UNIVERSIDADE FEDERAL DO MARANHÃO

Curso de Ciência da Computação

Weldson Amaral Corrêa

Integração e Modelagem de Dados Públicos de Pesquisa com foco na Extração de Conhecimento e Métricas de Produtividade

São Luís - MA

2018

Weldson Amaral Corrêa

Integração e Modelagem de Dados Públicos de Pesquisa com foco na Extração de Conhecimento e Métricas de Produtividade

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, **como parte dos requisitos necessários** para obtenção do grau de Bacharel em Ciência da Computação.

Universidade Federal do Maranhão

Curso de Ciência da Computação

Orientador: Prof. Dr. Geraldo Braz Junior

São Luís - MA

2018

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Núcleo Integrado de Bibliotecas/UFMA

Amaral Corrêa, Weldson.
Integração e Modelagem de Dados Públicos de Pesquisa com foco na extração de Conhecimento e Métricas de Produtividade / Weldson Amaral Corrêa. – 2018.
44 p.

Orientador(a): Geraldo Braz Junior.
Monografia (Graduação) – Curso de Ciência da Computação, Universidade Federal do Maranhão, UFMA, 2018.

1. Dados de produtividade. 2. Elaboração de métricas. 3. Extração de dados. I. Braz Junior, Geraldo. II. Título.

Weldson Amaral Corrêa

Integração e Modelagem de Dados Públicos de Pesquisa com foco na Extração de Conhecimento e Métricas de Produtividade

Monografia apresentada ao curso de Ciência da Computação da Universidade Federal do Maranhão, **como parte dos requisitos necessários** para obtenção do grau de Bacharel em Ciência da Computação.

Trabalho Aprovado em São Luís - MA, 24 de Janeiro de 2018:

Prof. Dr. Geraldo Braz Junior
Orientador

Prof. Dr. João Dallyson Sousa de Almeida
Examinador

Prof. Dr. Davi Viana dos Santos
Examinador

São Luís - MA
2018

Agradecimentos

A Deus, por permitir o cumprimento desta etapa em minha vida.

Agradeço ao Profº Dr. Geraldo Braz Junior, pela belíssima orientação, motivação e paciência, que certamente foram primordiais para a conclusão deste trabalho.

A minha família que sempre esteve ao meu lado. Aos meus colegas e amigos pelo companheirismo e incentivos nessa caminhada.

If I have seen further, it is by standing on the shoulders of giants.

Isaac Newton

Resumo

Esta monografia visa apresentar a modelagem de uma ferramenta para a extração de dados públicos, sobre as pesquisas realizadas pelos docentes da Universidade Federal do Maranhão. O trabalho irá descrever os processos de aquisição de dados nas múltiplas bases que foram utilizadas, bem é realizado o processo de armazenamento destes. Além disso, a ferramenta apresentada tem um modulo de visualização, que fornece informações sobre as pesquisas e pesquisadores desta universidade. A ferramenta apresentada pode ser utilizada como um meio de divulgação e promoção dos trabalhos dos docentes da universidade, com intuito de permitir a exploração e analise inteligente dessas informações.

Palavras-chaves: Produção cientifica, Integração de dados, Extração de informação, Universidade Federal do Maranhão.

Lista de abreviaturas e siglas

CAPES	Cordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CSV	Comma-Separeted Values
ENADE	Exame Nacional de Desempenho de Estudante
ETL	Extraction, Transformation, Load
HTML	HyperText Markup Language
ISI	Institute for Scientific Information
JSON	JavaScript Object Notation
JCR	Journal Citations Reports
MEC	Ministério da Educação
NTI	Núcleo de Tecnologia da Informação
PDF	Portable Document Format
SGBD	Sistema de Gerenciamento de Banco de Dados
SIGRH	Sistema Integrado de Gestão de Recursos Humanos
SNPG	Sistema Nacional de Pós-Graduação
SOAP	Simple Object Access Protocol
UFMA	Universidade Federal do Maranhão
XLS	Extensible Style Language
XML	Extensible Markup Language
WSDL	Web Services Description Language

Sumário

Lista de ilustrações

1 Introdução

A Universidade Federal do Maranhão - UFMA é um importante centro de produção intelectual no Estado do Maranhão. Não é difícil verificar que as pesquisas de ponta no estado são realizadas com participação completa ou parcial de pesquisadores que compõe o quadro de pesquisadores desta universidade.

As informações geradas pelas pesquisas realizadas na instituição estão em sua grande parte disponíveis através da internet. Esse grande conjunto de informação e de referências forma uma verdadeira massa de dados que necessitam, em certos momentos, de um grande esforço para filtragem.

Um exemplo prático, seria consultar a quantidade de periódicos que os pesquisadores da UFMA publicam por ano, ou a pergunta mais abrangente seria checar quais grupos de pesquisadores tem produção emergente e poderiam receber incentivos para se tornarem de excelência. A política de informação oficialmente estabelecida pelo governo brasileiro afirma ser a informação um recurso estratégico e propulsor do desenvolvimento, conclamando para a promoção do uso das novas tecnologias de comunicação nos campos econômico e social.

Mesmo sendo referência de pesquisa, a UFMA ainda não possui uma ferramenta que quantize a importância da pesquisa aqui realizada, e ainda promova-as de forma unificada, multilíngue e exploratória, de modo que permita a sua divulgação a agentes internos e externos a instituição e também a análise dessa informação.

A deficiência neste ponto faz com que abordagens não padronizadas sejam utilizadas através de divulgações não atualizadas de projetos de pesquisa no site da instituição, ou mesmo de outras instituições. A informação desatualizada pode ser uma barreira encontrada por muitos pesquisadores que poderiam cooperar com pesquisadores da UFMA e assim aumentar a qualidade e produtividade.

Outro ponto importante é que os gestores envolvidos no incentivo a produção não conseguem de maneira simples quantificar e qualificar a pesquisa realizada na UFMA. Normalmente um processo dispendioso e manual de checagem de currículos na plataforma lattes e em alguns casos, o contato direto com o pesquisador por falta de informações atualizadas ou determinantes para uma tomada de decisão.

Dentro desse contexto, este trabalho pretende apresentar uma ferramenta computacional que integre as múltiplas bases de dados que possuem informação de pesquisa e pesquisadores promovendo ao mesmo tempo uma interface que permita a divulgação da pesquisa de maneira colaborativa e a gestão de pesquisa no âmbito da Universidade,

embora não restrita a esta.

1.1 Objetivos

1.1.1 Geral

Fazer a extração dados de pesquisas dos docentes da UFMA, a partir de bases de dados públicas, para criação de métricas, que avaliem a produtividade dos docentes e setores desta universidade.

1.1.2 Específicos

- Coletar informações disponibilizadas de forma publica a cerca das pesquisa que são feita por docentes da UFMA.
- Extrair informações sobre a produtividade dos docentes a partir dos dados coletados.
- Disponibilizar através de um portal informações utilizando gráficos, sobre as pesquisas e pesquisadores da UFMA.

2 Fundamentação teórica

Atualmente, as universidades brasileiras possuem uma latente dificuldade para ter um controle sobre a produção científica de seu corpo docente. No entanto, grande parte desses dados, se encontram em plataformas online e gratuitas, tais como, a plataforma Lattes, que é um sistema de informação desenvolvido pelo Conselho Nacional de Desenvolvimento Científico e Tecnologico – CNPq, o Google Scholar ou Google Acadêmico que tem como finalidade fazer a indexação de produções de cunho científico, dentre outras.

A Universidade Federal do Maranhão, como todas as universidades federais, é um importante celeiro de produção científica dentro do estado que está localizada. As pesquisas feitas dentro do estado, são realizadas em sua maioria com participação completa ou parcial de pesquisadores que compõe o quadro de docentes desta universidade. E como parte das universidades brasileiras, existe uma deficiência de manter-se um controle sobre todas as produções feitas com participação de seus docentes.

Felizmente como já falado, grande parte, ou todas essas produções estão disponíveis em bancos de dados abertos. Segundo a (OPEN KNOWLEDGE INTENATIONAL, 2018) são considerados Dados Públicos ou abertos, quaisquer dados que possam ser usados, modificados e compartilhados por qualquer pessoa e para qualquer proposito. A partir dessas bases, que compartilham dados abertamente, tem-se que é possível mensurar a produtividade dos pesquisadores ali contida.

2.1 Dados e Indicadores Públicos de Pesquisa

2.1.1 SIGRH

A UFMA é uma das maiores universidades do estado, portanto, para que possa ter um funcionamento adequado, é natural que tenha muitos servidores a serviço desta. Para que todo esse contingente de servidores possam ser corretamente geridos, é necessário um sistema de informação para manter o cadastro destes, o qual é chamado de Sistema de Gestão de Recursos Humanos - SIGRH.

O SIGRH é portanto um sistema que permite que a UFMA, tenha um controle do cadastros de servidores. Este sistema é mantido pelo NTI, e oferece diversas funcionalidades que podem ser acessadas pelo público em geral, tais como, relatórios dos sobre colaboradores por funções, ativos e inativos, onde estão locados cada um dos servidores que fazem parte do corpo docente, entre outras.

Além desses serviços, o Núcleo de Tecnologia da Informação - NTI também fornece

acesso ao SIGRH via web service. O web service permite que outras aplicações se comunique de forma direta com o SIGRH, e é desta forma que serão integralizados os dados dos servidores ao portal apresentado. É importante destacar que alguns dos dados requisitados são considerados sigilosos, então para que a aplicação fizesse a aquisição dos dados, foi necessário a criação de um perfil de acesso.

2.1.2 Plataforma Lattes

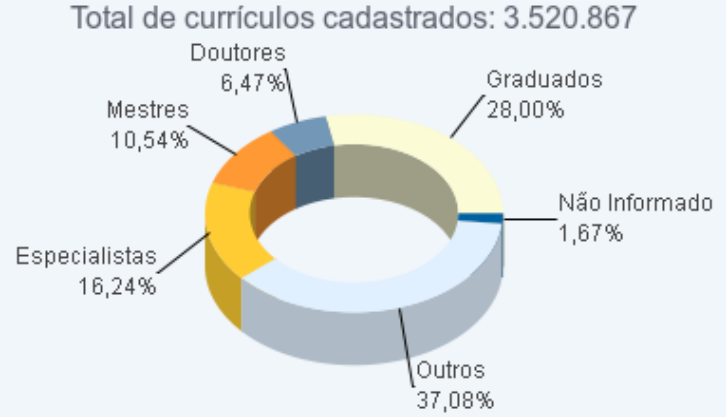
Segundo a (BRAS, 2003), a plataforma Lattes foi criada, com objetivo de integrar os sistema de informação das principais agencias de fomento do país da época, sendo resultado do esforço do Ministério da Ciência e Tecnologia - MCT, CNPq e CAPES/MEC. Esta plataforma recebeu o nome de Lattes, em homenagem ao cientista brasileiro Cesare M. Giulio Lattes, que é mundialmente conhecido por suas descobertas no campo da física.

