# Refinamento Colaborativo de Dados na Web baseado em *Social Coding*

**Helton Douglas A. dos Santos[1]*, Marcelo Iury S. Oliveira[2], Bernadette Farias Lóscio[1]**

[1]Centro de Informática – Universidade Federal de Pernambuco
Recife, PE – Brasil

[2]Unidade Acadêmica de Serra Talhada – Universidade Federal Rural de Pernambuco
Serra Talhada, PE – Brasil

{hdas,bfl}@cin.ufpe.br, marcelo.iury@ufrpe.br

***Abstract.*** *The Web has emerged as an important platform for sharing information, enabling the publishing and consumption of datasets from different domains. In this contexto, dataset refinement is a primary activity mainly related to data cleansing and enrichment. Usually, refinement is performed by publishers, although consumers often clean and enrich datasets in their consumption activities. However, in general, consumer's effort is lost, since most times the result of the refinement is not shared back with the publisher or other consumers. In this context, this work proposes a refinement strategy based on the principles of social coding to allow the refinement of datasets published on the Web in a collaborative way.*

***Resumo.*** *A Web tem emergido como um importante canal de compartilhamento de informações, habilitando a publicação e o consumo de conjuntos de dados. Neste contexto, o refinamento de conjunto de dados é uma atividade relacionada à limpeza e enriquecimento de dados. Normalmente, o refinamento é realizado pelos publicadores de dados, embora os consumidores também limpem e aprimorem conjuntos de dados em suas atividades de consumo. Porém, o esforço do consumidor é perdido, já que na maioria das vezes o resultado do refinamento não é compartilhado com o publicador ou outros consumidores. Assim, este trabalho propõe uma estratégia baseada nos princípios de social coding para permitir o refinamento de conjuntos de dados publicados na Web de forma colaborativa.*

## 1. Introdução

Atualmente, o aumento pelo interesse na publicação de dados na Web em diferentes formatos, com licença aberta ou privada, juntamente com o enorme volume de dados gerados pelas redes sociais, tem confirmado o potencial da Web como plataforma de compartilhamento de dados. De maneira geral, as atividades de publicação e consumo de dados na Web são desempenhadas por um conjunto de atores. Um ator pode agir como um publicador ou como um consumidor de dados. O primeiro entrega e produz dados de acordo com condições específicas. O segundo acessa e consome os dados, seja para realizar análises, construir visualizações, ou para gerar novos dados.

[0]Helton Santos e Marcelo Iury agradecem ao CNPq e CAPES pelo apoio financeiro.

Segundo o ciclo de vida dos dados na Web [Lóscio et al. 2015], a limpeza e o enriquecimento dos dados são atividades que fazem parte da fase de refinamento dos dados. Esta fase diz respeito à correção de erros nos conjuntos de dados publicados, como também à atualização e adição de novos dados. Dessa forma, o refinamento de dados na Web pode ser definido como um processo no qual são executados todos os procedimentos relacionados à limpeza e ao enriquecimento de dados.

Geralmente, o refinamento de conjuntos de dados na Web é realizado pelos publicadores de dados antes de efetuarem a publicação dos dados. Porém, consumidores de dados também realizam frequentemente procedimentos de limpeza e de enriquecimento nos conjuntos de dados a fim de melhorar sua qualidade antes da execução de suas atividades de consumo. No entanto, é importante notar que o resultado do refinamento realizado pelos consumidores, na maioria das vezes, não é compartilhado com os publicadores do conjunto de dados original e nem com outros consumidores interessados no mesmo conjunto de dados. Dessa forma, é muito comum que exista retrabalho, tanto por parte dos publicadores como por parte dos consumidores, uma vez que o resultado das atividades de refinamento não são compartilhados.

Em um cenário ideal, provedores e consumidores devem compartilhar seus dados limpos e enriquecidos, de tal forma que ambos poderão dar sua contribuição, sinalizando e corrigindo erros, bem como realizando o enriquecimento dos dados. Nesse contexto, o objetivo deste trabalho é propor uma estratégia baseada nos princípios de *open collaboration* e *social coding* para permitir o refinamento, de forma colaborativa, de conjuntos de dados publicados na Web, contribuindo para reduzir o retrabalho nas atividades de refinamento, bem como para melhorar a qualidade dos conjuntos de dados na Web.

*Open collaboration* é um sistema de produção que baseia-se em um conjunto distribuído de participantes, fracamente coordenados e orientados a objetivos, que interagem para criar um produto ou serviço [Levine and Prietula 2013]. Por sua vez, *social coding* é uma abordagem fundamentada nos princípios de *open collaboration*, destinada à implementação de projetos de software de maneira compartilhada, por meio de plataformas específicas como o Github[1] [Gousios et al. 2014].

