UNIVERSIDADE FEDERAL DE PERNAMBUCO
BACHARELADO EM SISTEMAS DE INFORMAÇÃO
CENTRO DE INFORMÁTICA
2015.2

Datafeed: Uma Ferramenta para coleta de *feedback* sobre Dados publicados na Web

Aluno: Helton Douglas Araújo dos Santos (hdas@cin.ufpe.br)
Orientadora: Bernadette Farias Lóscio (bfl@cin.ufpe.br)

Recife, 11 de janeiro de 2016

UNIVERSIDADE FEDERAL DE PERNAMBUCO
BACHARELADO EM SISTEMAS DE INFORMAÇÃO
CENTRO DE INFORMÁTICA
2015.2

Helton Douglas Araújo dos Santos

Datafeed: Uma Ferramenta para coleta de *feedback* sobre Dados publicados na Web

Trabalho apresentado à disciplina de Trabalho 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 Sistema de Informação.

Orientadora: *Profa. Dra. Bernadette Farias Lóscio*

Recife, 11 de janeiro de 2016

Deem graças ao Senhor, clamem pelo seu nome,
divulguem entre as nações o que ele tem feito.
1 Crônicas 16:8

Agradecimentos

Agradeço primeiramente a Deus, por ter me dado saúde e ter me ajudado desde o início desse curso.

A minha família, meus pais, meu irmão e a minha namorada, por ter me dado total apoio e incentivo para que eu pudesse chegar até aqui, me incorajando e me motivando ao longo dessa jornada.

A Profa. Bernadette, que me orientou neste trabalho, com toda paciência e sabedoria.

A todos os Professores, que de alguma forma, dedicaram um pouco do seu tempo a mim, por não somente terem me ensinado, mas por terem me feito aprender.

Ao Centro de Informática (CIn-UFPE), por ter sido um local de muito aprendizado e crescimento no âmbito profissional e acadêmico.

Aos meus amigos e irmãos na amizade, que fizeram parte da minha formação e vão continuar presentes em minha vida.

A todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

Resumo

O ecossistema de dados armazenados na Web é atualmente um recurso bastante acessado por pessoas que demonstram interesse por dados de domínios específicos, como os dados disponibilizados pelo governo. A publicação de dados na *Web* vem permitindo o compartilhamento de dados em larga escala, proporcionando acesso aos dados a uma grande variedade de público, com diferentes níveis de experiência, permitindo que qualquer pessoa possa livremente usá-los, reutilizá-los e redistribuí-los. Apesar do grande interesse na publicação e no consumo dos dados na Web, ainda existe um problema, relacionado à qualidade dos dados, que pode ter grande impacto no reúso dos dados oferecidos. Em muitos casos, a baixa qualidade é apenas uma consequência do uso de dados que possuem algum problema ou anomalia desde a sua fonte de origem. Porém, em outros casos, a baixa qualidade pode ser decorrente de falhas no momento da geração ou publicação dos dados na Web. Além disso, os dados podem apresentar outros problemas relativos à frequência de atualização ou à falta de informações que facilitem a compreensão ou a manipulação dos dados. Nesse contexto, torna-se muito importante obter um *feedback* dos consumidores de dados a fim de identificar possíveis falhas nos dados publicados na Web, bem como identificar a necessidade de publicação de novos dados.

Palavras-Chave: *Feedback*, Dados publicados na Web, Dados na Web, Dados abertos, Dados conectados, Consumo de Dados

Abstract

The data stored in the Web ecosystem is currently a resource often accessed by people who show interest in specific domain data, like the data provided by the government. The publication of data on the Web has been allowed large-scale data sharing, providing data access to a wide range of public, with varying levels of experience, allowing anyone to use it freely, reuse it and redistribute it. Despite the great interest in the publication and consumption of data on the Web, there is a problem related to the data's quality, which may have a great impact on the reuse of offered data. In many cases, low quality is simply a consequence of using data that has any problem or malfunction since its origin's source. But in other cases, the poor quality may be due to failures at the time of generation or publication of data on the Web. In addition, the data may have other problems related to the update frequency or lack of information that facilitate the data understanding and manipulation. In this context, it is very important to get feedback from data consumers to identify possible gaps in the data published on the Web, as well as to identify the need for new data publication.

Keywords: *Feedback, Data published on the Web, Data on the Web, Open Data, Linked Data, Data Consumption*

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

Neste capítulo, uma breve introdução do trabalho será feita, mostrando a motivação para o seu desenvolvimento, os objetivos, suas contribuições e uma descrição de como o documento está estruturado.

1.1 Motivação

Dados publicados na Web são, atualmente, um recurso bastante acessado por pessoas que demonstram interesse por dados de domínios específicos como, por exemplo, os dados governamentais. Desde o seu surgimento, a Web vem sendo o principal mecanismo de publicação e consumo de dados, o paradigma de WoT *(Web of Things*) juntamente com o movimento de OWP (*Open Web Platform*) só confirmam o poder dessa plataforma de compartilhamento. É importante ressaltar que o interesse na publicação de dados na Web não é algo novo [1, 2]. Porém, nos últimos anos, este interesse tem se caracterizado pela publicação de dados, provendo compartilhamento e a reutilização desses dados.

Apesar do grande interesse na publicação e no consumo dos dados na Web, ainda existem problemas relacionados à qualidade dos dados, que podem ter grande impacto no reuso dos dados oferecidos. Em muitos casos, a baixa qualidade é apenas uma consequência da forma como esses dados estão disponíveis em sua fonte de origem, em outros casos, essa baixa qualidade pode ser decorrente de falhas no momento da geração ou da publicação dos dados na Web. Além disso, os dados podem apresentar outros problemas relativos à frequência de atualização ou à falta de informações que dificultam a compreensão ou a manipulação dos dados.

Dessa forma, apenas disponibilizar o acesso aos dados não é suficiente. É necessário que os publicadores dos dados tenham uma atenção especial na qualidade dos dados publicados, a fim de que possam ser facilmente compreendidos e utilizados por consumidores de diferentes níveis de experiência, além de disponibilizar esses dados em formatos que possam ser facilmente processados por aplicações. Porém, a heterogeneidade dos dados e a falta de padrões para descrição e acesso aos conjuntos de dados, ainda tornam o processo de publicação, compartilhamento e consumo de dados uma tarefa complexa [3].

Na busca de contornar esses problemas, foram criadas as boas práticas para a publicação de dados na Web [4], às quais foram projetadas para atender às necessidades de profissionais de gestão da informação, desenvolvedores e grupos mais amplos, como os cientistas interessados na reutilização de dados de pesquisa na Web [3]. Contudo, ainda existem desafios a serem encarados pelos produtores de dados, um deles consiste em determinar o benefício e a relevância dos conjuntos de dados que são publicados na Web, bem como identificar anomalias dispersas nesses conjuntos de dados. Descrito nas boas práticas [4], o *feedback* é o principal meio de obtermos informações relevantes a respeito da experiência de uso dos dados por parte dos consumidores, ele permite que o consumidor tenha voz e contribua na identificação de anomalias, como erros de formatação e inconsistência dos dados. Além disso, o *feedback* permite que através dessas informações possamos determinar o benefício e a relevância dos conjuntos de dados publicados na Web.

1.2 Objetivos e Contribuições

O objetivo principal deste trabalho é o desenvolvimento de uma ferramenta de coleta de *feedback* para dados publicados na Web. A partir do *feedback* dos consumidores dos dados, é possível estabelecer uma forma de comunicação entre os provedores desses dados e seus consumidores. A comunicação entre essas duas partes ajudará na identificação de possíveis falhas e anomalias nesses conjuntos de dados. No decorrer dessa identificação, os provedores ou publicadores terão conhecimento dos problemas identificados e passarão a trabalhar em cima de possíveis correções, melhorando assim a qualidade desses dados. Dessa forma, um canal de comunicação entre publicadores de dados e seus consumidores é estabelecido, construindo um ambiente colaborativo e construtivo. A ferramenta proverá a coleta desses *feedbacks* e os disponibilizará de forma aberta para que publicadores e consumidores de dados possam acessá-los.

Dessa forma, podemos destacar as principais contribuições deste trabalho:

- Implementação de uma ferramenta que coletará e armazenará o *feedback* de consumidores de dados para conjuntos de dados publicados na Web;
- Disponibilização de uma *API* que permita a comunicação entre os portais de catalogação e publicação de dados e a ferramenta proposta;

- A disponibilização dos dados de *feedback* de forma aberta para que publicadores e consumidores de dados possam acessá-los e reutilizá-los.

1.3 Estrutura do documento

Os capítulos seguintes estão estruturados da seguinte forma: O Capítulo 2 aborda a contextualização deste trabalho, apresentando os conceitos de dados na Web, vocabulários utilizados na ferramenta e o conceito de *feedback* de conjuntos de dados publicados na Web. O Capítulo 3, por sua vez, contextualiza e descreve a ferramenta criada neste trabalho. O Capítulo 4 apresenta a implementação e avaliação da ferramenta. Por fim, o Capítulo 5 expõe a conclusão deste trabalho, além de sugestões para trabalhos futuros.

2 Contextualização

Neste capítulo serão abordados conceitos essenciais para o entendimento deste trabalho. Na Seção 2.1, apresentamos o conceito de dados na Web, seu ciclo de vida, quem são os publicadores e consumidores de dados, e o que são dados abertos. A Seção 2.2 apresenta o *Web Annotation Data Model*[1], modelo de anotação de *feedbacks* da ferramenta proposta. A Seção 2.3 descreve o vocabulário DUV[2], que por sua vez, foi utilizado como vocabulário padrão da ferramenta. Na Seção 2.4, é apresentado o vocabulário DCAT[3], vocabulário muito importante e bastante utilizado na catalogação de conjunto dados publicados na Web. Por fim, a Seção 2.5 descreve e conceitua o *feedback* de um conjunto de dados publicados na Web, bem como, sua importância.

2.1 Dados na Web

Nos últimos anos, a Web tem se tornado cada vez mais uma grande plataforma de compartilhamento e consumo de dados. Um enorme volume de dados vem sendo gerado e tornando-se disponível na Web. A publicação desses dados está trazendo vários benefícios para a sociedade e inúmeras aplicações buscam fazer o uso dessas fontes de dados disponíveis na Web, com o objetivo de gerar informações úteis e relevantes, e até mesmo com o intuito de gerar novos dados.

Com o cenário de crescimento na publicação desses dados, é importante observar alguns aspectos relacionados à proveniência e à qualidade dos dados como, por exemplo, disponibilizar metadados que permitam descrever o histórico de produção e publicação desses dados. Em particular, no contexto da Web, onde os dados são publicados a partir de diferentes provedores e cobrem diferentes domínios.

Devido a sua flexibilidade, a Web possibilita a publicação e o consumo de dados de maneira bastante simples como, por exemplo, não há a exigência de sistemas que controlem o acesso concorrente aos dados. Nesse contexto, os dados publicados na Web têm como característica a ausência completa ou parcial de um esquema que defina rigorosamente a estrutura dos dados a serem armazenados. Essa flexibilidade facilita o

[1] *http://www.w3.org/TR/annotation-model/*
[2] *http://www.w3.org/TR/vocab-duv/*
[3] http://www.w3.org/TR/vocab-dcat/

processo de publicação de dados, mas, em contrapartida, torna mais complexo o processo de consumo.

É importante ressaltar dois papéis que influenciam diretamente nesse cenário de publicação e consumo de dados: os provedores e os consumidores de dados. Os provedores ou publicadores tem como papel principal, publicar e disponibilizar os dados na Web. Os consumidores, por sua vez, que também podem ser provedores, são aqueles que fazem o uso desses dados, seja para a geração de informações úteis, seja para a geração de novos dados.

