deveriam personalizar seus softwares para os diferentes fornecedores de dados. A federação se mostrou útil em um cenário de falta de padronização dos serviços para permitir que novos softwares sejam criados para o usuário-final.

Além disso, como não existe padronização específica para recuperação de informações químicas, esta pesquisa se mostrou útil para criar novos componentes de software para tratamento de dados específicos, os quais podem ser intercambiados por programadores a partir da eficiência, de licenças comerciais e do custo computacional.

Duas abordagens de federação se mostraram possíveis. A primeira abordagem baseia-se na coleta e no armazenamento de dados no repositório local, com elevada ocupação de espaço de arquivos de patentes, nenhum conflito jurídico uma vez que as informações de patentes são de domínio público, e maior desempenho de recuperação. A segunda abordagem baseia-se na consulta em tempo de execução aos vários repositórios distribuídos, com nenhuma ocupação de espaço para arquivos de patentes, necessidade de licença para elevado consumo de banda, e menor desempenho de recuperação. Abordagens híbridas também são possíveis a partir da infraestrutura de dados química e da federação de serviços.

Em qualquer das abordagens, a principal contribuição desta pesquisa que pode ser citada é a preparação de modelo de arquitetura para a criação de diversos outros trabalhos, bem como desenvolvimento de softwares baseados no protótipo desenvolvido. Sendo assim através destes conjuntos de serviços federados foi possível experimentar a introdução de novas funcionalidades de forma que possibilite uma implementação transparente de clientes para consumo desse arcabouço de serviços.

Das classes de serviços prototipadas nesta pesquisa, a mais útil para o intercâmbio

de informações químicas é a de conversão. Essa classe se mostrou particularmente útil para garantir a uniformidade de busca de entidades químicas nas várias patentes, independentemente de idioma, origem da patente, e campo do conhecimento.

5.3 Trabalhos Futuros

Foram identificados três propostas para trabalhos futuros. Primeiramente desenvolver uma melhor interface homem-computador para serviços de busca de patentes. Por conseguinte desenvolver uma linguagem intermediaria para uso da própria API desenvolvida. Finalmente agregar outros serviços que não foram tratados nessa pesquisa.

Em um serviço federado de recuperação de informação, é natural que o usuário-final seja o primeiro agente beneficiado. Porém, pouca pesquisa foi desenvolvida sobre a interface homem-computador dos sistemas de busca de patentes. Com isso, são comuns muitas interfaces incompatíveis e excessivamente diferentes para os vários repositórios de patentes disponíveis.

Uma segunda oportunidade de pesquisa tem relação com linguagens intermediárias de busca, seleção e ordenação de resultados. É comum a adoção da linguagem SQL para este fim, porém, com dificuldades na criação e otimização dos compiladores ou interpretadores da linguagem em contextos muito específicos. Alguns outros programas e serviços possuem uma linguagem de interface baseadas em SQL, tal como a linguagem de programação Java, com a sua JPQL (ORACLE, 2011), e também o JIRA, que possui uma linguagem própria de busca em seus sistemas que é conhecida como JQL (RADIGAN, 2013). Estudos sobre linguagens de consulta mais otimizadas são úteis para infraestruturas de dados químicos e para bancos federados de patentes.

Esta pesquisa fez uso de apenas dois repositórios de patentes. Porém, um estudo aprofundado que inclua mais repositórios é importante para realizar avaliação de desempenho, identificar mais diferenças entre os vários repositórios, e principalmente para produzir um serviço mais abrangente.Um ponto de partida é federar pelo menos os maiores repositórios para que as buscas por informações químicas em patentes sejam mais úteis para os usuários-finais.

Referências

ALBERTS, D. et al. *Current Challenges in Patent Information Retrieval*. 1. ed. Springer-Verlag Berlin Heidelberg, 2011. (The Information Retrieval Series 29). ISBN 3642192300,9783642192302. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=872CAD325B6BE1F799A95858FE883B7C>. Citado na página 17.

ANASTÁCIO, I. *Location-Based Targeting and Ranking for Online Advertising*. 2009. 75 p. Dissertação (Master in Information Systems and Computer Engineering) — Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal, 2009. Citado na página 14.

BERNERS-LEE, T.; FIELDING, R.; FRYSTYK, H. *Hypertext transfer protocol–HTTP/1.0*. [S.l.], 1996. Citado na página 21.

BUYYA, R. High performance cluster computing. *New Jersey: F'rentice*, 1999. Citado na página 18.

CHEN, Y.-L.; CHIU, Y.-T. An IPC-based vector space model for patent retrieval. *Information Processing & Management*, [S.l], v. 47, n. 3, p. 309–322, 2011. Citado na página 10.

FIELDING, R. et al. *Hypertext transfer protocol–HTTP/1.1*. [S.l.], 1999. Citado na página 20.

GILES, C. L.; BOLLACKER, K. D.; LAWRENCE, S. Citeseer: an automatic citation indexing system. In: INTERNATIONAL CONFERENCE ON DIGITAL LIBRARIES, 3., 1998, Pittsburgh, PA, USA. *Proceedings...* New York, NY, USA: ACM, 1998. p. 89–98. Citado na página 10.

GOSLING, J. et al. *The Java language specification*. [S.l.]: Pearson Education, 2014. Citado na página 16.

HU, B.; SVENSSON, G. A case study of linked enterprise data. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 9., 2010, Shanghai, China. *Proceedings...* Heidelberg, Germany: Springer, 2010. p. 129–144. Citado na página 10.