O restante deste artigo está estruturado como se segue. Na Seção 2, são apresentados os trabalhos relacionados. Na Seção 3, é apresentada a solução proposta. Na Seção 4, é discutida a avaliação da solução proposta. Na Seção 5, são apresentadas as conclusões e sugestões para os trabalhos futuros.

## 2. Trabalhos Relacionados

Neste trabalho, o refinamento de dados pode ser definido como sendo a atividade direcionada a execução de procedimentos de limpeza de dados, como também de enriquecimento, realizando alterações e adições de dados (*e.g.*, dados e metadados), com o objetivo de melhorar a sua qualidade do conjunto de dados como um todo [Lóscio et al. 2015].

Na literatura, é possível encontrar diferentes definições para refinamento, enriquecimento e limpeza de dados. Apesar dessas atividades serem relacionadas, elas não devem ser tratadas da mesma forma. Além disso, entre os trabalhos que apresentam uma definição para esses termos, uma parcela significativa faz uso de definições superficiais,

[1] https://github.com/

não trazendo clareza quanto ao real significado e aplicação de refinamento, limpeza ou enriquecimento.

O trabalho [Rahm and Do 2000] define o conceito de limpeza de dados, como também aborda alguns problemas de qualidade que estão presentes em bases de dados. [Rahm and Do 2000] também descreve as principais fases do processo de limpeza de dados, que vai desde a análise dos dados até a realização da limpeza, fazendo o uso de técnicas e tecnologias usadas em *Data Warehouse*. [Maletic and Marcus 2000] também fornece uma visão geral da literatura sobre limpeza de dados. O principal objetivo desse trabalho é descrever e analisar métodos de detecção automática de erros em conjuntos de dados, como *data mining* e *clustering*. Além disso, o autor apresenta um experimento para investigar diferentes métodos de detecção de erros utilizando um conjunto de dados do mundo real. Por outro lado, [Chapman 2005] escreveu um livro abordando princípios e métodos da limpeza de dados. Ele afirma que limpeza de dados é o processo usado para determinar dados imprecisos ou incompletos, a fim melhorar a qualidade por meio da correção de erros encontrados.

Existem também trabalhos que relacionam o conceito de refinamento de dados com enriquecimento de dados. Contudo, na literatura, a maioria dos trabalhos que traz o enriquecimento de dados diz respeito ao enriquecimento semântico de conjuntos de dados. Por exemplo, [Clarke and Harley 2014] aborda o enriquecimento semântico como um processo direcionado à adição de *tags* semânticas em dados ou metadados a fim de facilitar a compreensão dos dados, como também auxiliar na descoberta. Similarmente, o trabalho proposto por [Fileto et al. 2015] fornece uma visão geral de propostas de enriquecimento semântico para dados em movimento (*i.e.*, dados que descrevem posições espaço temporais de objetos que podem ser capturados por sensores, como GPS), associando aos dados anotações semânticas que descrevem conceitos por meio de ontologias.

Além disso, [dos Santos et al. 2018] realizaram um mapeamento sistemático da literatura na área de publicação e consumo de dados na Web com o intuito de identificar lacunas de pesquisa na nesta área. Como resultado deste mapeamento, foi identificado a ausência de trabalhos que abordem o refinamento de dados na Web. De maneira semelhante, também não foram encontrados estudos relevantes na área de *Social Coding* que tratam ou direcionam aspectos relacionados ao refinamento de dados.

## 3. Refinamento Colaborativo de Conjuntos de Dados na Web

O refinamento colaborativo tem o objetivo de atribuir também ao consumidor de dados o papel de refinador, podendo solicitar a publicação da nova versão do conjunto de dados após a realização de alguma atividade de refinamento nos dados. O uso de uma estratégia colaborativa pode reduzir também o esforço dos publicadores no refinamento dos conjuntos de dados, visto que esta atividade passa agora a ser realizada também pelos consumidores. Isto também pode levar ao aumento da frequência de atualização dos conjuntos de dados publicados, pois os processos de atualização e correção dos conjuntos de dados, quando realizados por um único publicador, podem tornar-se demorados dependendo do volume de dados.

É importante ressaltar que a construção de uma estratégia para o refinamento colaborativo não é uma atividade trivial. Por exemplo, conflitos podem ocorrer quando dois ou mais consumidores refinam uma mesma versão de um conjunto de dados e tentam fa-

zer sua republicação. Dessa maneira, o versionamento e a resolução de conflitos é uma problemática decorrente do refinamento colaborativo e que serão tratadas neste trabalho.

Nas próximas subseções, detalharemos o funcionamento da estratégia proposta, incluindo todo processo de colaboração envolvendo atores e artefatos. Por fim, apresentaremos a nossa proposta para resolução de conflitos.

### 3.1. Estratégia Proposta

Este trabalho propõe uma estratégia para o refinamento colaborativo de conjunto de dados na Web, composta por um conjunto de processos e operações que são desempenhadas durante o ciclo de vida de dados da Web [Lóscio et al. 2015, da Silva 2019], envolvendo mais especificamente as etapas de publicação, refinamento, homologação e re-publicação. A estratégia proposta se baseia em princípios básicos de *Open Collaboration* e *Social Coding*.