Com a criação da plataforma tornou-se possível o controle do cadastro dos pesquisadores brasileiros num único local, o que facilitou o processo de controle, tanto para os pesquisadores que passaram a precisar fazer apenas um cadastro único, quanto para as agências do governo responsáveis pelas avaliações dos pesquisadores.

O projeto foi lançado em 1999 com aproximadamente 35 mil currículos cadastrados, e a partir de 2002 seu uso passou a ser obrigatório para todos os bolsistas de iniciação científica, mestrado e doutorado, orientadores credenciados, e outros, pois sem esse cadastro, estão impedidos de receber pagamentos, e fazer renovações.

No final do ano de 2016 a plataforma já contava com mais de 350 mil currículos cadastrados (CNPQ, 2016) conforme é possível ver na Figura 1.

Figura 1 – Estatísticas da Base de currículos da Platatorma Lattes.



Fonte: (CNPQ, 2016)

Como é possível observar houve uma grandes expansão da base de dados, do ano de lançamento até o ano de 2016, um crescimento de aproximadamente 1000%. O

Currículo Lattes conseguiu se consolidar nesse tempo, como um importante instrumento de armazenamento e divulgação de dados curriculares dos pesquisadores do Brasil.

2.1.3 Google Scholar

O Google é uma empresa criada por Sergey Brin, e Larry Page, sendo o resultado de um trabalho acadêmico realizados pelos dois enquanto estudantes da Universidade de Stanford. Inicialmente o projeto foi chamado de BackRub, e tinha como objetivo fazer a extração de imensos volumes de informação da internet, o algoritmo deveria fazer leitura de páginas web, e partir daí, fazer a verificação dos links que existiam numa página.

O grande diferencial desse projeto, foi a criação do algoritmo de *PageRank* (PAGE et al., 1999) que segundo (SILVA; GALANTE, 2008) é um número que determina o grau de relevância de uma página web na Internet, em outras palavras, quanto maior esse número, maior é a importância da página. Com isso observaram que o BackRub seria um projeto que conseguiria crescer de acordo com a Web, uma vez que o *PageRank* faz a análise das conexões da páginas, e quanto mais conexões na Web fossem criadas mais links poderiam ser avaliados pelo algoritmo. A partir decidiram chamar essa nova ferramenta de Google, que segundo (CARMONA, 2006) o "nome Google derivado da palavra googol, que designa a centésima potencia do número 10, ou o número fictício formado pelo número 1 seguido de cem zeros".

Foi perceptível assim a ambição dos empresário em alcançar uma gigante quantidade de informações na Web. E em 1998 a empresa foi oficialmente criada, tendo como missão segundo (LYER BALA, 2008), citado por (SOUZA, 2009) "organizar informações do mundo todo para que sejam úteis e disponíveis a todos[...]".

Desde que o Google estabeleceu sua missão, ele tem se tornado um importante conjunto de ferramentas com a finalidade de cumprir sua meta, passando a oferecer diversos serviços, dentre eles o Google Scholar ou Google Acadêmico em português que é voltado para a produção de cunho científico. O Google Acadêmico é uma plataforma que faz uma busca por informações acadêmicas/científicas como, artigos, monografias, revistas, resumos, dentre outros.

Segundo (SOUZA, 2009) muitos pesquisadores tem passado a utilizar o Google Acadêmico para recuperação de publicações científicas em detrimento ao próprio Google, uma vez que ao ser voltado para esse uso, os resultados das pesquisas tem um grau maior de precisão. Logo, é possível observar que esta ferramenta se tornou um referencial no segmento de recuperação de informações científicas.

Dentre as funcionalidades contidas no Google Acadêmico, devemos ressaltar a possibilidade de qualquer pesquisador criar um perfil na plataforma e desta forma ter seus artigos publicados, ligados diretamente a esse perfil. A plataforma além de fazer o

link com os artigos, também apresenta um quadro com o número de citações do autor conforme a Figura 2.

Figura 2 – Exemplo de citações de um perfil no Google Acadêmico.

Fonte: (GOOGLE. . . , 2018)

Diante de todos esses dados o que mais nos interessa nesse projeto é quantidade de citações de um autor, sendo que o índice de *citações* apresenta a quantidade total de citações do autor. O *índice h* é um método que consiste em fazer o relacionamento de do número de publicações científicas com o número de suas citações. E por fim o *índice i10* corresponde a quantidade de artigos que receberam pelo menos 10 citações.

Desta forma podemos concluir que o Google Acadêmico, oferece índices importantes para que se faça a medição de produtividade de um determinado autor.

2.1.4 Qualis

Segundo a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES: "Qualis é o conjunto de procedimentos utilizados [...] para estratificação da qualidade da produção intelectual dos programas de pós graduação". Em outras palavras o Qualis tem como objetivo medir qualidade das produções científicas, essa medição é feita de forma indireta, ou seja, para aferir a importância das produções, faz-se a análise da qualidade dos veículos de divulgação das produções, que são periódicos científicos e anais de eventos.

O processo de avaliação é feito anualmente, tendo em vista a relevância dos artigos para a sociedade, o nível científico, originalidade e escrita. Segundo (JUNIOR; CARLOS, 2016), os critérios de avaliação podem ser classificados como:

- **Periodicidade:** as revistas e periódicos que estão a mais tempo no mercado, tem um grau maior de credibilidade. A frequência das publicações garante também uma melhor classificação.
- **Qualidade do corpo editorial:** este item diz respeito a banca que faz a analise dos artigos que serão publicados no periódico.
- **Diversidade de origens do trabalho:** os periódicos com melhor apreciação são os que possuem desde autores institucionais a internacionais.
- **Difusão e popularidade da revista:** quanto mais conhecido o veículo melhor será, mais pessoas e de diferentes lugares fazem com que o periódico tenha uma boa classificação.
- **Indexação:** diz respeito ao quão fácil é recuperado qualquer informação, quando um usuário faz uma busca num sistema de informação. Uma boa indexação é obtida quando um periódico é acessível e de qualidade.

A classificação dos periódicos e eventos é feita conforme a área de avaliação, sendo que os indicativos de qualidade são: A1, o mais elevado, A2, B1, B2, B3, B4, B5 e C, que possui menos peso, para ter um melhor entendimento (JUNIOR; CARLOS, 2016) descreve os aspectos de classificação da seguinte forma:

- A1 e A2: Excelência internacional.
- B1 e B2: Excelência nacional.
- B3, B4 e B5: relevância média
- C: baixa relevância

É necessário destacar também que um mesmo periódico, pode receber classificações diferentes em áreas distintas. Apesar da classificação diferente, isso não pode ser considerado uma inconsistência, mas, o grau de relevância do periódico em questão conforme a área que está sendo considerada. Desta forma, não há a intenção de tornar essa classificação de qualidade de periódicos um número absoluto para todas as áreas.

A partir dos critérios que são adotados para classificação de artigos conforme o Qualis, podemos dizer que este é artificio que deve pode ser utilizado para a avaliação qualitativa das produções do pesquisadores brasileiros.

Os dados sobre os Qualis dos periódicos, são fornecidos através da Plataforma Sucupira. Segundo (CAPES, 2014) a plataforma que foi criada com objetivo de ser uma

Figura 3 – Plataforma Sucupira

Fonte: (PLATAFORMA. . . , 2018)

ferramenta para coletar informações, realizar análises e avaliações, e ser a base de referência do SNPG, Figura 3.

Para obter os índices Qualis, é necessário apenas escolher o período que se deseja recuperar. Caso tenha é necessidade, é possível recuperar os arquivos com critérios mais específicos, como, área de avaliação, ISSN, título ou classificação.

2.1.5 JCR

O JCR é outro importante índice para avaliação do alcance e qualidade das publicações científicas. Este índice é publicado anualmente pelo Institute for Scientific Information - ISI, sendo uma base reconhecida internacionalmente por avaliar periódicos indexados na plataforma Web of Science. O índice fornecido polo JCR, é chamado de Fator de Impacto, que é calculado a partir da soma das citações dos artigos publicados num período de dois anos, onde tanto a revista citada quanto a que faz a citação, devem estar incluídas no banco de dados do Web of Science.

A ideia da criação de uma medida de impacto de uma revista científica foi idealizada por Eugene Garfield, em 1995. E pouco tempo depois, tornou-se um importante parâmetro para avaliar a qualidade da produção científica. Este índice, não apenas ganhou importância no meio acadêmico, como passou a ter impacto gerencial, no que tange a concessão de bolsas de estudo e financiamento de projetos, inclusive no Brasil.

Podemos facilmente concluir portanto, que o Fator de Impacto JCR, constitui uma importante base para avaliarmos a qualidade das produções científicas.

2.2 Tecnologias, padrões e/ou serviços

2.2.1 Web service

Segundo (W3C, 2014) um Web Service é um sistema de software projetado para permitir interação entre máquinas através de uma rede, onde o protocolo utilizado para essa comunicação é o HTTP, com o padrão de mensagem SOAP, que será apresentado mais a frente. Em outras palavras, pode-se dizer que, um web service é utilizado para fazer a transferência de dados através de protocolos de comunicação conhecidos da Web, permitindo que essa interação ocorra independente de plataformas e linguagens de programação que dê suporte Web.

Entre os benefícios proporcionados pelo uso do web service, podemos destacar:

- Reutilização de código: uma vez que o serviço pode ser utilizado por diferentes plataformas, não se faz necessário escrever o código múltiplas vezes para cada plataforma.
- Mais segurança: o web service evita que a comunicação direta com a base de dados, desta forma os dados tem um pouco mais de proteção.
- Integração de sistemas: com o padrão de mensagens adotado pelo web service (SOAP), a comunicação entre sistemas é bastante simples, sem que se tenha a necessidade de saber detalhes sobre o funcionamento de cada sistema.

2.2.1.1 SOAP

O SOAP é um protocolo que foi elaborado com a finalidade de facilitar a chamada remota de funções via internet, permitindo assim que dois programas possam se comunicar de maneira tecnicamente muito semelhante à invocação de páginas da Web (ZARELLI, 2012).

