Automatização do Dicionário de Dados

Bruno Romeiro Comin[1], Marcos Felipe[2], Claudines Taveira Torres[3]

Curso de Tecnologia em Banco de Dados - Faculdade de Tecnologia de Bauru (FATEC)

Rua Manoel Bento da Cruz, nº 3-30 - Centro - 17.015-171 - Bauru, SP - Brasil

[1]bruno.comin@fatec.sp.gov.br, [2]omarkdev@gmail.com,
[3]claudines.torres@fatec.sp.gov.br

Abstract. *Having a high demand for projects with tight deadlines, we noticed that it is increasingly common to abandon the making of documentation that reports the steps taken by programmers and database analysts in their development. Given this type of neglect, this research aims to identify the importance of documentation in the projects and demonstrate some possible methods of automation to speed it up. It compares a small data dictionary generation framework with the traditional generation and concludes that in addition to saving time the tool helps to focus on development.*

Resumo. *Tendo uma alta demanda de projetos com prazos apertados, notamos que está cada vez mais comum o abandono da confecção de documentações que relatam os passos dados por programadores e analistas de banco de dados em seu desenvolvimento. Diante deste tipo de descaso, essa pesquisa tem como objetivo identificar a importância da documentação nos projetos e demonstrar alguns possíveis métodos de automatização para agilizá-la. Compara-se um pequeno framework de geração de dicionário de dados com a geração tradicional e conclui-se que além de economizar tempo a ferramenta ajuda a focar no desenvolvimento.*

1. Introdução

Segundo Monteiro (2017), antes de nos esquecermos de criar uma documentação bem elaborada, devemos nos fazer algumas perguntas: “Como você administra e gerencia aquilo que não conhece? Como você determina as regras de acesso para um atributo que você não sabe a origem e desconhece as regras de seu preenchimento? Como você vai adquirir qualquer vantagem competitiva utilizando dados que você nem conhece? Como você pode afirmar que seus dados possuem qualidade?”

Sabendo o quanto a documentação dos dados é importante, deve-se conhecer quais são as formas existentes de se fazer isso e qual delas mais se adapta à necessidade de cada caso em específico.

A qualidade de dados não é só obtida por inspeção e correção dos dados, que implicam custos decorrentes da qualidade pobre de dados. A qualidade de dados dentro de uma empresa é resultante de um projeto de qualidade inserido nos seus processos de negócio. Esse projeto de qualidade provê técnicas de qualidade conhecidas como: Planejamento-Execução-Análise e Definir-Medir-Analisar-Melhorar-Controlar os dados da empresa dentro de um contexto de negócio, conforme retrata English (2009).

Ainda citando English (1999), o autor definiu a qualidade de dados como a união de três aspectos indispensáveis: Apresentação – onde os dados devem estar

compreensíveis para aqueles que fazem o uso dele, exemplo de idioma correto. Valor – o valor do dado definido deve estar correto de acordo com às regras de qualidade relacionadas à acurácia, completude, precisão e atualidade. Definição – quesito onde será o foco do progresso desta pesquisa.

Batini e Scannapieco (2006) apresentam uma proposta onde atributos de qualidade de dados e os seus respectivos valores medidos são inseridos nos modelos de dados, sob a forma de novas entidades, a fim de diferenciá-los dos atributos da aplicação. As entidades de qualidade, juntamente com as entidades da aplicação, se relacionam por meio de novos relacionamentos introduzidos no modelo e compõem um modelo conceitual de QD. Esse modelo conceitual destes autores também é traduzido para um modelo relacional, representado pelo modelo de Entidade-Relacionamento (ER).

Analisando Resende (2017), nota-se que o *SQL Server* já possui uma ferramenta para auto documentação, onde o analista pode gerar um comentário especificando do que se trata uma coluna ou tabela, facilitando manutenções e intervenções futuras de terceiros.

No caso dessa pesquisa o foco estará na criação do dicionário de dados, que consiste em um documento legado que descreve com exatidão todas as tabelas e suas respectivas colunas do banco de dados. Esta documentação pode parecer simples, mas ela funciona da mesma forma que um mapa funciona para o mundo: se é a sua primeira vez naquele lugar e você não possui um guia, ficará perdido sem saber para onde ir.

2. Dicionário de Dados

Um dicionário de dados é uma planilha ou texto que centraliza informações sobre o conjunto de dados com o propósito de melhorar a comunicação entre todos os envolvidos no projeto. Segundo a IBM (s/d), um dicionário de dados são tabelas onde armazena-se a estrutura de um banco de dados relacional. Podemos concluir então que um dicionário de dados é um repositório que descreve, de forma estruturada, o significado, a origem, o relacionamento e o uso dos dados.