INFOCHEM. *ICEdit - chemical structure and reaction editing tool by InfoChem - infochem*. 2017. Disponível em: <http://www.infochem.de/products/software/icedit.shtml>. Citado na página 19.

JOSEFSSON, S. The base16, base32, and base64 data encodings. 2006. Citado na página 24.

KRALLINGER, M. et al. Overview of the chemical compound and drug name recognition (chemdner) task. In: *BioCreative challenge evaluation workshop*. [S.l.: s.n.], 2013. v. 2, p. 2. Citado na página 18.

LEE, D. Jxon: an architecture for schema and annotation driven json/xml bidirectional transformations. In: *Proceedings of Balisage: The Markup Conference*. [S.l.: s.n.], 2011. Citado na página 20.

MANNING, C. D. et al. *Introduction to information retrieval*. [S.l.]: Cambridge university press Cambridge, 2008. v. 1. Citado nas páginas 16 e 18.

MASIAKOWSKI, P.; WANG, S. Integration of software tools in patent analysis. *World Patent Information*, Elsevier, v. 35, n. 2, p. 97–104, 2013. Citado na página 19.

MILLER, M. A. Chemical database techniques in drug discovery. *Nature Reviews Drug Discovery*, Nature Publishing Group, v. 1, n. 3, p. 220–227, 2002. Citado nas páginas 17, 18 e 19.

OFFICE, E. P. Espacenet free access to 90 million patent documents worldwide. 2015. Disponível em: <http://documents.epo.org/projects/babylon/eponet.nsf/0/4E8744EB66E8F944C12577D600598EEF/\protect\T1\textdollarFile/espacenet_brochure_en.pdf>. Citado na página 20.

ORACLE. *JPQL Language Reference*. 2011. Disponível em: <http://docs.oracle.com/html/E13946_04/ejb3_langref.html>. Citado na página 44.

PASCHE, E. et al. Development and tuning of an original search engine for patent libraries in medicinal chemistry. *BMC Bioinformatics*, v. 15, n. 1, p. 1–9, 2014. ISSN 1471-2105. Disponível em: <http://dx.doi.org/10.1186/1471-2105-15-S1-S15>. Citado nas páginas 10, 16 e 17.

PASCHE, E. et al. Development of a text search engine for medicinal chemistry patents. *EMBnet.journal*, v. 18, n. B, 2012. ISSN 2226-6089. Disponível em: <http://journal.embnet.org/index.php/embnetjournal/article/view/545>. Citado na página 16.

PENCE, H. E.; WILLIAMS, A. *ChemSpider: an online chemical information resource*. [S.l.]: ACS Publications, 2010. Citado na página 17.

RADIGAN, D. *JIRA Query Language (JQL) Recap*. 2013. Disponível em: <http://blogs.atlassian.com/2013/03/jql-recap/>. Citado na página 44.

SCHMID, E.-G. *Datenbankgestützte Substanzbeschaffung in der forschenden Chemieindustrie–ein algorithmischer Optimierungsansatz*. 2010. Tese (Doutorado) — Universität Duisburg-Essen, Mercator School of Management-Fakultät für Betriebswirtschaftslehre» Technology and Operations Management» Wirtschaftsinformatik insbes. Business Intelligence, 2010. Disponível em: <http://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-24687/schmid_diss.pdf>. Citado na página 15.

SOUZA, R. R. et al. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. *Perspectivas em ciência da informação*, SciELO Brasil, v. 11, n. 2, p. 161–173, 2006. Citado nas páginas 11 e 20.

STEINBECK, C. et al. The chemistry development kit (cdk): An open-source java library for chemo-and bioinformatics. *Journal of chemical information and computer sciences*, ACS Publications, v. 43, n. 2, p. 493–500, 2003. Citado na página 16.

SUN, B. et al. Extraction and search of chemical formulae in text documents on the web. In: INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 16., 2007, Banff, Alberta, Canada. *Proceedings...* New York, NY, USA: ACM, 2007. p. 251–260. Citado na página 10.

USPTO. *PatentsView*. 2016. Disponível em: <http://www.patentsview.org>. Citado na página 19.

WEININGER, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, ACS Publications, v. 28, n. 1, p. 31–36, 1988. Citado na página 18.

WIPO. *STANDARD ST.9:RECOMMENDATION CONCERNING BIBLIOGRAPHIC DATA ON AND RELATING TO PATENTS AND SPCS*. New York: WIPO, 2013. Citado na página 26.

WIPO. Chemical structure search. *International Journal of Geographical Information Science*, 2016. Disponível em: <http://www.wipo.int/directory/en/urls.jsp>. Citado na página 15.

WIPO. *Directory of Intellectual Property Offices*. 2016. Disponível em: <http://www.wipo.int/directory/en/urls.jsp>. Citado na página 19.

WOLF, M.; WICKSTEED, C. Date and time formats. *W3C NOTE NOTE-datetime-19980827, August*, 1998. Citado na página 28.

Lista de abreviaturas e siglas

API	*Application Programming Interface*, Interface de programação para aplicações
JSON	*JavaScript Object Notation*
WS	*Web Service*, Serviço Web
XML	*eXtensible Markup Language*
HTTP	*Hyper Text Transfer Protocol*
WIPO	*World Intellectual Property Organization*
EPO	*European Patent Office*
SQL	*Structured Query Language*, Linguagem de Consulta Estruturada
EUA	Estados Unidos da América
USPC	*United States Patent Classification*
WIPO	*World Intellectual Property Organization*
OMPI	Organização Mundial da Propriedade Intelectual
NBER	*National Bureau of Economic Research*
W3C	*World Wide Web Consortium*