Entre as principais vantagens deste protocolo destacam-se:

- Simplicidade de implementação, teste e uso;
- Utilização dos padrões da Web para quase tudo, por exemplo, a comunicação é feita via HTTP com pacotes virtuais idênticos, mesmos protocolos de autenticação e encriptação, mesma forma de manutenção e implementação feita pelo próprio servidor Web.
- A descrição dos dados e funções são feitas em XML, o que faz do protocolo robusto e fácil de usar.
- Independente de sistema operacional.

- A utilização pode ser tanto de forma anônima como com autenticação (nome/senha)

As requisições feitas utilizando o SOAP têm três padrões: GET, POST e SOAP. Os padrões GET e POST são iguais aos feitos por navegadores de internet. Já o padrão SOAP são requisições semelhantes ao POST, com a diferença que essa requisição é feita em XML, tendo como vantagem permitir o uso de recursos mais avançados como passar estruturas e arrays. As respostas independente de qual padrão é utilizado é sempre em XML.

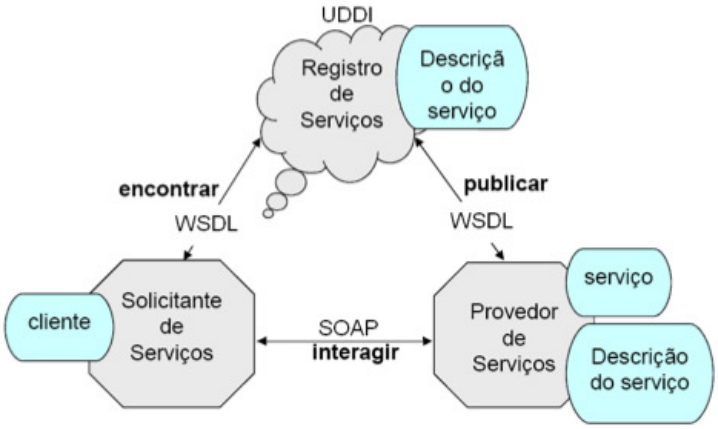
As vantagens das respostas serem em XML os dados podem ser descritos em tempo de execução, o que evita problemas causados por inadvertidas mudanças nas funções, uma vez que os objetos que são chamados têm a possibilidade de sempre validar so argumentos das funções, o que torna o protocolo mais seguro.

Outro padrão descrito pelo SOAP é o WSDL, que faz a descrição dos objetos e métodos disponíveis, utilizando páginas XML que são acessadas a partir da Web. Desta forma, a publicação de um serviço, é seguida da criação de um arquivo WSDL que auxilia como uma "documentação"para fazer a requisição de um serviço.

2.2.1.2 Mensagens entre Web Services

Para exemplificar a forma como se da a comunicação entre dois web services, um que é o solicitante de algum serviço, que pode ser chamado de cliente, e outro que é o provedor de serviços.

Figura 4 – Modelo de uma arquitetura de Web Services



Fonte: (MACEDO, 2012)

Inicialmente o cliente irá, enviar uma requisição ao provedor de serviço, e este vai retornar uma mensagem tendo o arquivo WSDL, que nada mais é que a descrição, de todos os serviços disponíveis, e o que é necessário para ter acesso aos serviços, ou seja,

quais os métodos, e parâmetros que devem ser passados para os métodos. Assim que o cliente, tem essas informações, ele pode solicitar qualquer serviço que estiver disponível, e como já explicado essa comunicação acontece através do protocolo SOAP.

2.2.2 Web Crawler

Segundo (RICOTTA, 2007), um web crawler que também pode ser chamado de spiders, ou robôs, é um programa ou script automático que navega por páginas da internet a procura de informações que sejam consideradas relevantes, tais como *tags*, links, palavras chaves, dentre outras. Para que essa varredura na internet seja realizada, geralmente o web crawler recebe uma lista de links iniciais, e ao fazer a leitura da página ele coleta novos links, e adiciona a sua lista para que possa navegar também por estes. Para que o web crawler não navegue pela mesma página duas vezes, ele mantêm uma lista de páginas já visitadas.

Os web crawlers navegam a internet utilizando o protocolo HTTP, para que possam recuperar os documentos dos servidores. As aplicações possíveis vão desde atualizar bases de dados de motores de busca, como por exemplo o Google, até fazer cópias de páginas (SOUZA, 2013). Uma outra utilização, pode ser a busca de um dado específico que se encontra numa página, como foi o caso da proposta apresentada, onde o web crawler deve buscar uma informação na página que é considerada relevante.

Cada robô tem uma estratégia diferente para decidir o que visitar, e de qual forma irá fazer isso. Geralmente, os robôs, iniciam com uma lista de visitação pré determinada de links, no qual, é realizada a navegação na página, descobertos novos links que são adicionados a lista de visitação. Para que não ocorra uma página sendo visitada duas vezes, é mantido também uma lista com as páginas web já visitadas, assim, cada vez que o robô irá visitar uma nova página, é feita uma consulta na lista de páginas visitadas.

2.2.3 Data Warehouse

Um Data Warehouse é um armazém de dados que tem como finalidade guardar dados de uma empresa e mantê-los disponíveis e acessíveis para consultas. Tais consultas devem fornecer resultados claros, para que possa se estabelecer estratégias que auxiliem na tomada de decisão de um empresa (NERI, 2004).

Segundo Poletto (2008) os dados que compõe um Data Warehouse são extraídos de Bancos de Dados Operacionais, a fim de centralizá-los e padronizá-los, por meio de ferramentas de ETL, para assim permitir que sejam usados no processo de tomada de decisões. Durante o processo da criação de um Data Warehouse, o processo de integração e padronização dos dados, é visto como uma parte crítica, uma vez que estes dados podem ser originais de diferentes fontes, podendo até mesmo, terem formatações diferentes.

2.2.3.1 Processo de ETL

No processo ETL, a primeira das três etapas corresponde a realizar a leitura e compreensão dos dados que se deseja extrair. A segunda etapa diz respeito a fazer a transformação dos dados, onde há muitas transformações possíveis, como filtragem dos dados, fazer a combinação de dados de origens diferentes, cancelar dados duplicados, padronizar unidades de medidas, dentre outras.

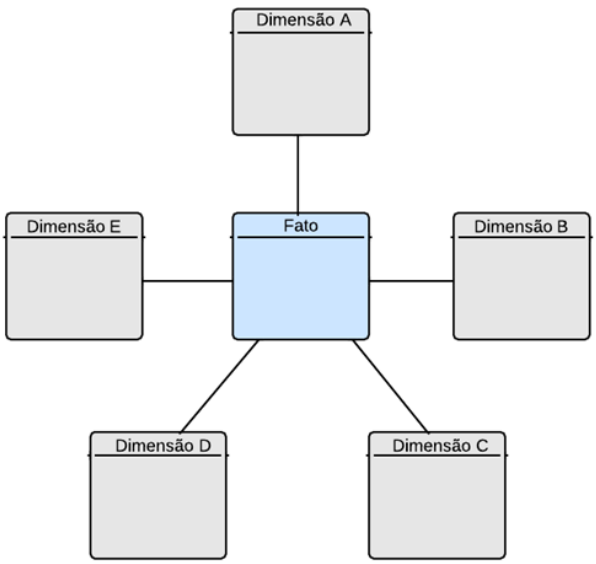
Por fim, a última das etapas a ser realizada é transportar os dados para o Data Warehouse, uma vez que os dados estão padronizados, este processo não deve apresentar dificuldades.

2.2.3.2 Modelo Estrela

Um Banco de Dados Multidimensional (CODD, 1970), ou dimensional como também pode ser chamado, tem uma modelagem conceitual, com objetivo de facilitar o processo de analise das informações da base de dados, ou seja, ele facilita as consultas aos dados em diferentes perspectivas.

Uma das formas que os bancos multidimensionais são apresentados é através do Modelo Estrela. O modelo consiste em ter uma única tabela de fatos(tabela principal) e chaves simples nas tabelas dimensões(tabelas auxiliares), onde cada dimensão é representada por uma única tabela.

Figura 5 – Modelo Estrela



Fonte: (ELIAS, 2014)

Os pontos positivos desse modelo são a eficiência, dada pelo número reduzido de junções nas pesquisas e pelas chaves simples, e facilidade de definir hierarquias.

3 Modelagem

A Universidade Federal do Maranhão está consolidada como um dos mais importantes campos de pesquisa do estado. Todo esse crescimento está baseado no tripé ensino, pesquisa e extensão, que são os pilares da universidade. Tais fatores são fundamentais para que a UFMA atinja uma boa avaliação nos testes de qualidade feitos continuamente pelo Ministério da Educação - MEC.

A universidade busca está cada vez mais inserida na comunidade através dos programas de extensão, que tentam levar um pouco do conhecimento científico para o cotidiano das pessoas que não tem acesso a universidade. E por fim, as pesquisas feitas na universidade tem um importante reconhecimento nacional, com artigos, periódicos e produções com qualis, sendo frequentemente produzidas pelos docentes.

Naturalmente todas as pesquisas feitas, geram uma enorme quantidade de produções científicas, as quais são publicadas em diversos eventos, revistas e congressos. Com publicações feitas em diversos locais, tem-se uma séria dificuldade em reaver todas essas informações. Tal trabalho tende a ser impossível de ser realizado manualmente, sem auxílio de uma ferramenta própria de pesquisa.

Uma das dificuldades encontradas por exemplo, seria encontrar todas as publicações de um docente, e fazer a classificação por ano que foi publicado. Outro problema encontrado, é manter-se o controle sobre as pesquisas mais recentes publicadas. Quando um veículo de comunicação faz uma reportagem sobre um determinado projeto, muitas vezes o projeto já foi encerrado, ou as informações não estão atualizadas.

Faz-se necessário também que seja feito todo esse controle para que a sociedade em geral, possa acompanhar o que é feito dentro da universidade. É de fundamental importância que tais dados sejam divulgados, para que se monitore o desempenho da universidade. Outro ponto importante seria a busca de parceria no meio privado, a fim de elevar a qualidade dos projetos atraindo mais investimentos.

A solução para as dificuldades citadas, seria a construção de um portal que agrupe todos esses dados. Inicialmente seria necessário manter um registro atualizado dos currículos lattes do docentes da universidade, para tanto é necessário fazer periodicamente requisições no webservice do CNPq.