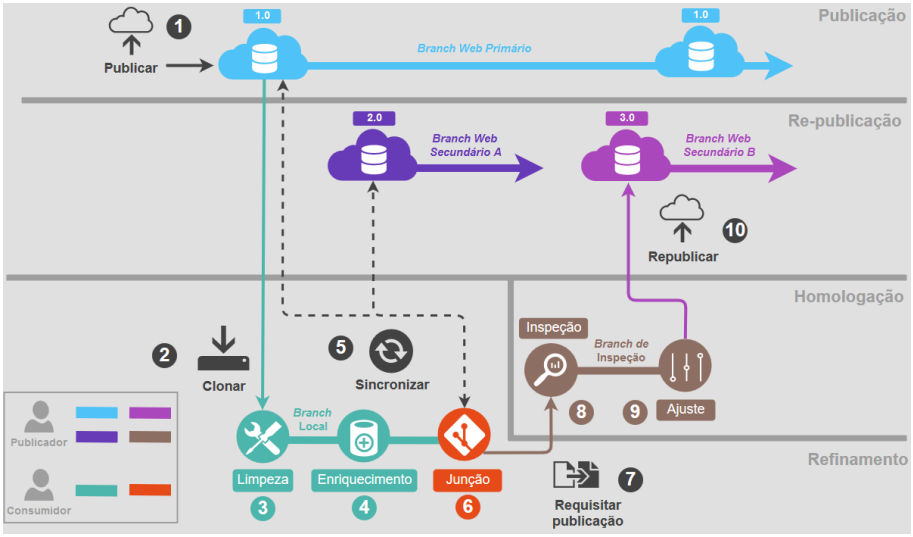


**Figura 1. Estratégia para refinamento colaborativo de conjuntos de dados na Web. Fonte: Autores**

A Figura 1 apresenta um cenário de utilização da estratégia proposta. Como pode ser visualizado, a estratégia de refinamento colaborativo faz uso de múltiplos *branches* e múltiplas operações.

### 3.2. Branches

O conceito de *branch* é bastante utilizado em sistemas de controle de versão (*e.g.*, CSV[2] e GitHub[1]). Eles representam cópias lógicas de um ou mais artefatos criadas para que modificações possam ocorrer de forma paralela, *e.g.*, implementar recursos e corrigir *bugs* [Gousios et al. 2014]. Nas plataformas de *Social Coding*, os *branches* são geralmente utilizados para implementar novos componentes de um software, onde instabilidades podem ocorrer no processo de implementação sem afetar os outros usuários.

Na estratégia proposta, a principal finalidade da utilização dos *branches* é limitar e organizar os diferentes estados que um conjunto de dados pode ter durante as etapas de

[2]https://savannah.nongnu.org/projects/cvs/

refinamento, como também gerenciar as diferentes versões geradas, habilitando assim o controle de versão dos conjuntos refinados.

Mais especificamente, são definidos quatro tipos de *branches* diferentes, sendo eles: (i) *Branch Web Primário*, que armazena a primeira versão de um conjunto de dados publicado na Web; (ii) *Branch Local*, que armazena localmente uma cópia de um conjunto de dados publicado na Web (*i.e.*, em computadores pessoais); (iii) o *Branch de Inspeção*, que armazena um conjunto de dados sujeito à inspeção e homologação; e (iv) *Branch Web Secundário*, armazena na Web a nova versão de um conjunto de dados refinado a partir de um *Branch Web Primário*.

Cada *branch* é criado por meio de operações específicas. Por exemplo, o *Branch Web Primário* é criado por meio da operação Publicar. Dessa forma, é possível isolar todas as alterações realizadas sobre um conjunto de dados armazenado no *Branch Local*, não refletindo na versão publicada no *Branch Web Primário*, por exemplo.

### 3.3. Operações

A estratégia proposta faz uso de um conjunto de atividades e operações que são destinadas a manipulação do conjunto de dados, seguindo um fluxo ordenado como apresentado na Figura 1. Cada atividade e operação possuem papéis e funções específicas que podem ser desempenhados tanto por publicadores como por consumidores.