2.1 Modelo padrão de dicionário de dados

Existem vários padrões de modelos de dicionário de dados a serem seguidos, mas todos eles possuem a mesma linha de raciocínio e acabam mostrando as mesmas informações com alguns detalhes mínimos que o diferem um do outro. Alguns bancos de dados disponibilizam ferramentas embutidas em seu sistema gerenciador para a geração automática do dicionário de dados, mas a grande maioria são ferramentas pagas ou muito complexas para serem utilizadas por pessoas que não são desenvolvedoras de sistema.

A *IBM* possui tabelas específicas para a geração do dicionário de dados em seu banco de dados DB2, já o *Microsoft SQL Server* possui um tipo de etiqueta invisível para comentar cada tabela e coluna existente em sua estrutura, o sistema de gerenciamento de banco de dados da *Oracle* trabalha com um ambiente separado internamente denominado *tablespaces*, onde todo o conteúdo já é indexado e registrado de uma forma automática restando apenas colocar os comentários personalizados, o MariahDB (derivado do *mySQL*) possui um ótimo gerenciador gratuito chamado de *phpMyAdmin*, onde é possível fazer uma geração quase que completa de um dicionário de dados.

2.2 Exemplo de dicionário de dados

Baseado no que pode ser visto no banco de dados original, Schneider (2011) criou esse dicionário de dados em um documento do *Microsoft Word*, utilizando recursos simples como a criação de tabelas e a formatação de fontes, com o intuito de facilitar a visualização de suas tabelas, podemos ver a imagem do resultado parcial na Figura 1:

TABELA: PESSOAS

	CAMPO	DESCRIÇÃO	TIPO	TAM	DEC
PK	PES_CPF	CPF DA PESSOA	VARCHAR	15	-
FK	CID_CEP	CIDADE E CEP DA PESSOA	INTERGER	-	-
	PES_NOME	NOME DA PESSOA	VARCHAR	100	-
	PES_RG	RG DA PESSOA	VARCHAR	15	-
	PES_ENDERECO	ENDERECO DA PESSOA	VARCHAR	100	-
	PES_EMAIL	EMAIL DA PESSOA	VARCHAR	100	-
	PES_TELEFONE	TELEFONE DA PESSOA	VARCHAR	15	-
	PES_CELULAR	CELULAR DA PESSOA	VARCHAR	15	-
	PES_SEXO	SEXO DA PESSOA	VARCHAR	1	-
	PES_DATANASCIMENTO	DATA DE NASCIMENTO DA PESSOA	DATE	-	-

TABELA: USUARIOS

	CAMPO	DESCRIÇÃO	TIPO	TAM	DEC
PK	USU_COD	CODIGO DO USUARIO	INTERGER	-	-
FK	PES_CPF	CPF DO USUARIO	VARCHAR	15	-
FK	NIV_CODIGO	NIVEIS DO CODIGO	INTERGER	-	-
	USU_SENHA	SENHA DO USUARIO	VARCHAR	20	-

TABELA: CIDADES

	CAMPO	DESCRIÇÃO	TIPO	TAM	DEC
PK	CID_CEP	CEP DA CIDADE	INTERGER	-	-
	CID_CIDADE	NOME DA CIDADE	VARCHAR	150	-
	CID_UF	UF DA CIDADE	VARCHAR	2	-

Figura 1. Exemplo de dicionário de dados simples.

Fonte: [http://bit.ly/exemplo-dd-fatec].

Nota-se que para cada tabela existe um nome mas não uma função; para as suas colunas internas estão identificados seus respectivos nomes, seus comentários personalizados (desenvolvidos pelo cientista de dados), tipos, seus tamanhos e se são dependentes de outras colunas. Conclui-se então que, com essa documentação em mãos, um segundo analista ou um desenvolvedor que inicialmente não fazia parte da equipe identificará rapidamente a utilidade e objetivo de cada pedaço daquele banco de dados.

3. Materiais utilizados

Com o objetivo de agilizar a documentação e facilitar manutenções e implementações posteriores, os autores dessa pesquisa irão comparar a velocidade de um *framework* desenvolvido especificamente para a geração do dicionário de dados automaticamente contra uma geração feita completamente de forma manual.