A atualização da base de dados do portal deve ser feita periodicamente, sendo o período mais aconselhável de uma semana. Uma vez que é um prazo razoável para que atualizações nos currículos sejam integralizadas ao portal, e tenha um prazo suficiente para tratamento de dados inconsistentes durante a atualização.

Além dessa constante atualização, é também necessário um tratamento nos dados antes de exibi-los. Há uma etapa de preparação dos dados, para remoção de quaisquer inconsistências. A partir daí, pode-se estabelecer métricas para analisar a produtividade dos docentes, que mais a frente podem ser utilizadas para traçar metas, e por conseguinte melhorar o desempenho da UFMA.

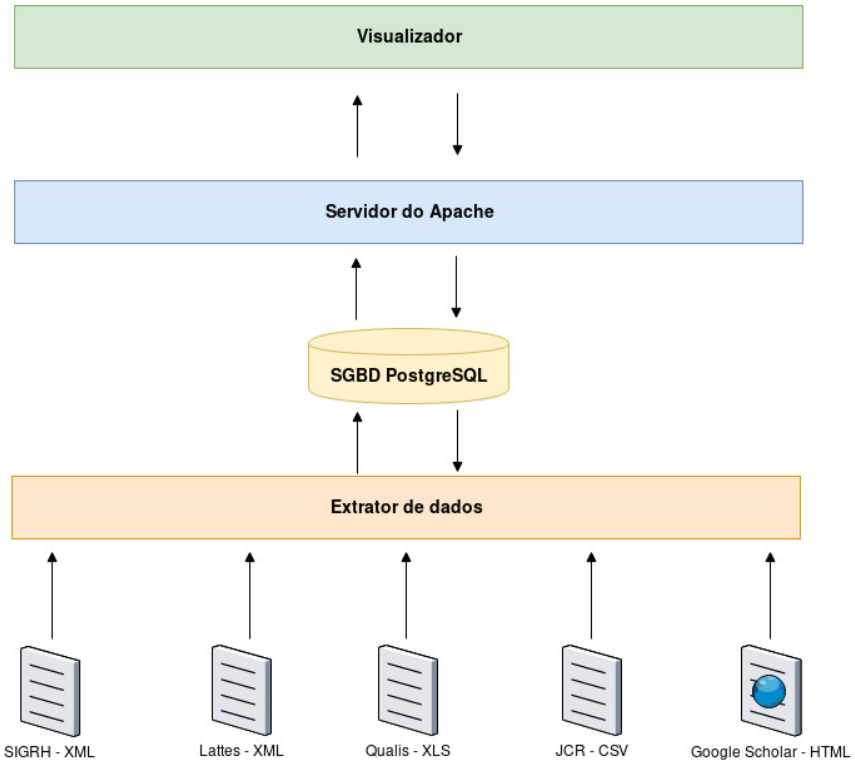
Nas seções seguintes, será feita uma explicação sobre a arquitetura geral do sistema, um resumo desde a aquisição dos dados, até como os dados são apresentados ao usuário. Após isso, é feita uma apresentação mais específica, que envolve os formatos que esses dados são obtidos, bem como as inconsistências encontradas, finalizando com uma explanação sobre o módulo de visualização.

Em seguida será apresentada, a modelagem do banco de dados da aplicação, com as devidas explicações da modelagem, e as considerações finais, onde serão explicitadas as vantagens e limitações da proposta apresentada.

3.1 Arquitetura geral

Esta seção descreve a arquitetura, fazendo a explicação do processo de aquisição dos dados, como é feito o tratamento, a modelagem do banco de dados e construção do módulo de visualização, conforme a Figura 6.

Figura 6 – Arquitetura Geral do Sistema.



Fonte: acervo do autor.

O usuário do sistema interage diretamente com o módulo de visualização, no qual são apresentados os gráficos, que são extraídos a partir dos dados que foram extraídos. A camada de visualização faz a interação diretamente com o servidor da aplicação, e este por sua vez faz a consulta dos dados no SGBD que é utilizado. Para fazer a obtenção desses dados existe a camada do módulo de extração, reponsável por realizar a comunicação, com webservice, e fazer a leitura de diversos formatos de arquivos.

Pode-se dizer de forma simplificada, que o passo inicial é buscarmos quem são os docentes da universidade a partir do webservice do SIGRH, após, termos a informação de quem são os docentes ativos, estamos aptos a buscar os dados que nos interessam sobre estes. Tendo uma lista com os dados de quem deve-se recuperar informações, pode-se assim acessar a plataforma Lattes, e o Google Acadêmico, a procura do perfil destes docentes.

Agora que temos a lista de docentes junto com o perfil de cada um, acrescentamos as bases do Qualis e JCR, que são bases reconhecidas por avaliar a qualidade das publicações do meio científico, e portanto, temos dados não só quantitativos, mas, qualitativos sobre a produção intelectual de cada um dos docentes.

E por fim, tem-se a obtenção dos dados do Google Acadêmico, onde é feita a leitura da própria página HTML, para fazer a importação para a base de dados. Cada módulo usado para realizar esse processo é explicado nas subseções a seguir com mais detalhes.

3.1.1 Módulo Extrator

O módulo extrator faz a interação com diversas fontes que contenham dados públicos. Como não existe um modelo para receber esses dados, foi necessário a implementar estratégias especificas para cada uma das fontes escolhidas, o módulo funciona de acordo com a Figura 7.

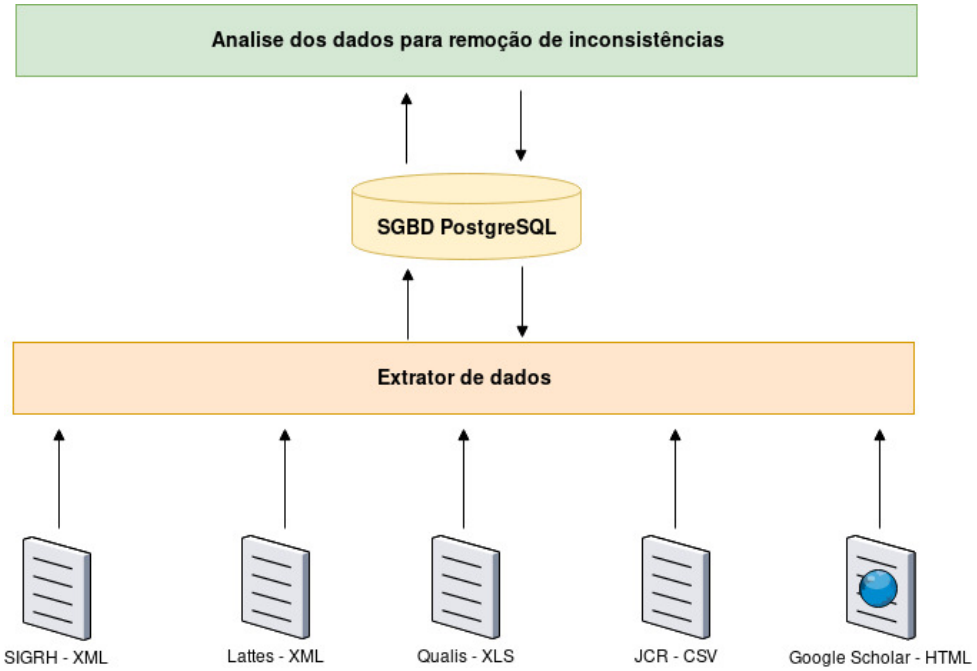
No quadro apresentado, cada uma das bases está descrita juntamente com o formato que o extrator deve ler. Como é possível observar cada uma das bases que foram utilizadas fornece um arquivo em seu próprio formato. Os múltiplos formatos que devem ser lidos fazem com que o trabalho de importação seja mais custoso.Além disso, a implementação do importador, requer mais tempo, já que não é possível fazer um importador genérico.

3.1.1.1 Integração com o SIGRH

A primeira informação que deve ser recuperada para que o sistema possa funcionar corretamente, é buscar a lista de docentes ativos da UFMA. Esta informação é crítica para todo o funcionamento correto do sistema. Felizmente essas informações podem ser recuperadas através do web service do SIGRH, conforme a Figura 8.

O extrator faz a requisição da lista de dados ao webservice do SIGRH passando suas credenciais, após a verificação que o extrator tem as credencias necessárias, o webservice

Figura 7 – Arquitetura do Módulo Extrator.



Fonte: acervo do autor.

Figura 8 – Integração com o SIGRH.

Fonte: acervo do autor.

retorna uma lista com os docentes, no formato XML. O arquivo é então lido e os dados são integrados a base de dados.

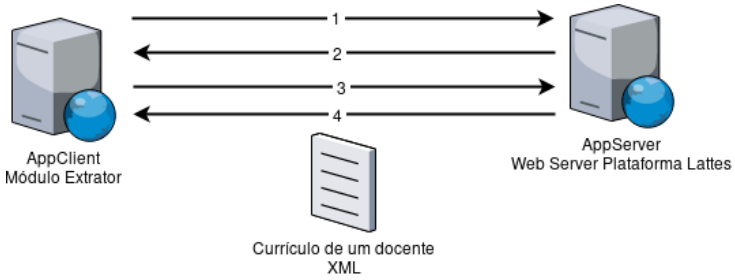
Depois que os dados são inseridos no Data Warehouse, é necessário que estes sejam analisados para correção de inconsistências, como por exemplo, nomes errados, CPF inválido, e outros. Caso tenha algum desses erros, é necessário fazer novamente a requisição da lista de docentes ativos, e somente quando os dados estiverem sem erros, essa etapa é considerada concluída.

3.1.1.2 Integração com a Plataforma Lattes

A Plataforma Lattes possui um web service que permite que aplicações externas se comunique e façam requisição de currículos cadastrados na plataforma. O extrator se comunica com o webservice da plataforma, fazendo a requisição dos currículos referentes aos docentes da UFMA.

O módulo extrator, possui um cliente web service, o qual foi chamado de AppClient

Figura 9 – Interação entre os web services.