**Operação Publicar:** A operação Publicar pode ser vista como uma função *publicar* que, a partir de um conjunto de dados *DW1*, cria e retorna um novo *Branch Web Primário* $B_{web}$ armazenando o conjunto de dados *DW1*:

$$B_{web} = publicar(DW1)$$

Ao realizar a operação Publicar, é criado um novo *Branch Web* que armazena o conjunto de dados em sua versão inicial (versão 1.0), tornando-se disponível na Web para o acesso e o consumo. Para que um conjunto de dados possa passar pelos processos de limpeza e enriquecimento é necessário primeiro a execução da operação Clonar. Esta operação tem o objetivo de gerar uma cópia de um conjunto de dados publicado no *Branch Web* e armazená-lo em um *Branch Local* que é criado pela operação, permitindo assim conservação dos dados originais. A operação Clonar pode ser definida como:

**Operação Clonar.** A operação Clonar pode ser vista como uma função *clonar* que, a partir de um *Branch Web* $B_{web}$, cria e retorna um novo *Branch Local* $B_{local}$ armazenando uma cópia *DL1* do conjunto de dados *DW1*:

$$B_{local} = clonar(B_{web})$$

Os processos de limpeza e enriquecimento são compostos por procedimentos que tem o objetivo de corrigir erros dentro do conjunto (*e.g.*, correção de ortografia) ou de enriquecer o conjunto (*e.g.*, adição de dados). As operações e os procedimentos de limpeza e enriquecimento podem ser desempenhados sobre cópias de conjunto de dados criadas e armazenadas no *Branch Local* pela operação Clonar. Cada procedimento executado sobre um conjunto de dados gera uma entrada em um *log* de refinamento de conjunto de dados

Com o uso do *log* de refinamento, é possível localizar uma alteração realizada no conjunto de dados, como também identificar valores inseridos e substituídos. Um *log* de

refinamento de um conjunto de dados *C* pode ser descrito como um conjunto $C.L_{ref} = \{Lproc_1, Lproc_2, ..., Lproc_n\}$, no qual cada $Lproc_i$ é uma entrada no *log* e descreve um procedimento de refinamento que foi executado sobre o conjunto de dados *C*.

De modo geral, um $Lproc_i$ pode ser definido como {*t, id, a, v, va*}. Em particular, *t* representa o *timestamp* da realização do procedimento, *id* representa o identificador do registro, *a* representa o atributo do conjunto de dados modificado, *v* representa o novo valor inserido e *va* representa valor anterior substituído, quando houver. Exceções a esse formato são os procedimentos de enriquecimento resultam em entradas diferentes, que registram informações a respeito a estrutura do conjunto de dados ou de seus metadados. O *log* de refinamento pode ser implementado através de um arquivo tabular ou mesmo um arquivo JSON.

A execução dos procedimentos pode ser auxiliada por ferramentas específicas de refinamento de dados, que por sua vez podem gerar os *logs* de forma automática. O *log* de refinamento é essencial para as operações Sincronizar e para a atividade de Junção.

**Operação Sincronizar.** A operação Sincronizar pode ser vista como uma função *sincronizar* que, a partir do *log* de refinamento *logDL2* de uma nova versão refinada *DL2* do conjunto de dados *DL1*, armazenado no *Branch Local* $B_{local}$, verifica se existem novas versões publicadas na Web, analisando os *logs* de refinamento de ambos, *logDL2* e *logDW2*, e retornando um conjunto de procedimentos não conflitantes $P = \{p_1, p_2, ..., p_n\}$ e um conjunto de conflitos $C = \{c_1, c_2, ..., c_n\}$. Dois procedimentos $p_i$ e $p_j$ são considerados conflitantes quando atuam em um mesmo conteúdo do conjunto de dados, por exemplo um mesmo registro e atributo, e em versões diferentes de um mesmo conjunto de dados. Cada par de procedimentos conflitantes $< p_i, p_j >$ corresponde a um conflito *c*:

$$P, C = sincronizar(DL2)$$

A operação Sincronizar recupera o *log* de refinamento da versão mais recente de um conjunto de dados publicado na Web e o compara com o *log* de refinamento do conjunto de dados do *Branch Local*. A sincronização dos *logs* de refinamento é realizada por meio do confronto dos procedimentos realizados no conjunto de dados local com o conjunto publicado na Web verificando a existência de conflitos. Como já mencionado, o *log* de refinamento é capaz de fornecer informações precisas sobre as alterações realizadas no conjunto de dados por meio de procedimentos de limpeza e enriquecimento desempenhados, como também local da alteração e o valor inserido ou removido.

A atividade de junção compreende a resolução dos conflitos resultantes da operação Sincronizar. Como parte da atividade de junção, são executados os procedimentos de refinamento sobre o conjunto de dados local. Após a execução dos procedimentos e realização da atualização, uma nova versão do conjunto de dados local (*DL3*) é gerada juntamente com o novo *log* de refinamento *logDL3*. Para que um conjunto de dados possa ser inspecionado e homologado, é necessário que o consumidor requisite a republicação do conjunto refinado através da operação Requisitar Publicação.