2.1.1 Ciclo de Vida

O processo de publicação e consumo de dados na Web envolve várias fases que vão desde o planejamento até o refinamento dos dados publicados. Esse conjunto de fases que compõem o processo de publicação e consumo dos dados é chamado de ciclo de vida dos dados na Web.

O ciclo de vida representado abaixo foi proposto por Lóscio et al. [3], sendo uma instância do *Abstract Data Lifecrycle Model (ADLM)* proposto por Möller [5]. Segundo Müller [5], *ADML* é um modelo genérico para representação de ciclo de vida de dados e metadados, o qual foi derivado a partir de uma coleção de modelos do ciclo de vida para domínios centrados em dados. De acordo com esse modelo genérico, o ciclo de vida deve ser composto pelas seguintes fases: desenvolvimento de ontologia, planejamento, criação, arquivamento, refinamento, publicação, acesso, uso externo, *feedback* e término. Embora esse modelo abstrato tenha definido todas essas fases, no ciclo de vida para dados publicados na Web, não há a cobertura de todas essas definições. Segundo Lóscio et al. [3], a criação de ontologias é considerada uma atividade independente e por isso não foi incluída. O arquivamento e o término não foram considerados, pois uma vez que o dado foi publicado na Web ele sempre estará disponível.

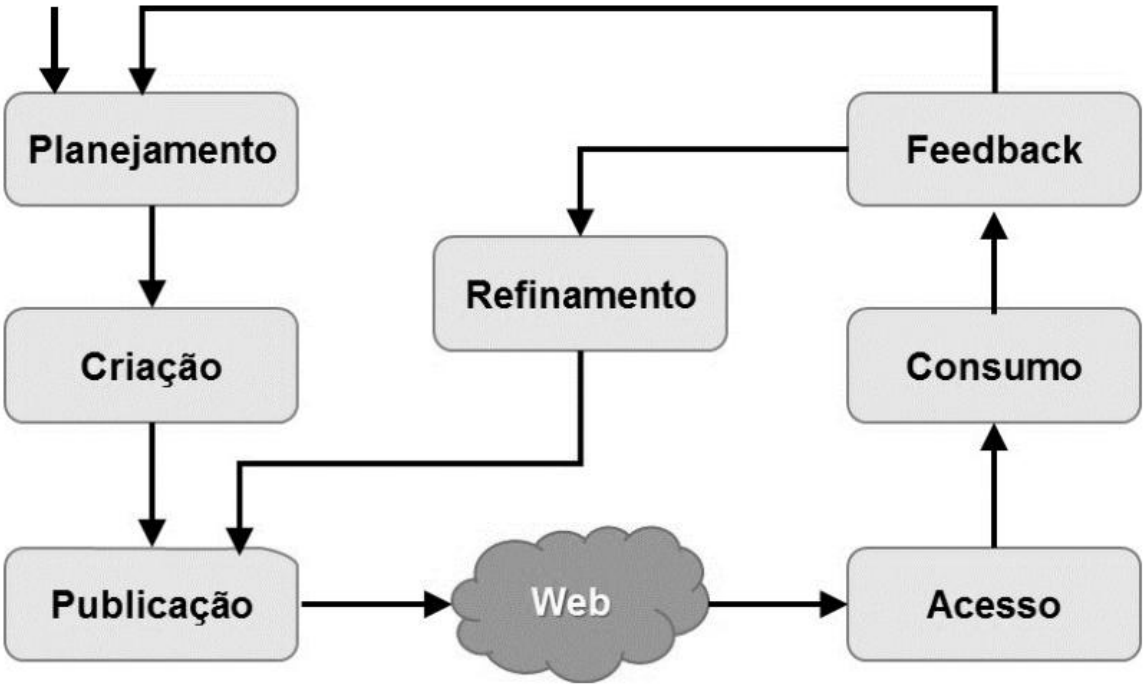

Figura 1. Ciclo de vida de dados na Web

Fonte: [3]

A Figura 1 representa o ciclo de vida de dados na Web, onde podemos visualizar cada fase descrita no modelo [3]. A seguir serão brevemente descritas cada fase do ciclo de vida dos dados na Web:

- Planejamento: É a primeira fase do ciclo de vida, é nela que surge a intenção de publicar os dados, no qual se estende até a seleção dos dados que serão publicados.
- Criação: Está fase diz respeito ao momento em que os dados são criados, passando pela etapa da extração dos dados até a transformação para o formato adequado.
- Publicação: É a fase na qual os dados serão disponibilizados de forma pública na Web. Ferramentas de catalogação de dados, como o *CKAN*[4], são utilizadas nessa fase.

[4] *http://ckan.org/*

- Acesso: Esse é o momento do ciclo de vida em que os consumidores ganham acesso aos dados. Isso ocorre através da divulgação de que os dados foram liberados.
- Consumo: Denota o momento em que os dados estão sendo consumidos, seja para a criação de novos dados, seja para a geração de informação. Está fase está diretamente relacionada aos consumidores de dados, que podem ser desenvolvedores interessados em criar uma aplicação que faça o uso daqueles dados.
- *Feedback*: Está é uma fase de fundamental importância, ela consiste na coleta de informações passadas pelos consumidores sobre os dados e metadados que foram utilizados por eles.
- Refinamento: É a última fase do ciclo, nela são aplicadas todas as atividades relacionadas às adições ou atualizações dos dados já publicados, e essas manutenções poderão ser feitas a partir de *feedbacks* coletados na fase anterior.

2.1.2 Consumidores de dados

A publicação de dados na Web possibilita que seus consumidores utilizem e reutilizem os dados de acordo com seu interesse. Considerando o ecossistema de dados abertos, os consumidores exercem um papel que merece um grande destaque, pois a maioria dos dados publicados na Web são destinados ao seu consumo. Os consumidores são aqueles que consomem ou utilizam os dados para gerar informações, bem como novos dados [3]. Há vários perfis de consumidores de dados no ecossistema Web, abaixo estão listados alguns deles:

- Desenvolvedor de software: São profissionais da área de tecnologia da informação que desenvolvem aplicações com o objetivo de consumir dados na Web.A partir dos dados coletados na Web, podem ser desenvolvidas aplicações que geram informações que serão disponibilizadas para usuários finais.
- Empresas ou Organizações: Existem empresas que utilizam dados publicados na Web com objetivo de coletar informações que ajudem na tomada de decisão empresarial. Umas das categorias de dados em que há um grande interesse por

parte dos empresários são os dados financeiros. Esses dados podem ser transformados em informações valiosas, as quais podem impactar diretamente na decisão estratégica da organização.

- Sociedade: A sociedade é um dos perfis de consumidores de dados mais comum. Pessoas que tem interesse em uma determinada categoria de dados e acessam portais de dados abertos em busca desses dados são exemplos reais desse perfil de consumidor.

2.1.3 Publicadores de dados

O forte crescimento do ecossistema de dados na Web foi um reflexo do aumento da intenção de se publicar dados. Dessa forma, um dos papéis que mais se destacam na publicação de dados são os próprios publicadores, o seu papel principal é publicar e fornecer os dados com qualidade, e que sejam facilmente compreendidos pelos consumidores, não esquecendo também publicação em um formato que seja facilmente processado por sistemas e aplicações. No contexto da publicação de dados, dois perfis de publicadores se destacam:

- Governo: A publicação de dados por parte do governo vem se destacando com o movimento de transparência de informações incentivado por parte dos governantes no mundo. São inúmeros portais no Brasil e no mundo que proveem dados governamentais na Web. O portal brasileiro de dados abertos[5] só confirma o grande número de conjuntos de dados governamentais publicados na Web, como também o portal da transparência do governo federal brasileiro[6].
- Desenvolvedor de software: Os desenvolvedores de aplicações além de consumir também exercem o perfil de publicadores. Um dos seus objetivos é construir uma aplicação que colete os dados brutos, transformando-nos em informações e disponibilizando-os em uma forma mais simples e legível para os consumidores. No contexto brasileiro, nos últimos anos houve um grande aumento no número de aplicativos que consomem os dados publicados na Web. Esse aumento ocorreu, principalmente, graças aos eventos chamados *hackathons,* cujo objetivo é reunir

[5] *http://dados.gov.br/*
[6] *http://www.portaltransparencia.gov.br/*

profissionais ligados ao desenvolvimento de software com o intuito de desenvolver aplicações que atendam a um fim específico, e que sejam inovadoras e de interesse da sociedade.

2.1.4 Dados abertos

Definido pela *Open Knowledge Foundation (OKF)*[7], dados abertos são dados que podem ser livremente utilizados, reutilizados e redistribuídos por qualquer pessoa, sem qualquer restrição legal, tecnológica ou social [11]. A OKF é uma das maiores incentivadoras do movimento de dados abertos, através do Guia de Dados Abertos[8], é discutido aspectos legais, sociais e técnicos dos dados abertos.

No cenário de dados abertos, tanto a publicação quanto o consumo de dados são tarefas fundamentais. Um dado é considerado aberto quando apresenta as seguintes características [21]:

- Disponibilidade e Acesso: Os dados precisam estar disponíveis por inteiro e devem estar em formato conveniente e modificável.
- Reutilização e Redistribuição: Os dados precisam ser fornecidos em condições de reúso e redistribuição podendo ser combinados com outros dados.
- Participação Universal: Todos podem usar, reusar e redistribuir os dados sem restrições de áreas, pessoas ou grupos.

Seguindo o movimento dos dados abertos, governos de diversos países estão usando a Web como meio para publicação de dados e informações sobre suas administrações. Esses dados, denominados Dados Abertos Governamentais, podem ser facilmente encontrados nos chamados Portais de Dados Abertos, os quais oferecem uma interface mais amigável para o acesso aos dados. Com o intuito de chegar a um consenso dos requisitos necessários para se caracterizar uma base de dados abertos, o grupo de trabalho, *Open Government Working Group*, elaborou oito princípios para Dados Abertos Governamentais [22]:

[7] *http://okfn.org/*
[8] *http://opendatahandbook.org/guide/pt_BR/*

- Completos: Todos os dados devem estar disponíveis e não limitados. Um dado público é o dado que não está sujeito a limitações válidas de privacidade, segurança ou privilégios de acesso.
- Primários: Os dados devem estar em formato bruto, sem agregação ou modificação.
- Atuais: Os dados devem ser publicados tão rapidamente quanto necessário para preservar o seu valor.
- Acessíveis: Os dados devem ser acessíveis pelo maior número possível de usuários e para o maior número possível de finalidades.
- Processáveis por máquinas: Os dados devem ser razoavelmente estruturados para permitir processamento automatizado.
- Não discriminatórios: Os dados devem ser disponíveis para todos, sem necessidade de cadastro.

A publicação dos dados é parte chave do processo de abertura de dados, ela mantém a interface de quem disponibiliza os dados com as pessoas que irão utilizá-los. Os dados devem ser disponibilizados nos chamados catálogos de dados, os quais são ferramentas ou serviços responsáveis pela gestão e publicação de dados e metadados na Web. No Brasil, a principal referência na disponibilização de conjuntos de dados é o Portal Brasileiro de Dados Abertos[9], portal utilizado pelo governo brasileiro que tem o objetivo de disponibilizar de forma aberta, dados e informações públicas. Esse portal utiliza o software CKAN[10] como ferramenta de catalogação dos dados, o qual é muito utilizada nesse contexto.

O CKAN é um software livre que provê um portal de dados, ele permite a exposição de catálogos de dados, bem como funções para publicação, armazenamento e gerenciamento dos conjuntos de dados. Esse software foi inicialmente desenvolvido pela OKF, mas atualmente, é desenvolvido e mantido pela comunidade CKAN. Além do portal brasileiro, ele também é usado nos principais portais de outros países, como Reino Unido, Estados Unidos e Holanda [24]. O CKAN conta com uma *API* para acesso automático, visualização dos dados em qualquer dispositivo, visualização de *dashboards* pré-elaboradas e faz o uso do conceito de Software como Serviço (SaaS).

