DW-CGU: Integração dos Dados do Portal da Transparência do Governo Federal Brasileiro

Eduardo de Paiva[1,2], Kate Revoredo[1], Fernanda Baião[1]

[1]Departamento de Informática Aplicada – Universidade Federal do Estado do Rio de Janeiro (UNIRIO)
Av. Pasteur 458, Urca, 22290-240, Rio de Janeiro –RJ - Brasil

[2]Controladoria Geral da União (CGU)
SAS, Quadra 01, Bloco A, Edifício Darcy Ribeiro, Brasília-DF - Brasil

{eduardo.paiva,katerevoredo,fernanda.baiao}@uniriotec.br

Abstract. *Government transparency portals are tools for democracy consolidation and development. However, the mere provision of information does not assure a unified view of all data. Since the presented data derives from several independent Information Systems, many integration problems arise. The goal of this paper is to develop a data integration solution to the Brazilian Federal Government transparency portal and its main contribution is the proposal of an architecture able to control all stages of the data integration process within a corporate portal. The obtained results show the analysis potential that the integrated data can offer. The validation of the proposed architecture evaluated the results of the data integration process not only before the solution deployment, but also throughout the system operation phase. The paper evidences that data integration involves issues, which go beyond technological challenges and it needs semantic data interpretations.*

Resumo. *Os portais de transparência governamental são ferramentas de consolidação e desenvolvimento da democracia. Porém, a simples disponibilização de informações não garante uma visão unificada dos dados. O fato de os dados apresentados serem oriundos de Sistemas de Informação independentes traz problemas de integração. O objetivo desse trabalho é desenvolver uma solução de integração de dados para o Portal da Transparência do Governo Federal e sua principal contribuição é a proposta de uma arquitetura capaz de controlar todas as fases do processo de integração dos dados de um portal coorporativo. Os resultados obtidos evidenciaram o potencial de análise que esses dados integrados oferecem. A validação da arquitetura proposta avaliou os resultados do processo de integração de dados não apenas antes da implantação da solução, como também durante toda a fase de operação do sistema. O trabalho evidencia que a integração de dados envolve questões que vão além dos desafios tecnológicos, carecendo de interpretações semânticas dos dados.*

PAIVA, E.; REVOREDO, K.; BAIÃO; F.;
DW-CGU: Integração dos Dados do Portal da Transparência do Governo Federal Brasileiro
iSys – Revista Brasileira de Sistemas de Informação, Rio de Janeiro, vol. 9, No. 1, pp. 06-32, 2016

1. Introdução

A popularização da Internet tem ajudado a tornar os portais de transparência governamental importantes ferramentas de consolidação e desenvolvimento da democracia [Agner 2008]. Tais portais facilitam o acesso aos dados governamentais, propiciando maior poder de controle aos cidadãos. Esses sites permitem que as pessoas possam acompanhar o emprego dos recursos públicos, bem como a atuação dos seus representantes políticos.

Um dos fatores que viabilizou a institucionalização desses instrumentos de transparência foi o alto grau de informatização dos governos. Atualmente, praticamente todas as atividades realizadas pelo poder público são controladas por algum sistema de informação [Jardim 2004]. Essa automatização dos processos governamentais gera grandes volumes e variedades de dados que são utilizados como fontes de informação de transparência pública. Outro fator que impulsionou a difusão dos portais de transparência pública no Brasil foi a publicação da lei complementar 131 [Brasil. Lei nº 131 2009], que determina a disponibilização, em tempo real, de informações pormenorizadas sobre a execução orçamentária e financeira da União, dos Estados, e dos Municípios.

Porém, o simples fato de se disponibilizar as informações em sites de transparência não garante que a população terá condições de acompanhar efetivamente as atividades governamentais, uma vez que os sistemas são, em sua maioria, independentes e mantidos por instituições distintas, o que acarreta dificuldades na consolidação de informações provindas de diversas fontes. Muitas vezes o mesmo assunto é representado de forma distinta nos diferentes sistemas ou, mesmo quando os dados são apresentados conjuntamente, tem-se a impressão de que as informações são diferentes, tornando inviável a comparação de dados de fontes distintas, prejudicando assim a transparência.