Para validação destes testes serão utilizadas três ferramentas imprescindíveis que foram escolhidas por possuírem a opção de instalação e utilização de forma gratuita, pela popularidade de propagação no ano vigente deste artigo e pela familiaridade dos autores com elas. Essas ferramentas são: A linguagem de programação *Hypertext Preprocessor* (PHP) para o desenvolvimento do *framework* de geração automática dos dicionários de dados, o banco de dados MariaDB utilizado como exemplo de banco de dados de onde as informações serão consolidadas para a geração do dicionário e o *Microsoft Excel* que armazenará o dicionário de dados.

3.1 PHP

Segundo Praciano (2013), PHP é uma linguagem com um conjunto de instruções que são interpretadas sequencialmente, como o funcionamento de um algoritmo básico. Estas instruções servem para automatizar tarefas que poderiam ser feitas por mãos humanas, uma a uma.

Apesar de ser comumente utilizado para desenvolvimento de sites comuns da web e muitas vezes confundido com uma linguagem que tem apenas este propósito, o PHP também serve para desenvolver programas complexos de forma que os resultados possuem muito mais portabilidade do que um sistema desenvolvido em outra linguagem de programação, visto que uma aplicação web desenvolvida em PHP pode funcionar em qualquer computador, *smartphone* ou *tablet* com um navegador de internet.

3.2 MariaDB

O *mySQL* foi originalmente desenvolvido pela empresa *MySQL AB*, fundada por *David Axmark, Allan Larsson e Michael "Monty" Widenius*. Segundo a empresa HostDime (2017), sua primeira versão apareceu em 1995 e foi criado para uso pessoal, evoluindo em pouco tempo para um banco de dados empresarial e tornando-se o *software* de código aberto mais popular do mundo, sendo vendido em 2008 por US$ 1 bilhão para a *Sun Microsystems*.

Após muitos processos na justiça contra a comissão européia, a Oracle conseguiu comprar seu maior concorrente, a *Sun Microsystems*, em 2009, com a aprovação da Comissão Europeia. Os desenvolvedores originais do *mySQL* não gostaram da aquisição da Oracle e não concordavam com as novas políticas de trabalho, o que causou uma divisão da equipe de desenvolvimento onde a parte que abandonou a empresa saiu para criar o MariaDB.

Ainda segundo a HostDime (2017), segundo vários testes de desempenho comparativos, que são chamados de *benchmarks*, foi comprovado que o MariaDB pode ter 5% a mais de velocidade quando comparado com o *mySQL*, o que pode parecer pouco mas isso para grandes empresas com um grande conjunto de dados pode ser um grande número. O próprio otimizador de consulta interno também foi bastante modificado, gerando bastante melhorias em relação ao *mySQL*.

3.3 Microsoft Excel

Criado pela *Microsoft* em 1987, o Excel é o *software* de planilhas eletrônicas mais conhecido por todas as pessoas com acesso a informática, hoje é a ferramenta mais utilizada por empresas que querem um controle intuitivo de seus recursos.

Segundo Meyer (2013), uma planilha eletrônica é um programa de computador que utiliza linhas e colunas para realização de cálculos ou apresentação de dados. Cada célula dentro dessas tabelas possui uma identificação única que é formada pela junção da letra identificadora da coluna com o número identificador da linha, por exemplo: se uma determinada célula está localizada no encontro da linha 3 com a coluna C, a célula terá o nome de C3.

4. Banco de dados

Segundo Ferré (2016), o modelo de entidade de relacionamento, popularmente conhecido simplesmente pela sigla MER, serve para descrever todas as entidades existentes em um domínio, incluindo como se relacionam e as características de cada uma delas.

Como modelo de testes, um banco de dados fictício foi criado para servir de exemplo. Nota-se que segundo o MER na Figura 2, os nomes foram feitos propositalmente de uma forma que não fique tão fácil identificar qual a função correta de cada tabela ou coluna, justamente para demonstrar a utilidade de um dicionário de dados.