[9] *http://dados.gov.br/*
[10] *http://ckan.org/*

Além do CKAN, também existe o Socrata, um software pago, também utilizado na catalogação de dados e que tem como seu principal diferencial a construção de visualizações mais elaboradas para o portal, como formatação condicional, gráficos e mapas. Assim como o CKAN, o Socrata também utiliza uma arquitetura baseada em SaaS e provê uma API para acesso automático. O site oficial do Socrata[11], disponibiliza toda a documentação necessária para o desenvolvimento de aplicações e uso dos dados nos portais que utilizam seu software. O Socrata é utilizado em portais de várias cidades dos Estados Unidos, como Chicago e Nova York.

2.2 Web Annotation Data Model

O *Web Annotation Data Model*[12] é um modelo de anotação para dados na Web. Seu objetivo é fornecer um modelo de descrição e formato padrão para permitir que anotações sejam compartilhadas entre sistemas [6], criando formas de associações entre partes distintas da informação. São consideradas anotações, comentários sobre compartilhamento de fotos ou vídeos, menções de recursos da Web, sejam em redes sociais ou em outras plataformas. Essa interoperabilidade pode existir tanto no compartilhamento, como também na migração de anotações privadas entre dispositivos e plataformas.

O modelo de anotação de dados na Web provê um *framework* extensível e interoperável para expressar anotações que podem ser facilmente compartilhadas entre plataformas e satisfazer exigências complexas de requisitos. Uma anotação é considerada um conjunto de recursos conectados, o qual geralmente possui um corpo (*body*) e um objetivo (*target*). Um anotação transporta o conteúdo do seu corpo, que por sua vez está ligado a um objetivo, fazendo com que a natureza exata desta relação mude de acordo com a intenção da anotação. Isso resulta em uma perspectiva de um modelo baseado em três partes, descritos na Figura 2.

Uma anotação é um modelo único e consistente, que pode ser usado por todas as partes interessadas, e pode possuir propriedades descritivas adicionais. Abaixo serão apresentados três exemplos de um simples caso de uso descrito em [6], onde o exemplo descreve uma requisição *http* que utiliza o método *post*, esse método representa um

[11] *https://www.socrata.com/*
[12] *http://www.w3.org/TR/annotation-model/*

comentário sobre uma determinada página na Web. O exemplo está descrito em três formatos. O primeiro formato está representado em *JSON-LD*, que é um formato para dados conectados na Web. O segundo formato está descrito em *Turtle*, que é uma representação de dados em *RDF*. Por fim o terceiro formato está representado na forma de diagrama.

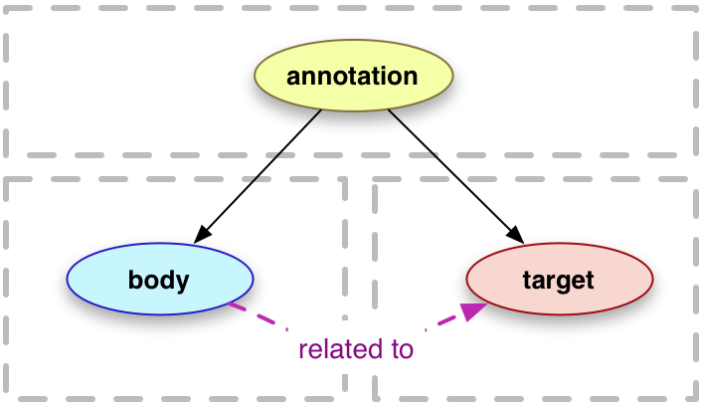

Figura 2. Modelo de anotação para dados na Web

Fonte: [6]

- Termos utilizados nos exemplos:
  - *@id*: Propriedade que identifica a anotação, ela possui uma *URI* que a identifica.
  - *@type*: Relacionamento entre a anotação e sua classe, que possui como *URI* "*rdf:type*".
  - *Annotation*: Classe para anotações na Web. Uma classe de anotação deve ser associada com uma anotação usando *@type*.
  - *Body*: Relacionamento entre a anotação e seu corpo (*body*).
  - *Target*: Relacionamento entre uma anotação e seu objetivo (*target*).

```
{
  "@id": "http://example.org/anno1",
  "@type": "Annotation",
  "body": {"@id": "http://example.org/post1"},
  "target": "http://example.com/page1"
}
```

Figura 3. Formato 1: Modelo de anotação simples em JSON-LD

Fonte: [6]

```
<http://example.org/anno1> a oa:Annotation ;
    oa:hasBody <http://example.org/post1> ;
    oa:hasTarget <http://example.com/page1> .
```

Figura 4. Formato 2: Modelo de anotação simples em Turtle

Fonte: [6]

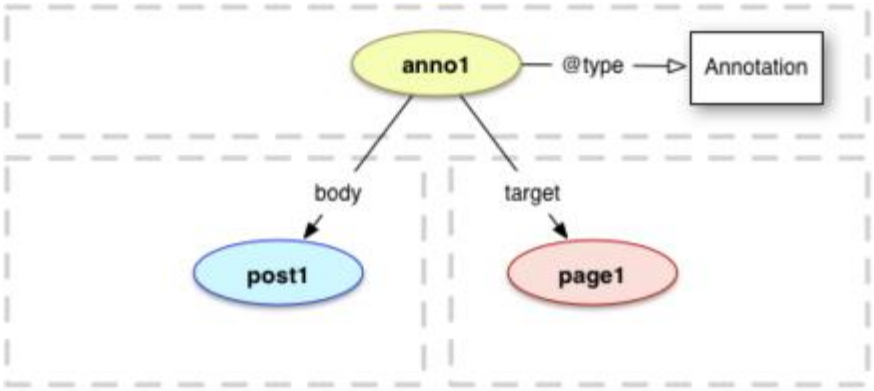

Figura 5. Formato 3: Modelo de anotação simples em forma de Diagrama

Fonte: [6]

2.3 Dataset Usage Vocabulary

Os conjuntos de dados publicados na Web são acessados e experimentados por inúmeros consumidores, com vários tipos de perfil. Visto que pouca informação sobre a experiência de uso dos consumidores é passada aos publicadores, foi criado o *Dataset*

Usage Vocabulary (DUV)[13], o qual pode ser usado para descrever experiências, citações e *feedback* sobre os conjuntos de dados a partir de uma perspectiva humana.

O vocabulário DUV é utilizado para preencher um nicho que ajuda a padronizar o modo como o uso dos conjuntos de dados publicados na Web poderá ser compartilhado. Esse vocabulário recomenda e exige que os publicadores de dados forneçam um mecanismo que colete o uso de dados de consumidores, na forma de *feedback*, citação e correção de dados [7].

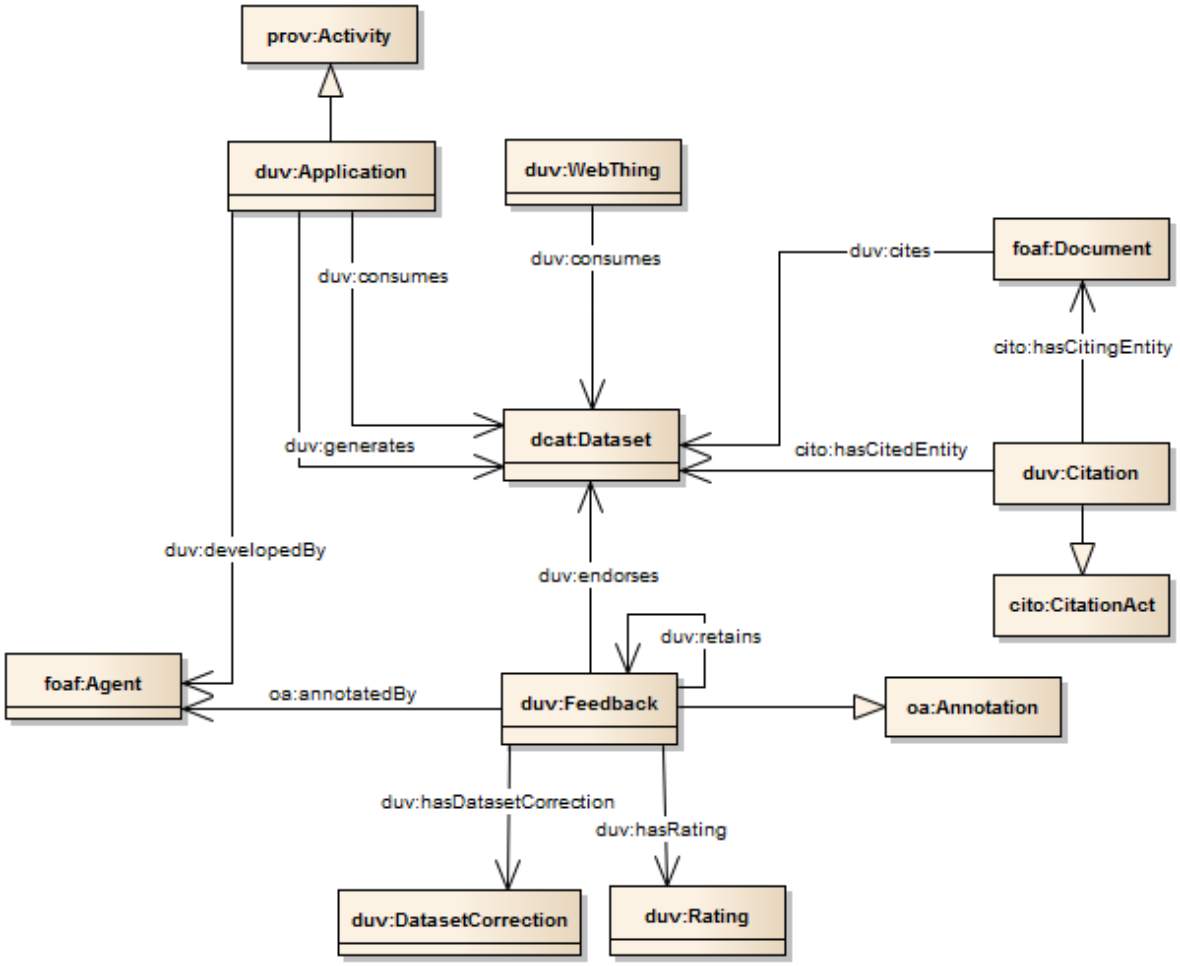

Figura 6. Diagrama DUV

Fonte: [7]

[13] *http://www.w3.org/TR/vocab-duv/*

O DUV é um vocabulário que reutiliza e estende classes e propriedades existentes de outros vocabulários, como o DCAT[14], para apoiar as citações, experiências e o *feedback* dos consumidores para conjuntos de dados. A Figura 6 apresenta um diagrama com a visão geral do vocabulário DUV, com todas as suas classes, propriedades e relações.

2.4 Data Catalog Vocabulary

O vocabulário de catalogação de dados (DCAT) foi projetado para facilitar a interoperabilidade entre catálogos de dados publicados na Web [8]. Ao utilizar o DCAT para a catalogação de dados, os publicadores poderão facilitar a descoberta desses dados e permitir que aplicativos possam consumir facilmente os metadados de vários catálogos publicados.

Os dados coletados a partir de um conjunto de dados podem vir em vários formatos, como *XML* e *Json* [8]. O DCAT não permite fazer quaisquer suposições ou especificações sobre o formato dos conjuntos de dados descritos em um catálogo. Para fornecer mais informações específicas sobre um conjunto de dados, outros vocabulários complementares podem ser usados em conjunto com o DCAT, como o *DCMI Metadata Terms*[15], que é utilizado na descrição de alguns metadados da classe *Catalog* do DCAT. O DCAT pode ser aplicável em muitos contextos, incluindo o *RDF*, que é acessível em *endpoints SPARQL*, e também pode ser serializado em *XML* ou em *Turtle*. O exemplo abaixo está descrito em *Turtle* e fornece uma visão geral de como o DCAT pode ser usado para representar um catálogo e também descrever um conjunto de dados publicado na Web. A Figura 7 apresenta a catalogação de três datasets, um deles é descrito na Figura 8. Seguem também (Tabela 1) os prefixos utilizados neste exemplo.