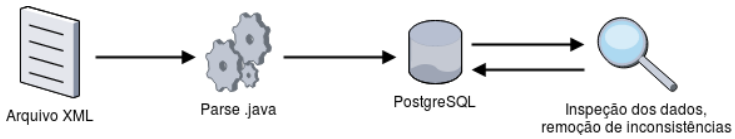


Fonte: acervo do autor.

para facilitar a explicação. 1 - AppCliente envia uma requisição para o servidor web service do Lattes, que será chamado de AppServer. 2 - O AppServer retorna então como resposta um arquivo WSDL, o qual descreve os métodos, nomes de parâmetros, endereço do serviço, como é o formato do arquivo de entrada e o formato de arquivo de saída, ou seja, tudo o que é necessário para utilizar o serviço. De posse dessa informação. 3 - então o AppClient envia uma requisição, para que o AppServer execute a operação necessária. 4 - assim que o AppServer processa a requisição ele retorna um arquivo XML com os dados existentes nos currículos dos docentes.

Uma vez que o extrator recebe os arquivos XML, ele pode fazer a leitura e importação dos mesmo para a base de dados, conforme a Figura 10.

Figura 10 – Integração com a Plataforma Lattes



Fonte: acervo do autor.

O arquivo XML retornado pelo web service da Plataforma Lattes, contém todos os dados que foram cadastrados pelos docentes tais como:

- Resumo CV
- Formação acadêmica
- Área de atuação
- Trabalhos em eventos
- Artigos publicados
- Orientações

- Produções técnicas

O currículo em XML, possui outros dados além dos que foram listados, pois, o cadastro da plataforma lattes, permite o preenchimento de diversos campos. Devido a grande quantidade de dados, é inevitável que em alguns momentos existam inconsistências, como valores errados, exemplo um campo onde está escrito a palavra NULL. Erros como esse são facilmente tratados, uma vez que esse padrão se repete, é necessário apenas, remover a palavra e deixar o campo vazio.

Por outro lado há erros que são mais difíceis de serem tratados, como por exemplo, um título de artigo escrito de forma incorreta, esse é um erro que não tem como ser tratado, pois, a quantidade de artigos é demasiadamente grande, para que seja possível verificar um por um. A correção para esse tipo de problema, é o próprio docente dono do currículo atualizar os dados corretamente na Plataforma Lattes, então, quando for feita uma nova importação, o nome será corrigido.

É importante ressaltar que os currículos devem ser constantemente atualizados, isso significa que semanalmente, o extrator deve fazer a importação dos currículos. Isso é necessário para que a base esteja sempre com os dados mais atuais possíveis.

3.1.1.3 Integração com a Base Qualis CAPES

Os dados da Base Qualis devem ser obtidos manualmente através da Plataforma Sucupira que é mantida pela CAPES. O próprio administrador do sistema deve ir na plataforma e pegar os arquivos relativos as avaliações Qualis, os arquivos são retornados no formato XLS. O módulo de extração faz a leitura automática do arquivo, e importa para da base de dados, conforme mostra a Figura 11.

Figura 11 – Integração com o Qualis

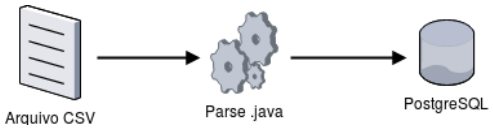
Fonte: acervo do autor.

A leitura do arquivo recebido é feita de forma automática pelo extrator, fazendo a importação dos dados para o banco de dados, e depois da importação ser concluída, é feita uma analise dos dados para remoção de inconsistências caso tenham. Diferentemente, dos currículos que devem ser atualizados frequentemente, os dados do Qualis, devem ser atualizados apenas uma vez no ano, ou quando a CAPES atualizar os qualis do periódicos.

3.1.1.4 Integração com o JCR

O processo de integração com o Fator de Impacto do JCR é semelhante ao Qualis, os índices JCR são encontrados na internet no formato de PDF, e por estarem descritos como uma listagem de periódicos, podem ser facilmente convertido para um arquivo CSV. Esse passo de conversão do formato do arquivo é feito devido a semelhança para o extrator ler o formato CSV com os outros arquivos XML e XLS, o que acontece de acordo com a Figura 12.

Figura 12 – Integração com o JCR



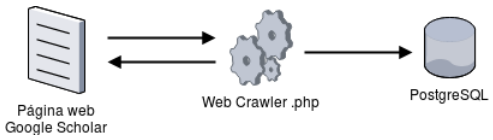
Fonte: acervo do autor.

O arquivo com os dados do JCR, são lidos pelo extrator e importados para a base de dados, sendo necessária essa importação uma vez por ano, já que assim como o Qualis, esses índices são atualizados anualmente.

3.1.1.5 Integração com o Google Acadêmico

Por fim, e não menos importante, são obtidos os índices de citações dos docentes da UFMA fornecidos pelo Google Acadêmico. Como essa plataforma não oferece nenhum meio de interação com a base de dados, foi necessário a criar um web crawler. O webcrawler faz parte do extrator e faz a navegação no Google Acadêmico a procura dos perfis dos docentes da UFMA, conforme a Figura 13.

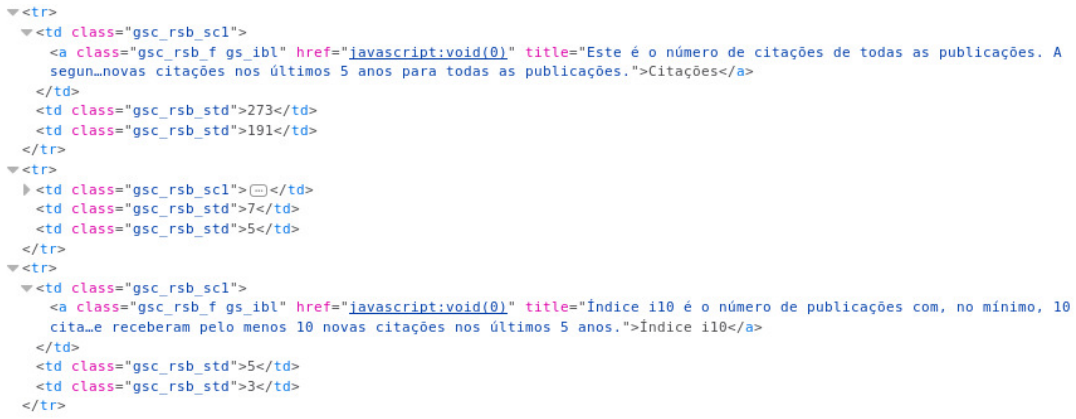
Figura 13 – Integração com o Google Acadêmico



Fonte: acervo do autor.

Para o extrator pesquisar corretamente os perfis dos docentes, é passado a ele uma lista de nomes que devem ser consultados. Quando a página é acessada, o web crawler deve procurar se existe um link para o perfil do docente que está sendo procurado, se não houver, ele passa para o próximo nome. Caso, seja encontrado um link para o perfil, a página do perfil é acessada. A informação que se deseja encontrar está marcado como uma tag, Figura 14.

Figura 14 – Estrutura HTML de um perfil do Google Acadêmico.



Fonte: (GOOGLE. . . , 2018)

As informações que serão procuradas na estrutura HTML, estão na primeira ocorrência das <td class="gsc_rsb_std» dentro de uma tag <tr>, o primeiro número encontrado é o *índices de citações*, seguido pelo *índice h* e *índice i10*. É importante destacar que nem todos os docentes possuem um perfil na plataforma. Além disso, a política da plataforma limita o web crawler a fazer aproximadamente apenas 300 requisições por dia.

Diferentemente dos outros dados importados, o Google Acadêmico, não permite a atualização dos dados de todos os docentes de uma única vez, além de que por se tratar de informações, da internet, são dados, que a todo momento podem ter atualizações, por isso o extrator deve fazer uma busca todos os dias de partes dos docentes.

3.1.2 Visualizador

O processo de visualização de dados é utilizado para facilitar a visualização destes. Grandes volumes de dados são melhor interpretados visualmente, que de forma textual, então quanto maior a quantidade de dados, mais eficaz se faz a utilização de alguma estratégia de visualização (PIRES, 2015).

O sistema de visualização dos dados, está modelado em forma de hierarquia, onde o menor nível será o detalhamento de docente, e os demais níveis são compostos por setores. É importante ressaltar que um setor pode conter ou ser contido por outro setor, sendo que o maior nível é a própria UFMA, que contém todos os setores.

Na modelagem proposta, os setores podem ser, laboratórios de pesquisa, coordenações, departamentos, pró-reitorias, centros, dentre outros. Onde, os laboratórios de pesquisas e coordenações, fazem parte de um departamento, este por sua vez faz parte de um centro de ensino, e os centros fazem parte da UFMA.

Na proposta apresentada, um docente pode fazer parte diretamente de um ou mais

setores, por exemplo, um docente, que faz parte de um laboratório de pesquisa, e também está vinculado a uma coordenação. E indiretamente o docente está ligado, aos setores que contêm os que ele está vinculado diretamente, como por exemplo, um docente que tem vinculo direto com uma coordenação, também estará vinculado indiretamente ao departamento que a coordenação faz parte.

A forma que a modelagem foi elaborada, está preparada para ser acrescentado qualquer outro tipo de setor, que as estatísticas de produtividades podem ser extraídas. Uma vez que, os dados são extraídos dos docentes que fazem parte do setor, e todos os setores possuem docentes.

O portal é apresentado ao usuário, a partir de uma visão geral, partindo para uma visão mais específica, ou seja, ao entrar no portal, são apresentadas aos usuário as métricas de produtividade de toda a UFMA, e este tem a possibilidade de pesquisar de forma mais detalhada, que são os setores que compõe a UFMA, chegando até verificar a produtividade de um docente em específico.

3.1.3 Gerador de Relatórios

O modelo apresentado possui um banco de dados com diversas informações que podem ser úteis aos docentes da universidade, a partir dessa ideia, considerou-se elaborar um módulo que permitisse a um docente criar relatórios simples, a partir dos dados armazenados.

Existem cinco tipos de relatórios que são possíveis gerar:

- Congresso por local
- Produção de periódico
- Projeto
- Orientação
- Produção técnica

Com a opção de gerar relatórios comparando a produção entre setores, ou docentes que fazem parte de um mesmo setor.

Apesar de não permitir a geração de relatórios tão elaborados, como por exemplo, gerar um relatório com os setores que apresentam um maior número de produções, com Qualis A1 e A2, este módulo permite acompanhar de forma precisa por exemplo, qual docente de um setor, tem uma maior quantidade de projetos, ou orientações por exemplo.

Uma funcionalidade incorporada ao Gerador de Relatório, foi a possibilidade de salvar o mesmo, para posterior consulta, ou ainda, exportar o relatório criado para o formato de PDF, para utilizar os relatórios fora do ambiente da tela do computador.

