UNIVERSIDADE FEDERAL DE PERNAMBUCO
GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO
CENTRO DE INFORMÁTICA
2016.2

Geração de Dados Abertos sobre Discentes da UFPE

Priscila Thyfani Rodrigues Gomes de Melo
Trabalho de Graduação

Recife, 13 de janeiro de 2017

Priscila Thyfani Rodrigues Gomes de Melo

Geração de Dados Abertos sobre Discentes da UFPE

Trabalho apresentado ao Programa de Graduação em Sistemas de Informação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Sistemas de Informação.

BANCA EXAMINADORA:

_____

Profª Bernadette Farias Lóscio (Orientadora)

_____

Prof. Márcio Lopes Cornélio (Avaliador)

Recife, 13 de janeiro de 2017

“Pois o Senhor é quem dá sabedoria;
de sua boca procedem o conhecimento e o discernimento.”
Provérbios 2:6

AGRADECIMENTOS

Agradeço, primeiramente, a Deus por me conceder esta oportunidade e por estar comigo em todos os momentos me sustentando com sua graça.

Agradeço a meu lindo e amado esposo, Wanderson, por todo apoio e compreensão durante toda essa jornada no CIn e, principalmente, neste projeto que consumiu ainda mais do meu tempo com ele. Tô quase terminando, amor. Aos meus pais, Waudirene e Uilson, que sempre estiveram comigo nos momentos alegres e também nos mais difíceis que enfrentei, me apoiando durante o curso e fizeram o que puderam para me manter firme e viva. Andressa (Bel, Joana, ou qualquer nome que mamãe nos chame), vovós Eliude, Ivanize e Rosa, vovôs Izaias e Benedito, obrigada por tudo. Tio Toninho por todo o apoio sempre que preciso, amo o senhor! Cris e Waguinho, sem vocês eu não teria nem chegado a tempo de fazer o segundo dia da segunda fase do vestibular, fui a última a entrar. E obrigada também por me acolherem e me permitirem conviver com vocês, Thamy e Thays.

Aos meus amigos do CIn Ray, minha grande amiga e parceira em todas as horas; Feo pela amizade, apoio e trabalhos que fizemos juntos; Ruan, Sandrelle, Jorge, Chrystian, Bia, Gerson, Carlinhos, Huguinho, Bruno, Elson, Annem, obrigada por tudo. Yasmine e Mike, meus primos cuidem bem de Danielzinho. Enfim, agradeço a todos pelos momentos de diversão e ajuda mútua durante esses anos. Sentirei saudade.

Aos todos do NTI que se dispuseram a ajudar no que fosse preciso, em especial a Domingos que sempre me atendia, respondia minhas solicitações e tirava minhas dúvidas sobre as tabelas e os dados do SIG@. A Nadja por me dar a oportunidade de entrar no NTI, Shirley, Daniel e a todos da CIG pelo aprendizado e amizade. Aos meus amigos João, Milena, Suely, Ivonete, Luana e a todos os outros que são muitos, graças a Deus. A minha equipe de projeto, Filipe e Wilker, foi difícil terminamos, mas não acaba por aqui, podemos melhorar ainda mais este projeto.

A Berna pela maravilhosa orientação. Sua disponibilidade para nos guiar e tirar nossas dúvidas foram fundamentais para realização deste trabalho. Que possamos ter a oportunidade de continuar este projeto e ajudar a melhorar a educação através da publicação de Dados Abertos. Foi um privilégio trabalhar com você.

E a todos que por ventura tenha esquecido de citar, obrigada, pois ninguém conquista nada sozinho.

RESUMO

Quando Tim Berners-Lee criou a World Wide Web o objetivo era compartilhar informação. Com o passar do tempo, a Web passou a ser uma das principais formas de compartilhamento de dados utilizada em todo o mundo. Nesse contexto de compartilhamento, surgiu o conceito de Dados Abertos, contribuindo ainda mais para facilitar o compartilhamento de dados na Web. Diante disto, eclodiu a necessidade dos Governos disponibilizarem seus dados para garantir transparência e permitir que seus dados fossem utilizados por pessoas, organizações ou outras entidades para gerar benefícios para o país por meio de análises e criação de aplicações. A Lei de Acesso à Informação (LAI) foi criada para normalizar o direito constitucional de qualquer pessoa solicitar e receber dos órgãos públicos, informações públicas que estão sob o poder deles. Como as instituições públicas têm uma grande quantidade de dados armazenados em seus BDs tradicionais, percebeu-se a necessidade de encontrar uma forma de mapear os dados que estão nessas bases para uma estrutura mais adequada para a publicação na Web. Nesse contexto, este trabalho tem como objetivo a geração de dados abertos a partir de dados do SIG@ (Sistema de Gestão Acadêmica) da UFPE. Para isso, foram analisados dados disponíveis no banco de dados relacional do SIG@ a fim de extrair apenas aqueles considerados dados públicos e que poderiam ser abertos. A partir dessa análise, foram geradas visões de extração dos dados, as quais foram, posteriormente, publicadas em formato aberto. É importante notar que este trabalho de graduação faz parte de um projeto maior que engloba outros dois trabalhos de graduação, os quais tiveram como objetivo a publicação automática de dados na Web e a criação de um portal para permitir a realização de análises e geração de visualizações a partir dos dados abertos do SIG@, respectivamente.

Palavras-chave: Dados Abertos, Banco de Dados Relacional, Dados na Web, Publicação de Dados, Lei de Acesso à Informação.

ABSTRACT

When Tim Berners-Lee created the World Wide Web the goal was to share information. Over time, the Web has become one of the main forms of data sharing used around the world. In this context of sharing, the concept of Open Data emerged, contributing even more to facilitating the sharing of data on the Web. In view of this, broke out the need for governments to make their data available to ensure transparency and allow their data to be used by individuals, organizations or Other entities to generate benefits for the country through analysis and creation of applications. The Law on Access to Information (LAI) was created to standardize the constitutional right of any person to request and receive from public bodies, public information that is under their power. As public institutions have a large amount of data stored in their traditional DB's, it has become necessary to find a way to map the these data to a more adequate structure for publication on the Web. In this context, this work has the objective of generating open data from the SIG@ (Academic Management System) of UFPE. For this, we analyzed data available in the relational database of SIG@ in order to extract only those considered public data that could be opened. Based on this analysis, we generate data extraction visions, which were later published in an open format. It is important to note that this work is part of a larger project that aim at the automatic publication of data on the Web and the creation of a portal to allow the analysis and generation of visualizations of open data of the SIG@.

Keywords: Open Data, Relational Database, Data on the Web, Publication of Data, Law on Access to Information.

LISTA DE FIGURAS

Figura 1: Publicação de Dados na Web 13
Figura 2: Comparação entre a publicação de dados na Web.....15
Figura 3: Representação do Banco de Dados.....15
Figura 4: Passos para Publicação de Dados do SIG@ 23
Figura 5: Visão Exportada do Esquema Censo-Discente.....26
Figura 6: Visão Exportada do Esquema Notas.....28
Figura 7: Visão Exportada do Esquema Oferta-Componente 31
Figura 8: Modelo Entidade-Relacionamento: Unificação das Visões Exportadas 33
Figura 9: Arquivo JSON com metadados e consulta.....39
Figura 10: Indicador Desistências no Curso de Ciência da Computação.....42

LISTA DE TABELAS

Tabela 1: Descrição das entidades e relacionamentos da Visão Exportada Censo-Discente 27
Tabela 2: Descrição das entidades e relacionamentos da Visão Exportada Notas...29
Tabela 3: Descrição das entidades e relacionamentos da Visão Exportada Oferta-Componente.....32
Tabela 4: Descrição dos *Datasets*.....35
Tabela 5: Descrição dos metadados que descrevem os *datasets* 37
Tabela 6: Frequência de Atualização das Tabelas.....40
Tabela 7: Frequência de Atualização dos *Datasets* 41

SUMÁRIO

1 INTRODUÇÃO 10
1.1 Contextualização e Motivação 10
1.2 Objetivos 12
1.3 Estrutura do Documento 12
2 REFERENCIAL TEÓRICO 13
2.1 Contexto dos Dados na Web 13
2.2 Bancos de Dados e Publicação de Dados na Web 14
2.3 Desafios para Publicação de Dados na Web 16
2.3.1 Metadados 16
2.3.2 Licença de Dados 17
2.3.3 Proveniência dos Dados 17
2.3.4 Versionamento 18
2.3.5 Qualidade dos Dados 18
2.3.6 Identificação 19
2.3.7 Mecanismos de Acesso 19
2.3.8 Formatos 20
2.3.9 Vocabulários 20
2.3.10 *Feedback* 21
2.4 Considerações Finais 21
3 PUBLICAÇÃO DE DADOS ABERTOS DO SIG@-UFPE 22
3.1 Metodologia para extração e publicação dos dados do SIG@ 22
3.2 Criação das Visões Exportadas 25
3.3 Criação das Visões Publicadas 32
3.3.1 Definição de *Datasets* 34
3.3.2 Criação dos *Datasets* 36
3.4 Avaliação dos *datasets* 41
3.5 Considerações finais 42
4 CONCLUSÃO E TRABALHOS FUTUROS 43
5 REFERÊNCIAS 45

1 INTRODUÇÃO

Neste capítulo, será apresentada uma contextualização a fim de auxiliar no entendimento do tema e do motivo da realização do trabalho. Além disso, são descritos os objetivos e a estrutura do documento.

1.1 Contextualização e Motivação

A *World Wide Web* foi proposta por Tim Berners-Lee no início dos anos 1990. A ideia surgiu a partir da identificação de uma necessidade: o compartilhamento de informação. Através da observação das pessoas no ambiente em que trabalhava, ele percebeu a dificuldade que elas tinham em ter acesso à informação. Geralmente, as informações que eram necessárias para executar uma atividade estavam em computadores diferentes. Por isso, era necessário se deslocar até outros computadores e fazer *login* para ter acesso a elas.

Além disso, muitos destes computadores tinham sistemas operacionais diferentes com programas diferentes, então era necessário que o indivíduo aprendesse a utilizar cada um deles. Berners-Lee percebeu que poderia usar a Internet, que já conectava milhares de computadores, e uma tecnologia emergente chamada hipertexto, para mitigar estas duas dificuldades [1]. Lee criou a Web e as três tecnologias necessárias para o seu funcionamento: a primeira, uma linguagem de marcação para a Web chamada HTML (*HyperText Markup Language*); a segunda, é uma forma de identificação de recursos na Web chamada URI (*Uniform Resource Identifier*); e, por último, a criação de um protocolo de aplicação utilizado para requisições na Web que é o HTTP (*HyperText Transfer Protoco*l).

Com a popularização da Web, surgiu o interesse em permitir o acesso a informações públicas a partir dela, a fim de garantir o direito constitucional de acesso à informação principalmente em países que têm a democracia como forma de governo. Segundo De maneira geral, a Publicação dos Dados Governamentais não só aproxima o cidadão do estado, como também fortalece a democracia por meio do aumento da confiança gerada pela transparência do que está sendo realizado pelo governo do país, fortalecendo a democracia [2].