**Operação Requisitar Publicação.** A operação Requisitar Publicação pode ser vista como uma função *requisitar* que, a partir de um conjunto de dados refinado e atualizado *DL3*, o *log* do refinamento contendo os procedimentos e alterações realizadas sobre

ele *logDL3* e o ator do refinamento $A_{consumidor}$, cria e retorna um *Branch de Inspeção* destinado à inspeção e homologação do conjunto de dados $B_{inspecao}$:

$$B_{inspecao} = requisitar(DL3, logDL3, A_{consumidor})$$

Ao executar esta operação, um novo *Branch de Inspeção* é criado, e nele é armazenado o conjunto de dados refinado e atualizado pela atividade de Junção (*DL3*), juntamente com o seu *log* de refinamento *logDL3*. Por fim, a criação do *Branch de Inspeção* é necessária para que a atividade de inspeção e ajuste seja desempenhado de forma isolada.

O processo de inspeção tem o objetivo de analisar o conjunto de dados e verificar a consistência dos dados e metadados, definindo o destino do conjunto de dados, seja para aceitação ou rejeição. Após a verificação, o conjunto de dados pode passar por alguns ajustes por parte do publicador. Os ajustes podem ser necessários como complemento do refinamento e tratam de mudanças adicionais aplicadas sobre o conjunto de dados inspecionado. Por exemplo, o publicador ao inspecionar o conjunto de dados pode se deparar com algum problema onde ele mesmo pode realizar o ajuste necessário, não necessitando da interferência do consumidor. O ajuste também pode ser motivado por uma comunicação extra entre o consumidor e o publicador buscando um melhor entendimento do refinamento realizado, podendo conduzir a ajustes adicionais. Por fim, após a inspeção e possíveis ajustes, o conjunto de dados é homologado e pronto para a republicação na Web

**Operação Operação Re-publicar.** A operação Re-publicar pode ser vista como uma função *republicar* que, a partir de um conjunto de dados homologado *DL3*, um *log* de refinamento *logDL3* contendo os procedimentos e alterações realizadas sobre ele, o ator do refinamento $A_{consumidor}$ e o ator da republicação $A_{publicador}$ cria e retorna um novo *Branch Web Secundário* $B_{websec}$ armazenando uma nova versão do conjunto de dados *DW3*:

$$B_{websec} = republicar(DL3, logDL3, A_{consumidor}, A_{publicador})$$

Esta operação cria um novo *Branch Web Secundário* e armazena o conjunto de dados homologado *DL3* juntamente com o seu *log* de refinamento *logDL3*. Nesta operação, deve ser informado o consumidor que realizou o refinamento do conjunto, como também o publicador que está realizando a republicação. Após a republicação, o conjuntos de dados republicado torna-se disponível na Web para o acesso, podendo ele ser acessado por outros consumidores, como também passar por um novo refinamento se necessário. Dessa forma, um novo ciclo de vida pode ser iniciado, como também uma nova interação da estratégia de refinamento, contribuindo assim para o aumento da qualidade dos conjuntos de dados publicados na Web.

### 3.4. Resolução de Conflitos Refinamento

A estratégia de refinamento colaborativo define que consumidores que modifiquem as cópias privadas dos conjuntos de dados, *i.e.*, cada consumidor modifica apenas seu *Branch Local* de um conjunto de dados. Esse isolamento permite que dois ou mais consumidores trabalhem em paralelo. No entanto, conflitos podem emergir devido ao trabalho concorrente, e se tornam mais complexos à medida que as mudanças crescem sem serem integradas e à medida que novas operações de refinamento são realizadas no conjunto de dados.