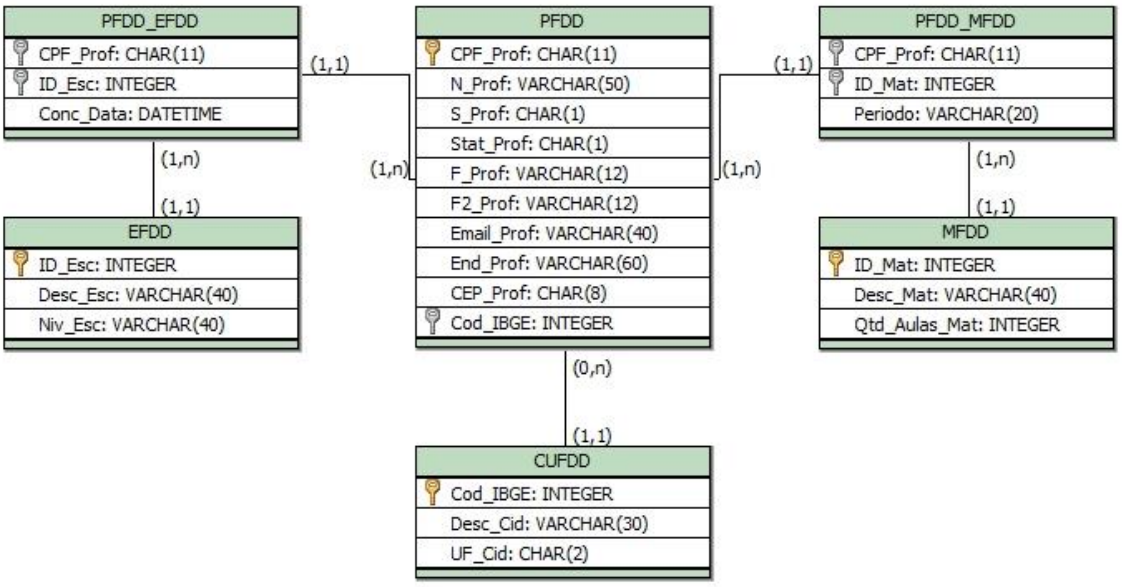

Figura 2. Modelo de Entidade de Relacionamento.

Analisando a imagem pode-se ver o quão seria difícil dizer corretamente qual a definição de cada tabela e suas respectivas colunas sem uma documentação bem descritiva e condizente com a última versão do banco de dados construído.

5. Resultados – Geração manual

Supondo que um programador entre em um projeto em andamento e este projeto tenha como base o banco de dados demonstrado na Figura 2; para começar a programar utilizando os recursos do banco de dados ele precisaria saber exatamente para que serve cada tabela, qual a função de cada coluna, qual o tamanho limite que ele pode enviar para cada coluna, se elas pertencem à outra tabela e outros detalhes.

Utilizando recursos do *Microsoft Excel*, uma geração de dicionário de dados pode ser feita para suprir todas essas requisições e demonstrar com exatidão cada detalhe de cada “pedaço” do banco de dados selecionado. Apesar de ser uma forma árdua e completamente manual, é utilizada em grande escala pois o resultado pode ser bem satisfatório se trabalhado junto com um bom planejamento no projeto, conforme podemos ver na Figura 3 a seguir:

Tabela: PFDD - Função: Armazenar o cadastro de cada um dos professores				
Coluna	**Tipo**	**Nulo**	**Ligações de**	**Comentários**
CPF (*Primária*)	char(11)	Não	-	Somente números do CPF como identificador
N_Prof	varchar(50)	Não	-	Nome completo do professor
S_Prof	char(1)	Sim	-	Sexo do professor: H (Homem) ou M (Mulher)
Stat_Prof	char(1)	Não	-	Status de atividade: I (Inativo) ou A (Ativo)
F_Prof	varchar(12)	Não	-	Telefone primário do professor
F2_Prof	varchar(12)	Sim	-	Telefone secundário do professor
Email_Prof	varchar(40)	Não	-	Email do professor
End_Prof	varchar(60)	Sim	-	Endereço do professor
CEP_Prof	char(8)	Sim	-	CEP do endereço do professor (sem traços)
Cod_IBGE (*Extrangeira*)	integer	Sim	Tabela CUFDD	Número de relacionamento com a Cidade/UF

Tabela: CUFDD - Função: Armazenar o cadastro de cidades vinculadas aos seus respectivos Estados				
Coluna	**Tipo**	**Nulo**	**Ligações de**	**Comentários**
Cod_IBGE (*Primária*)	integer	Não	-	Números do código IBGE da cidade
Desc_Cid	varchar(30)	Não	-	Nome da cidade cadastrada
UF_Cid	char(2)	Não	-	Sigla da UF vinculada (Ex.: SP)

Tabela: EFDD - Função: Armazenar as possíveis escolaridades e seus graus de instruções agregados				
Coluna	**Tipo**	**Nulo**	**Ligações de**	**Comentários**
ID_Esc (*Primária*)	integer	Não	-	Identificador da escolaridade (Auto-Incremento)
Desc_Esc	varchar(40)	Não	-	Descrição (Ex.: Engenharia de Software)
Niv_Esc	varchar(40)	Não	-	Nível (Ex.: Pós Graduação)