Diante deste contexto, o Brasil criou a Lei de Acesso à Informação (LAI), que entrou em vigor em 16 de maio de 2012. A LAI busca garantir o direito constitucional do cidadão de ter acesso a qualquer informação pública e não sigilosa de qualquer órgão público.

Atualmente, os órgãos estão adotando duas práticas para disponibilizar os dados. A primeira é através do Sistema Eletrônico do Serviço de Informação ao Cidadão (e-SIC), sistema pelo qual é possível solicitar informações públicas aos órgãos governamentais. A segunda é por meio da disponibilização dos dados em formato aberto na Web, criando portais onde o cidadão pode ter acesso aos dados.

A OKI (*Open Knowledge Internacional*[1]) traz uma definição para o termo "aberto" dentro do contexto de conhecimento. Um conhecimento é aberto quando se torna um bem comum, onde qualquer pessoa pode acessá-lo, utilizá-lo, modificá-lo e compartilhá-lo para qualquer finalidade, desde que preservem sua proveniência e sua abertura [2].

De acordo com este conceito, atualmente a melhor forma de as pessoas terem acesso a Dados Abertos é por meio da disponibilização deles na Web. Entretanto, disponibilizar um dado na Web, não implica apenas em criar um arquivo e disponibilizá-lo para *download*. A publicação de dados na Web, deve seguir padrões para que os mesmos sejam estruturados, de maneira a promover o seu compartilhamento e sua reutilização.

Para isso, os dados devem ser disponibilizados em formatos que possam ser interpretados facilmente por quem os vê e que possam ser processados por aplicações que tragam benefícios para a sociedade. Entretanto, muitas vezes a publicação é feita de forma estática com povoamento manual de tabelas e manipulação dos dados para gerar os arquivos de dados em formato aberto.

Diante da necessidade de publicação de dados na Web de forma estruturada, surgiu a necessidade de fazer um mapeamento de dados já disponíveis em Bancos de Dados Relacionais para Conjuntos de Dados na Web. Nesse contexto, este trabalho aborda tal necessidade e apresenta um estudo de caso que ilustra o mapeamento de dados relacionais para dados na Web. Especificamente, foram usados dados do Sistema de Gestão Acadêmica (SIG@) da UFPE. O SIG@ possui uma enorme quantidade de dados. O Banco de Dados, onde estão armazenadas as

[1] https://okfn.org/

informações do SIG@, é relacional, possibilitando um bom estudo de caso. O SIG@ possui mais de 700 tabelas, e para este estudo reduziu-se o domínio às tabelas relacionadas à vida acadêmica dos alunos do Centro de Informática da UFPE (CIn/UFPE).

1.2 Objetivos

Este trabalho tem como objetivo geral coletar os dados do SIG@, referentes aos alunos da UFPE, e disponibilizá-los em formato aberto para que a comunidade possa utilizá-los para a realização de análises, bem como permitir a criação de aplicações para visualização dos dados. Dessa forma, espera-se contribuir para a melhoria da gestão acadêmica.

Como objetivos específicos deste trabalho destacam-se:

- Análise das tabelas relacionais disponíveis no banco de dados do SIG@ a fim de fazer a modelagem dos conjuntos de dados a serem publicados.
- Definição dos mapeamentos entre o banco de dados relacional de origem e os conjuntos de dados a serem publicados.
- Criação de *scripts* para a geração automática dos conjuntos de dados a serem publicados na Web.

1.3 Estrutura do Documento

Após este capítulo de *Introdução*, no *Referencial Teórico* é apresentada a contextualização do trabalho, a fim de fornecer o conhecimento necessário para o entendimento do trabalho proposto. O capítulo seguinte, *Publicação de Dados Abertos do SIG@-UFPE*, é composto pelo estudo propriamente dito, desde o planejamento até análise e resultados obtidos. E, por fim, *Conclusões e Trabalhos Futuros*, o último capítulo, discute as conclusões e alguns dos desafios enfrentados, bem como contribuições deste trabalho e indicação dos futuros.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta os principais conceitos que embasam este projeto, para que o leitor compreenda melhor o que foi realizado neste trabalho.

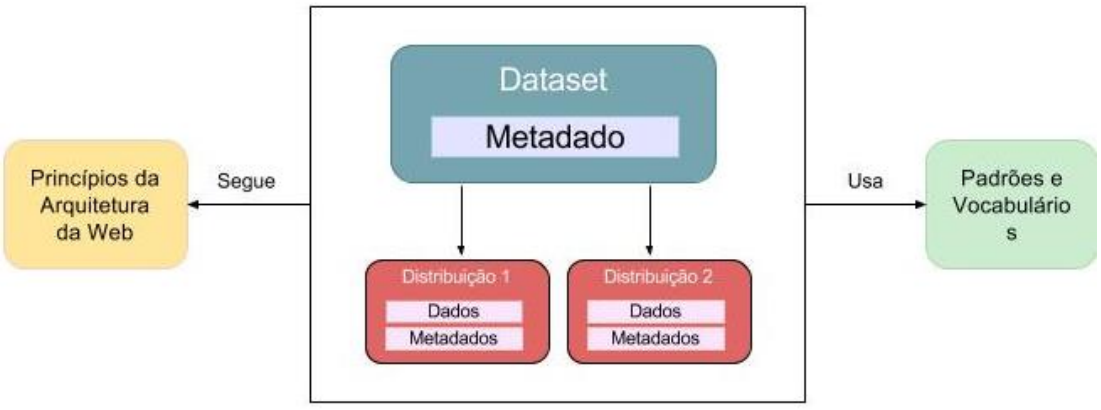
2.1 Contexto dos Dados na Web

No contexto de Dados Abertos, e dentro das especificações do DCAT[2], *Dataset* ou Conjunto de Dados é uma coleção de dados disponível para *download* em formato aberto para ser acessado e utilizado por qualquer pessoa.

Atualmente há uma lista de 35 Melhores Práticas que foram criadas por um grupo de trabalho do W3C, o *Data on the Web Best Practices* (DWBP)[3], para que os provedores de dados publiquem *datasets* que sejam úteis para os consumidores . A maioria delas são aplicadas em *Datasets* e Distribuições.

A Figura 1 apresenta o contexto de publicação dos dados na Web.

Figura 1: Publicação de Dados na Web



Fonte: Adaptada de [*World Wide Web Consortium. Data on the Web Best Practices. Data on the Web Best Practices*, 2016.]

Um *dataset* deve ser composto pelos dados e por metadados que o explicam para que os consumidores dos dados possam entender o contexto no qual o *dataset*

[2] https://www.w3.org/TR/vocab-dcat/
[3] https://www.w3.org/TR/dwbp/

está inserido. Distribuições são a forma física dos dados. A publicação deve ser feita em muitas distribuições para que os *datasets* possam ser utilizados por diversos consumidores. Além disso, para publicar dados, deve-se seguir os princípios da arquitetura da Web e utilizar padrões e vocabulários recomendados para este fim.

Os dados devem ser disponibilizados em diversos formatos para que aumente a probabilidade das pessoas acessarem os dados no formato que elas necessitam. Desta forma o compartilhamento de dados será realizado em grande escala possibilitando que uma vasta quantidade de consumidores absorvam esses dados, criem suas aplicações e façam suas análises, cada um para a finalidade desejada.

Além de disponibilizar os dados, é preciso disponibilizar os metadados que descrevem em detalhes os *datasets*, permitindo que os consumidores saibam a origem e os principais usuários dos dados.

2.2 Bancos de Dados e Publicação de Dados na Web

Segundo Lóscio [5], as plataformas mais utilizadas para a publicação e consumo de dados são os Sistemas de Banco de Dados (SBD) e a Web. Os SBDs possuem funcionalidade com foco no compartilhamento de dados, como atualização, recuperação de dados, suporte a acesso concorrente e simultâneo. Além de permitir consultas e manipulação dos dados, sem conhecimento da estrutura de armazenamento físico dos dados.

A Web na versão 1.0 era *read-only*, pois permitia apenas que as pessoas visualizassem informações colocadas por empresas e outras organizações. A versão 2.0 *read-write* possibilitou interação entre usuários como, por exemplo, um comentário em um blog ou rede social. Tim Berners-Lee juntamente com o *World Wide Web Consortium (W3C)*[4] idealizaram a versão 3.0, conhecida como Web Semântica, que viabiliza a interação não só entre humanos, como também entre máquinas com uma linguagem interpretável por ambos [6].

Este trabalho está diretamente relacionado à versão 3.0, pois quando os Dados Abertos são publicados pelos provedores, os consumidores devem conseguir entender o conjunto de dados e as aplicações implementadas por eles devem conseguir processar os dados.

[4] https://www.w3.org/

Tanto os Sistemas de Bancos de Dados convencionais, quanto Dados Publicados na Web apresentam mecanismos de armazenamento e compartilhamento de dados. Mas, existem diversas diferenças entre as duas plataformas, as quais estão descritas na Figura 2.

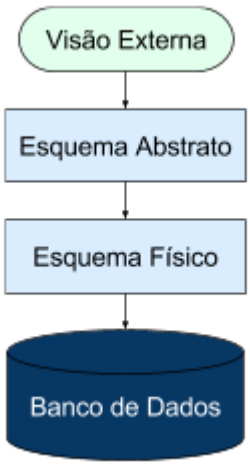
Figura 2: Comparação entre a publicação de dados na Web

Característica	Bancos de Dados Convencionais	Publicação de dados na Web
Organização dos Dados	Estruturado	Estruturado, semiestruturado ou não-estruturado
Esquema	Fixo	Flexível ou ausente
Mecanimos de Acesso	Interfaces Padronizadas, Drivers e Linguagens Declarativas	Requisições HTTP e Serviços em formato *Web Service* ou *Rest Service*
Controle de Acesso	Controlado	Sem controle de acesso para consumo, mas controlado para publicação
Tipos de Consulta	Multiatributos, faixa de valores, agregações e junções	Dependem da forma de publicação de dados
Controle de Concorrência	ACID	BASE ou inexistente
Escalabilidade	Implementação Complexa	Dependem do modelo de concorrência adotado

Fonte: Lóscio (2015, p. 41].

Um ponto de destaque entre as diferença apresentadas diz respeito ao “Esquema”. Um esquema é uma descrição do Banco de dados de acordo com um modelo [6] podendo ser abstrato ou físico. A Figura 3 demonstra a ligação dos esquemas com o Banco de dados.

Figura 3: Representação do Banco de Dados



Fonte: Elaborada pela autora.

O Esquema Abstrato descreve superficialmente a estrutura e operações do banco, como entidades e relacionamentos, utilizando um modelo de dados como referência, como um Esquema Relacional ou um Esquema Orientado a Objetos. O Esquema Físico contém a descrição interna e detalhes de como é feito o armazenamento físico dos dados. Também existe a possibilidade de personalizar um Esquema por meio de uma Visão Externa, que descreve apenas uma parte do Banco, a qual o usuário ou grupo de usuários tem interesse [7].