A transparência não é uma qualidade binária. De acordo com Araújo et al. (2010), o grau de transparência pode ser avaliado por níveis de maturidade. Os Portais de Transparência dos diversos entes da Administração Pública possuem diferentes estágios de maturidade, mas todos carecem de recursos que ofereçam melhores formas de apresentação dos dados.

Apesar da área de transparência pública ser um campo de pesquisa e práticas relativamente novo, já existem muitos estudos que tratam de tal assunto, como pode ser visto em [Prado et al. 2012] e [Prado e Loureiro 2008]. Porém, essas iniciativas se propõem a avaliar o grau de transparência e satisfatibilidade dos portais, sem apresentar uma solução específica para os problemas levantados.

Nesse contexto, Nazario *et al.* (2012) fazem uma avaliação do Portal da Transparência do Governo Federal Brasileiro. Apesar desse estudo concluir que a maioria dos critérios avaliados sejam classificados como satisfatórios, alguns critérios foram considerados deficientes. Os atributos mal avaliados foram: *Clear* (que avalia se a informação é compreensível para o grupo alvo) e *Convenient* (que avalia se a informação atende às necessidades do usuário).

Sendo assim, apesar dos trabalhos apontarem carências não preenchidas pelos portais, ainda existe uma lacuna com relação a iniciativas que propõem formas de

melhorar a qualidade das informações apresentadas nos portais de transparência existentes. Nesse sentido, Rommel, Carvalho *et al.* (2013) e Rommel, Carvalho *et al.* (2014) introduzem uma forma para aumentar o grau de informatividade do site de transparência do governo federal através de uma metodologia para a formulação de preços de referência dos produtos comprados e apresentados no Portal. Para a formulação desses preços, são utilizados basicamente os dados de um único sistema, sendo que a única interação com informações oriundas de outro *software* é para buscar a descrição de um código específico, sem a intenção de fornecer uma visão integrada dos dados.

Beluzo (2015) já propõe um integrador de dados da execução orçamentária para os dados do Governo de São Paulo. Porém, o escopo do trabalho se restringe a dados orçamentários, não contemplando diversas outras informações apresentadas nos sites de transparência pública, que podem contextualizar e esclarecer o cidadão no entendimento das atividades governamentais de forma integrada.

Atualmente, os portais de transparência apresentam informações oriundas de diversas fontes distintas, de forma individualizada, sem apresentar nenhum tipo de relacionamento de informações recebidas de sistemas diferentes, dificultando assim, uma visão sistêmica dos dados.

Esse fato evidencia a necessidade de ações que favoreçam a interoperabilidade dos dados, pois assim consegue-se um ganho informacional muito maior do que a simples apresentação dos dados de forma individualizada. Por exemplo, o processo licitatório das compras realizadas pelo governo é controlado por um determinado sistema corporativo, enquanto os pagamentos dessas compras (em favor das empresas vencedoras dos certames) são feitos em outro sistema; logo, o fato dessas informações não estarem devidamente conectadas prejudica a transparência e dificulta o acompanhamento do processo como um todo, por parte do cidadão.

Assim, a questão de pesquisa que esse artigo aborda é como disponibilizar, de forma integrada, as informações disponíveis no portal de transparência do governo federal brasileiro, oriundas dos diversos sistemas coorporativos. Uma arquitetura para a integração desses dados é proposta, a fim de propiciar uma visão consolidada e harmoniosa das informações publicadas. O trabalho também propõe um processo de validação para avaliar a eficácia da solução e identificar possíveis erros.

O artigo apresenta a hipótese de que, se os dados forem devidamente tratados e interligados, é possível se obter uma visão integrada das informações governamentais, mesmo que esses dados sejam oriundos de fontes distintas.

Para testar a hipótese, foi proposta a utilização de um *data warehouse* que integra os dados dos diversos sistemas coorporativos do governo brasileiro. A ideia é a utilização das informações compartilhadas pelos diferentes sistemas a fim de dar uma visão integrada dos assuntos apresentados no portal. As principais limitações para essa tarefa são o grande volume e variedade de dados, a falta de documentação sobre os sistemas fornecedores dos dados e a diferença de granularidade e de frequência de atualização das diversas fontes de dados.