3.2 Modelagem de dados

A modelagem estrela utilizada no Data Warehouse foi planejada para dar uma maior eficiências nas consultas realizadas à base. A tabela escolhida como a principal, que também é chamada de tabela fato, foi a tabela currículos, que é uma tabela que contem dados básicos sobre o docente, como nome, nome em citações, número de citações no Google Acadêmico, dentre outros.

As tabelas dimensões, ou auxiliares, são tabelas que complementam as informações de um docente, como, formações, orientações, projetos, e outras. Salientando que, as tabelas auxiliares que possuem a chave para a tabela principal, desta forma a representação que um docente pode ter várias formações, orientações, projetos e etc., é perfeitamente representável por este modelo, como pode ser visto na Figura 15.

3.3 Considerações finais

Nesta seção serão considerados, aspectos críticos da proposta apresentada, isto é, problemas que são uma ameaça ao correto funcionamento do portal.

3.3.1 Manutenção

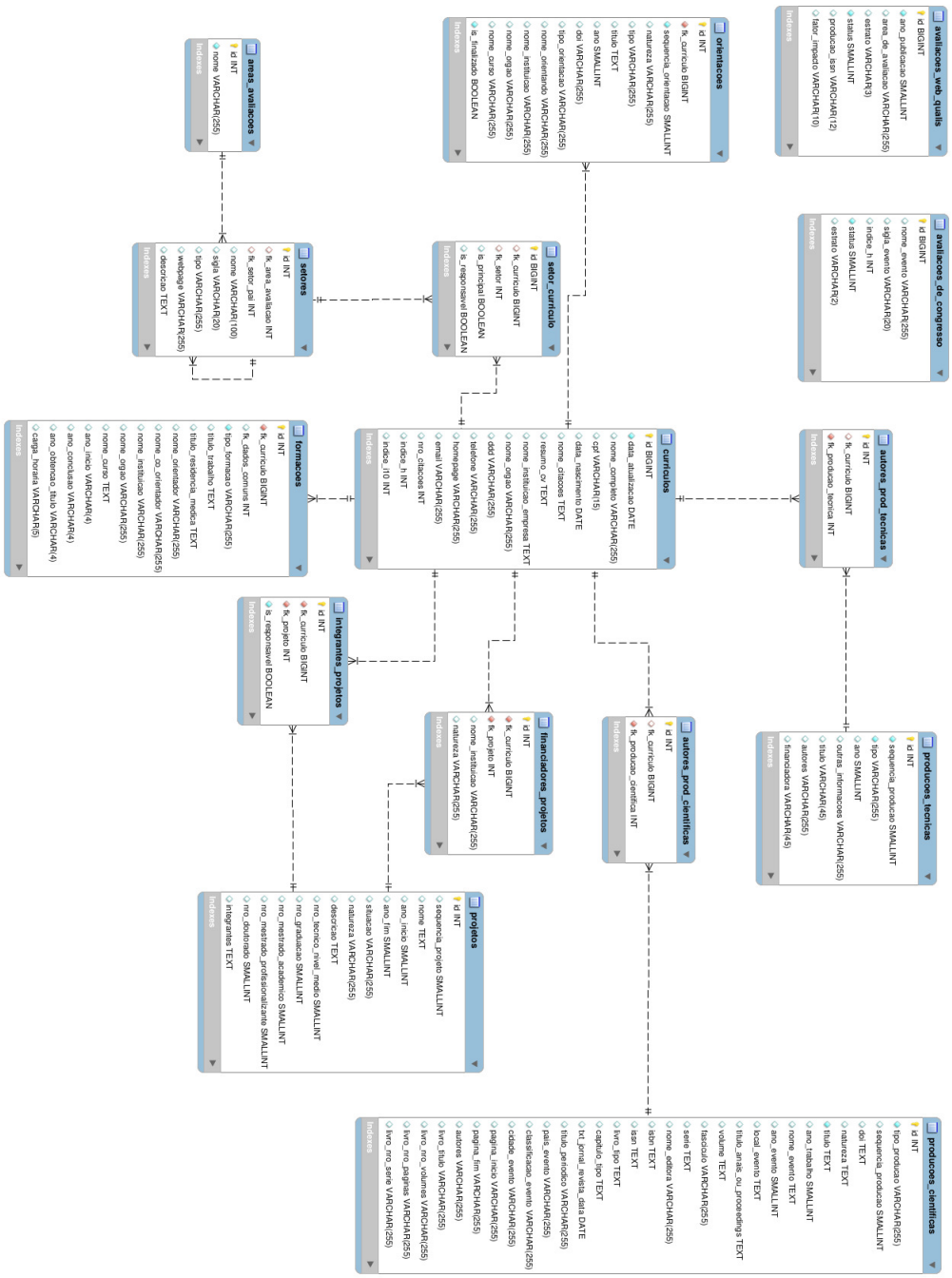
A proposta apresentada para o módulo extrator, tem alguns pontos que impactam diretamente, na longevidade dessa solução e que devem ser expostos.

O primeiro ponto que devemos considerar, diz respeito a quantidade de fonte de dados, com as quais o banco é alimentado, ao trabalharmos com 5 bases diferentes, onde cada uma requer uma estratégia diferente de aquisição de dados, a manutenção do extrator passa a ser um ponto extremamente critico, já que qualquer mudança em uma das bases terá um impacto neste.

O web crawler que retira informações do Google Acadêmico, é o modulo do extrator que está sujeito, já que este se baseia em fazer leitura de uma página da internet, a busca de uma *tag* específica. No momento que a página sofrer qualquer alteração, será necessário pensar numa nova estratégia de busca.

Um ponto crítico da modelagem apresentada, é a presença de erros nos dados, oriundos principalmente do currículo Lattes, que por ser um arquivo que tem bem mais dados que os demais, apresenta mais chances de ter inconsistências nos dados. Como já

Figura 15 – Modelagem do banco de dados



Fonte: acervo do autor.

discutido, um dos erros que ocorrem são campos com valores indevidos, com a palavra "NULL", ou outros erros pontuais semelhantes. Para esses casos, é necessário pensar numa estratégia que o extrator automaticamente desconsidere esses dados.

3.3.2 Adição de novas bases

O módulo de extração trabalha de forma independente em cada uma das bases, ou seja, a alteração no componente que insere os dados do currículo Lattes por exemplo, não interfere em nada os outros componentes.

Partindo desse ponto de vista, é possível afirmar que a adição de uma nova base de dados não iria causar nenhum impacto na extração das bases já presentes. Desta forma, é possível extrair dados de quantos lugares forem necessários, que o extrator se comportará bem às adições de fontes de dados. É válido lembrar que quantos mais bases forem utilizadas maior pode ser o esforço necessário na manutenção.

4 Resultados

Está seção visa apresentar os resultados obtidos com os modelos proposto. Toda a implementação do sistema se encontra disponível no endereço http://dev.nca.ufma.br caso queira realizar testes e consultas.

4.1 Extrator

Como resultante do modelo extrator descrito no trabalho, foi construído um módulo de extração que retira informações de bases públicas, e faz a importação para um Data Warehouse.

A ferramenta de extração consegue trabalhar com diferentes formatos de arquivos, tais como XML, XLS, CSV, e Páginas Web(HTML). As informações provenientes da Plataforma Lattes, são as que mais estão sujeitas a ocorrerem erros, pois, o pesquisador preenche essas informações sem um padrão definido, o que ocasionou em dados duplicados, ou ainda dados que vinham com caracteres indesejados.

Erros de duplicidades podem continuar a ocorrer uma vez que esses dados são de responsabilidades dos docentes, preencherem corretamente e não podem ser conferidos manualmente devido o grande esforço necessário, no entanto erros como um campo com dado que não pode acontecer como "NULL", foram corrigidos.

4.2 Visualizador

Como resultado da modelagem de um visualizador apresentado no trabalho, foi construído um portal a fim de apresentar de forma simplificada e detalhada os trabalhos que são feitos pelos docentes da UFMA. A visão foi divida de forma hierárquica conforme o modelo proposto, onde é apresentado inicialmente a visão geral da UFMA, e a partir daí, é possível ter uma visão mais detalhada, passando pelos diversos setores e chegando ao nível de maior detalhamento, que são as produções de um único docente.

Os gráficos apresentados, se dividem em cinco categorias, que são: qualis, publicações, orientações, Projetos e produções técnicas. Todas as visões, tanto da UFMA(Imagens acima), quanto os setores, e docentes, são apresentados os gráficos referentes as produções destes. A medida que a visão vai ficando mais específica, mais detalhes são apresentados, como por exemplo, na apresentação de um setor, são listados os docentes, que fazem parte deste, bem como alguns índices referentes ao grupo de docentes.

Figura 16 – Gráficos das produções dos docentes da UFMA com Qualis.

Fonte: (PORTAL. . . , 2018)

Figura 17 – Gráficos das publicações dos docentes da UFMA.

Fonte: (PORTAL. . . , 2018)

Figura 18 – Gráficos das orientações feitas pelos docentes da UFMA.

Fonte: (PORTAL. . . , 2018)

Figura 19 – Gráficos das projetos realizados pelos docentes da UFMA.

Fonte: (PORTAL. . . , 2018)

Figura 20 – Gráficos das produções técnicas dos docentes da UFMA.

Fonte: (PORTAL. . . , 2018)

Figura 21 – Listagem, e índices de docente que fazem parte de um setor

Fonte: (PORTAL. . . , 2018)

E por fim, temos a apresentação dos docentes que é nível de maior detalhamento, que são todos os dados públicos, extraídos das bases utilizadas neste trabalho.

Figura 22 – Dados pessoais de um docente, resumo cv, contatos, e índices: citações, h, i10 e JCR

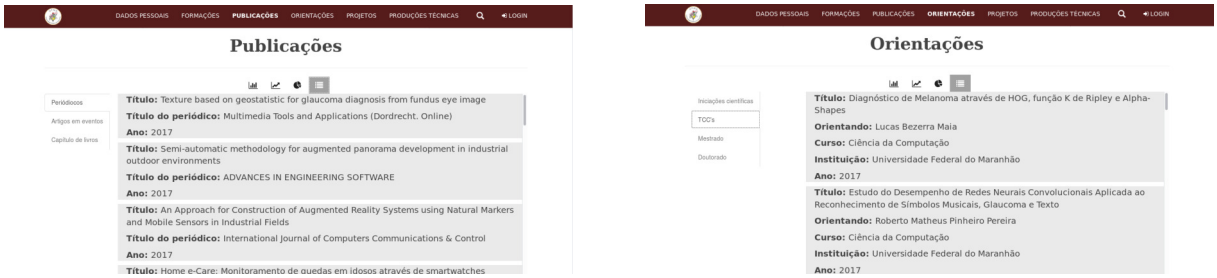
Fonte: (PORTAL. . . , 2018)

Figura 23 – Descrição de todas as formações informadas pelo docente.