Como na Web, muitas vezes o Esquema é flexível ou nem existe, um grande desafio da publicação em muitas instituições é como extrair os dados de um Banco Relacional totalmente estruturado, e colocá-los em uma estrutura que permita a publicação dos dados na Web de forma eficiente.

2.3 Desafios para Publicação de Dados na Web

A Web é tratada como uma Plataforma Aberta de compartilhamento de Dados. O movimento *OWP*[5] *(Open Web Platform)* permite que qualquer pessoa implemente uma aplicação para a Web sem precisar de aprovação ou pagamento de taxa [4].

Entretanto, a flexibilidade da Web traz desafios que precisam ser superados para que haja qualidade e segurança no compartilhamento, publicação e reutilização dos dados.

Lee et al. 2015 [8] apresenta alguns estudos de caso, de onde são extraídos alguns desafios para a publicação de dados a partir de situações reais. Alguns desafios existentes no documento são abordados neste trabalho e apresentados nesta sessão.

2.3.1 Metadados

São dados utilizados para descrever outros dados. Eles podem tanto descrever um conjunto de dados como um todo, quanto um dado específico do conjunto. Os Metadados devem ser facilmente compreendidos por usuários e estar

[5] *https://www.w3.org/wiki/Open_Web_Platform*

em formato processável por máquina. Eles podem seguir um padrão, como o que foi definido pelo DCAT[6], ou não seguir estrutura nenhuma. Seguindo ou não uma estrutura, os Metadados devem ser flexíveis por causa da diversidade na forma como os dados são publicados. Mas, ainda não há uma concordância sobre quais são os Metadados que devem estar presentes nas publicações dos *datasets*.

2.3.2 Licença de Dados

Uma Licença de Dados é como um contrato que autoriza a utilização do dado. Segundo a *Open Knowledge International*[7] (OKI), no caso de Dados Abertos, a Licença serve para garantir que "qualquer indivíduo tenha a liberdade de usá-lo, reutilizá-lo e redistribui-lo". Ou seja, é uma concessão formal para a utilização dos dados. Algumas licenças foram criadas para este fim como a *General Public License* (GPL), *Open Data Commons* e o conjunto de licenças da *Creative Commons* (que possui apenas as licenças CC-Zero, CC-BY e CC-BY-SA consideradas abertas, pois as outras violam os princípios de Dados Abertos).

A legislação de cada país também pode impor restrições, como a proibição de divulgação de dados pessoais como CPF, pois a publicação desses dados iria ferir a constituição e a LAI. A licença deveria fazer parte do *dataset* para possibilitar o processamento por máquinas e garantir o cumprimento dos princípios de Dados Abertos e suas restrições legais na publicação. Porém, como não há uma forma genérica para este fim, ainda há uma complexidade na verificação automática dessas licenças.

2.3.3 Proveniência dos Dados

Proveniência são as informações que descrevem a origem dos dados e seus usuários. Capturar e disponibilizar informações de proveniência são importantes porque, caso o provedor dos dados não tenha contato com o consumidor, é bom que esteja documentado, de forma estruturada, o processo de criação e origem dos dados, permitindo o processamento e verificação.

[6] https://www.w3.org/TR/vocab-dcat/
[7] https://okfn.org/

O W3C possui alguns documentos que auxiliam na definição da proveniência e troca de informações num ambiente heterogêneo como a Web, os quais estão disponíveis no PROV-primer[8]. Contudo, ainda há uma falta de concordância e detalhes da proveniência dos dados que devem ser capturados.

2.3.4 Versionamento

Os dados que são publicados na Web, geralmente, sofrem alterações e/ou atualizações. Por este motivo, é importante que haja um controle de versão e que este esteja disponível ao usuário para que ele saiba quais versões de um *dataset* estão disponíveis e quais as diferenças entre elas. Além disso, é necessário que as informações de versionamento estejam condizentes com a realidade, para que o usuário saiba exatamente quando será necessário acessar os dados novamente e/ou fazer atualizações nas aplicações criadas com esses dados.

O grande desafio é criar um conceito de versionamento para Dados Abertos que seja aceito pela comunidade responsável e padronizar a forma de descrição e publicação das informações de versionamento.

2.3.5 Qualidade dos Dados

Quando uma aplicação é criada utilizando Dados Abertos, geralmente coletados em portais disponíveis na Web, o consumidor verifica se os dados estão organizados, se está faltando informação, se há dados inconsistentes, entre outras possibilidades. Isto acontece porque muitas instituições não se preocupam com a qualidade dos dados que estão disponibilizando.

A qualidade é definida pelo “grau de adequação ao uso”. Atualmente há alguns trabalhos como os de [Wang e Strong 1996] e [Zaveri et al. 2013], que se preocupam com a disponibilização e avaliação da qualidade da informação através de critérios de descrição e avaliação [19].

O grande desafio é disponibilizar informações de qualidade de forma estruturada e em variados formatos. Atualmente, ainda não há uma solução

[8] https://www.w3.org/TR/prov-primer/

amplamente aceita pela comunidade sobre dimensões, métricas e forma de avaliação da qualidade dos dados.

2.3.6 Identificação

No contexto de Dados na Web, a identificação deve ser baseada em URIs, mas não qualquer URI, pois há alguns que não acessam recurso na Web. Para acessar recursos da Web, deve-se usar a URI HTTP, caso contrário o recurso não poderá ser encontrado.

Para a publicação de dados faz-se necessário criar boas URIs para identificação do recurso associado. Em um trecho presente em Lóscio [4], a autora apresenta considerações de Sauermann para criar boas URIs: I) Usar URIs HTTP para tudo, tornando-as passíveis de serem “dereferenciadas”; II) Evitar URIs com detalhes de implementação ou do ambiente em que estão publicadas. A URI http://www.bad.url.org:8080/~fulano/index.php é um exemplo do que deve ser evitado, pois possui detalhes sobre o autor, a porta 8080 usada em seu ambiente de publicação e o script implementado em PHP necessário para sua execução e III) Manter as URIs estáveis e persistentes, pois a alteração das URIs pode levar à quebra de links já estabelecidos, criando um problema para a localização de recursos.

A principal dificuldade é o uso de URIs persistentes na identificação de recursos na Web. Nesse contexto, o grande desafio é garantir que os recursos sempre poderão ser utilizados, sem nenhum tipo de interferência, por exemplo. Além disso, existem questões em aberto relacionadas à identificação de *datasets* e o versionamento deles.

2.3.7 Mecanismos de Acesso

A forma mais utilizada para publicação de dados é converter os dados em formatos abertos e disponibilizá-los para *download* em páginas HTML. Mas há várias formas de acesso aos dados, como o acesso a todos os dados do *dataset* ou a apenas algum específico por meio de uma requisição.

Uma questão importante de discussão é como ter um mecanismo de acesso que seja dinâmico e frequentemente atualizado, para mitigar o problema de o provedor ter que fazer manualmente a atualização dos conjuntos de dados. APIs têm sido criadas com esta finalidade, mas ainda é um grande desafio.

2.3.8 Formatos

Para garantir que os dados disponibilizados serão úteis para os consumidores, uma boa prática é disponibilizar os *datasets* no maior número de formatos abertos possível. Pois se o *dataset* não estiver no formato que o consumidor deseja, pode tornar-se inútil para ele.

Muitos *datasets* são publicados em formatos semi-estruturados, mas a ausência de metadados estruturais e alguns erros na estrutura dos dados precisam de intervenção do consumidor para só depois dos erros corrigidos, os dados serem processáveis de forma automática pelas máquinas e aplicações.

Os *datasets* devem ser publicados em formatos que não gerem obstáculos no processamento dos mesmos. Formatos como CSV, JSON e XML são indicados para publicação de dados, mas é importante evidenciar que a publicação em formatos variados, pode resultar em redundância e também inconsistência devido ao aumento dos dados armazenados.

2.3.9 Vocabulários

A interoperabilidade semântica é um grande desafio quando os dados se originam de fontes diferentes. Há uma preocupação quanto à apresentação dos conceitos que formam o domínio. Por isso o reuso, criação e descrição dos vocabulários tem um papel fundamental na publicação de dados. Eles auxiliam tanto no processo de especificação de metadados no que se refere às informações de proveniência e de qualidade, como também da descrição de esquemas dos dados.

Segundo [Garshol 2004], os vocabulários são classificados de acordo com o nível de formalismo. Podem ser:

- Vocabulários controlados: Lista de termos onde cada termo é um nome que corresponde a um conceito específico;

- Taxonomias: Classificação que organiza os termos de um vocabulário controlado em hierarquia;
- Thesaurus: Permitem classificar um termo como preferido ou mais importante em relação a outro termo do vocabulário
- Ontologias: Diferente dos anteriores que possuem relacionamentos fixos, as ontologias permitem relacionamento entre quaisquer dados.

A escolha do vocabulário a ser utilizado é uma tarefa difícil, principalmente quando não são capazes de descrever o domínio, precisando assim criar um novo vocabulário ou estender um existente.

2.3.10 *Feedback*

A publicação de Dados Abertos envolve diversos níveis de consumidores no que diz respeito a *expertise*. O nível vai do iniciante, que está começando a desenvolver e precisa de dados estruturados para criar suas aplicações, até o nível avançado, que já possui experiência em desenvolvimento e tratamento de dados em formato aberto.

Por isso, faz-se necessário que o provedor tenha um *feedback* dos consumidores quanto à qualidade dos dados, estrutura e formato dos dados publicados.

A Publicação de Dados Abertos não é recente, mas há alguns anos esta prática tem crescido bastante em todo o mundo.

2.4 Considerações Finais

Neste capítulo foram apresentados conceitos fundamentais para o entendimento e realização deste trabalho. Tanto o contexto quanto os desafios enfrentados para a publicação dos dados na web, contribuíram para a compreensão da necessidade de seguir boas práticas onde além do fornecimento de dados confiáveis, a utilização de metadados para descrever os *datasets*, e a publicação em diversas distribuições e formatos são fundamentais para que consumidores utilizem o *dataset* criado em suas aplicações e análises.

3 PROCESSO PARA PUBLICAÇÃO DE DADOS ABERTOS DO SIG@-UFPE

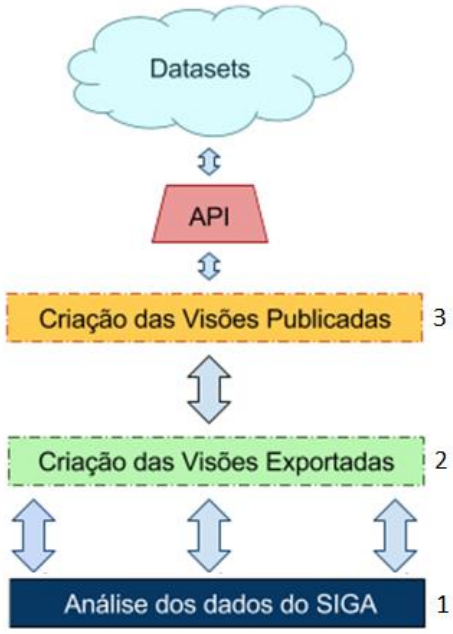
Para este estudo, utilizamos um conjunto de dados reais extraídos a partir do banco de dados do Sistema de Gestão Acadêmica (SIG@) da Universidade Federal de Pernambuco (UFPE). O SIG@ é um sistema que abrange diversas áreas da gestão acadêmica, possuindo mais de 700 tabelas de dados. Por isso, decidimos focar, inicialmente, nas tabelas relacionadas à vida acadêmica do aluno. Pois, durante entrevistas realizadas com profissionais do NTI e gestores acadêmicos, um problema muito citado foi a dificuldade em se conseguir informações relacionados ao desempenho dos alunos, como “Índice de reprovação em disciplinas”, “Índice de Evasão”, entre outros indicadores que são necessários para a melhoria da gestão acadêmica.