Esse artigo apresenta as principais técnicas utilizadas no processo de integração dos dados, bem como a arquitetura desenvolvida para operacionalizar essa tarefa. O

restante desse trabalho está organizado da seguinte forma: A seção 2 faz uma compilação dos conceitos de integração de dados apresentados pelos principais autores dessa área, a seção 3 demonstra a proposta de integração dos dados do Portal da Transparência do Governo Federal Brasileiro. Já as seções 4 e 5 descrevem os resultados e as formas de validação utilizadas, respectivamente. A última seção se destina à apresentação das conclusões do trabalho.

2. Revisão teórica de integração de dados

Apesar de ainda ser uma área pouco explorada no campo da transparência pública, a teoria de integração de dados apresenta-se bem desenvolvida como disciplina. Esta seção apresenta uma compilação de alguns conceitos formulados pelos principais autores dessa área.

Segundo Lenzerini (2002), a integração de dados consiste na combinação de diferentes fontes para prover ao usuário uma visão unificada desses dados. Porém, tal combinação não costuma ser simples. Hull (1997) cita algumas questões fundamentais que devem ser tratadas ao lidar com essa heterogeneidade de dados: (i) como fornecer uma visão integrada da sobreposição de conjuntos de dados de múltiplas bases de dados; (ii) como suportar atualizações; (iii) como identificar e especificar o relacionamento entre duas ou mais instâncias de dados replicados; e (iv) como manter dados replicados "sincronizados".

Para tratar de tais questões e também dos outros desafios da unificação de informações, foi desenvolvida toda uma área de estudo. Essa seção tem o objetivo de apresentar uma revisão teórica de alguns trabalhos referentes à integração de dados. Ela está dividida em subseções para facilitar a distinção dos conceitos elencados. A subseção 2.1 trata das diferentes formas de mapeamento dos esquemas. As abordagens de integração são tratadas na subseção 2.2. As subseções 2.3 e 2.4 tratam das questões de integração de esquemas e de conflitos de integração, respectivamente. Finalmente, a subseção 2.5 aborda as dificuldades da integração semântica.

2.1. Formas de mapeamento

A atividade de integração de dados consiste da definição de um esquema de dados único, através do qual é possível acessar informações provenientes de fontes de dados distintas [Lenzerini 2002]. Ou seja, cada fonte possui os dados representados e armazenados de acordo com seus esquemas; porém, é necessária a definição de um esquema de reconciliação que trate essa diversidade e forneça uma visão única dos dados. Nesse contexto, o mapeamento de esquemas descreve o relacionamento entre o esquema mediador e os esquemas das fontes de dados [Doan *et al.* 2012].

Lenzerini (2002) propôs duas abordagens de integração: *global-as-view* (GAV) e *local-as-view* (LAV). A primeira requer que o esquema unificador seja definido em função das fontes de dados, ou seja, o GAV define o esquema mediador como um conjunto de visões sobre cada fonte de dados [Doan *et al.* 2012]; já o LAV requer que o esquema unificador seja expresso de forma independente, e cada fonte seja definida em função dele.

Na abordagem GAV, para cada componente do esquema unificador é escrita uma consulta sobre os esquemas locais. Logo, para cada relação global R do esquema

unificador, é definida uma consulta para obter as tuplas de R a partir das tuplas das relações armazenadas nas fontes de dados locais [Alves Costa e Salgado 2005]. Sendo assim, no mapeamento GAV, cada componente do esquema unificador tem uma correspondência com os esquemas das fontes de dados. Nesse caso, todo elemento no esquema unificador está associado a uma visão sobre a fonte de dados e tal visão especifica como obter os dados do esquema unificador, utilizando para isso consultas sobre os esquemas locais.

Na abordagem LAV, ao invés de se escrever consultas que definem como os componentes do esquema integrador são obtidos, são feitas consultas que descrevem como se obtém a extensão das fontes de dados a partir do esquema integrador. Nessa forma de mapeamento, as consultas são feitas diretamente no esquema integrador e, portanto, pressupõe-se um esforço anterior no sentido de mapear exatamente como cada um dos conceitos locais será representado no esquema unificador.

Para exemplificar a diferença entre as formas de mapeamento GAV e LAV, pode-se considerar uma situação com 3 fontes de dados distintas (fonte 1, fonte 2 e fonte 3) e um esquema unificador que apresente essas fontes já integradas. No mapeamento GAV, essas fontes não são copiadas em nenhum outro repositório. Dessa forma, a integração dos dados se dá através de consultas especificamente montadas que leem os diversos bancos de dados a fim de apresentar uma visão consolidada desses dados, ou seja, a integração se dá em um nível lógico, e não físico.