Tabela: MFDD - Função: Armazenar as várias matérias que um professor pode lecionar				
Coluna	**Tipo**	**Nulo**	**Ligações de**	**Comentários**
ID_Mat (*Primária*)	integer	Não	-	Identificador da matéria (Auto-Incremento)
Desc_Mat	varchar(40)	Não	-	Descrição (Ex.: Matemática Discreta)
Qtd_Aulas_Mat	integer	Não	-	Aulas daquela matéria por dia (Ex.: 4)

Tabela: PFDD_EFDD - Função: Vincular os professores às suas escolaridades concluídas				
Coluna	**Tipo**	**Nulo**	**Ligações de**	**Comentários**
CPF_Prof (*Extrangeira*)	char(11)	Não	PFDD	Somente números do CPF como identificador
ID_Esc (*Extrangeira*)	integer	Não	EFDD	Identificador da escolaridade
Conc_Data	datetime	Não	-	Data de conclusão dessa escolaridade

Tabela: PFDD_MFDD - Função: Vincular os professores às matérias que ele pode lecionar				
Coluna	**Tipo**	**Nulo**	**Ligações de**	**Comentários**
CPF_Prof (*Extrangeira*)	char(11)	Não	PFDD	Somente números do CPF como identificador
ID_Mat (*Extrangeira*)	integer	Não	MFDD	Identificador da matéria
Periodo	varchar(20)	Não	-	Ex.: Matutino, Vespertino ou Noturno

Figura 3. Dicionário de dados documentando as funções dos dados.

Apesar de ficar visualmente fácil de se entender os dados do banco selecionado, qualquer atualização no banco de dados, seja ela de exclusão, inserção ou mesmo alteração, requer uma atualização manual da planilha no *Microsoft Excel* e este empecilho torna essa opção de criação do dicionário de dados muito custosa para ser executada em um ambiente de produção, fazendo com que os cientistas de dados optem por não desenvolver uma documentação adequada.

Analisando essa fraqueza da forma manual, uma ferramenta para criação automática do dicionário de dados foi desenvolvida para demonstrar a facilidade e ganho de performance na atualização dos dados e otimização da geração inicial.

6. Resultados – Apresentação do *Framework*

O Data Dictionary foi desenvolvido em dois projetos que o consiste, sendo: API e *Front*. A API consiste em toda integração com os bancos de dados, é responsável por formatar os dados e apresentar para o usuário, foi desenvolvido utilizando-se das tecnologias PHP 7.2 e *Laravel Framework*. Já o *Front*, que é a tela apresentada para o usuário e que faz a comunicação com a API, foi desenvolvido nas tecnologias HTML 5, CSS 3, *JavaScript* e VueJS 2.

Com o projeto instalado e configurado, basta acessá-lo e a tela com a lista de projetos será apresentado, nessa tela será escolhido o projeto que irá ser manipulado, além disso, você pode criar um novo projeto ou deletar um projeto já existente.

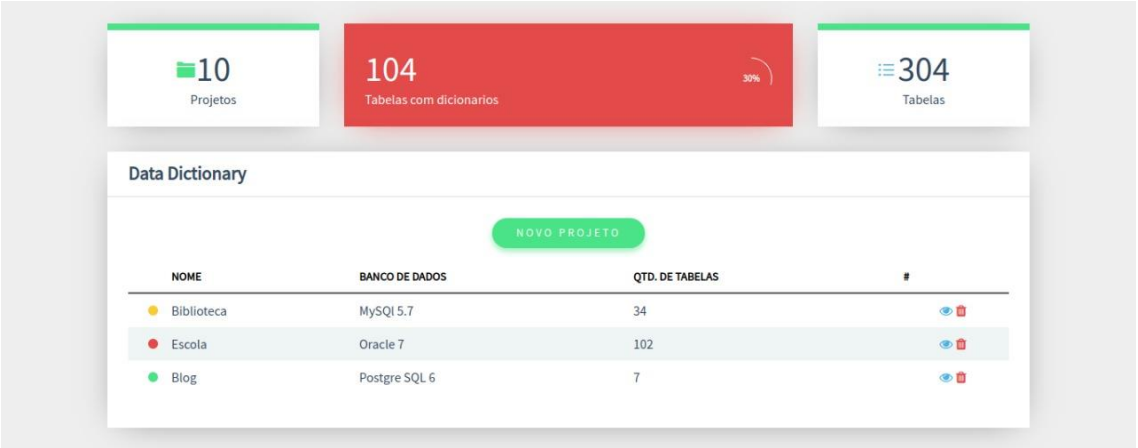

Figura 4. Aparência do sistema em funcionamento.