Os dados coletados a partir do SIG@ podem ser divididos em duas categorias de dados: Dados Sigilosos e Dados Públicos. Este estudo consiste em Publicação de Dados Abertos dos Dados Públicos que podem ser úteis para a Gestão Acadêmica da UFPE. Assim, foi necessário separar e extrair apenas os dados que podem ser divulgados na Web. Para isso, foram criadas visões das tabelas do Banco de Dados relacionadas à vida acadêmica do aluno. A seguir, a metodologia adotada para a publicação dos dados do SIG@ será apresentada.

3.1 Metodologia para extração e publicação dos dados do SIG@

A Figura 4 representa os passos utilizados para a publicação dos dados do SIG@ em formato em aberto, os quais são descritos a seguir.

Figura 4: Passos para Publicação de Dados do SIG@



Fonte: Elaborada pela autora.

Passo 1: Análise dos Dados do SIG@

Um Sistema de Gestão armazena dados sobre diversas atividades realizadas por uma organização, os quais podem ser sigilosos ou não. Sendo assim, o primeiro passo para a publicação dos dados do SIG@ em formato aberto consistiu em analisar os dados para decidir quais são de fato públicos e podem ser disponibilizados como Dados Abertos.

Devido a enorme quantidade de atividades geridas pelo SIG@, para este trabalho resolveu-se reduzir o escopo para as tabelas relacionadas à vida acadêmica do Aluno. Essas tabelas estão distribuídas em três esquemas: Censo Discente, Notas e Oferta. A partir dos esquemas, foi feita a análise e classificação dos dados em Aberto e Sigiloso, pois na UFPE não havia um documento de Classificação dos Tipos de Informação.

Além disso, fez-se necessário estudar a LAI (Lei de Acesso à informação)[9] e realizar uma pesquisa em Portais de Dados Abertos de instituições com estrutura

[9] http://www.acessoainformacao.gov.br/

semelhante à da UFPE, como o da UFRN[10] a fim de definir quais os dados contidos nas tabelas poderiam ser publicados.

Também foi feito um estudo dos esquemas relacionais do SIG@ a fim de compreender a forma como os dados foram organizados e os relacionamentos existentes com o objetivo de facilitar a seleção dos dados que seriam publicados.

Passo 2: Criação das Visões Exportadas

Após a análise dos dados, foi criado um conjunto de Visões Exportadas, de tal maneira que cada Visão Exportada representasse uma porção dos dados a ser publicada em formato aberto. Para isso, foi necessário criar uma *visão* de cada tabela do banco de dados, que estava relacionada aos dados dos alunos, modificando a nomenclatura de alguns atributos, a fim de manter a segurança dos dados, pois alguns deles dizem respeito à uma informação restrita, bem como para facilitar a compreensão dos metadados. Por exemplo: na tabela original o nome do atributo é “ps_nm”, ao modificá-lo na visão exportada passa a ser “nome_pessoa”.

Nos casos em que a chave-primária do banco original é sigilosa, foi necessário definir uma outra chave-primária a ser usada na visão exportada. Digamos que o número do RG, “numero_RG”, seja uma chave-primária da tabela “Pessoa” no banco original. Como o número do RG é um dado sigiloso, deve-se definir um outro atributo da tabela como chave-primária da tabela “Pessoa_Exportada” nas Visões Exportadas do banco de dados.

Passo 3: Criação das Visões Publicadas

Após a criação das visões exportadas, o próximo passo consistiu em utilizar engenharia reversa partindo do esquema lógico utilizado nos esquemas do SIG@ e nas Visões Exportadas, para um Esquema Conceitual e de mais alto nível. Essa transformação foi necessária porque os *datasets* normalmente são desnormalizados, ou seja, os dados relacionados a um mesmo conceito, em geral, são disponibilizados em conjunto.

A seguir, as etapas 2 e 3 da metodologia adotada serão detalhadas.

[10] http://dados.ufrn.br/

3.2 Criação das Visões Exportadas

Os SIG@ possui mais de 700 tabelas. Sendo assim o NTI utiliza visões internas para gerar esquemas de acordo com a necessidade. Por exemplo: O MEC(Ministério da Educação) todo ano publica o Censo-discente. Por isso, todas as IFES(Instituições de ensino superior) enviam um relatório com as informações necessárias. Então o banco do SIG@ um esquema relacional na forma de visão com as tabelas necessárias para gerar o Censo-discente.

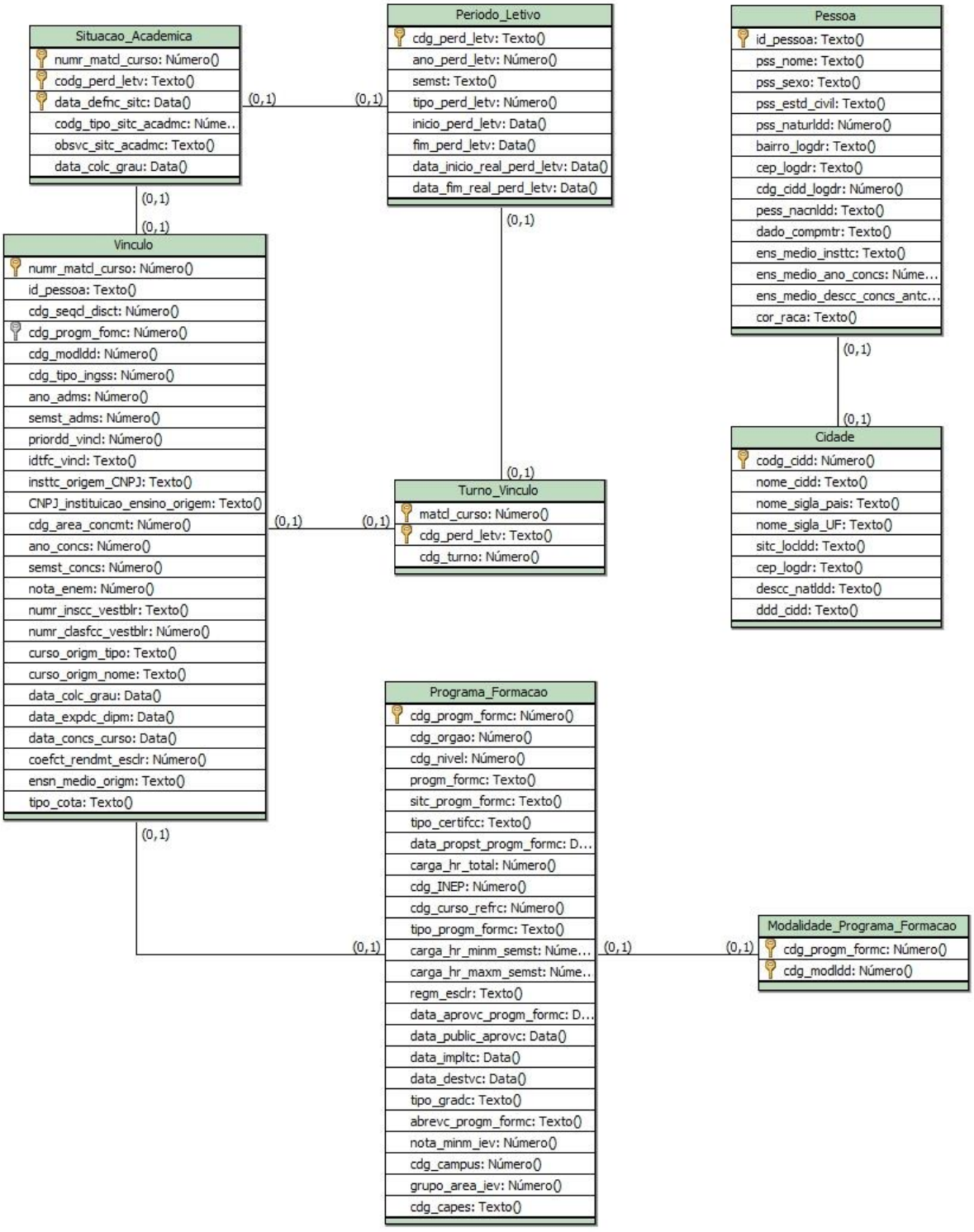
Diante disto, a criação das Visões Exportadas refere-se à criação de uma Visão Exportada para cada esquema relacional que contém informações sobre os alunos da UFPE. Na criação dessas visões, todas as tabelas presentes nos esquemas relacionais do Banco de Dados original são mantidas. Entretanto, todas as tabelas e os atributos são renomeados e só permanecem os atributos que podem ser publicados.

Como o esquema relacional do banco de dados do SIG@ é sigiloso, as imagens não foram colocadas neste documento. Porém, as visões exportadas de cada esquema são apresentadas nas Figuras 5, 6 e 7.

A Figura 5 representa a Visão Exportada do esquema Censo-Discente. É importante ressaltar que a tabela Pessoa tinha como chave-primária um atributo cujos valores não poderiam ser publicados, pois tratavam-se de dados sigilosos. A fim de substituir a chave primária original, foi criada uma nova chave_primária para esta tabela, denominada “id_pessoa”. O novo valor para a chave foi inserido como a partir de um procedimento que percorre as tuplas a fim de coletar as informações definidas para criação da chave “id_pessoa”.

Na visão exportada, todas as tabelas que se relacionam com a tabela Pessoa têm agora como chave estrangeira essa nova chave_primária (“id_pessoa”). Isto é feito também no esquema Notas.

Figura 5: Visão Exportada do Esquema Censo-Discente



Fonte: Elaborada Pela autora.

A Tabela 1 descreve cada tabela presente no Esquema da Visão Exportada Censo-Discente.