Fonte: (PORTAL. . . , 2018)

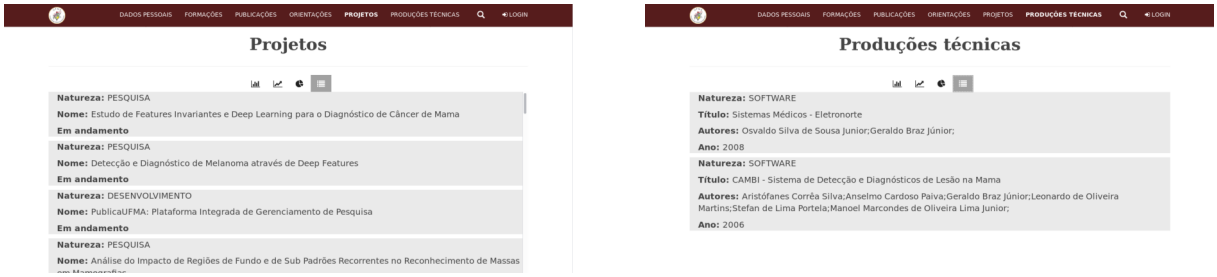
Conforme apresentado, a visualização parte de uma visão geral, que engloba toda as produções da UFMA, passando por uma visão intermediária, que são os setores, onde é adicionado a informação dos docentes que fazem parte deste, chegando ao nível com maior

Figura 24 – Publicações e orientações apresentam descrição textual no nível de docente



Fonte: (PORTAL. . . , 2018)

Figura 25 – Projetos e produções técnicas apresentam descrição textual no nível de docente.



Fonte: (PORTAL. . . , 2018)

detalhamento que é a visão de um docente. Como resultante, o visualizador está preparado para qualquer adição, de setores, ou docentes, e continuará atendendo as necessidades.

4.3 Geração de relatórios

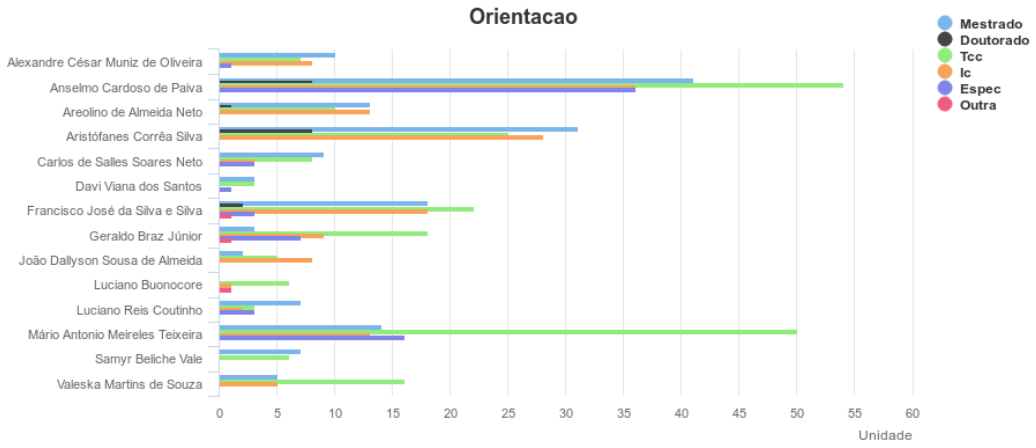
O módulo de relatório que foi descrito no capítulo anterior, se mostrou uma ferramenta eficaz para comparação, e análise da produtividade dos pesquisadores que fazem parte da UFMA, como é possível ver na Figura 26.

Para exemplificar, foi gerado um relatório com a quantidade de orientações dos docentes do Programa de Pós Graduação em Ciência da Computação. Das opções de gráfico disponíveis, que são em linha, barra e pizza, foi escolhido a opção de barra.

4.4 Extração de outras métricas

O modelo apresentado para modelagem do Data Warehouse onde são armazenados os dados, mostrou-se bem eficaz para analisar a produtividade dos docentes, permitindo a criação de diversas métricas de avaliação. Embora que no visualizador, e no gerador de relatórios, sejam apresentados apenas algumas consultas simples, é possível extrair métricas muito mais complexas a partir dos dados extraídos.

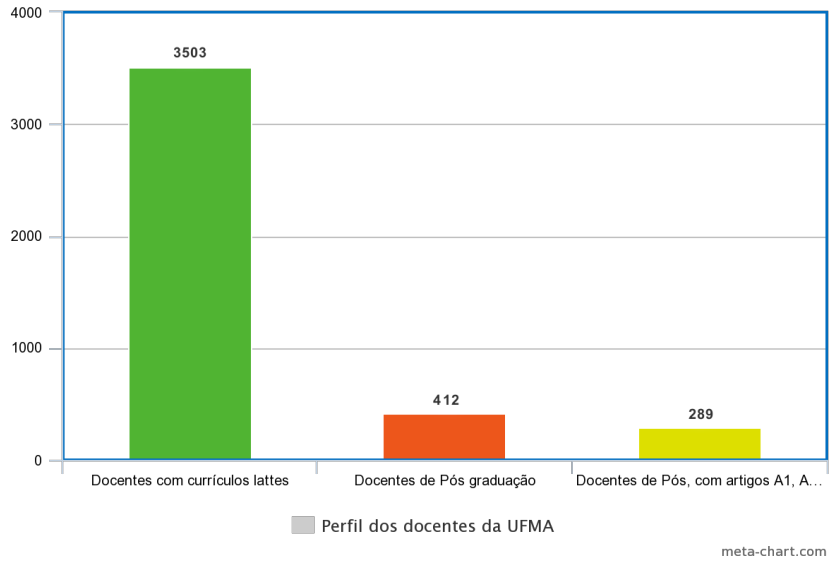
Figura 26 – Gráfico criado pelo Gerador de relatório



Fonte: (PORTAL. . . , 2018)

Para testar a modelagem do Data Warehouse, foram criadas outras métricas para avaliar o perfil dos docentes da UFMA, como é possível conferir na Figura 27, onde é possível observar quantos docentes possuem currículo na Plataforma Lattes, quantos estão vinculados a programas de pós graduação, e quantos estão vinculados a programas de pós graduação e possuem pelo menos um artigo de Qualis A1, A2 ou B1 nos últimos 4 anos.

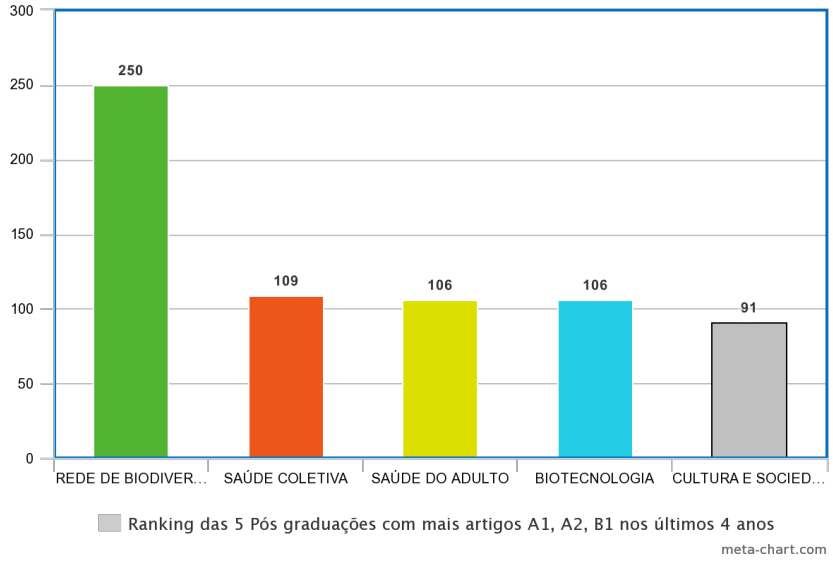
Figura 27 – Perfil dos docentes da UFMA



Fonte: (PORTAL. . . , 2018)

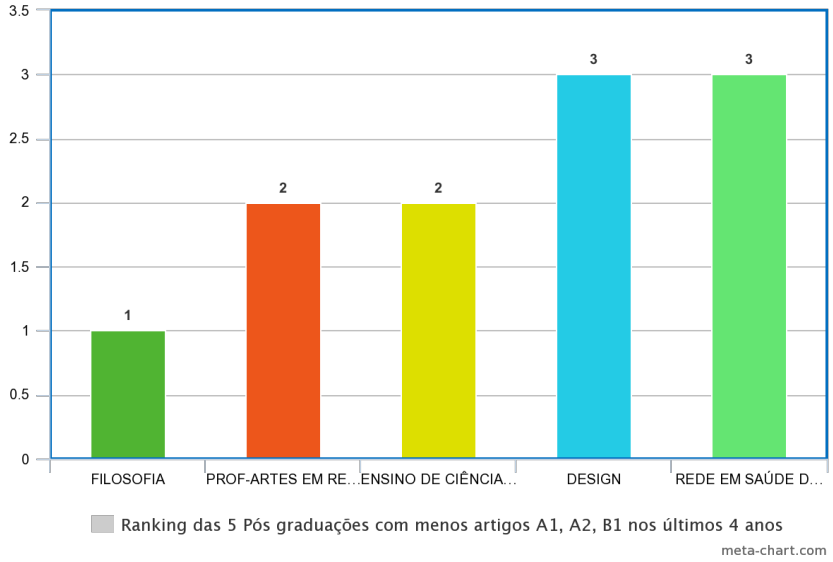
Para avaliar os programas de pós graduação, foi definido como critério quais programas possuem mais artigos com Qualis A1, A2 ou B1, limitando-se aos últimos 4 anos, a partir daí foi elaborado um gráfico com os 5 melhores desempenhos na Figura 28, e com os 5 piores desempenhos na Figura 29.

Figura 28 – Ranking das 5 pós graduações com mais artigos Qualis A1, A2 ou B1 nos últimos 4 anos.



Fonte: (PORTAL. . . , 2018)

Figura 29 – Ranking das 5 pós graduações com menos artigos Qualis A1, A2 ou B1 nos últimos 4 anos.