6.1 Criando um novo projeto

Para criar um novo projeto, deve ser clicado no botão "Novo Projeto" e preencher as informações pedidas, que são: Driver, Host, Porta, Usuário e Senha, logo após preencher, basta clicar no botão Criar que o projeto será criado.

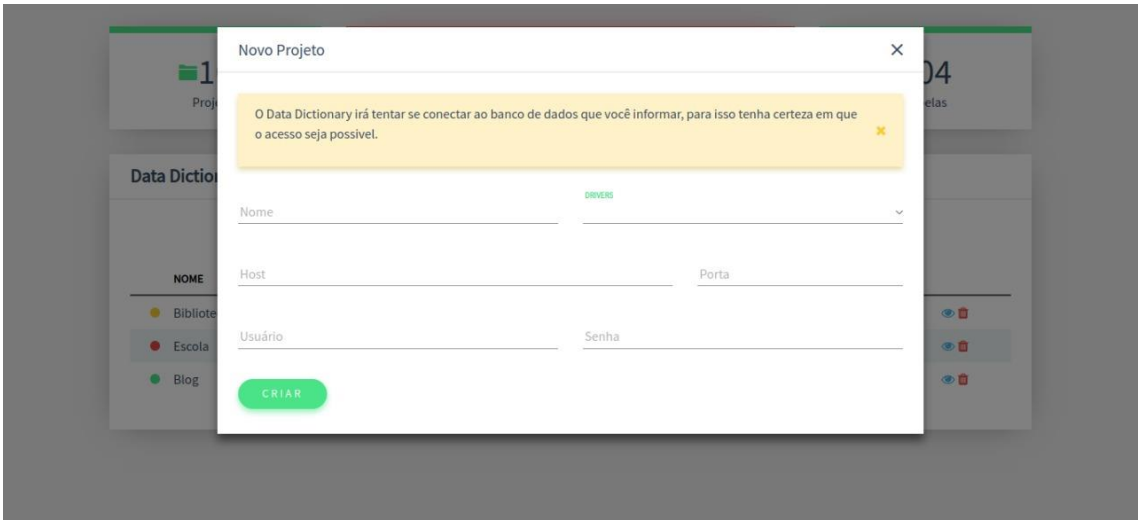

Figura 5. Adicionando os dados de um novo projeto.

6.2 Deletando um projeto existente

Na lista de projetos, na última coluna da tabela, tem um botão vermelho com um ícone de um lixeira, ao clicar nele, um modal irá aparecer perguntando se tem certeza que deseja excluir, após confirmar, o projeto será excluído.

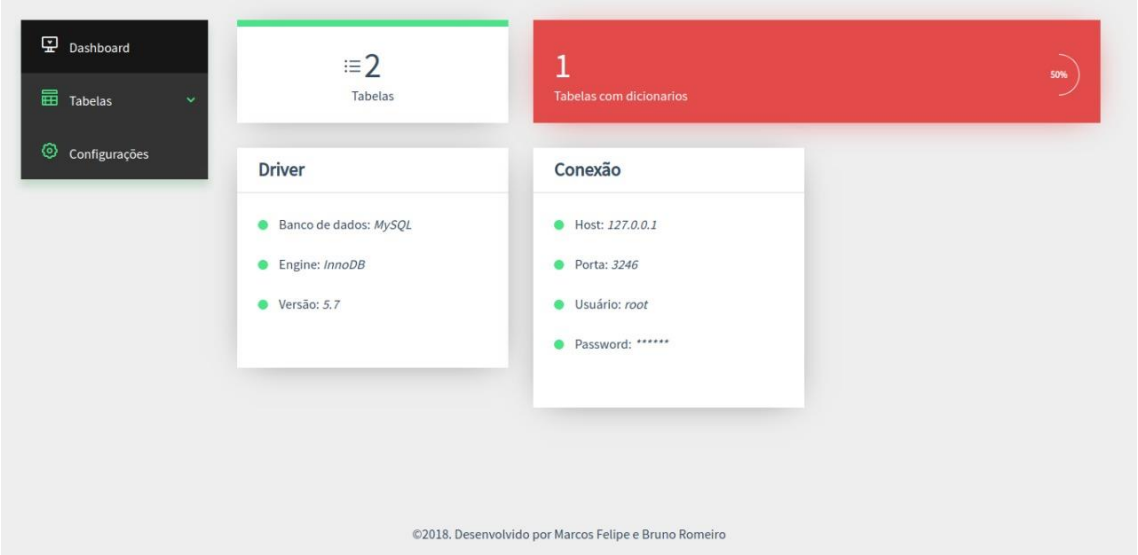

Figura 6. Verificando andamento de dicionários existentes.