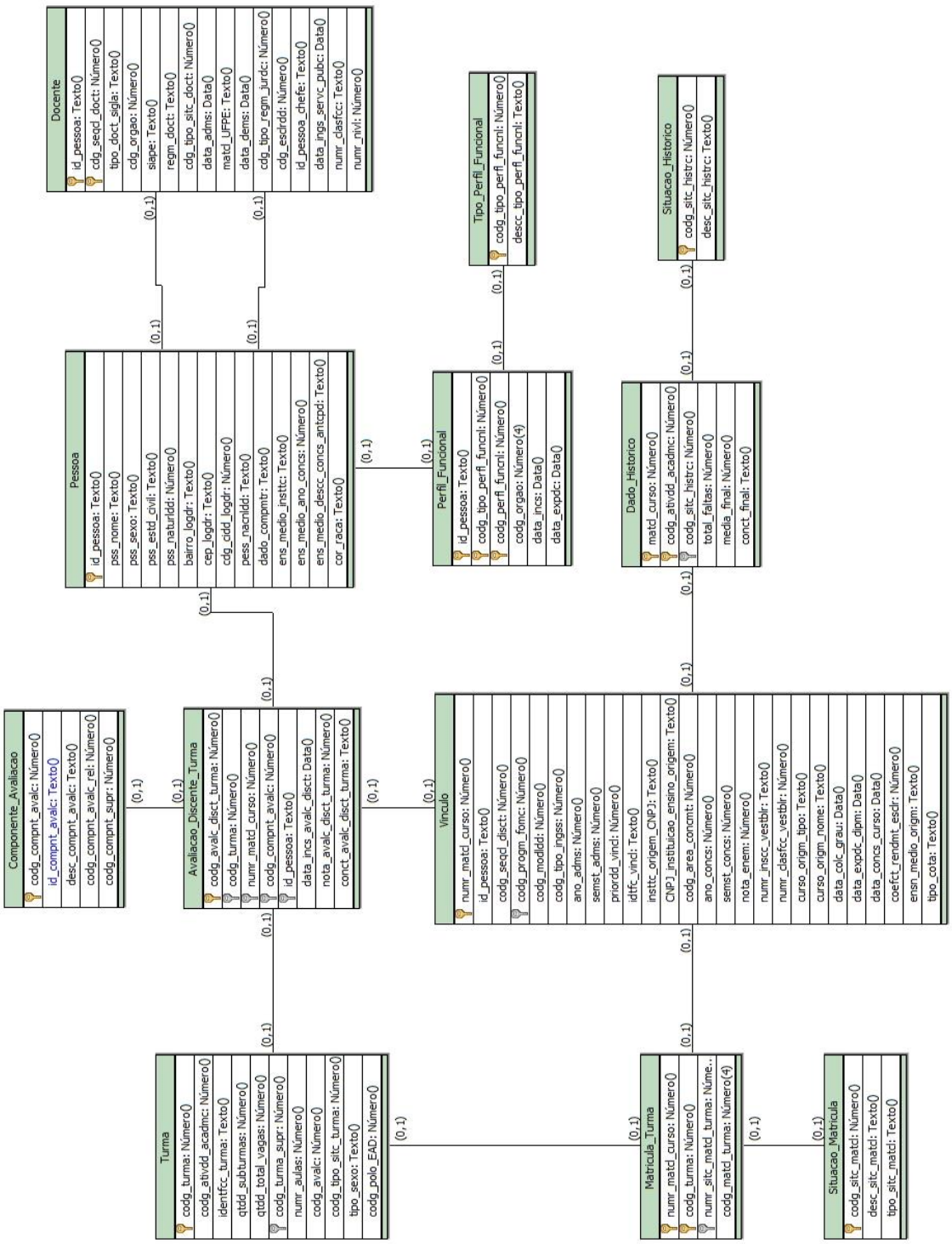
Tabela 1: Descrição das entidades e relacionamentos da Visão Exportada Censo-Discente

Tabela	Descrição
Situação_Acadêmica	Situação do aluno no período(Exemplo: Matriculado, Trancado, entre outros).
Vínculo	Tabela que contém os dados dos alunos
Período Letivo	Tabela com informações sobre o Período que o aluno está cursando ou cursou (Exemplo: 2016.2).
Turno_Vínculo	Turno em que o aluno estuda.
Programa_Formação	Informações sobre o curso do aluno.
Modalidade_Programa_Formação	Modalidade do curso que pode ser presencial, à distância, sequencial, entre outros. Há cursos que tem mais de uma modalidade.
Pessoa	Contém informações sobre as pessoas (Alunos, Docentes, Servidores).
Cidade	Complemento de informações sobre a pessoa, como cidade natal, cidade onde a pessoa mora, entre outras.

Fonte: Elaborada pela autora.

A Figura 6 representa a Visão Exportada do Esquema Notas, que contém informações sobre as turmas, discentes, docentes e avaliações.

Figura 6: Visão Exportada do Esquema Notas



Fonte: Elaborada pela autora.

A Tabela 2 descreve cada tabela presente no Esquema da Visão Exportada do esquema Notas.

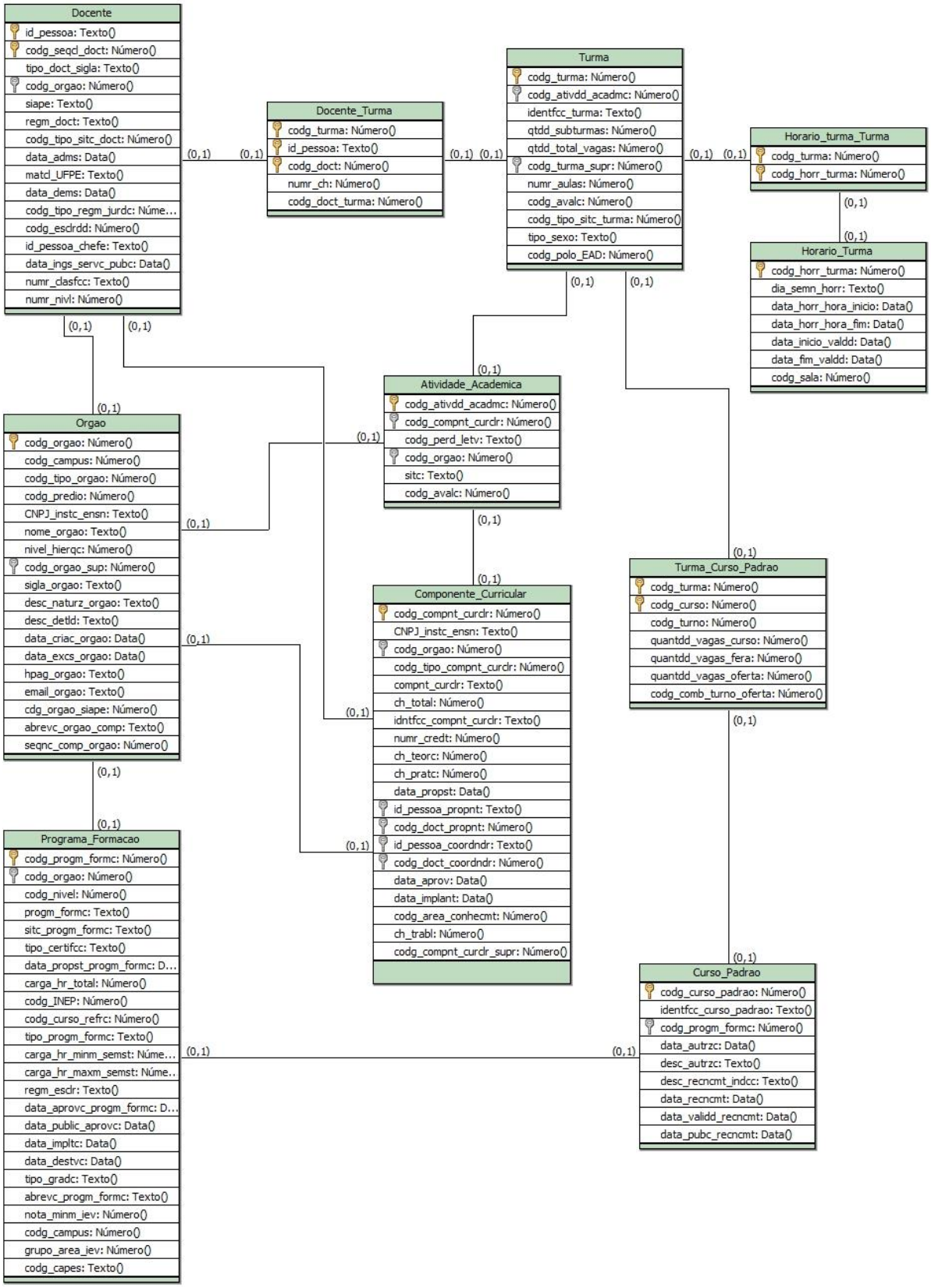
Tabela 2: Descrição das entidades e relacionamentos da Visão Exportada Notas

Tabela	Descrição
Turma	Contém dados da turma sem dados sobre alunos. Relaciona-se principalmente com atividade acadêmica do aluno e com docentes que lecionam na turma.
Matrícula_Turma	Dados sobre a matrícula de um aluno em uma turma.
Situação_Matrícula	Descrição da situação do aluno na matrícula em uma turma (confirmada, cancelada com ônus, falta de vaga).
Componente_Avaliação	Exercício escolar, ou seja, avaliações da turma,
Avaliação_Discente_Turma	Relacionamento entre o aluno, a turma e o exercício escolar.
Vínculo	Dados sobre o aluno.
Pessoa	Tabela que contém dados gerais sobre as pessoas da UFPE.
Perfil_Funcional	Descrição da função que a pessoa tem.
Tipo_Perfil_Funcional	Tipo de perfil (discente, docente, desbloqueio) no sistema.
Dado_Histórico	Notas finais e situação dos alunos na atividade acadêmica.
Situação_histórico	Descrição da situação dos alunos nas turmas (reprovado, aprovado, reprovado por falta, entre outros).
Docente	Dados sobre os docentes como data de admissão, nível, classificação, entre outros

Fonte: Elaborada pela autora.

A Figura 7, na página seguinte, representa o esquema da visão exportada de Oferta-Componente. A tabela Docente também sofreu alteração na chave-primária composta no que diz respeito à inserção da chave-primária-parcial id_pessoa.

Figura 7: Visão Exportada do Esquema Oferta-Componente



Fonte: Elaborada pela autora.

A Tabela 3 descreve cada tabela presente no Esquema da Visão Exportada do esquema Oferta-Componente.