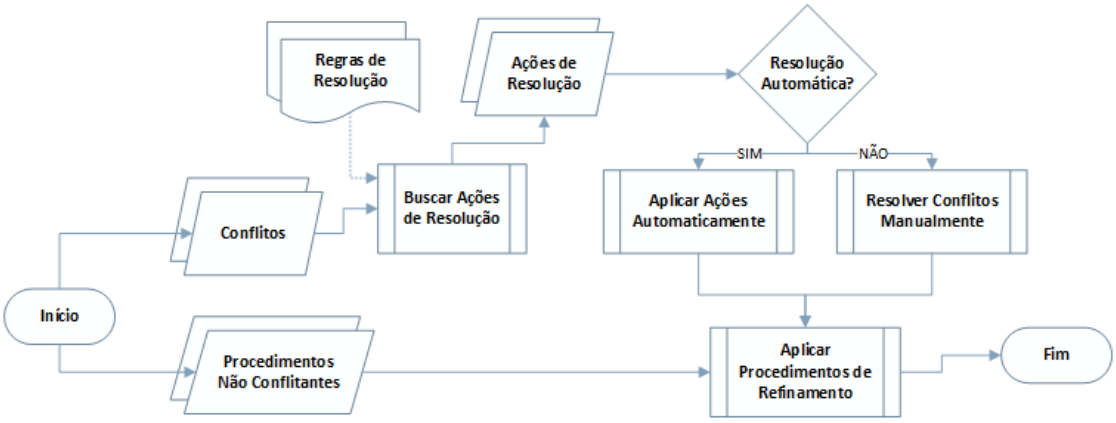


**Figura 2. Fluxograma - Junção de conjuntos de dados. Fonte:Autor**

A fim de solucionar os conflitos resultantes do refinamento colaborativo de um mesmo conjunto de dados, foi proposta a abordagem para resolução de conflitos apresentada na Figura 2 e descrita a seguir. A solução proposta faz uso de heurísticas para facilitar a resolução de conflitos. Estas regras definem heurísticas para resolução de conflitos mais frequentes. Por exemplo, dado um conflito $c1 =< p_1, p_2 >$, tal que $p1$ é um *procAddCampoVazio* realizado no conjunto de dados local que tem como entrada $timestamp = 1528111295$, $i = 1$, $a =$ "$area$", $v =$ "100" e $p2$ é um *procAddCampoVazio* realizado na versão mais recente do conjunto de dados publicada na Web que tem como entrada $timestamp = 1528810818$, $i = 1$, $a =$ "$area$", $v =$ "200", observa-se que os procedimentos foram executados em um mesmo registro e em um mesmo atributo. Uma regra de resolução de conflito pode analisar os procedimentos da seguinte forma: Se os procedimentos $p1$ e $p2$ são *procAddCampoVazio*, então selecione o procedimento mais recente. Dessa forma, a ação de resolução descarta o procedimento mais antigo, permanecendo o mais recente.

Cada ação de resolução pode resolver um ou mais conflitos de forma automática, sem interferência do usuário. Por outro lado, nem todos os conflitos podem ser resolvidos por meio das regras de resolução. Para os casos restantes, a atuação do consumidor é necessária para resolver o conflito. Alguns dos conflitos são facilmente resolvidos. Esta resolução manual pode inclusive ser assistida por uma ferramenta gráfica que forneça funcionalidades para comparação dos elementos dos elementos conflitantes e permita ao usuário selecionar qual alteração realizada é a correta ou mais adequada. Ocasionalmente, porém, o conflito é tão difícil de resolver que o usuário não pode fazê-lo sozinho de tal maneira que as mudanças feitas em ambas as versões mescladas sejam levadas em consideração.

## 4. Avaliação

Para a avaliação da estratégia de refinamento proposta neste trabalho, optou-se pela execução de um experimento no qual um conjunto de participantes (estudantes de pós-graduação) realizaram o processo de refinamento em três cenários distintos.

No primeiro cenário, o refinamento foi realizado apenas por um único participante, atuando no papel de publicador, não sendo assim usada a estratégia de refinamento colaborativo. Em especial, o participante com o papel de publicador possui experiência ativa

na publicação de dados na Web, como também o conhecimento de todo o processo convencional de publicação. No segundo cenário, o refinamento foi realizado por dois participantes, um consumidor e um publicador, sendo aplicada a estratégia colaborativa. Após realizar as operações de limpeza e enriquecimento, o consumidor requisitou a publicação do refinamento realizado, anexando o conjunto de dados refinado e o *log* de refinamento. Por sua vez, o publicador verificou que há uma nova requisição de republicação e inspecionou o conjunto de dados com o auxílio do *log* de refinamento anexado. Por fim, o publicador validou o conjunto de dados e realizou a republicação da nova versão. O terceiro cenário se diferencia do segundo no que tange ao refinamento ter sido realizado por dois consumidores e um publicador. Porém, um dos consumidores ao requisitar a publicação do conjunto refinado, realizou a sincronização do *log* de refinamento a fim de atualizar a sua versão do conjunto por meio do processo de Junção.

Para viabilizar a realização do refinamento, foi desenvolvido um conjunto de serviços para Refinamento Colaborativo[3] que possibilitam gerenciar as operações Requisitar Publicação, Sincronizar e Republicar, como também o processo de Inspeção. Esses serviços foram incorporados à solução DWMS proposta por [Oliveira et al. 2018]. O DWMS é uma coleção de serviços, organizados em módulos, que permitem aos usuários compartilhar conjuntos de dados na Web.

Também foi desenvolvido um conjunto de dados com base em dados[4] da Empresa de Manutenção e Limpeza Urbana (Emlurb). Em particular, o conjunto de dados usado na avaliação continha 100 registros. Nos quais, foram introduzidos 20 erros, tais como erros referenciais, valores falsos e valores abreviados.

Para execução do experimento, foram selecionados 4 participantes, sendo que um dos participantes assumiu o perfil de publicador e os outros três de consumidor. Em especial, o participante com o perfil de publicador possui experiência na publicação de dados na Web. Dessa forma, foi possível garantir que de fato o participante com perfil de publicador tem o conhecimento de todo o processo convencional de publicação de dados na Web.