6.3 Acessando um projeto

Na lista de projetos, na última coluna da tabela, tem um ícone de um olho, ao clicar nesse ícone será acessado o projeto.

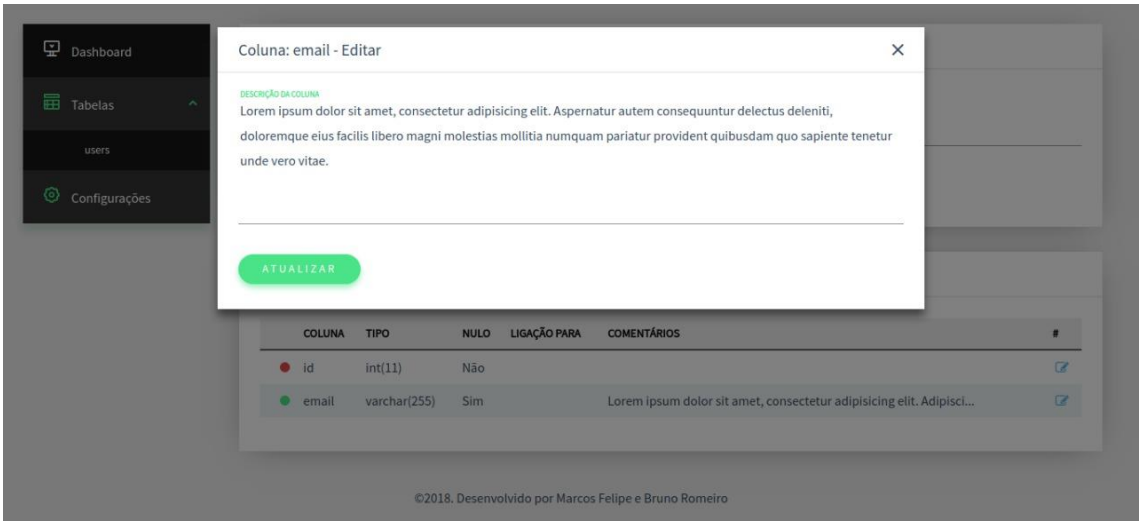

Figura 7. Editando a descrição de uma tabela.

6.5 Preenchendo um dicionário de dados

Para preencher o dicionario de dados de um banco, primeiro você precisa acessá-lo, no lado esquerdo do Data Dictionary, terá um menu com uma opção "Tabelas", ao clicar nessa opção, outras opções aparecem, sendo as tabelas que esse projeto contém, basta clicar em uma cima de uma tabela para acessar o dicionario de dados dessa tabela.

Após acessado, pode ser possível adicionar uma descrição a essa tabela e para salvar essa descrição, basta clicar no botão atualizar.

Para adicionar observações as colunas, logo na tabela de colunas, na última coluna terá um ícone de edição, ao clicar, um modal irá aparecer com um campo texto, se essa coluna já estiver alguma observação, esse campo vem preenchido com essa observação, caso contrário, vem vazio. Após perencher esse campo, ao clicar em Salvar, essas observações serão salvas na coluna.