Tabela 1. Prefixos utilizados no vocabulário DCAT

Prefixo	***Namespace***
dcat	http://www.w3.org/ns/dcat#
dct	http://purl.org/dc/terms/

[14] *http://www.w3.org/TR/vocab-dcat/*
[15] *http://purl.org/dc/terms/*

dctype	http://purl.org/dc/dcmitype/
foaf	http://xmlns.com/foaf/0.1/
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
skos	http://www.w3.org/2004/02/skos/core#
vcard	http://www.w3.org/2006/vcard/ns#
xsd	http://www.w3.org/2001/XMLSchema#

```
:catalog
    a dcat:Catalog ;
    dct:title "Imaginary Catalog" ;
    rdfs:label "Imaginary Catalog" ;
    foaf:homepage <http://example.org/catalog> ;
    dct:publisher :transparency-office ;
    dct:language <http://id.loc.gov/vocabulary/iso639-1/en>  ;
    dcat:dataset :dataset-001  , :dataset-002 , :dataset-003 ;
    .
```

Figura 7. DCAT: Catálogo

Fonte: [8]

```
:dataset-001
    a dcat:Dataset ;
    dct:title "Imaginary dataset" ;
    dcat:keyword "accountability","transparency" ,"payments" ;
    dct:issued "2011-12-05"^^xsd:date ;
    dct:modified "2011-12-05"^^xsd:date ;
    dcat:contactPoint <http://example.org/transparency-office/contact> ;
    dct:temporal <http://reference.data.gov.uk/id/quarter/2006-Q1> ;
    dct:spatial <http://www.geonames.org/6695072> ;
    dct:publisher :finance-ministry ;
    dct:language <http://id.loc.gov/vocabulary/iso639-1/en>  ;
    dct:accrualPeriodicity <http://purl.org/linked-data/sdmx/2009/code#freq-W>  ;
    dcat:distribution :dataset-001-csv ;
    .
```

Figura 8. DCAT: Dataset

Fonte: [8]

2.5 Feedback para Dados na Web

O feedback é uma ferramenta importante na comunicação crítica e sugestiva entre dois indivíduos, ou mais. Essa crítica ou sugestão pode ser positiva ou não, de âmbito profissional ou pessoal, mas sempre vista de forma construtiva. A comunicação pode ser

realizada através de uma resposta escrita, um elogio, uma análise ou até mesmo uma opnião direta ao solicitante. Ele também é considerado como uma ferramenta de aprendizado e aprimoramento nos meios organizacionais, e que também pode ser aplicado nas demais áreas. Dentre essas áreas, a publicação de dados na Web vem utilizando essa ferramenta como um meio de comunicação entre publicadores e consumidores de dados.

Com a utilização do feedback, os publicadores de dados terão a garantia de que os dados publicados atenderão aos requisitos dos consumidores. Consumidores poderão relatar experiências de uso no consumo dos dados, bem como suas dificuldades na utilização, e também poderão dar sugestões de melhorias para um melhor aproveitamento dos dados que estão publicados.

Após coletado, o *feedback* auxiliará os provedores de dados a melhorarem a qualidade dos dados publicados. Ao descrever suas experiências de uso dos dados, os consumidores auxiliam e contribuem para a publicação de novos dados e para a manutenção desse ecossistema. Com o intuito do compartilhamento de usabilidade e de colaboração, o *feedback* deve estar disponível de forma pública, para que outros consumidores possam analisá-lo. Um ambiente colaborativo, onde consumidores podem compartilhar opiniões e esclarecer dúvidas sobre os dados publicados é de grande valia para os publicadores, pois eles terão como definir se seu conjunto de dados está suprindo as necessidades dos consumidores.

A importância do *feedback* é caracterizada pela sua capacidade de transmitir informações ao publicador, sobre como os seus dados ou conjuntos de dados publicados estão sendo utilizados pelos consumidores. Um *feedback* pode relatar erros relacionados aos dados, com isso, os publicadores poderão fazer correções e melhorias em seus conjuntos de dados, criando um ambiente colaborativo, de correções e melhorias contínuas, agregando qualidade aos dados. Com a ausência do *feedback*, publicadores de dados não poderão saber se os seus dados foram consumidos com sucesso, ou se há algum impedimento para o consumo.

O *feedback* é referenciado nas boas práticas de dados na Web[16] como uma ferramenta que ajuda os publicadores a garantir que os dados satisfaçam as necessidades de consumo dos usuários [4]. Um dos fatores de maior importância em relação ao *feedback* é disponibilizar aos consumidores voz ativa, descrevendo

[16] *http://www.w3.org/TR/dwbp/*

experiências de uso. É importante que o *feedback* seja compartilhado com outros consumidores de forma pública, dando suporte a um ambiente colaborativo entre os usuários.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct:  <http://purl.org/dc/terms#> .
@prefix oa:   <http://www.w3.org/ns/oa#>  .
@prefix duv: <http://www.w3.org/ns/duv#> .
@prefix : <http://example.org#> .

:laufer
   a duv:Person  ;
   foaf:givenName "Laufer"  ;
   foaf:mbox <mailto:laufer@example.org> ;
.

:dataset-03312004
   a dcat:Dataset ;
   dct:title "Mars Quarterly Temperature Plot" ;
.

:comment1
   a duv:Feedback ;
   oa:hasBody "Written in MS-DOS text format." ;
   oa:hasTarget :dataset-03312004 ;
   oa:annotatedBy :laufer ;
.