Todo o processo de avaliação foi realizado em uma sala de reunião fechada. Primeiramente, foi exposta a estratégia para o refinamento colaborativo com intuito de guiar os participantes no desempenho de todo o processo. Foram explicados os procedimentos de refinamento que poderiam ser utilizados, tanto os procedimentos de limpeza, quanto os de enriquecimento. Em seguida, foi apresentado o *log* de refinamento que registra cada procedimento de refinamento realizado.

O DWMS foi instanciado e executado em um servidor local e cada participante teve acesso por meio de um navegador Web convencional. Em cada cenário, os participantes tiveram 20 minutos para realizar as atividades. É importante destacar que neste experimento foi utilizado o conjunto de dados apenas no formato CSV.

### 4.1. Hipóteses e Critérios de Avaliação

A seguir são apresentadas as hipóteses avaliadas pelo experimento.

[3]Detalhes de implementação e arquitetura desse serviço estão disponível em https://blind-review.org/

[4]http://dados.recife.pe.gov.br/dataset/central-de-atendimento-de-servicos-da-emlurb-156/resource/8d4b73c8-d1e1-4efc-97a3-086a682b93b2

- **H1**: A estratégia para o refinamento colaborativo de dados na Web reduz o esforço do publicador no refinamento de conjuntos de dados.
- **H2**: Quando consumidores acessam uma nova versão de um conjunto de dados que foi refinado anteriormente por outro consumidor, há reaproveitamento de trabalho.
- **H3**: Quando um conjunto de dados é refinado pelo consumidor por meio da estratégia colaborativa, há aumento na qualidade do conjunto de dados.

Para avaliar a eficiência da estratégia, foi realizada uma pesquisa qualitativa com cada participante. Ao final de todo o experimento pediu-se que os participantes respondessem um questionário composto de afirmações relacionadas à estratégia proposta. Assim, a partir das notas atribuídas, obteve-se as métricas destinadas à validação das hipóteses H1 e H2. Por meio das métricas, calculamos a média de cada item do questionário de avaliação, i.e., somamos as notas e dividimos pelo número de participantes.

Para avaliar a qualidade do conjunto de dados, utilizou-se os critérios de completude e corretude [Wang and Strong 1996]. Para cada cenário, foi avaliado a qualidade do conjunto de dados da versão inicial e da versão refinada. Assim, ao final de cada cenário, identificou-se quais procedimentos de refinamento foram realizados com base no *log* de refinamento. Verificou-se também a quantidade de procedimentos de limpeza e de enriquecimento que foram realizados e contabilizamos quantos erros foram corrigidos com os procedimentos realizados ao final de cada cenário, assim como o que foi enriquecido. Dessa forma, foi possível mensurar a corretude e completude do conjunto de dados antes e depois de cada cenário e assim direcionar à validação da hipótese H3.

### 4.2. Resultados

De forma geral, os participantes indicaram resultados positivos em relação à diminuição do esforço relacionado às atividades de refinamento proporcionada pela estratégia. Em particular, o participante que atuou como publicador afirmou que mesmo com as atividades adicionais de inspeção e republicação, os consumidores já haviam corrigido a maior parte dos erros, diminuindo assim seu esforço.

Os participantes apontaram que a estratégia colaborativa também permitiu que o trabalho gasto no refinamento de um conjunto de dados possa ser reaproveitado por meio da republicação deste conjunto. Os mesmos reconheceram o benefícios que essa colaboração pode trazer a novos consumidores, inclusive se sentiram mais motivados a contribuir. Contudo, os participantes também afirmaram que a utilização da estratégia colaborativa não evita necessariamente a realização de novas atividades de refinamento. Ainda existem casos que consumidores podem ou devem incrementar o refinamento já realizado anteriormente, seja com o enriquecimento ou com ajustes de limpeza que julgaram necessários.

Seguindo com os resultados, foram avaliados também a qualidade do conjunto de dados refinado em cada cenário (Figura 3). Por meio do *log* de refinamento gerado por cada participante, foram identificados quais erros contidos na versão inicial do conjunto de dados foram corrigidos a fim de avaliar a corretude do conjunto de dados obtido ao final do refinamento.

Como apresentado na Figura 3, no Cenário 1, o participante que desenvolveu o papel de publicador ao realizar o refinamento do conjunto de dados realizou somente 4

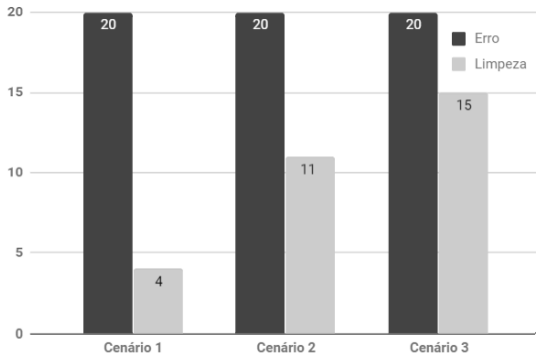


**Figura 3. Distribuição de número de erros inseridos e erros corrigidos por cenário. Fonte: Autor**

procedimentos de limpeza, corrigindo, assim, apenas 20% dos erros contidos. Por outro lado, no Cenário 2, foram realizados, no total, 11 procedimentos de limpeza que corrigiram 11 dos 20 erros contidos no conjunto de dados, totalizando 55% de erros corrigidos. Comparando os Cenários 1 e 2, pode-se observar que houve um aumento expressivo no número de procedimentos de limpeza realizados.