Já na abordagem LAV, a integração é feita através da criação de um repositório central, onde os dados das diversas fontes (fonte 1, fonte 2 e fonte 3) são carregados, de forma que essa carga seja feita com todos os problemas de integração tratados. Logo, nesse caso, os dados já consolidados ficam carregados fisicamente em um repositório integrador.

2.2. Abordagens para integração de dados

A heterogeneidade de fontes é um fator que torna a integração de dados uma atividade cada vez mais necessária. Sendo assim, várias propostas de solução para esse tipo de problema foram desenvolvidas. Ziegler e Dittrich (2007) categorizam as principais abordagens de integração de dados de acordo com seu nível. As categorias definidas são apresentadas a seguir:

- Integração manual: os usuários interagem diretamente com todos os sistemas de informação relevantes. A integração de dados é feita de forma manual. Nessa abordagem, os usuários precisam lidar com diferentes interfaces e linguagens de consulta. Os usuários também precisam ter informações detalhadas sobre a localização e a representação lógica e semântica dos dados.
- Interface comum do usuário: o usuário é suprido com uma interface comum que provê uma visão uniforme dos dados. Nesse caso, os dados continuam separados, porém, o usuário consegue acessá-los através de uma interface única.
- Integração por aplicação: são utilizadas aplicações de integração que acessam as fontes de dados e retornam um resultado integrado para o usuário. Ziegler e Dittrich (2007) consideram que essa solução é prática para sistemas com poucos

componentes, mas ela se torna muito custosa à medida que se aumenta o número de interfaces e formatos de dados que devem ser homogeneizados e integrados.

- Integração por *middleware*: o *middleware* fornece funcionalidades reutilizáveis que geralmente são utilizadas para resolver aspectos específicos do problema de integração. Segundo Barbosa (2001), o principal objetivo do *middleware* é liberar o usuário de ter que conhecer detalhes sobre todas as fontes de dados e ter que interagir com cada uma delas individualmente.
- Acesso uniforme aos dados: uma integração lógica de dados é realizada no nível de acesso a dados. Nessa abordagem são utilizadas aplicações globais que fornecem visões unificadas dos dados distribuídos. Nesse caso, os dados não são replicados em um novo repositório, eles permanecem em seus repositórios locais, sendo que uma camada virtual provê a visão unificada.
- Área de armazenamento comum: Na área de armazenamento comum é executada uma integração física. Os dados são transferidos das diversas fontes de dados e integrados em um repositório comum.

Ziegler e Dittrich (2007) ressaltam ainda que existem soluções de integração de dados que podem ser classificadas em mais de uma das categorias citadas acima.

2.3. Integração de esquemas

Independente da abordagem empregada no processo de integração de dados, é necessária a realização da integração dos esquemas. As técnicas de integração de esquemas são consideradas como uma das operações básicas requeridas pelo processo de integração de dados [Bernstein e Melnik 2004]. Essas técnicas oferecem a possibilidade de unificar a representação de vários esquemas dentro de um esquema global [Arfaoui e Akaichi 2015].

O objetivo da integração de esquemas é produzir um conjunto de correspondência entre diferentes esquemas [Doan *et al.* 2012], a fim de se obter um esquema integrador.

Batini, Lenzerini e Navathe (1986) e Mello (2002) apresentam um processo para integração de esquemas composto por 4 passos: Pré-integração, comparação de esquemas, conformação de esquemas, e junção e reestruturação.

Na etapa de pré-integração é feita uma análise dos esquemas locais com o intuito de se definir a política de integração. Nessa fase é definida a estratégia para integração, assim como a quantidade de esquemas a serem integrados ao mesmo tempo. A coleta de informações adicionais para a integração também é feita nessa etapa.

Na comparação de esquemas são feitas análises e comparações para se determinar as correspondências e possíveis conflitos entre os esquemas de banco de dados.

Quando conflitos são identificados, faz-se um esforço para resolvê-los, a fim de que a fusão dos vários esquemas seja possível. Essa atividade ocorre na fase de conformação de esquemas. Batini, Lenzerini e Navathe (1986) ressaltam que a resolução automática de conflitos geralmente não é viável, dado que é necessária uma

estreita relação entre projetistas e usuários antes que qualquer acordo de integração seja firmado.