:comment2
   a duv:Feedback;
   duv:hasRating "3 Star";
   oa:hasBody "Linked Data Rating";
   oa:hasTarget :dataset-03312004;
.
```

Figura 9: Feedback DUV (Turtle)

Fonte: [7]

A Figura 9 apresenta um exemplo de um *feedback* descrito no vocabulário de uso dos conjuntos de dados (DUV)[17]. Um *feedback* descreve o uso dos conjuntos de dados, onde esse *feedback* é anotado por um agente ou um indivíduo. Nesse contexto, um *feedback* também possui dois tipos de motivação, as quais definem de fato o motivo do *feedback*, ou seja, ele pode ser motivado por uma correção ou por uma avaliação. Por fim um *duv:feedback* é modelado por uma anotação[18], possuindo um corpo (*body*), que será

[17] *http://www.w3.org/TR/vocab-duv/*
[18] *http://www.w3.org/TR/2015/WD-annotation-model-20151015/*

definido pela motivação do *feedback*, e um alvo (*target*) que referenciará o conjunto de dados a quem está destinado o *feedback*.

A Figura 10 representa um diagrama com uma visão geral do *duv:Feedback* e seus relacionamentos.

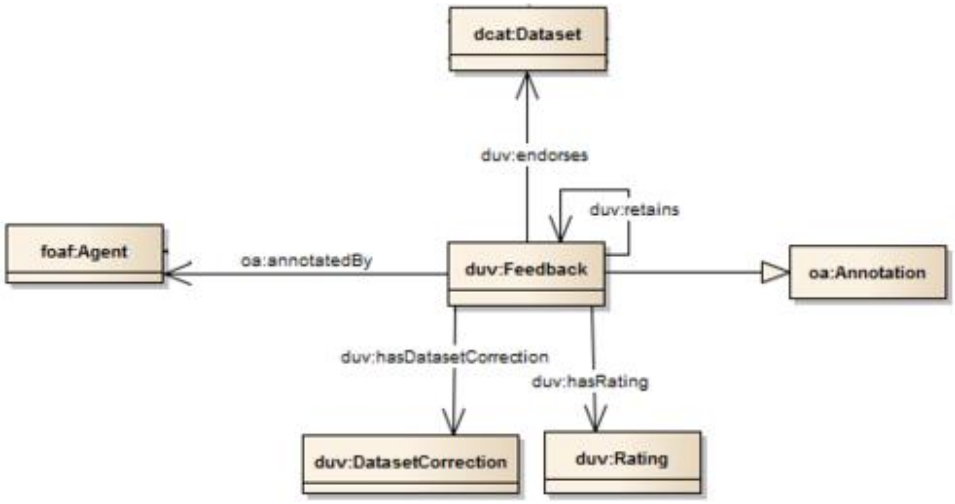

Figura 10: Feedback DUV (Diagrama)

Fonte: [7]

O DUV é um vocabulário que reutiliza e extende classes de outros vocabulários, bem como propriedades existentes que apoiam o *duv:Feedback*. Como mostra a Figura 10, o *feedback* é apoiado por boa parte desse vocabulário, onde, através do *duv:DatasetCorrection*, é possível cobrir problemas relacionados à um conjunto de dados, e pelo *duv:Rating*, classificá-lo de acordo com o possível benefício desse *dataset*. Por fim, o registro da anotação do *feedback* é coberto pela classe *foaf:Agent*, onde é armazenado dados do indivíduo que anotou o *feedback*.

3 A Ferramenta Datafeed

Neste capítulo será apresentada a ferramenta Datafeed que foi desenvolvida neste trabalho, bem como seus casos de uso e o vocabulário descrito para construí-la.

3.1 O Datafeed

A ferramenta desenvolvida nesse projeto, denominada *Datafeed*, tem como funções principais a coleta de *feedback* de conjuntos de dados publicados na Web e a disponibilização dos mesmos de forma aberta, através de um serviço Web. O *Datafeed* foi criado com o objetivo de se adequar há vários portais de dados na Web. Munido de uma interface de comunicação que funciona como um *plugin*, ele pode ser utilizado na maioria dos portais disponíveis no ecossistema de dados na Web.

O consumidor de dados poderá utilizar a ferramenta como um meio de comunicação com o publicador. Com o uso da ferramenta, o consumidor poderá fazer anotações de *feedback* diretamente no portal. As anotações podem ser relacionadas à classificação dos conjuntos de dados utilizando uma classificação em estrelas, bem como anotações relacionadas a correções a respeito de alguma anomalia encontrada nos conjuntos de dados. O *Datafeed* possibilita a coleta do *feedback, bem como o* seu armazenamento em um banco de dados, disponibilizando-os para que sejam visualizados diretamente no portal de catalogação de dados.

O publicador de dados poderá também visualizar as últimas anotações de *feedback* disponibilizadas no portal, como também poderá fazer a coleta diretamente na ferramenta. Com a coleta realizada, o publicador poderá analisar a usabilidade dos seus consumidores, transformando informações em conhecimento, para assim tomar decisões à respeito dos conjuntos de dados, seja para corrigir anomalias ou problemas na publicação.

3.2 Casos de uso

Nesta seção serão apresentados os casos de uso relacionados ao *Datafeed*. A Figura 11 apresenta o diagrama de casos de uso, ilustrando o relacionamento entre os autores e seus casos de uso.

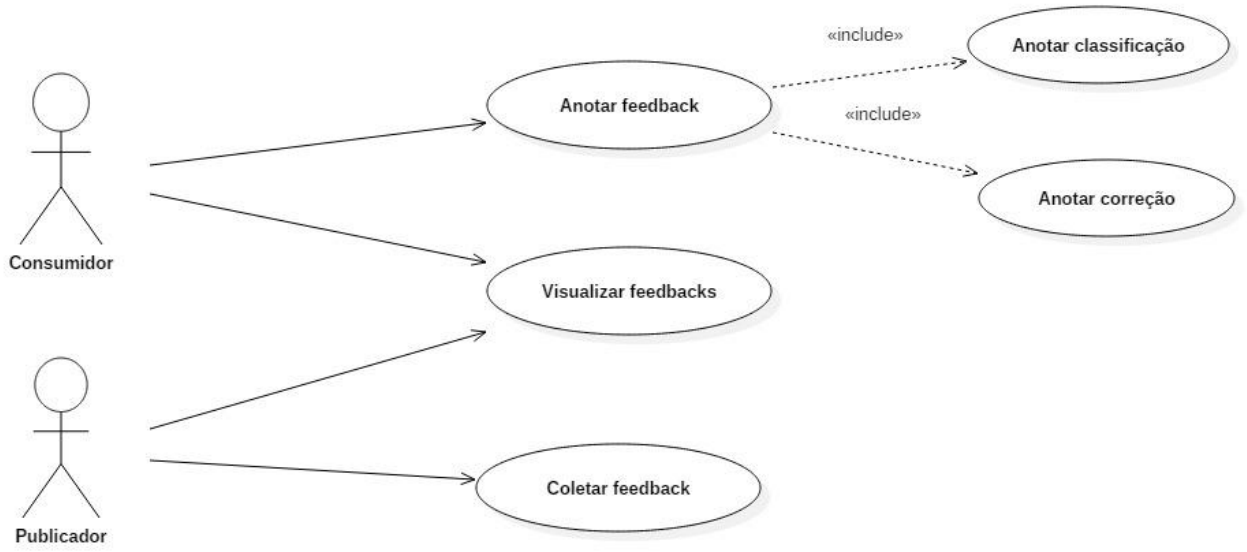

Figura 11: Diagrama de casos de uso

- **Anotar *feedback***

Descrição	O consumidor poderá anotar um *feedback* através da ferramenta.
Ator	Consumidor de dados na Web.
Pré-condições	● Ter acessado a opção para adicionar o *feedback*.
Fluxo de eventos	● Escolher a motivação do *feedback*: classificação ou correção. ● Preencher os dados relacionados ao *feedback*. ● Adicionar o *feedback*.
Fluxo alternativo	● O consumidor poderá se identificar ao anotar seu *feedback*, não sendo obrigatório essa identificação.

- **Anotar classificação**

Descrição	O consumidor poderá classificar um conjunto de dados em estrelas através do *feedback*.
Ator	Consumidor de dados na Web.
Pré-condições	● Ter acessado a opção para adicionar o *feedback*. ● Ter escolhido classificação como motivação.
Fluxo de eventos	● Escolher a quantidade de estrelas relacionadas à classificação do conjunto de dados. ● Adicionar o *feedback*.
Fluxo alternativo	● O consumidor poderá se identificar ao classificar um conjunto de dados, não sendo obrigatória essa identificação.

- **Anotar correção**

Descrição	O consumidor poderá anotar uma correção relacionada à alguma anomalia à respeito de um conjunto de dados pelo *feedback*.
Ator	Consumidor de dados na Web.
Pré-condições	● Ter acessado a opção para adicionar o *feedback*. ● Ter escolhido correção como motivação.
Fluxo de eventos	● Comentar a respeito da anomalia encontrada no conjunto de dados. ● Adicionar o *feedback*.
Fluxo alternativo	● O consumidor poderá se identificar ao anotar uma correção, não sendo obrigatória essa identificação.

- **Visualizar *feedbacks***

Descrição	O consumidor e o publicador de dados poderão visualizar os últimos *feedbacks* no portal.

Atores	● Consumidor de dados na Web. ● Publicador de dados na Web.
Pré-condições	● Está na seção de *feedbacks* do portal.
Fluxo de eventos	● Abrir a seção de *feedbacks* do portal.
Fluxo alternativo	

- **Coletar *feedbacks***

Descrição	O publicador de dados poderá fazer a coleta de todos os *feedbacks* relacionados a um determinado conjunto de dados, fazendo uma requisição ao serviço Web da ferramenta.
Ator	Publicador de dados na Web.
Pré-condições	● Criar um serviço Web que se comunicará com a ferramenta através de requisições *http*.
Fluxo de eventos	● Fazer a requisição *http* enviando como parâmetro o identificador do conjunto de dados.
Fluxo alternativo	● Se o conjunto de dados relacionado ao identificador não estiver na base de dados da ferramenta, ela retornará vazio.

3.3 Vocabulário DUV aplicado à Ferramenta Datafeed

A Ferramenta Datafeed utiliza o vocabulário DUV[19] para organizar e formatar as anotações de *feedback* dentro do sistema. Abaixo, está descrito todo o vocabulário DUV, acrescido de três propriedades RDF: *dct:hasRating*, a qual foi inserida para armazenar a média das classificações associadas a um *dataset*; *dct:dateSubmitted*, que foi inserida para armazenar a data de submissão de uma anotação de *feedback*; *oa:motivatedBy*, que por sua vez, foi inserida para indicar as razões pela qual a anotação de *feedback* foi criada. A Tabela 2 mostra os prefixos utilizados no vocabulário DUV.

[19] *http://www.w3.org/TR/vocab-duv/*

Tabela 2. Prefixos utilizados no vocabulário DUV

Prefixo	***Namespace***
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
foaf	http://xmlns.com/foaf/0.1/
dcat	http://www.w3.org/ns/dcat#
dct	http://purl.org/dc/terms#
oa	http://www.w3.org/ns/oa#
duv	http://www.w3.org/ns/duv#
:	http://example.org#

- **Classe: *Agent***

RDF Class:	**foaf:Agent**
Definição	Um agente (e.g. pessoa).
Subclasses	Person
rdfs:isDefinedBy	http://xmlns.com/foaf/spec/#term_Agent
Label	Agent

- **Classe: *Person (Subclass of Agent)***

RDF Property:	**foaf:Person**
Definição	Pessoa que anota um *feedback*.
rdfs:isDefinedBy	http://xmlns.com/foaf/spec/#term_Person
Label	Person

Propriedade: *giveName*

RDF Property:	**foaf:giveName**
Definição	O nome dado de alguma pessoa.
Domínio	foaf:Agent
Range	rdfs:Literal

Propriedade: *mbox*

RDF Property:	**foaf:mbox**
Definição	Uma caixa email pessoal associada à um proprietário.
Domínio	foaf:Agent
Range	rdfs:Literal

- **Classe: *Rating***

RDF Class:	**duv:Rating**
Definição	Métrica utilizada para avaliar um conjunto de dados.
rdfs:isDefinedBy	http://www.w3.org/ns/duv
Label	Rating

- **Classe: *Dataset***

RDF Class:	**dcat:Dataset**
Definição	Uma coleção de dados publicados por uma única fonte, e disponibilizados para serem acessados e baixados em vários formatos.
rdfs:isDefinedBy	http://www.w3.org/ns/dcat
Label	Dataset
rdfs:subClassOf	dctype:Dataset

Propriedade: *identifier*

RDF Property:	**dct:identifier**
Definição	Um identificador exclusivo do conjunto de dados.
Domínio	dcat:Dataset
Range	rdfs:Literal
Nota de uso	O identificador pode ser utilizado como parte da URI do conjunto de dados.

Propriedade: *hasRating*

RDF Property:	**dct:hasRating**
Definição	Uma classificação fornecida como parte do *feedback*.
Range	duv:Rating
rdfs:isDefinedBy	http://www.w3c.org/ns/duv
Nota de uso	A classificação pode ser utilizada associada a um conjunto de dados, agregando um valor de usabilidade.

- **Classe: *Annotation***

RDF Class:	**oa:Annotation**
Definição	Informação sobre um recurso na Web ou associação entre recursos.
rdfs:isDefinedBy	http://xmlns.com/foaf/spec/#term_Agent
Label	Annotation

- **Classe: *Feedback***

RDF Class:	**dcat:Dataset**
Definição	Feedback de um conjunto de dados. Expressa se o conjunto de dados foi proveitoso ou não, por exemplo.
rdfs:isDefinedBy	http://www.w3.org/ns/duv
Label	Feedback
rdfs:subClassOf	oa:Annotation

Propriedade: *dateSubmitted*

RDF Property:	**dct:dateSubmitted**
Definição	Data de submissão do recurso.
Range	rdfs:Literal
Nota de uso	A data em que o *feedback* foi submetido.

Propriedade: *annotatedBy*

RDF Property:	**oa:annotatedBy**
Definição	Recurso de *feedback* que identifica o agente responsável pela criação da anotação.
Range	foaf:Agent
Label	annotatedBy
Nota de uso	Pessoa que anota um *feedback* poderá se identificar através dessa instância.

Propriedade: *motivatedBy*

RDF Property:	**oa:motivatedBy**
Definição	A relação entre uma anotação e uma motivação, indicando as razões pelas quais a anotação foi criada.
Range	oa:Selector

Label	motivatedBy
Nota de uso	Uma correção pode ser definida como uma motivação para a anotação do *feedback*.

Propriedade: *hasTarget*

RDF Property:	**oa:hasTarget**
Definição	O relacionamento entre oa:Annotation e o corpo. O corpo é algo sobre o oa:hasTarget da anotação.
Range	dcat:Dataset
Label	hasTarget
Nota de uso	Conjunto de dados alvo do *feedback*.

Propriedade: *hasBody*

RDF Property:	**oa:hasBody**
Definição	O relacionamento entre oa:Annotation e o corpo. O corpo é algo sobre o oa:hasTarget da anotação.
Range	rdfs:Literal
Label	hasBody
Nota de uso	Dependendo da motivação, pode ser um comentário de correção para um conjunto de dados.

A Figura 12 mostra uma aplicação do vocabulário DUV na Ferramenta Datafeed, em um cenário real. Neste exemplo, um consumidor de dados anota dois *feedbacks* distintos, ambos relacionados a um único conjunto de dados. O primeiro *feedback* tem como motivação uma correção, em que o consumidor se queixa de um erro ortográfico no título do *dataset*. O segundo *feedback* é motivado por uma avaliação, onde seu corpo possuí a contagem da avaliação em estrelas dada pelo consumidor.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct: <http://purl.org/dc/terms#> .
@prefix oa: <http://www.w3.org/ns/oa#>  .
@prefix duv: <http://www.w3.org/ns/duv#> .
@prefix : <http://example.org#> .

:consumer-001
   a foaf:Person  ;
   foaf:givenName "Helton Santos"  ;
   foaf:mbox "hdas@cin.ufpe.br" ;
.

:dataset-001
   a dcat:Dataset ;
   dct:identifier "http://www.dadosabertosbrasil.com.br/index.php?p=dataset&id=1575&dtId=27" ;
   duv:hasRating "3" ;
.

:comment-001
   a duv:Feedback ;
   dct:dateSubmitted "09-11-2015 23:27:00" ;
   oa:hasBody "Título com erro ortográfico." ;
   oa:hasTarget :dataset-001 ;
   oa:annotatedBy :user-001 ;
   oa:motivatedBy "CORRECTION" ;
.

:comment-002
   a duv:Feedback ;
   dct:dateSubmitted "10-11-2015 23:27:00" ;
   oa:hasBody "3" ;
   oa:hasTarget :dataset-001 ;
   oa:annotatedBy :user-001 ;
   oa:motivatedBy "RATING" ;
.
```

Figura 12: Aplicação do vocabulário DUV na Ferramenta Datafeed

4 Implementação e Avaliação

Nesta seção, serão apresentados alguns detalhes sobre a implementação da ferramenta Datafeed, linguagens de programação utilizadas, arquitetura do sistema, gerencia de configuração, banco de dados, *frameworks* e bibliotecas auxiliares. Apresentaremos também como é feita a integração com os portais de dados na Web, e também a implantação no portal Dados Abertos Brasil[20], que foi utilizado na avaliação.