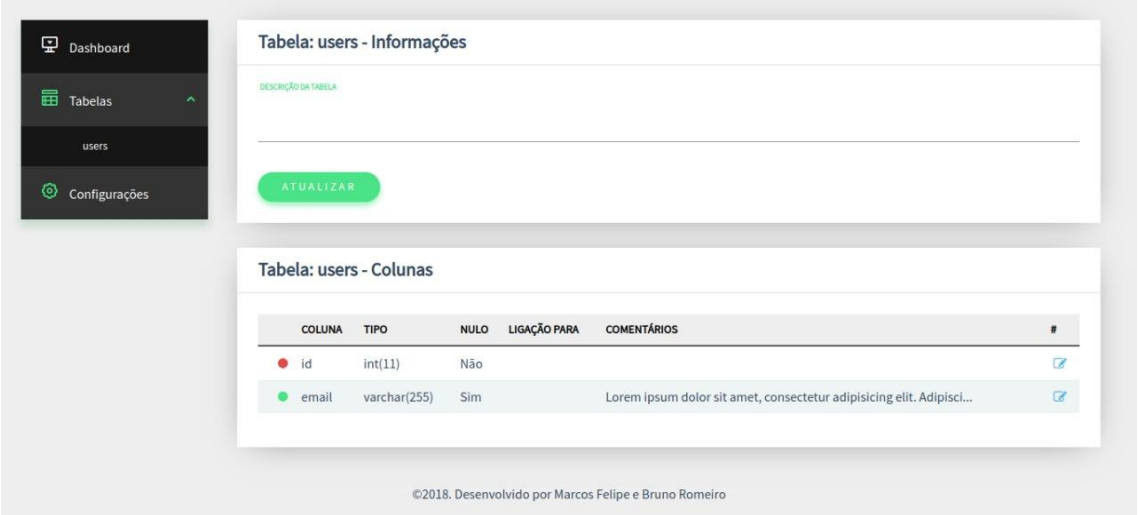

Figura 8. Tela de criação dos dicionários.

7. Conclusão

Baseando-se na experiência em que o grupo teve para desenvolver e validar os dados desta pesquisa, afirma-se que a necessidade de uma boa documentação auto-explicativa é evidente e facilita qualquer manipulação vinda de desenvolvedores e/ou analistas que não participaram do desenvolvimento inicial, seja ela preventiva ou corretiva.

Com os prazos apertados nota-se que muitos desenvolvedores e analistas optam por não fazer uma documentação eficiente, o que prejudica abundantemente os novos integrantes da equipe designada para o projeto.

Analisando esta pesquisa, nota-se como um dicionário de dados pode ser intuitivo mesmo para uma pessoa leiga na área de banco de dados e que não faz parte do projeto de desenvolvimento. Apenas ao olhar o dicionário de dados, entende-se que existe uma coluna com um determinado nome e ela possui uma função de armazenar a informação que está descrita na mesma linha.

Se comparado o *framework* de geração automática com a geração tradicional feita manualmente no *software* de planilhas eletrônicas, *Microsoft Excel*, conclui-se que a geração pelo *framework* pode ser relativamente mais rápida, intuitiva e de fácil manutenção. Uma geração automatizada já vem com a estrutura pronta, não sendo necessário iniciar todo o trabalho do zero, digitando os nomes das colunas, os tipos, e outros dados que já podem ser importados do banco de dados, exigindo apenas a descrição de cada coluna.

Define-se, então, a importância de um dicionário de dados para aprimorar a velocidade dos projetos e facilitar o trabalho de um administrador de banco de dados.

8. Referências

Batini, Carlo, Scannapieco, Maria (2006), “Data Quality Concepts, Methodologies and Techniques”, Springer-Verlag Berlin Heidelberg, p. 19-68.

English, Larry Page (1999), “Improving Data Warehouse and Business Information Quality”, Indianapolis, Indiana, Wiley Publishing Inc, p. 304-311.

English, Larry Page (2009), “Information Quality Applied – Best practices for improving business information, processes and systems”, Indianapolis, Indiana, Wiley Publishing Inc, p. 57-245.

Ferré, Rodrigo (2016), “O que é MER e como é aplicado em meu site?”, <https://www.next4.com.br/blog/o-que-e-mer-modelo-de-entidade-relacional-e-como-e-aplicado-no-meu-site/>. Maio.

Hostdime (2017), “10 razões para migrar do MySQL para o MariaDB”, <http://blog.hostdime.com.br/painel/10-razoes-para-migrar-o-mysql-para-mariadb/>. Fevereiro.

IBM (s/d), “Tabelas do dicionário de dados”, <https://www.ibm.com/support/knowledgecenter/pt-br/SSLKT6_7.6.0/com.ibm.mt.doc/configur/r_data_dictionary_tables.html>.

Meyer, Maximiliano (2013), “O que é o excel?”, <https://www.aprenderexcel.com.br/2013/tutoriais/o-que-e-excel/>. Julho.

Monteiro, Dani (2017), “Documentação da metadados”, <http://db4beginners.com/blog/documentacao-de-metadados/>. Dezembro.

Praciano, Elias (2013), “Explicando o PHP para iniciantes”, <https://elias.praciano.com/2013/10/o-que-e-php-um-texto-para-iniciantes/>. Outubro.

Resende, Dirceu (2017), “SQL Server - Como documentar o banco de dados e seus objetos”, <https://www.dirceuresende.com/blog/sql-server-como-documentar-o-banco-de-dados-e-seus-objetos-tabelas-procedures-colunas-utilizando-extended-property>. Outubro.

Schneider, Acácio (2011), “Dicionário de Dados do Sistema de Informática”, <http://projeto3osemestre.blogspot.com/2011/05/dicionario-de-dados-do-sistema-de.html>. Maio.