Na fase junção e reestruturação, já existe um esquema integrador intermediário. Nesse momento, o esquema intermediário é analisado e, se necessário, é reestruturado para se atingir algumas qualidades desejáveis. De acordo com Dang e Feldmann (2010) e Batini, Lenzerini e Navathe (1986), o esquema global pode ser testado com relação aos seguintes critérios de qualidade:

- Completude e corretude: O esquema global deve conter todos os conceitos presentes nos esquemas locais de forma semanticamente correta.
- Minimalidade: um conceito representado em mais de um esquema local deve ser representado uma única vez no esquema global.
- Entendibilidade: O esquema integrado deve ser de fácil entendimento para projetistas e usuários finais. Isso implica que, dentre as várias possíveis representações, deve-se optar pelo modelo que permita o melhor entendimento.

2.4. Conflitos de Integração

As técnicas de integração de esquema oferecem a possibilidade para unificar a representação de vários esquemas em um esquema global [Arfaoui e Akaichi 2015]. No entanto, durante essa integração podem ser detectados conflitos entre os diferentes esquemas [Doan *et al.* 2012]. Dang e Feldmann (2010) e Batini, Lenzerini e Navathe (1986) citam os conflitos de nome e de estrutura.

Conflito de nome

Uma fonte óbvia de heurísticas para a conciliação de esquemas é a correspondência baseada na comparação entre os nomes dos elementos, com a ideia de que os nomes transmitam o verdadeiro significado semântico dos elementos [Doan *et al.* 2012].

No entanto, nem sempre essa comparação de nomes fornece uma relação direta entre as estruturas dos esquemas comparados. Segundo Naiman e Ouksel (1995), os conflitos de nome são aqueles que se referem aos relacionamentos entre nomes de atributos ou instâncias. Batini, Lenzerini e Navathe (1986) e Bellström (2006) classificam os conflitos de nome em dois tipos:

- Homonímia: quando um nome referencia dois ou mais conceitos [Bellström 2006].
- Sinonímia: quando dois ou mais nomes se referem a um mesmo conceito [Bellström 2006].

Conflito de estrutura

Os conflitos de estrutura ocorrem quando conceitos do mundo real são modelados usando diferentes construtores em esquemas distintos [Lee e Ling 1995]. Muitas vezes, os conflitos de estrutura ocorrem em decorrência de diferentes requisitos no mesmo domínio de dados, o que leva a diferentes definições de atributos, tipos de dados e etc. [Dang e Feldmann 2010]. Batini, Lenzerini e Navathe (1986) citam quatro tipos de heterogeneidades de estrutura:

- Conflito de tipo: quando o mesmo conceito é representado por diferentes níveis em esquemas diferentes. Por exemplo, um conceito é representado como uma entidade em um esquema e como um atributo em outro.
- Conflito de dependência: quando um grupo de conceitos relacionados entre si possuem diferentes dependências em diferentes esquemas. Por exemplo, o relacionamento conjugal entre um homem e uma mulher é de 1:1 em um esquema, mas de m:n em outro que contabiliza o histórico de casamentos.
- Conflito de chave: chaves diferentes são designadas para o mesmo conceito em esquemas diferentes.
- Conflito de comportamento: surge quando diferentes condições de inserção/deleção estão associadas à mesma classe de objeto em esquemas diferentes. Por exemplo, em um esquema o departamento pode existir sem empregados, e em outro esquema a exclusão do último empregado de um departamento implica na exclusão desse departamento. Esse conflito só pode existir quando o modelo de dados permite a representação das propriedades comportamentais dos objetos.

2.5. Integração semântica

Ziegler e Dittrich (2004) citam que a integração é muito mais do que somente um problema estrutural ou técnico. Segundo Ziegler e Dittrich (2004), tecnicamente não é difícil conectar diferentes sistemas de bancos de dados relacionais. O mais complicado é integrar dados descritos por diferentes modelos, e pior ainda são os problemas causados por dados com semânticas heterogêneas. Relações com semânticas heterogêneas (que implicitamente parecem ser iguais) conduzem a resultados errados e conclusões divergentes.

Portanto, Ziegler e Dittrich (2007) ressaltam que são necessárias semânticas explícitas e precisas para os dados a serem integrados, afim de que os resultados obtidos sejam corretos e significativos.