Tabela 3: Descrição das entidades e relacionamentos da Visão Exportada Oferta-Componente

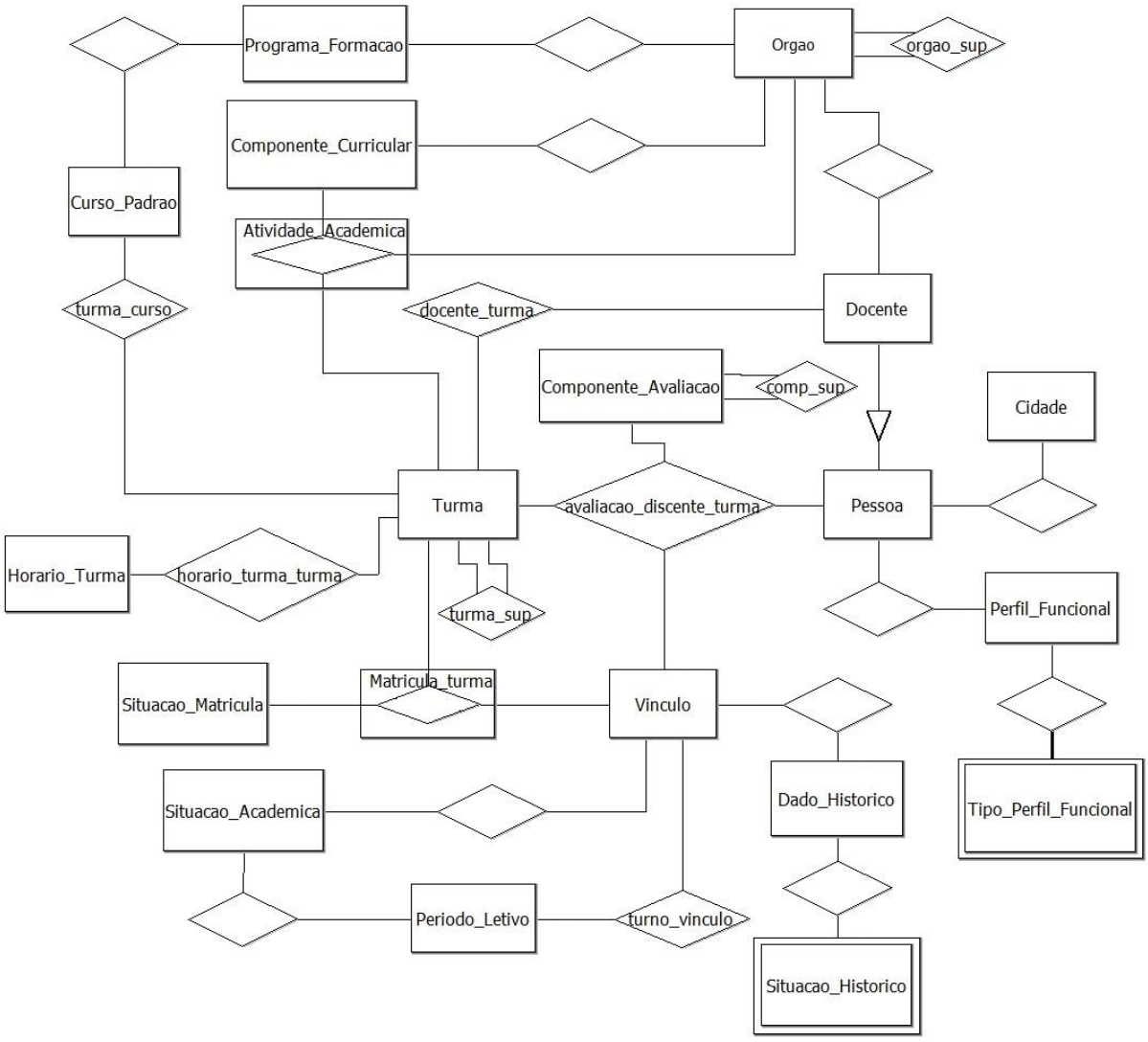
Tabela	Descrição
Docente	Dados sobre os docentes como data de admissão, nível, classificação, entre outros.
Docente_Turma	Dados sobre o(os) docente(s) em uma turma.
Turma	Contém dados da turma sem dados sobre alunos. Relaciona-se principalmente com atividade acadêmica do aluno e com docentes que lecionam a turma.
Horário_Turma_Turma	Relacionamento entre os horários possíveis para a turma e a turma propriamente dita.
Horário_Turma	Contém dias e horários padrões das turmas.
Atividade_Academica	Junção de componente curricular, curso que está ofertando e em qual período letivo.
Componente_Curricular	Dados sobre disciplina.
Orgao	Órgão que oferta as atividades acadêmicas e as turmas.
Programa_Formação	Dados sobre o curso oferecido na UFPE.
Turma_Curso_Padrao	Relacionamento que conecta a Turma ao Curso_Padrão (com informações do Mec).
Curso_Padrao	É uma tabela com informações do do MEC sobre o curso, como autorização e código.

Fonte: Elaborada pela autora.

3.3 Criação das Visões Publicadas

Além dos esquemas lógicos, apresentados na seção anterior, foi criado um esquema conceitual a fim de auxiliar no processo de definição das Visões Publicadas, ou seja, a criação dos *datasets* propriamente ditos. Para a criação do esquema conceitual foi adotado o Modelo Entidade-Relacionamento (ER). O diagrama correspondente é apresentado na Figura 8.

Figura 8: Modelo Entidade-Relacionamento: Unificação das Visões Exportadas



Fonte: Elaborada pela autora.

O diagrama ER foi criado a partir da junção dos três esquemas relacionais apresentados anteriormente, por meio de uma engenharia reversa para um único diagrama. Isso foi necessário porque algumas tabelas estão presentes em mais de um esquema, onde possui relacionamentos com tabelas diferentes. Por isso a junção de todas as tabelas e relacionamentos em um único diagrama trouxe uma

visualização e compreensão dos esquemas com suas tabelas (entidades) e relacionamentos durante o processo de desnormalização.

Em geral, os dados publicados na Web são desnormalizados ao contrário dos bancos de dados relacionais que seguem regras de normalização, ou seja, tabelas são divididas em tabelas menores a fim de evitar redundâncias e anomalias de atualização [17]. A partir do modelo conceitual, iniciou-se a análise dos *datasets* que deveriam ser criados a fim de disponibilizar os dados do SIG@ em formato aberto.

3.3.1 Definição de *Datasets*

Durante a análise do modelo, percebeu-se que não faz muito sentido criar um conjunto de dados para cada entidade, pois algumas entidades fazem parte de um mesmo tema e, portanto, seus dados deveriam ser publicados em conjunto. Dessa forma, algumas tabelas (entidades, entidades associativas e relacionamentos) foram agrupadas em um mesmo *dataset*. Por meio de uma simples observação, é possível identificar entidades que se relacionam com diversas outras, como Turma e Vínculo, por exemplo. Estas são fortes candidatas a se tornar um conjunto de dados, pois estão interligadas a outras em um tema central. A maioria das entidades foi agrupada em um único *dataset*, apenas Órgão que ficou em um único *dataset* pois algumas aplicações precisam só desta informação. Então, torna-se desnecessário ter outras informações agrupadas.

A definição dos *datasets* foi baseada no processo de desnormalização do esquema relacional e auxiliada pelo modelo ER apresentado na Figura 8 deste trabalho, que trouxe uma ampla visualização de todas as tabelas e relacionamentos. O mapeamento do diagrama ER para dados na Web foi realizado da seguinte forma:

- Entidades fracas, como Situação_Histórico, foram colocadas no mesmo *dataset* das entidades fortes;
- Entidades que tinham informações sobre temas relacionados foram agrupadas no mesmo *dataset*. Por exemplo: Vínculo, Situação_Academica, Periodo_Letivo, Turno_Vinculo, Dado_Historico, Situacao_Historico, Siga_Avaliacao_Discente_Turma e Componente_Avaliacao foram agrupadas no datasetAluno, porque possuem informações sobre o aluno.

- Entidades associativas ou Relacionamentos com cardinalidade M:N poderiam ser colocados em um dos dois *datasets* e referenciar o outro.

A Tabela 4 apresenta os *datasets* criados através do processo de desnormalização, a descrição de cada um deles e as tabelas que o compõem.

Tabela 4: Descrição dos *Datasets*

Dataset	**Descrição**	**Tabelas**
Pessoa	Refere-se a todas as pessoas que compõem a UFPE, sejam docentes, alunos ou técnicos-administrativos.	Pessoa, Cidade, Perfil Funcional, Tipo Perfil Funcional
Aluno	O *dataset* Aluno contém todas as informações de vínculo do aluno desde a matrícula até a colação de grau.	Vinculo, Situacao_Academica, Periodo_Letivo, Turno_Vinculo, Dado_Historico, Situacao_Historico, Siga_Avaliacao_Discente_Turma, Componente_Avaliacao
Docente	Contém informações sobre os Docentes da UFPE.	Docente, Docente_Turma
Órgão	Composto por coordenações de curso, diretorias de centros entre outros.	Órgão
Curso	Traz informações sobre os cursos que são ofertados na UFPE.	Programa_Formacao, Modalidade_Programa_Formacao, Curso_Padrao
Turma	Neste conjunto de dados, há informações que descrevem as turmas dos cursos de graduação da UFPE.	Turma, Matricula_Turma, Situacao_Matricula, Turma_Curso_Padrao -Horario_Turma_Turma -Horario_Turma -Componente_Curricular -Atividade_Academica

Fonte: Elaborada pela autora.

3.3.2 Criação dos *Datasets*

Para disponibilizar os *datasets*, além de defini-los também foi necessário criar as consultas que seriam utilizadas para o povoamento das visões publicadas, ou seja, os *datasets*.

Cada *dataset* requer uma consulta SQL que é colocada em um arquivo JSON juntamente com os metadados que o descrevem. A Tabela 5 descreve os metadados que são utilizados na criação dos *datasets* deste trabalho.

Tabela 5: Descrição dos metadados que descrevem os *datasets*

Nome	Descrição
datasetTitle	Nome escolhido para o *dataset*
sqlQuery	Consulta SQL que será executada para a extração dos dados no Banco de Dados de origem que geram o *dataset.*
collectionIdColumnName	Se refere ao nome do atributo que será utilizado como identificador único de cada tupla no conjunto de dados.
updateFrequency	Descreve a frequência de atualização do *dataset.* Exemplo: Semanal.
creator	Nome do criador do *dataset.*
keywords	Palavras-chave para encontrar o *dataset.*
publisher	Nome do provedor de dados.
contactPoint	Forma de contato com o provedor de dados. Exemplo: email ou número de telefone.
period	Período temporal que o conjunto de dados abrange.
spacialCoverage	Área associada ao recurso descrito pelos metadados. A descrição do metadado pode ser em forma de coordenada ou uma descrição da localização. Exemplo: Sul-Brasil.
theme	Tema ao qual o *dataset* se refere.
language	Língua em que os dados do *dataset* estão escritos.
dateTimeFormats	Formato dos dados do tipo data.
description	Descrição do *dataset.*

Fonte: Elaborada pela autora

Além das consultas e metadados, foi necessário também criar um identificador único para a tabela pessoa, o qual substituiria o CPF, que não pode ser publicado. Para isso foi criada uma *procedure* que gera o novo id da tabela Pessoa e o substitui nas tabelas que também tem o CPF como chave-primária ou chave-

estrangeira. A *procedure* não pode ser publicada por ser informação interna, mas o resultado poderá ser conferido nos dados publicados na aplicação *OpenAcademy* [14].

As consultas para publicação dos dados têm nomenclaturas restritas do banco original, e por isso os scripts não podem ser apresentados neste trabalho. Para ter acesso aos dados, foi necessário fazer uma solicitação formal por meio de um ofício que está no Apêndice 1 deste documento.

Além disso, no Apêndice 2 está o termo de confidencialidade assinado pelos membros da equipe, no qual nos comprometemos a não divulgar informações restritas do banco. Por isso a Figura 9 exemplifica como os arquivos JSON foram criados com os metadados e as consultas que geram os *datasets*.