No Cenário 3, foram realizados 15 procedimentos de limpeza sobre o conjunto de dados, corrigindo 15 dos 20 erros inseridos, *i.e*, 75% dos erros. O elevado indicador de correção se deu pelo fato de 2 consumidores terem realizado colaborativamente o refinamento do conjunto de dados. Ao refinar o conjunto de dados colaborativamente, os participantes somaram esforços. Um dos participantes realizou 8 procedimentos de limpeza, enquanto o outro realizou mais 7. É importante destacar que não houve conflito nos procedimentos realizados, pois foram realizados para a correção de diferentes erros, que, por sua vez, estavam inseridos em diferentes locais no conjunto de dados.

Por outro lado, devido a não realização de procedimentos de enriquecimento sobre o conjunto de dados pelos participantes do experimento, não foi possível avaliar a completude do conjunto de dados refinado ao final de cada cenário. Porém, os indicadores mostram que a utilização da estratégia para o refinamento colaborativo aumentou consideravelmente a qualidade dos conjuntos de dados, como ilustrado nos Cenários 2 e 3.

## 5. Conclusão e Trabalhos Futuros

Este trabalho propõe uma nova abordagem para a realização de refinamento de dados na Web a partir do trabalho colaborativo de consumidores e publicadores de dados. A estratégia proposta faz uso de um conjunto de operações e *branches* para criação, homologação e manutenção de versões refinadas de conjuntos de dados publicado na Web.

Realizando a análise dos resultados obtidos concluiu-se que a estratégia para o refinamento colaborativo de conjuntos de dados na Web se mostrou eficiente nos cenários em que ela foi aplicada (Cenário 2 e 3), e em todo o experimento realizado as hipóteses elencadas foram validadas. Dessa forma, de acordo com os experimentos realizados, o uso da estratégia de refinamento colaborativa contribui para reduzir o retrabalho, o esforço do publicador, como também favorece o aumento da qualidade dos dados publicados na

Web.

Como trabalhos futuros, destacam-se o problema da resolução de conflitos que ocorre no processo de junção dos conjuntos de dados e a definição de heurísticas de resoluções de conflitos de forma que esta operação seja realizada automaticamente. Além disso, está prevista a realização de novos experimentos com mais participantes, com uma maior variedade de conjuntos de dados a serem refinados, e com diferentes faixas de duração, a fim de comparar o refinamento realizado em ambos. A partir dos novos experimentos, serão coletadas informações úteis para a realização de melhorias na estratégia proposta.

## Referências


Chapman, A. D. (2005). *Principles of data quality*. GBIF.

Clarke, M. and Harley, P. (2014). How smart is your content? using semantic enrichment to improve your user experience and your bottom line. *Science*, 37(2):41.

da Silva, K. M. (2019). Um modelo de ciclo de vida dos dados na web. Master's thesis, Universidade Federal de Pernambuco, Centro de Informática, Curso de Pós-Graduação em Ciências da Computação, Recife.

dos Santos, H. D. A., Oliveira, M. I. S., Glória de Fátima, A., da Silva, K. M., Muniz, R. I. V. C. S., and Lóscio, B. F. (2018). Investigations into data published and consumed on the web: a systematic mapping study. *Journal of the Brazilian Computer Society*, 24(1):14.

Fileto, R., Bogorny, V., May, C., and Klein, D. (2015). Semantic enrichment and analysis of movement data: probably it is just starting! *SIGSPATIAL Special*, 7(1):11–18.

Gousios, G., Pinzger, M., and Deursen, A. v. (2014). An exploratory study of the pull-based software development model. In *Proceedings of the 36th International Conference on Software Engineering*, pages 345–355. ACM.

Levine, S. S. and Prietula, M. J. (2013). Open collaboration for innovation: Principles and performance. *Organization Science*, 25(5):1414–1433.

Lóscio, B. F., Oliveira, M. I. S., and Bittencourt, I. I. (2015). Publicação e Consumo de Dados na Web: Conceitos e Desafios. *Tópicos em Gerenciamento de Dados e Informações (Mini Cursos - SBBD 2015)*, d:39–69.

Maletic, J. I. and Marcus, A. (2000). Data cleansing: Beyond integrity analysis. In *Iq*, pages 200–209.

Oliveira, L. E. R., Oliveira, M. I. S., Santos, W. C. d. R., and Lóscio, B. F. (2018). Data on the web management system: a reference model. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, page 2. ACM.

Rahm, E. and Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33.