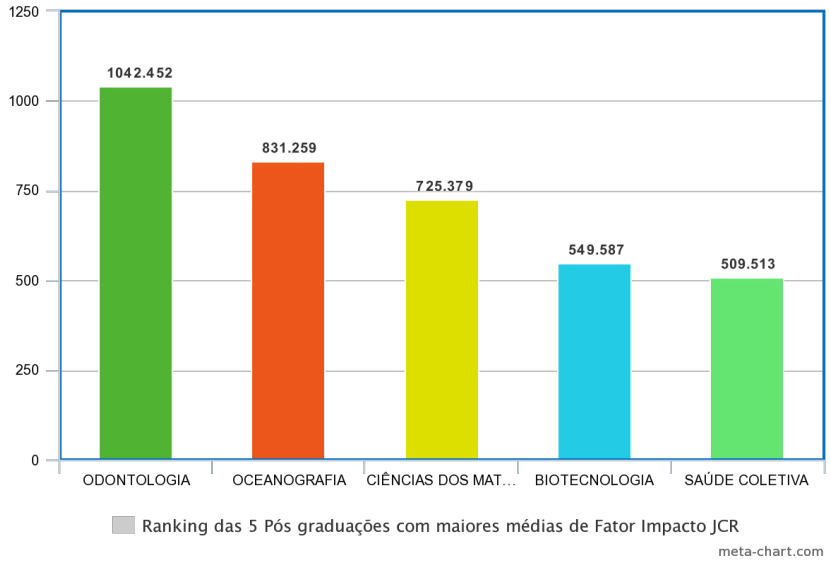


Fonte: (PORTAL. . . , 2018)

Outra métrica estabelecida para analise dos cursos de pós graduações, foi verificar quais programas possuem as melhores médias de Fator de Impacto JCR, que pode ser conferido na Figura 30.

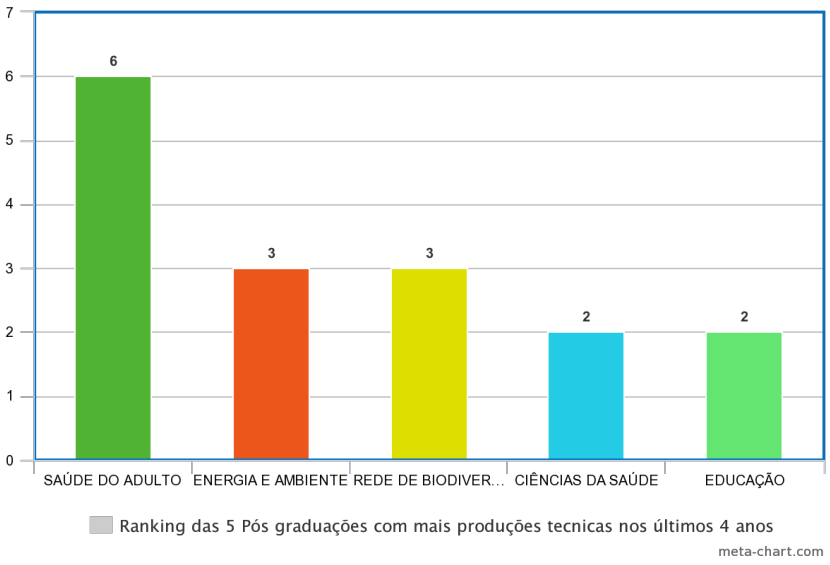
E por fim, foi buscado o ranking dos programas de pós graduação com maior números de produções técnicas nos últimos 4 anos. (Figura31).

Figura 30 – Pós graduações com melhores médias de Fator de Impacto JCR



Fonte: (PORTAL. . . , 2018)

Figura 31 – Programas de pós graduações com maior quantidade de produções técnicas nos últimos 4 anos



Fonte: (PORTAL. . . , 2018)

Os testes realizados, provaram que é possível estabelecer diversos tipos de métricas para avaliar a produtividade dos docentes, e setores da UFMA.

5 Conclusão

Este trabalho apresentou uma ferramenta para fazer a extração automática de dados das bases, SIGRH, Plataforma Lattes, Qualis, JCR, e Google Acadêmico. Ainda neste trabalho, foi apresentada, a modelagem do Data Warehouse que armazena os dados extraídos das bases citadas. Foi discutido também os módulos que permitem fazer a visualização e gerar relatórios a cerca da produtividade dos docentes da UFMA.

Foi desenvolvido neste trabalho uma plataforma que é composta, por uma ferramenta de extração de dados, provenientes de cinco bases diferentes, onde esses dados são armazenados num Data Warehouse. A modelagem do Data Warehouse, provou-se ser adequada a proposta do trabalho, pois, permite que sejam extraídos diversos índices de produtividade, tanto os que são contemplados, nas ferramentas de visualização e geração de relatórios, quanto para consultas mais complexas, que visam a extração de conhecimento para criar programas de incentivos a pesquisa.

Dentre as dificuldades encontradas durante o desenvolvimento deste trabalho algumas devem ser destacadas. O desenvolvimento do extrator concentrou as maiores dificuldades na extração de dados da Plataforma Lattes e Google Acadêmico. O currículo lattes, não exige uma padronização, cada pesquisador preenche livremente os dados inseridos, em outra palavras, cada pesquisador pode colocar a mesma informação de diferentes formas. Outra dificuldade citada é o extrair dados do Google Acadêmico, para fazer a obtenção dos índices de citações fornecidos nessa plataforma, é necessário que o docente possua um perfil na mesma, o que não acontece com a maioria dos docentes, outro fator, é a limitação encontrada em relação as requisições diárias possíveis, o que causa uma morosidade no processo de importação desses dados.

Como trabalhos futuros propõe-se que o extrator faça a descoberta e importação de produções científicas, feitas pelos docentes da UFMA, no Google Acadêmico que ainda não foram indexadas pelos autores, na plataforma Lattes. Outra proposta, é ajustar o Data Warehouse para que os docentes possam ser agrupados por área de interesse. E por fim, adicionar mais opções de filtros ao Gerador de Relatórios, para ser possível pesquisar utilizando critérios mais específicos, por exemplo índices de produções, e médias de orientações.

Referências

BRAS, P. O. Organização do currículo–plataforma lattes curriculum vitae organization–the lattes software platform. *Pesqui Odontol Bras*, SciELO Brasil, v. 17, n. Supl 1, p. 18–22, 2003. Citado na página 13.

CAPES. *Plataforma Sucupira*. 2014. Disponível em: <http://www.capes.gov.br/avaliacao/plataforma-sucupira>. Acesso em: 16 jan. 2018. Citado na página 16.

CARMONA, T. *Dominando os recursos do Google*. [S.l.]: São Paulo: Digerati Books, 2006. Citado na página 14.

CNPQ. *Painel Lattes*. 2016. Disponível em: <http://estatico.cnpq.br/painelLattes/>. Acesso em: 16 jan. 2018. Citado na página 13.

CODD, E. F. A relational model of data for large shared data banks. *Communications of the ACM*, ACM, v. 13, n. 6, p. 377–387, 1970. Citado na página 21.

ELIAS, D. *Dimensões e Fatos no contexto do Business Intelligence*. 2014. Disponível em: <https://canaltech.com.br/business-intelligence/dimensoes-e-fatos-no-contexto-do-business-intelligence-bi-18710/>. Acesso em: 14 jan. 2018. Citado na página 21.

GOOGLE Acadêmico. 2018. Disponível em: <https://scholar.google.com.br/citations?user=F_w458IAAAAJ&hl=pt-BR&oi=ao>. Acesso em: 16 jan. 2018. Citado 2 vezes nas páginas 15 e 29.

JUNIOR, V.; CARLOS, N. Procedimentos utilizados pela capes: periódicos qualis. 2016. Citado 2 vezes nas páginas 15 e 16.

LYER BALA, D. T. Engenharia reversa da máquina da google. Revista Havard Business Review, São Paulo, v. 86, n. 4, p. 27–36, abril 2008. Citado na página 14.

MACEDO, D. *Web Services*. 2012. Disponível em: <http://www.diegomacedo.com.br/web-services/>. Acesso em: 14 jan. 2018. Citado na página 19.

NERI, F. Tecnologia e projeto de data warehouse. *Editora Érica*, 2004. Citado na página 20.

OPEN KNOWLEDGE INTENATIONAL. *The Open Definition*. 2018. Disponível em: <http://opendefinition.org/>. Acesso em: 7 jan. 2018. Citado na página 12.

PAGE, L. et al. *The PageRank citation ranking: Bringing order to the web*. [S.l.], 1999. Citado na página 14.

PIRES, L. B. Integração, visualização e análise de informações eleitorais usando bancos de dados analíticos e fontes heterogêneas de grande. 2015. Citado na página 29.

PLATAFORMA Sucupira. 2018. Disponível em: <https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf>. Acesso em: 16 jan. 2018. Citado na página 17.

POLETTO, A. S. R. d. S. *Um modelo para projeto e implementação de bancos de dados analítico-temporais.* Tese (Doutorado) — Universidade de São Paulo, 2008. Citado na página 20.

PORTAL do Pesquisador da UFMA. 2018. Disponível em: <dev.nca.ufma.br/PortalPesquisador/public/>. Acesso em: 18 jan. 2018. Citado 6 vezes nas páginas 35, 36, 37, 38, 39 e 40.

RICOTTA, F. C. M. *Como os search engines funcionam.* 2007. Citado na página 20.

SILVA, G. R. da; GALANTE, R. de M. Um mecanismo de detecção de versões de páginas web para melhoria do desempenho do algoritmo de pagerank. 2008. Citado na página 14.

SOUZA, F. A. S. d. *Recuperação da informação na Web: uma análise da ferramenta de busca Google Acadêmico.* Dissertação (B.S. thesis) — Biblioteconomia, 2009. Citado na página 14.

SOUZA, P. H. *Um sistema de coleta de dados de fontes heterogêneas baseado em computação distribuída.* 2013. Citado na página 20.

W3C. *Web Services Architeture.* 2014. Disponível em: <https://www.w3.org/TR/ws-arch/#whatis>. Acesso em: 16 jan. 2018. Citado na página 18.

ZARELLI, G. B. *Como funciona o SOAP - Protocolo Simples de Acesso a Objetos.* 2012. Disponível em: <http://helpdev.com.br/2012/03/22/como-funciona-o-soap-protocolo-simples-de-acesso-a-objetos/>. Acesso em: 16 jan. 2018. Citado na página 18.