Figura 9: Arquivo JSON com metadados e consulta

```
,
    "matcl_curso",
    essoa AS \"id_pessoa:Pessoa\",V.matcl_curso AS \"matriculaCurso\",V.codg_seqcl_disct AS \"codigoDiscente\",V.codg_progm_fomc AS \"
    rso\",V.codg_modldd AS \"codigoModalidade\",V.codg_tipo_ings AS \"codigoTipoIngresso\",V.ano_adms AS \"anoAdimissao\",V.semst_adms
    V.priordd_vincl AS \"prioridadeVinculo\",V.idtfc_vincl AS \"idVinculo\",V.instcc_origem_CNPJ AS \"cnpjInstituicaoSuperiorOrigem\"
    o AS \"cnpjInstituicaoUFPE\",V.codg_area_concmt AS \"codigoAreaConhecimento\",V.ano_concs AS \"anoConclusao\",V.semst_concs AS \"
    a_enem AS \"notaEnem\",V.numr_inscc_vestblr AS \"inscricaoVestibular\",V.numr_clasfcc_vestblr AS \"classificacaoVestibular\"
    ipoCursoOrigem\",V.curso_origm_nome AS \"cursoOrigem\",V.data_colc_grau AS \"dataColacaoGrau\",V.data_expdc_dipm AS \"
    data_concs_curso AS \"dataConclusaoCurso\",V.coefct_rendmt_escir AS \"coeficienteRendimentoEscolar\",V.ensn_medio_origm AS \"
    o_cota AS \"tipoCota\",SA.matcl_curso AS \"matriculaCurso\",SA.codg_perd_letv AS \"codigoPeriodoLetivo\",SA.data_defnc_sitc AS \"
    A.codg_tipo_sitc_acadmc AS \"codigoTipoSituacaoAcademica\",SA.obsvc_sitc_acadmc AS \"observacaoSituacaoAcademica\",SA.data_colc_grau
    .codg_perd_letv AS \"codigoPeriodoLetivo\",PL.ano_perd_letv AS \"anoPeriodo\",PL.semst AS \"semestre\",PL.tipo_perd_letv AS \"
    icio_perd_letv AS \"inicioPeriodoLetivo\",PL.fim_perd_letv AS \"fimPeriodoLetivo\",PL.data_inic_real_perd_letv AS \"
    vo\",PL.data_fim_real_perd_letv AS \"dataRealFimPeriodoLetivo\",TV.matcl_curso AS \"matriculaCurso\",TV.codg_perd_letv AS \"
    codg_turno AS \"codigoTurno\",ADT.codg_avalc_disct_turma AS \"codigoAvaliacaoDiscenteTurma\",ADT.codg_turma AS \"codigoTurma\"
    riculaCurso\",ADT.codg_compnt_avalc AS \"codigoComponenteAvaliacao\",ADT.data_incs_avalc_disct AS \"dataInclusaoAvaliacaoDiscente\"
    ma AS \"notaAvaliacaoDiscenteTurma\",ADT.conct_avalc_disct_turma AS \"conceitoAvaliacaoDiscenteTurma\",CA.codg_compnt_avalc AS \"
    \",CA.id_compnt_avalc AS \"idComponenteAvaliacao\",CA.descc_compnt_avalc AS \"descricaoComponenteAvaliacao\"
    cnd AS \"Codigo_Componente_Avaliacao_Relacionado\",DH.matrl_curso AS \"matriculaCurso\",DH.codg_ativdd_acadmc AS \"
    ",DH.codg_sitc_histrc AS \"codigoSituacaoHistorico\",DH.total_faltas AS \"totalFaltas\",DH.media_final AS \"mediaFinal\"
    eitoFinal\",SH.codg_sitc_histrc AS \"codigoSituacaoHistorico\",SH.desc_sitc_histrc AS \"descricaoSituacaoHistorico\" FROM Vinculo V
    Academica SA ON V.matcl_curso = SA.matcl_curso FULL OUTER JOIN Turno_Vinculo TV ON TV.matcl_curso = SA.matcl_curso FULL OUTER JOIN
    odg_perd_letv = TV.codg_perd_letv FULL OUTER JOIN Avaliacao_Discente_Turma ADT ON ADT.matcl_curso = V.matcl_curso FULL OUTER JOIN
    N CA.codg_compnt_avalc = ADT.codg_compnt_avalc FULL OUTER JOIN Dado_Historico DH ON DH.matcl_curso = V.matcl_curso FULL OUTER JOIN
    SH.codg_sitc_histrc = DH.codg_sitc_histrc WHERE ROWNUM < 10",

    ni",
    uno, Histórico, Período Letivo",

    ufpe.br",
    sde 1970.",
    , Brasil",

    L,JSON",
    e dados que contém informações sobre os alunos e seu vínculo com o curso e a universidade."
```

Fonte: Elaborada pela autora

A frequência de atualização é a temporalidade definida para atualização das tabelas do banco. O conhecimento sobre a frequência de atualização das tabelas é fundamental para garantir que os dados disponibilizados nos *datasets* estejam atualizados, auxiliando também o versionamento dos *datasets*.

A Tabela 6 apresenta as tabelas do SIG@ utilizadas neste trabalho e sua respectiva frequência de atualização.

Tabela 6: Frequência de Atualização das Tabelas

Tabela	Frequência de Atualização
Pessoa	Mensal
Vínculo	Semestral
Programa Formação	Por demanda
Turno Vínculo	Semestral
Situação Acadêmica	Semestral
Período Letivo	semestral
Cidade	Por demanda
Modal Programa Formação MPF	Por demanda
Docente	Mensal
Docente Turma	Semestral
Turma	Semestral
Horário Turma-Turma	Semestral
Horário Turma	Semestral
Turma Curso Padrão	Por Demanda
Curso Padrão	Por Demanda
Atividade Acadêmica	Semestral
Componente Curricular	Por Demanda
Órgão	Mensal
Componente Avaliação	Por demanda
Avaliação Discente Turma	Semestral
Perfil Funcional	Mensal
Tipo Perfil Funcional	Por demanda
Matrícula Turma	Semestral
Situação Matrícula	Semestral
Historico	Semestral
Situacao_Historico	Semestral

Fonte: Elaborada pela autora.

A Tabela 7 apresenta a frequência de atualização dos conjuntos de dados que foi baseada na menor frequência de atualização das tabelas que compõem o *dataset,* descritas anteriormente.

Tabela 7: Frequência de Atualização dos Datasets

Dataset	Frequência de Atualização
Pessoa	Mensal
Aluno	Semestral
Docente	Mensal
Órgão	Mensal
Curso	Por Demanda
Turma	Semestral

Fonte: Elaborada pela autora.

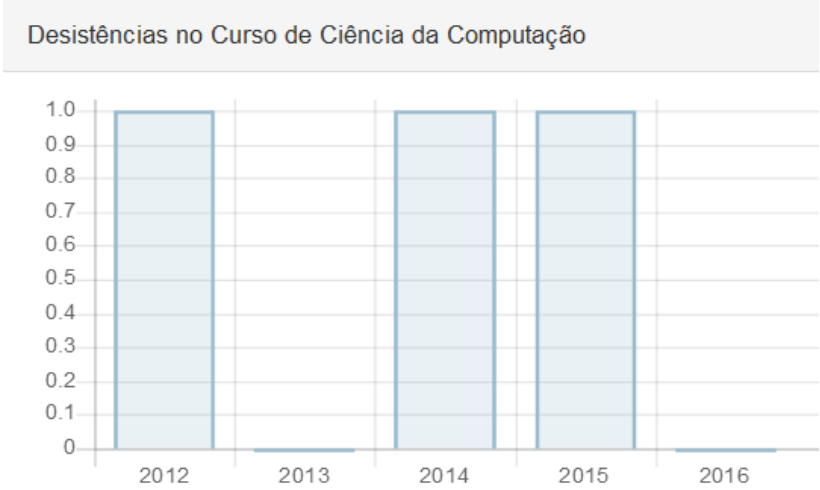
3.4 Avaliação dos *Datasets*

O resultado da definição e criação dos *datasets,* ou seja, os arquivos JSON contendo as consultas e os metadados, deve ser utilizado por uma API, que é a responsável por executar a consulta no banco de dados e gerar a informação para uma aplicação Web como um portal por exemplo.

Para validação dos *datasets* criados neste trabalho, a API DoWMaker[SANTOS, W.C.R, 2017] foi responsável por executar as consultas de criação dos *datasets*. A função da ferramenta era receber uma requisição feita por uma aplicação web e gerar as informações no formato solicitado.

A aplicação Web que realizou as requisições, foi a *OpenAcademy* [OLIVEIRA, F. E., 2017], com o objetivo de gerar indicadores para auxiliar a Gestão Acadêmica do CIn. O usuário seleciona as informações que deseja, a *OpenAcademy* requisita as informações por meio de consulta para a DoWMaker que usa as consultas para gerar os dados presentes nos *datasets* criados por este trabalho. A Figura 10 apresenta um indicador gerado pela ferramenta OpenAcademy.

Figura 10: Indicador Desistências no Curso de Ciência da Computação



Fonte: *OpenAcademy*

O gráfico demonstra a quantidade de alunos que desistiram do curso de Ciência da Computação de 2012 até 2016.

3.5 Considerações finais

Este capítulo apresentou o processo de criação dos *datasets* com os dados do SIG@ a serem publicados em formato aberto, incluindo desde a definição da metodologia utilizada até a criação dos scripts de publicação dos dados. Na Seção 3.1, foi apresentada a metodologia utilizada com o passo a passo de como criar *datasets* utilizando dados de um banco relacional. As seções subsequentes abrangem o processo de definição dos *datasets* e criação dos scripts de publicação.

4 CONCLUSÃO E TRABALHOS FUTUROS

O principal objetivo deste trabalho foi fazer a geração de dados abertos a partir de dados armazenados no banco de dados do SIG@-UFPE. Para isso, foi necessário fazer o mapeamento de esquemas relacionais para conjuntos de dados a serem publicados na Web. Especificamente, foram mapeados dados sobre a vida acadêmica dos Alunos da UFPE. Esse domínio foi escolhido devido à necessidade de gerar indicadores que auxiliem Diretores de Centro, Coordenadores de Ensino, Coordenadores de Curso dentre outros, a melhorar a qualidade de ensino nas disciplinas dos Cursos da UFPE.

Para tal, este trabalho foi dividido em Análise dos Dados do domínio escolhido, a definição das Visões Exportadas e Definição das Visões de Publicação. Ao final do processo, foram gerados os scripts para criação dos conjuntos de dados. Cada script consiste de uma consulta SQL que materializa os dados do *dataset* e uma descrição em JSON dos metadados do conjunto de dados. A partir dos scripts, os conjuntos de dados podem ser automaticamente gerados.

Os maiores desafios enfrentados durante o desenvolvimento deste trabalho foram: I) A definição do domínio a ser abordado no trabalho, devido à grande quantidade de dados existente no Banco de Dados do SIG@; II) A definição de quais dados de cada tabela que poderiam ser publicados, uma vez que na UFPE ainda não existe um documento de classificação dos dados que podem ser abertos; e III) A definição de uma metodologia a ser usada para a publicação.