4.1 Arquitetura do sistema

A ferramenta foi desenvolvida para ser utilizada em qualquer portal que compartilhe e publique dados na Web. Para que esse requisito fosse cumprido, foi necessário que a ferramenta se subdividisse em duas partes: o servidor e a API (*Application Programming Interface*), que é a *interface* de comunicação entre sistemas. Toda codificação da ferramenta está armazenada no GitHub[21], pois utilizamos o *Git*[22] como *framework* de gerência de configuração.

A Figura 13 apresenta uma visão geral da ferramenta, desde a comunicação com o portal, até a persistência no banco de dados.

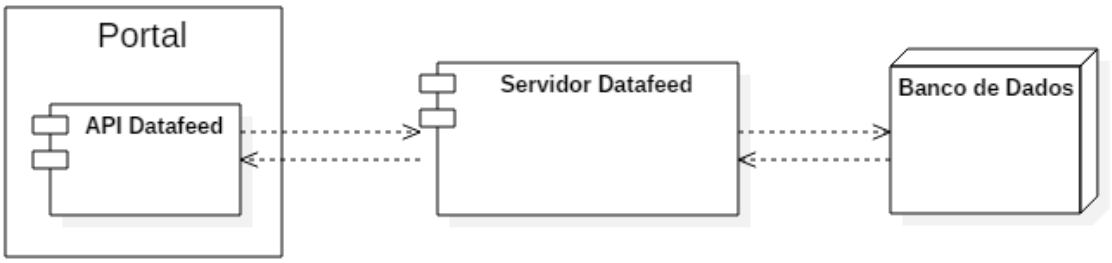

Figura 13: Visão geral da Ferramenta

[20] *http://www.dadosabertosbrasil.com.br/index.php*
[21] *https://github.com/heltonsantos/feedback-dpw*
[22] *https://git-scm.com/*

4.1.1 Servidor Datafeed

O servidor Datafeed é a parte da ferramenta que é executada em um servidor de aplicação independente, sendo considerado o coração da ferramenta, onde fica toda a comunicação com o banco de dados, regras de negócio e serviços remotos. Essa parte da ferramenta foi desenvolvida no padrão *Java EE7*[23], em que foi utilizado a linguagem *java* como padrão para o desenvolvimento. No desenvolvimento e codificação fizemos o uso de vários *frameworks*, dentre os principais estão: *Hibernate*[24], *Jboss Resteasy*[25], *Google Inject*[26], *Jackson Annotations*[27].

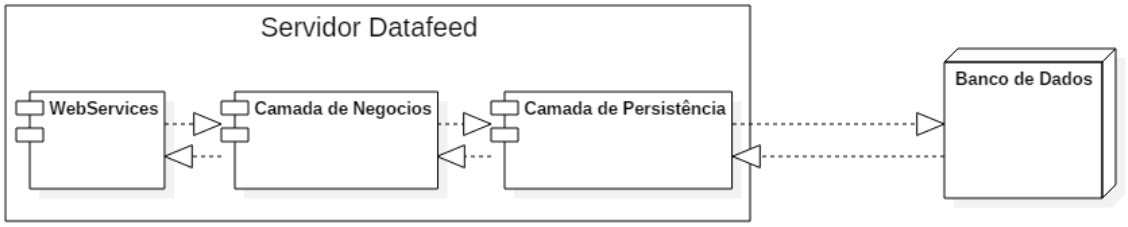

Figura 14: Servidor Datafeed

O diagrama da Figura 14 descreve como está organizado o servidor Datafeed. A primeira camada está composta de *webservices* que fazem a comunicação entre o servidor e a *API*. A camada de negócios é responsável pela implementação das regras de negócio relacionadas à ferramenta. A camada de persistência faz a conexão e a persistência dos dados no servidor de banco de dados. Toda essa estrutura é executada, atualmente, sobre o servidor de aplicação WildFly[28], na sua versão 8.2.1.Final.

[23] *http://www.oracle.com/technetwork/java/javaee/overview/index.html*
[24] *http://hibernate.org/*
[25] *http://resteasy.jboss.org/*
[26] *https://github.com/google/guice*
[27] *https://github.com/FasterXML/jackson-annotations*
[28] *http://wildfly.org/*

4.1.2 *API* Datafeed

A *API* é uma interface de comunicação que liga o portal ou aplicação provedora de dados e o servidor Datafeed. Essa comunicação com o servidor é realizada através de serviços Web, que foram construídos sob o padrão RESTful. A *API* também permite a construção dinâmica de toda a estrutura gráfica (*html* e *css*) utilizada para a coleta do *feedback*. Para isso foi utilizado a biblioteca Jquery[29], que ajudou na criação das estruturas *html* dentro dos portais, o que torna essa ferramenta dinâmica.

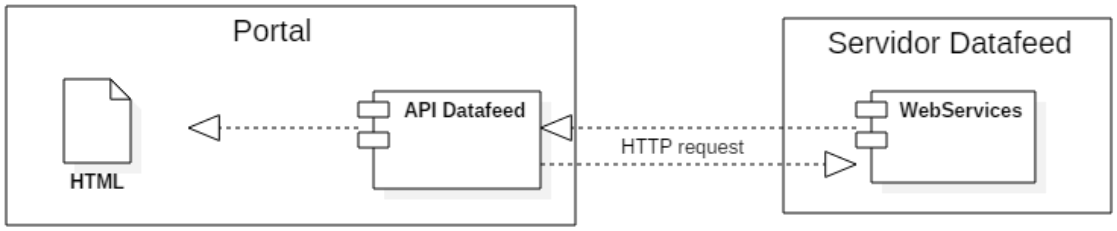

Figura 15: API Datafeed

Na Figura 15, estão representas as duas principais funcionalidades da *API*, a criação da estrutura *html* dentro do portal e a comunicação via requisições *http* com os *webservices* do servidor Datafeed. Para a criação da folha de estilo (*css*) do documento *html*, foi utilizado o *framework* Bootstrap[30], juntamente com o *html*, esse *framework* é utilizado para agilizar a criação da interface gráfica da *API*.

A estrutura do banco de dados foi criada no modelo relacional, espelhando o vocabulário da ferramenta, conforme está na Figura 16. Os dados são armazenados em três tabelas: *dataset*, *feedback* e *person*. Cada tabela possui seus atributos conforme está no vocabulário Datafeed, com as mesmas nomenclaturas. Um *dataset* pode ter várias anotações de *feedback*, uma anotação de *feedback* só poderá ter uma pessoa associada ou nenhuma, pois esse relacionamento não é obrigatório.

[29] *https://jquery.com/*
[30] *http://getbootstrap.com/*

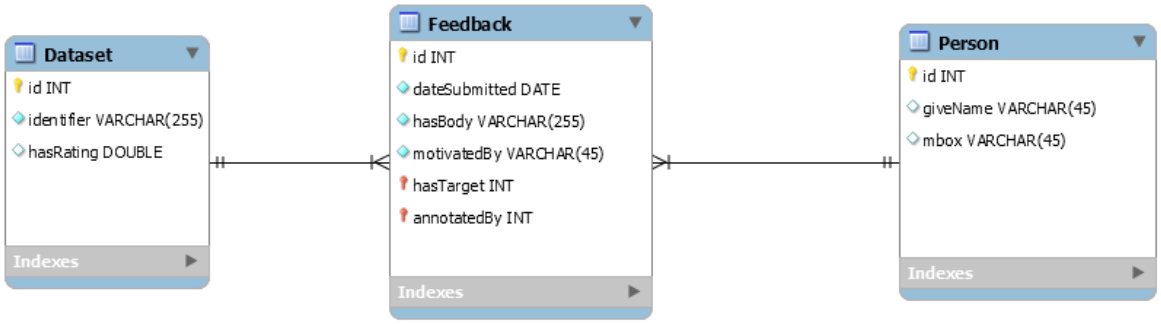

Figura 16: Modelo lógico do banco de dados do Datafeed

O SGBD (Sistema de Gerenciamento de Banco de Dados) escolhido foi o MySQL[31], pela sua praticidade e eficiência.

4.1.3 *Interface* de coleta de *feedback*

Pensando em uma maneira de prover aos consumidores uma forma de coletar o *feedback* relacionado aos conjuntos de dados, foi criado um serviço Web (*webservice*) aberto, sobre o padrão RESTful. Nesta seção mostraremos como acessar o serviço através de uma aplicação em *javascript*, em que foi utilizada a biblioteca Jquery[32] para auxiliar nas requisições ao serviço Web da ferramenta.

O serviço Web foi criado para receber uma requisição *http* juntamente com o método *get* para buscar os dados. Ele recebe como parâmetro o identificador do conjunto de dados, ou seja, sua *uri* de acesso na Web. A *uri* deverá estar codificada sobre o padrão *percent-encoding*, como também chamado *url encoding*. O serviço retorná um objeto *Json* estruturado conforme o vocabulário descrito na seção anterior.

Para acessar o serviço é necessário seguir os seguintes passos:

- Fazer uma requisição *http* através do método *get* para a *url "http://datafeed-hdas.rhcloud.com/datafeed/rest/open/getDatasetFeedback?identifier="*, passando como parâmetro a *string* representando o identificador do *dataset* codificado.

[31] *https://www.mysql.com/*
[32] *https://jquery.com/*

- Após fazer a requisição o servidor Datafeed poderá retorna os seguintes dados:
    - Se o *dataset* possuir *feedbacks* a serem coletados, será retornado um objeto *Json* com os dados para a coleta.
    - Se o *dataset* não possuir *feedbacks*, a ferramenta retornará um objeto vazio.

Na Figura 17 podemos observar um documento *html* que possui toda codificação necessária para acessar o serviço Web aberto do Datafeed. Esse é um exemplo simples de como acessar e fazer a coleta do *feedback*.

```
<!DOCTYPE html>
<html>
    <head>
    <meta charset="UTF-8">
    <title>Coletar Feedback</title>
        <script src="https://ajax.googleapis.com/ajax/libs/jquery/1.11.3/jquery.min.js"></script>
        <script type="text/javascript">
        $(function () {
            $("#button").click(function () {
                var url = "http://datafeed-hdas.rhcloud.com/datafeed/rest/open/getDatasetFeedback?identifier=";

                var identifier = encodeURIComponent($("#identifier").val());

                $.get(url + identifier, function(data,status,xhr){

                        console.log(JSON.stringify(data));
                        $("#datafeed").empty();
                        $("#datafeed").append("<p>"+ JSON.stringify(data) +"</p>");

                }).fail(function(jqXHR,htmlError,error){
                    console.log(jqXHR.responseText);
                });
            });
        });
        </script>
    </head>
    <body>
        <div id="form">
            <h2>Insira o identificador do dataset</h2>
            <label for="identifier">Identifier: </label>
            <input type="text" id="identifier" size="100" maxlength="255"/><br>

            <button id="button">Coletar</button>
        </div>
        <div id="datafeed"></div>
    </body>
</html>
```

Figura 17: Documento html/javascript para a coleta de feedbacks

4.2 Avaliação

Como prova de conceito e avaliação da ferramenta, a implantação do Datafeed foi realizada no portal Dados Abertos Brasil[33], portal esse que faz parte de um projeto de pesquisa desenvolvido por pesquisadores do CIn/Universidade Federal de Pernambuco. O objetivo principal deste portal é prover um catálogo de todos os conjuntos de dados abertos disponíveis em portais de dados abertos governamentais do Brasil.

Nesta integração, implantamos a API Datafeed[34] dentro do portal. Para que a *API* seja executada, é necessário colocar no portal uma *div html*, com o identificador igual à "datafeed", conforme a Figura 18. Dessa forma, a *api* reconhecerá que dentro desta *div* é o local onde ela irá construir toda estrutura *html* para a coleta dos *feedbacks*. Para que a ferramenta seja carregada, é necessário chamar a função de início da *API*, passando como parâmetro a *uri* do conjunto de dados a ser acessado.

```
<div id="feedback">
  <div id="datafeed" class="container reset_datafeed"></div>
</div>
<!-- Carrega a API do DataFeed -->
<?php
  $pageURL = $_SERVER["SERVER_NAME"] . $_SERVER["REQUEST_URI"];
  echo ("<script>loadDatafeed('$pageURL');</script>");
?>
```

Figura 18: Chamada da API Datafeed

A Figura 18 representa o código de chamada da API Datafeed dentro do portal Dados Abertos Brasil. O código que chama o método de execução da *api* está escrito na linguagem *php*, pois é a linguagem nativa desse portal. Cada vez que esse método é chamado, a *api* faz uma requisição *http* ao servidor Datafeed para buscar os dados de *feedback* relacionados ao conjunto de dados instanciado.

O carregamento da ferramenta resultará na aba Feedback dentro do portal, especificamente, na página onde o conjunto de dados é apresentado. É nessa aba que fica toda a parte de coleta, proveniência e visualização dos *feedbacks* anotados pelos consumidores de dados. A Figura 19 mostra a implantação da ferramenta dentro do portal.

[33] *http://www.dadosabertosbrasil.com.br/*
[34] *https://github.com/heltonsantos/feedback-dpw/tree/master/datafeed_api*

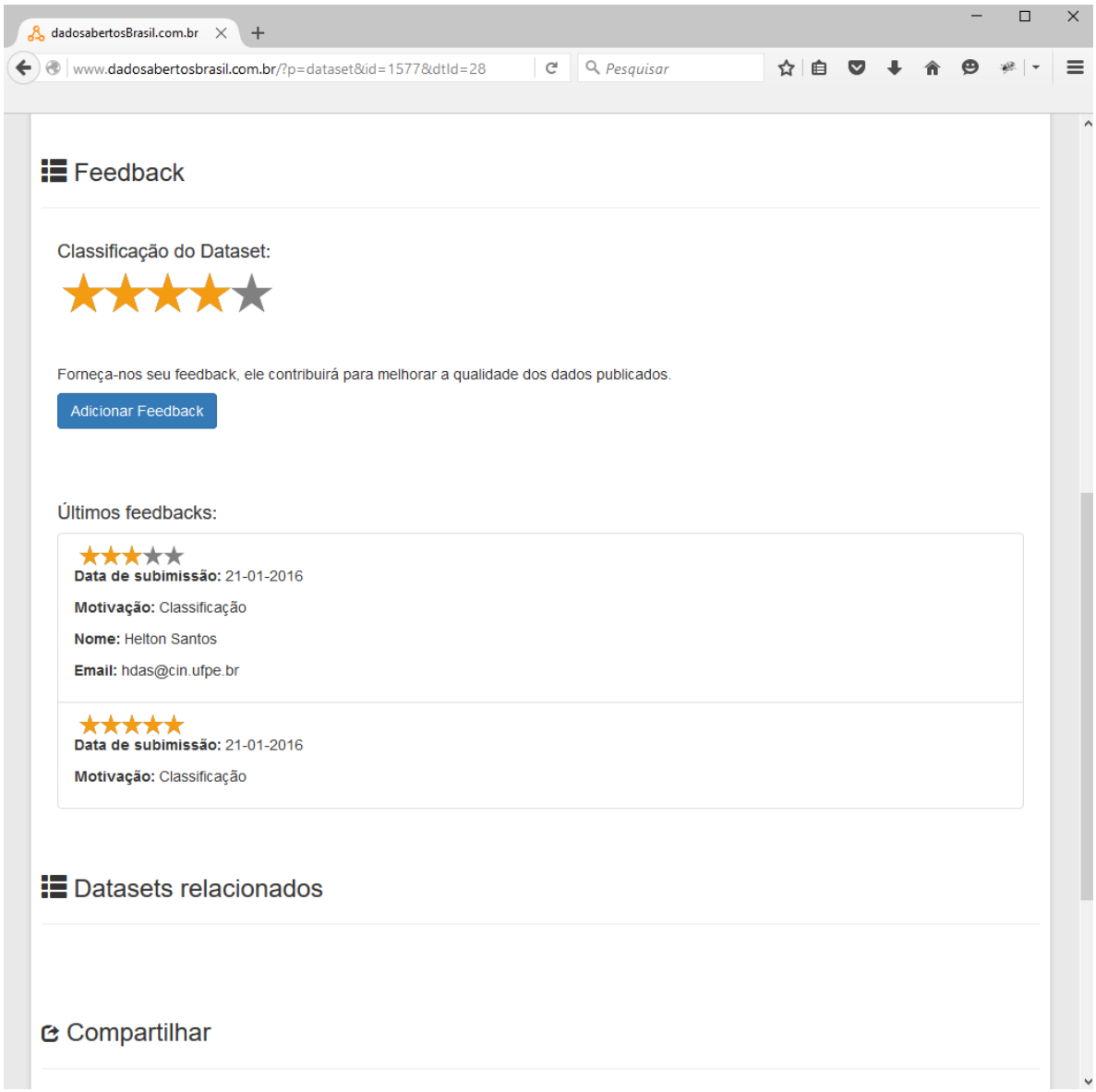

Figura 19: Datafeed executado no portal Dados Abertos Brasil

Para que um consumidor adicione um *feedback*, é necessário que ele clique no botão “Adicionar Feedback” conforme mostra a Figura 19. Após ter clicado, o consumidor escolherá qual o motivo da adição do *feedback*, seja por uma correção, ou por uma classificação relacionada ao conjunto de dados. Se a motivação for uma classificação, a ferramenta abrirá um formulário para que o consumidor possa inserir uma classificação em estrelas sobre o conjunto de dados, numa escala de zero a cinco estrelas, conforme mostra a Figura 20. Caso a motivação seja uma correção, o formulário mostrará um

combobox para que o consumidor escolha qual a anomalia ou dificuldade encontrada, e também terá um campo para que ele possa prover seu comentário sobre a correção necessária para o *dataset*, como mostra a Figura 21. Nos dois casos, o consumidor poderá ou não se identificar, informando seu nome e *email* para que seus dados possam ser anexados ao seu *feedback*. Para a visualização das anotações de *feedback* cadastradas, a ferramenta deixa disponível as últimas anotações para o conjunto de dados instanciado, mostrando também a data de cadastro do *feedback* e quem o anotou.