Este trabalho, juntamente com os de Filipe Epifânio de Oliveira [14] e Wilker Cavalcante do Rego Santos [15], faz parte da etapa inicial do projeto OpenUFPE, que visa publicar conjuntos de Dados Abertos de toda a UFPE. Logo, há uma série de ações futuras que auxiliarão a publicação no cenário onde os dados são armazenados em Banco de Dados Relacional. Entre elas, destacam-se:

- Criação de um documento de classificação de dados da UFPE, para garantir que os dados publicados estarão de acordo com a LAI;
- Estender o domínio do mapeamento realizado para todos os dados da UFPE;

- Criação de uma Metodologia de Mapeamento que possa ser utilizada por outras instituições públicas que possuam um cenário semelhante com o deste trabalho.

Por fim, este trabalho contribuiu não só para aumentar o número de publicações sobre o tema, mas para tornar mais conhecido os benefícios da publicação de Dados Abertos.

5 REFERÊNCIAS

[1] World Wide Web Foundation. History of the Web. Disponível em: <http://webfoundation.org/about/vision/history-of-the-web/>. Acesso em: 13 de outubro de 2016.

[2] MAZONI, Marcos Vinicius Ferreira. **Dados Abertos para a Democracia na era Digital**. Brasília: Fundação Alexandre Gusmão, 2011. 84 p. Disponível em: <http://funag.gov.br/loja/download/817-Dados_Abertos_para_a_Democracia_na_Era_Digital.pdf>. Acesso em: 13 de outubro de 2016.

[3] WORLD WIDE WEB CONSORTIUM. **Data on the Web Best Practices. Data on the Web Best Practices, 2016.** Disponível em: <https://www.w3.org/TR/dwbp/>. Acesso em: 19 de novembro de 2016.

[5] LÓSCIO, Bernardette Farias; OLIVEIRA, Marcelo Iury S.; BITTENCOURT, Ig Ilbert. **Tópicos em Gerenciamento de Dados e Informações**. Petrópolis: SBC, 2015. 95 p. Disponível em: <http://dexl.lncc.br/sbbd2015/anais/ShortCourses.pdf>. Acesso em: 12 de novembro de 2016.

[5] AGHAEI, Sareh; NEMATBAKHSH, Mohammad Ali; FARSANI, Hadi Khosravi. **Evolution of the World Wide Web: From Web 1.0 to Web 4.0**. University of Isfahan. Isfahan, 2012. 10 p. Disponível em: http://www.ftsm.ukm.my/ss/Book/EVOLUTION%20OF%20WWW.pdf. Acesso em: 11 de novembro de 2016.

[6] BAPTISTA, Cláudio. **Banco de Dados**. Universidade Federal do Maranhão. São Luís, 2009. p. 1-14. Disponível em: http://www.dsc.ufcg.edu.br/~baptista/cursos/BDadosI/Capitulo1.pdf Acesso em: 22 de novembro de 2016.

[7] BITTENCOURT, Rogério Gonçalves. **Aspectos Básicos de Bancos de Dados**. Florianópolis, 2004. 65 p. Disponível em: <https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/BD%20-%20Aspectos%20Basicos.pdf>. Acesso em: 01 de dezembro de 2016.

[8] LEE, Deirdre; LÓSCIO, Bernardette Farias; ARCHER, Phil. **Data on the Web: Best Practices, Use Cases & Requirements**. Massachusetts: W3C, 2015. Disponível em: <https://www.w3.org/TR/dwbp-ucr/>. Acesso em: 20 de novembro de 2016.

[9] ALVES, R. P. G. **SBCEO: uma Ontologia para a Descrição dos Eventos da Sociedade Brasileira de Computação**. Universidade Federal de Pernambuco. Recife, p. 50. 2016.

[10] PETNews. Disponível em: <http://www.dsc.ufcg.edu.br/~pet/jornal/maio2011/materias/recapitulando.html>. Acesso em: 8 de dezembro de 2016.

[11] Portal Brasileiro de Dados Abertos. Disponível em: <http://dados.gov.br/dados-abertos/>. Acesso em: 16 de setembro de 2016.

[12] Dados Abertos UFRN. Disponível em: <http://dados.ufrn.br/>. Acesso em: 8 de dezembro de 2016.

[13] e-SIC. Sistema Eletrônico do Serviço de Informação ao Cidadão. Disponível em: <http://esic.cgu.gov.br/sistema/site/index.html?ReturnUrl=%2fsistema>. Acesso em: 16 de setembro de 2016.

[14] OLIVEIRA, F. E. **OpenAcademy: Uma ferramenta de Gestão Acadêmica com Dados Abertos Conectados da Universidade Federal de Pernambuco**. (Trabalho acadêmico). Universidade Federal de Pernambuco. Recife, 2017.

[15] SANTOS, W.C.R. **DoWMaker: Uma Ferramenta para Publicação Automática de Dados na Web** (Trabalho acadêmico). Universidade Federal de Pernambuco. Recife, p. 42. 2017.

[16] TENNISON, Jeni. **CSV on the Web: A Primer**. Massachusetts: W3C, 2016. Disponível em : <https://www.w3.org/TR/tabular-data-primer/>. Acesso em: 10 de dezembro de 2016.

[17] NAVATHE, Shamkant B; ELMASRI, Ramez. **Fundamentals of Database Systems**. Boston: Pearson Education, Inc., 2003. 1029 p. Disponível em: <http://www.uoitc.edu.iq/images/documents/informatics-institute/Competitive_exam/Database_Systems.pdf>. Acesso em: 12 de janeiro de 2017.

[18] SILVA, R.I.V.C. **OpenSBC: Um Portal de Dados Abertos Conectados sobre os Eventos da Sociedade Brasileira de Computação.** Universidade Federal de Pernambuco. Recife, p. 65. 2016.

[19] LÓSCIO, Bernardette Farias; TRAVASSOS, Walter; SOUZA, Damires. **QualityStamp – Avaliando a Qualidade de Linked Datasets para Aplicações de Domínio Específico**. Curitiba: SBBD, 2014. 6 p. Disponível em: <http://www.inf.ufpr.br/sbbd-sbsc2014/sbbd/proceedings/artigos/pdfs/2.pdf>. Acesso em: 12 de dezembro de 2016.

Apêndice 1: Ofício de solicitação de acesso aos dados do SIG@.

Centro de Informática – CIn – UFPE
Universidade Federal de Pernambuco

Of. nº 172/16-CIn

Recife, 07 de dezembro de 2016.

Senhor Diretor,

Venho por meio deste, solicitar ao Núcleo de Tecnologia da Informação(NTI) acesso à alguns dados do SIGA para uso no Trabalho de Conclusão de Curso de 3 (três) alunos do curso de Sistemas de Informação do Centro de Informática(CIn).

As Tabelas do Banco de Dados as quais os alunos precisarão de acesso são: Pessoa, Vinculo, Programa Formação, Turno Vinculo, Situação Acadêmica, Periodo Letivo, Cidade, Modalidade Programa Formação, Docente, Docente Turma, Turma, Horário Turma-Turma, Horário Turma, Turma Curso Padrão, Curso Padrão, Atividade Acadêmica, Componente Curricular, Órgão, Componente Avaliação, Avaliação Discente Turma, Perfil Funcional, Tipo Perfil Funcional, Matricula Turma, Situação Matricula. Os alunos que usarão os dados são: Filipe Epifanio de Oliveira, Priscila Thyfani Rodrigues Gomes de Melo e Wilker Cavalcante do Rego Santos. Sendo Priscila Thyfani Rodrigues Gomes de Melo a aluna responsável por pegar o Pendrive com os dados no NTI.

Uma cópia deste oficio está sendo enviada à Diretoria de Gestão Acadêmica (DGA-PROACAD), como órgão responsável pelos dados, conforme especificação da própria DGA. Também está anexado a este oficio um Termo de Confidencialidade, onde os alunos se comprometem a utilizar os dados apenas para o TCC e não torná-los públicos sem autorização.

Sem mais no momento, renovamos nossos cordiais cumprimentos.

Atenciosamente,

Prof. José Augusto Suruagy Monteiro
Vice-Diretor
Centro de Informática
CIn/UFPE

Ao Núcleo de Tecnologia da Informação – NTI/UFPE
(Com cópia: PROACAD/DGA)

Apêndice 2: Termo de Confidencialidade para acesso aos dados.

Termo de Confidencialidade

Nós, Filipe Epifanio de Oliveira (CPF: 053.378.684-38), Priscila Thyfani Rodrigues Gomes de Melo(CPF: 096.519.124-99) e Wilker Cavalcante do Rego Santos (CPF: 103.144.124-76), assumimos, por este instrumento legal, o compromisso de manter **confidencialidade e sigilo** sobre todas as informações contidas nas Tabelas do Banco de Dados do SIGA(Sistema de Gestão Acadêmica) da UFPE as quais teremos acesso para realização do Trabalho de Conclusão de Curso(TCC) no Curso de Sistemas de Informação da UFPE. As tabelas são: Pessoa, Vinculo, Programa Formação, Turno Vinculo, Situação Acadêmica, Período Letivo, Cidade, Modalidade Programa Formação, Docente, Docente Turma, Turma, Horário Turma-Turma, Horário Turma, Turma Curso Padrão, Curso Padrão, Atividade Acadêmica, Componente Curricular, Orgão, Componente Avaliacão, Avaliação Discente Turma, Perfil Funcional, Tipo Perfil Funcional, Matrícula Turma, Situação Matricula.

Por este Termo de Sigilo e Confidencialidade, comprometemo-nos a:

1. Não utilizar as informacões contidas nas tabelas a que tivermos acesso, para gerar benefício próprio exclusivo e/ou unilateral, presente ou futuro, ou para uso de terceiros;
2. Não efetuar nenhuma gravacão ou cópia da documentacão confidencial as quais tivermos acesso relacionado ao conteúdo desse Trabalho de Conclusão de Curso;
3. Não nos apropriar para nos mesmos ou para outrem de material confidencial e/ou sigiloso que venha a ser disponibilizado aos participantes desse Trabalho de Conclusão de Curso;
4. Não repassar o conhecimento das informacões confidenciais, as quais tivermos acesso.

A vigência da obrigacão de confidencialidade e sigilo, assumida pelas partes interessadas deste Trabalho de Conclusão de Curso, por meio deste termo, terá a validade enquanto a informacão não for tornada de conhecimento público por qualquer outra pessoa, ou mediante autorizacão escrita, concedida qualquer um de nós pelas partes interessadas neste termo.

Pelo não cumprimento do presente Termo de Confidencialidade e Sigilo, fica o abaixo assinado ciente de todas as sancões judiciais que poderão advir.

Recife, 06 de dezembro de 2016

Filipe Epifânio de Oliveira

Filipe Epifanio de Oliveira

Priscila Thyfani R. G. de Melo

Priscila Thyfani Rodrigues Gomes de Melo

Wilker Cavalcante do Rego Santos

Wilker Cavalcante do Rego Santos