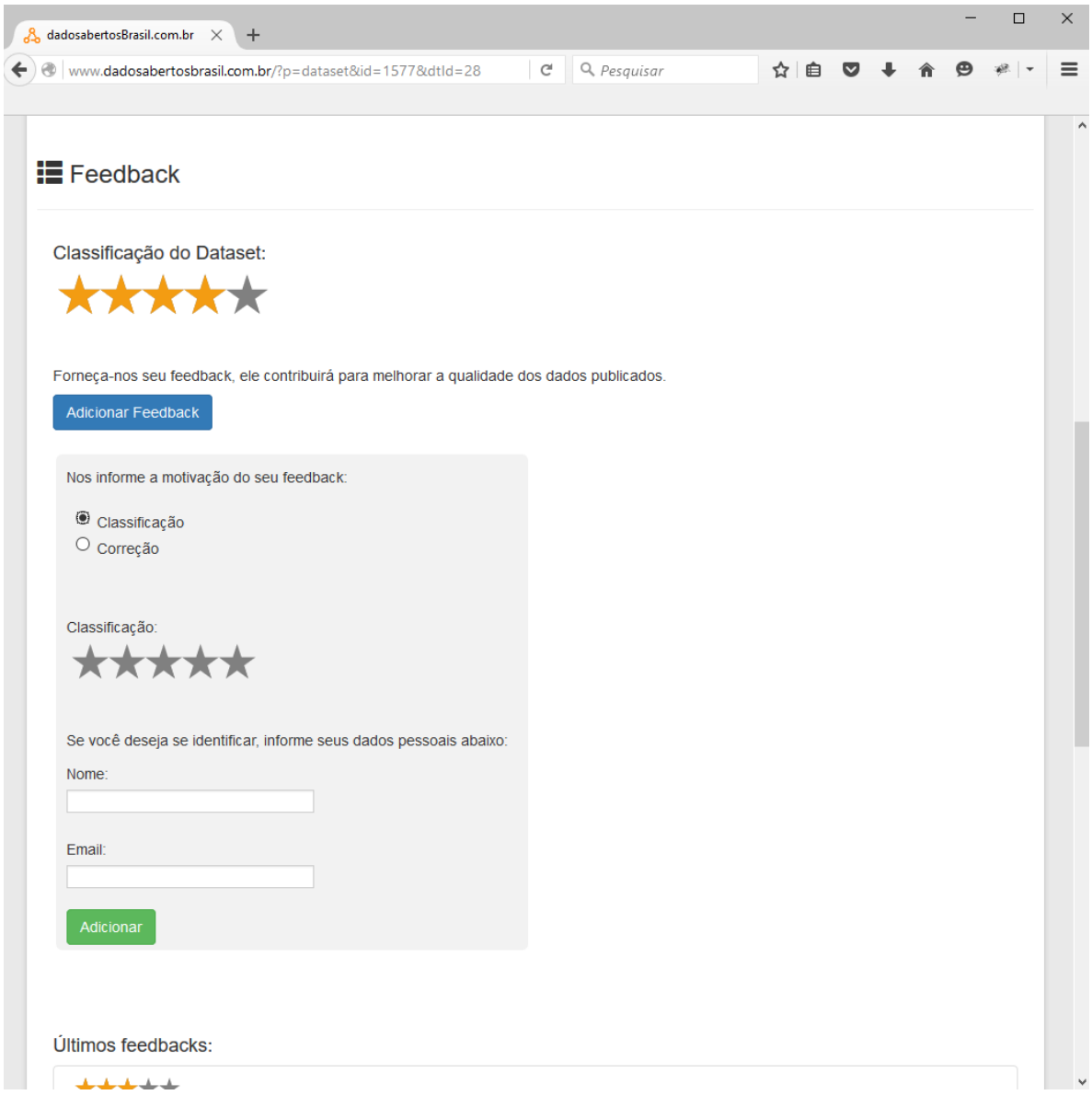

Figura 20: Adicionando um feedback de classificação no portal

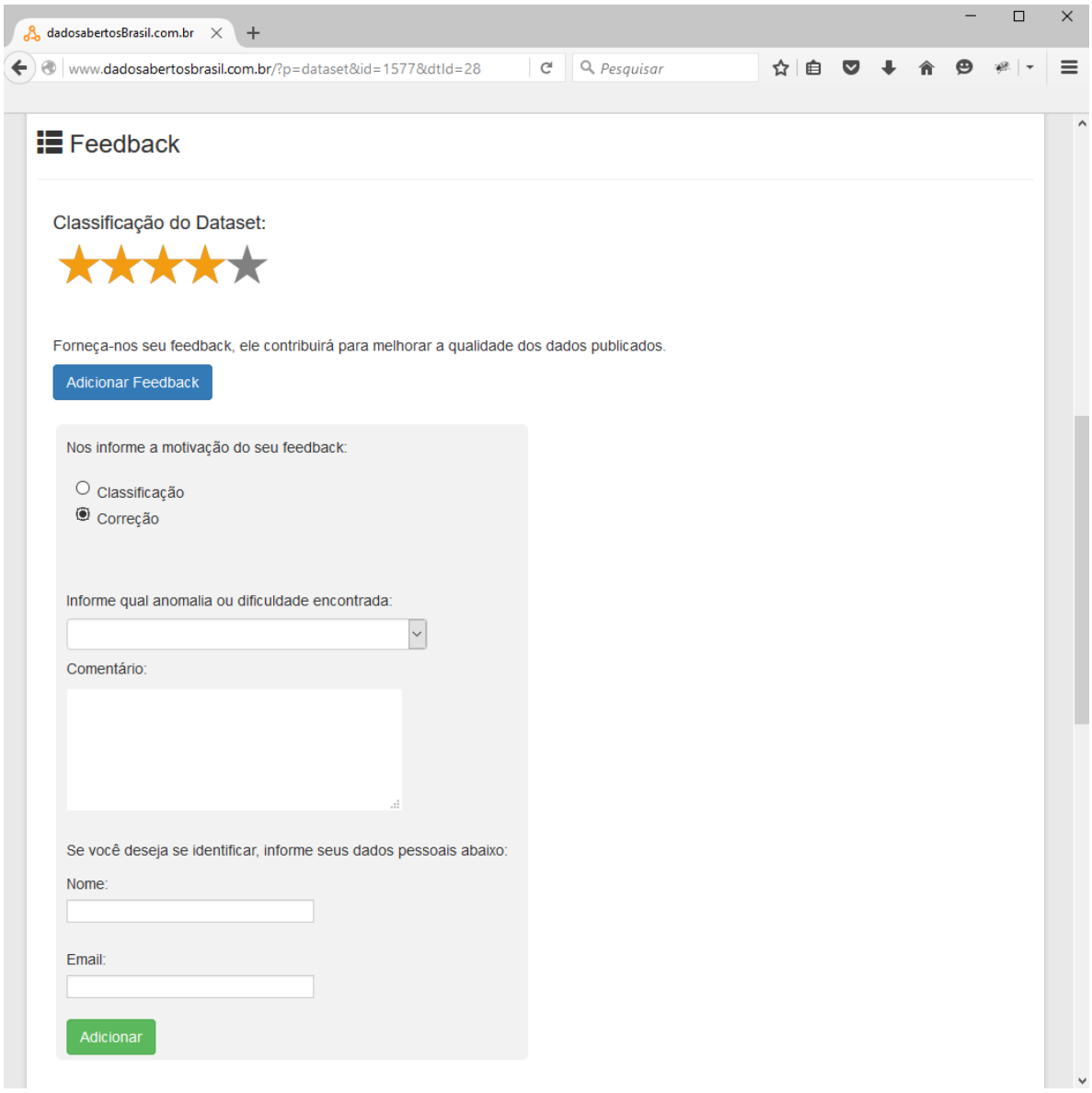

Figura 21: Adicionando um feedback de correção no portal

As principais anomalias descritas no formulário de correção foram identificadas após termos realizado uma pesquisa com possíveis consumidores de dados. Esta pesquisa foi composta de três perguntas, onde o propósito era identificar aspectos relevantes que deveriam ser considerados para a aquisição de *feedbacks* dos consumidores de dados publicados na Web. Dentre as perguntas, a primeira teve o objetivo de listar as principais anomalias ou dificuldades encontradas pelos consumidores de dados. A segunda pergunta, buscou identificar informações que deveriam estar disponíveis como parte do *feedback*. Por fim, a última pergunta almejou conhecer quais

os mecanismos de avaliação seriam mais interessantes para avaliar um conjunto de dados de maneira coerente e correta.

O formulário da pesquisa foi construído com o auxílio da ferramenta *Google Forms*[35] e foi divulgado através de alguns mecanismos de comunicação específicos, dentre eles, o *email*, o *Facebook,* o *Hangouts* e até mesmo através do *Whatsapp*. O público alvo da pesquisa foram os profissionais e estudantes da área de tecnologia da informação, por estarem mais próximos da realidade encontrada nos conjuntos de dados na Web. As respostas da pesquisa foram registradas anônimamente e armazenadas em uma planilha juntamente com a sua data de submissão.

[35] https://www.google.com/forms/about/

5 Conclusão e Trabalhos Futuros

Este trabalho foi dividido essencialmente em duas partes: a criação da ferramenta de coleta e compartilhamento de *feedback* para conjuntos de dados publicados na Web; e a implantação desta ferramenta no portal Dados Aberto Brasil.

Antes da construção da ferramenta, foi necessário realizar um estudo sobre *feedback*, como ele está descrito no vocabulário DUV, bem como sua importância na publicação de dados dentro do ecossistema de Dados na Web. Não podemos deixar de mencionar que o *feedback* também está especificado nas boas práticas de publicação de dados na Web, isso mostra o crescimento e o valor que o *feedback* do consumidor possui para com o conjunto de dados.

A ferramenta construída, denominada Datafeed, provê aos consumidores de dados uma forma de compartilhar experiências de uso nos conjuntos de dados publicados na Web. Para a análise do *feedback*, a ferramenta também disponibilizou uma interface de comunicação para a coleta de *feedback*, a qual retorna os dados no formato Json. Um dos requisitos não funcionais de grande valia de ferramenta é sua capacidade de ser implantada na maioria dos portais que publicam dados na Web. A API Datafeed foi toda construída em *javascript*, linguagem utilizada mundialmente em sites e aplicações publicadas na Web.

Finalmente, a implantação no portal Dados Abertos Brasil ocorreu com sucesso. Atualmente a ferramenta já está em produção e pode ser acessada pelo portal. Com essa implantação, tomamos conhecimento da real capacidade da ferramenta em se adaptar em grande parte dos portais e aplicações que disponibilizam dados na Web. A ferramenta foi de grande valia para o portal, pois ela auxiliou na descrição do benefício dos conjuntos de dados publicados no portal.

Dentre as principais contribuições realizadas neste trabalho, podemos destacar:

- Construção e implementação da Ferramenta Datafeed
- Implantação da ferramenta no portal Dados Abertos Brasil
- Disponibilização da ferramenta para ser implantada em portais de publicação de dados na Web

Este trabalho foi concebido como parte inicial de um estudo ainda maior na área de análise do benefício de conjunto de dados na Web. Dessa forma, algumas limitações deste trabalho são conhecidas e futuras ações são esperadas. Dentre elas, podemos destacar as seguintes:

- Propor uma estratégia para o cálculo do benefício dos conjuntos de dados publicados na Web, bem como definir uma abordagem para determinar o benefício ou relevância desses conjuntos de dados.
- Implementar melhorias na ferramenta, aumentando sua compatibilidade com os portais e aplicações provedoras de dados na Web.
- Modificar o armazenamento dos dados da ferramenta, de uma base de dados relacional, para uma base de dados RDF[36], modelo padrão para o intercâmbio de dados na Web.

Por fim, este trabalho contribuiu para o desenvolvimento de uma abordagem de coleta de *feedback* para conjuntos de dados publicados na Web. Com o forte crescimento da publicação de dados, a importância de se ter informações sobre a experiência de uso dos consumidores se torna indispensável. Dessa forma, esperamos que o uso desta ferramenta possibilite a criação de um importantíssimo canal de comunicação entre consumidor e publicador de dados, melhorando assim a qualidade e os benefícios dos dados que são publicados.

[36] *http://www.w3.org/RDF/*

6 Referências

[1] Berners-Lee, T., Connolly, D., e Swick, R. R. (1999). **Web architecture: Describing and exchanging data.** Recuperado em 25 de outubro de 2015, do <http://www.w3.org/1999/04/WebData>.

[2] Abiteboul, S., Buneman, P., e Suciu, D. (2000). **Data on the Web: from relations to semistructured data and XML.** Morgan Kaufmann.

[3] Lóscio, B. F., Oliveira, M. I. S., Bittencourt, I. I., 2015. **Publicação e Consumo de Dados na Web: Conceitos e Desafios.** Dados na Web, Minicurso SBBD, 2015.

[4] Lóscio, B. F., Burle, C., Calegari, N. **Data on the Web Best Practices. W3C Second Public Working Draft (2015).** Recuperado em 22 de outubro de 2015, do <http://www.w3.org/TR/dwbp/>.

[5] Möller, K., 2013. **Lifecycle models of data-centric systems and domains: The abstract data lifecycle model.** Semantic Web 4, 1, 67-88.

[6] Sanderson, R., Ciccarese, P., Young, B. **Web Annotation Data Model. W3C Working Draft(2015).** Recuperado em 22 de outubro de 2015, do <http://www.w3.org/TR/annotation-model/>.

[7] Lóscio, B. F., Stephan, E. G., Purohit, S. **Dataset Usage Vocabulary (DUV). W3C Second Public Working Draft (2015).** Recuperado em 1 de novembro de 2015, do <http://www.w3.org/TR/vocab-duv/>.

[8] Maali, F., Erickson, J. **Data Catalog Vocabulary. W3C Recommendation (2014).** Recuperado em 1 de novembro de 2015, do <http://www.w3.org/TR/vocab-dcat/>.

[9] Gama, K. S., Lóscio, B. F. **Towards Ecosystems based on Open Data as a Service.** In 17th International Conference on Enterprise Information Systems (ICEIS), Barcelona, Spain, 2014.

[10] Lóscio B. F., Batista M., Souza D., Salgado A. C. (2012). **Using information quality for the identification of relevant Web Data Sources: a proposal.** Proceedings of the

14th International Conference on Information Integration and Web-based Applications & Services, pages 36-44.

[11] Open Knowledge Foundation. **What is Open?**. Recuperado em 8 de dezembro de 2015, do <https://okfn.org/opendata>.

[12] Oliveira, L., Lóscio, B. F. **Uma Abordagem para Captura de Informações sobre Aplicações que fazem uso de Dados Abertos.** III Simpósio Brasileiro de Tecnologia da Informação (SBTI 2014), Maceió, Alagoas, 2014.

[13] Wang, R. Y., Strong, D. M., 1996. **What Data Quality Means to Data Consumers.** Recuperado em 10 de dezembro de 2015, do <http://www.jstor.org/stable/40398176?origin=JSTOR-pdf&seq=1#page_scan_tab_contents>.

[14] Oliveira, H. R., Tavares, A. T., e Lóscio, B. F. (2012). **Feedback-based data set recommendation for building linked data applications.** Proceedings of the 8th International Conference on Semantic Systems, pages 49-55.

[15] Albertoni R., Guéret, C.. F., Isaac, Antoine. **Data Quality Vocabulary, 2015.** Recuperado em 15 de dezembro de 2015, do <http://www.w3.org/TR/vocab-dqv/>.

[16] Jacobs, I., e Walsh, N. (2004). **Architecture of the World Wide Web, Volume One.** 15 December 2004. W3C Recommendation, Recuperado em 17 de dezembro de 2015, do <http://www.w3.org/TR/webarch/>.

[17] Berners-Lee, T., Connolly, D., Swick, R. R. (1999). **Web architecture: Describing and exchanging data.** Recuperado em 17 de dezembro de 2015, do <http://www.w3.org/1999/04/WebData>.

[18] Sauermann, L., Cyganiak, R., Völkel, M. (2011). **Cool URIs for the semantic web.** Recuperado em 19 de dezembro de 2015, do <http://www.w3.org/TR/cooluris/>.,

[19] Wang, R. Y., e Strong, D. M. (1996). **Beyond accuracy: What data quality means to data consumers.** Journal of management information systems, volume 12, issue 4, march 1996, pages 5-33.

[20] Klyne, G., e Carroll, J. J. (2006). **Resource description framework (RDF): Concepts and abstract syntax.** Recuperado em 19 de dezembro de 2015, do <http://www.w3.org/TR/rdfconcepts>.

[21] Open Data Handbook. **O que são dados abertos?**. Recuperado em 14 de janeiro de 2016, do <http://opendatahandbook.org/pt_BR/what-is-open-data/index.html>.

[22] Tauberer, J. (2007). **8 Principles of Open Government Data**. Recuperado em 14 de janeiro de 2016, do <http://opengovdata.org/>.

[23] W3C. **Catalog**. Recuperado em 14 de janeiro de 2016, do <http://www.w3.org/TR/vocab-dcat/#Class:_Catalog>.

[24] CKAN. **About**. Recuperado em 14 de janeiro de 2016, do <http://ckan.org/about/>.