Segundo Ziegler e Dittrich (2004), na área de banco de dados, a semântica pode estar relacionada com a interpretação que as pessoas têm dos dados e dos esquemas com relação as suas interpretações do mundo em um determinado contexto.

Ouksel e Sheth (1999) afirmam que, na integração de dados, a semântica do mundo real deve ser mapeada para o modelo computacional, e que esse mapeamento envolve a interpretação humana do significado e uso desses dados.

A integração semântica é a tarefa de agrupar, combinar ou completar dados de diferentes fontes, levando em conta a semântica de dados de forma explícita e precisa, a fim de evitar que dados semanticamente incompatíveis sejam mesclados.

Ziegler e Dittrich (2004) ainda citam que um pré-requisito para resolver a ambiguidade semântica é a utilização de metadados explícitos para elicitar todas as suposições e informações de contexto implícitas.

Logo, o problema da integração de dados é muito mais que uma questão de incompatibilidades tecnológicas ou de modelagem de dados. Existem questões mais profundas que estão relacionadas ao próprio significado dos dados e com o entendimento do contexto que está sendo representado. Sendo assim, é importante se ter

em mente que integrar dados não se resume a tratar diferenças de tecnologia e alinhar modelos de dados. Essa atividade também está diretamente relacionada com o conhecimento do negócio que está sendo tratado.

3. Integração dos dados do Portal da Transparência do Governo Federal Brasileiro

Nesta seção será apresentada a arquitetura proposta para operacionalizar o processo de integração dos dados do Portal da Transparência do governo federal brasileiro, bem como as soluções a serem implementadas diante dos principais problemas ligados à unificação dos dados do portal.

A seção está dividida em 6 subseções. A primeira subseção descreve o contexto em que o Portal está inserido. A subseção 3.2 apresenta a arquitetura de integração dos dados do Portal. A integração de esquemas dos dados do Portal é tratada na subseção 3.3. Na subseção 3.4 são demonstrados os conflitos encontrados, bem como as soluções utilizadas. A subseção 3.5 aborda a questão da integração semântica dos dados do Portal.

3.1. Cenário do Portal da Transparência

O portal da transparência do Governo Federal Brasileiro[1] é uma iniciativa da Controladoria Geral da União e tem o objetivo de apresentar os dados governamentais de forma a estimular o controle social e fortalecer a democracia no Brasil.

O portal reúne dados de uma série de sistemas governamentais brasileiros, com propósitos distintos e que funcionam de forma independente. Cada solução é mantida por órgãos específicos, cuja única responsabilidade com relação à transparência é disponibilizar os dados, no formato original, para o mantenedor do Portal.

A Figura 1 apresenta os temas tratados atualmente pelo portal da transparência. Cada retângulo cinza representa um fornecedor distinto de informação e cada retângulo interno branco representa um assunto. Conforme pode ser observado, atualmente o portal possui 14 fornecedores de informação distintos e trata de 24 assuntos diferentes, sendo que esse escopo aumenta constantemente devido a demandas do governo e da própria sociedade. Esta diversidade de assuntos confere grande abrangência ao portal.

Conforme apresentado em [Araújo *et al.* 2010], o conceito de transparência é composto por uma série de características, as quais podem ser agrupadas em cinco grupos distintos: acessibilidade, usabilidade, informatividade, entendimento e auditabilidade. A simples disponibilização dos dados, conforme ocorre atualmente no portal, propicia apenas a satisfação das características do grupo acessibilidade, deixando uma lacuna com relação aos demais grupos. A apresentação desses dados de forma integrada não irá assegurar a plena satisfação de todas as características que compõem o conceito de transparência, mas é um pré-requisito para a evolução do grau de transparência do portal.

Um exemplo negativo da apresentação dos dados de forma não consolidada pode ser visto no caso dos convênios[2]. Em uma determinada área do portal são

[1] http://transparencia.gov.br/

apresentados todos os convênios celebrados pelo Governo Federal. Em outra parte do portal são apresentados todos os pagamentos realizados, inclusive os pagamentos dos convênios firmados. Porém, como essas informações são oriundas de sistemas diferentes, elas acabam ficando sem qualquer tipo de conexão e são apresentadas em áreas distintas do portal, tornando muito mais difícil a interligação destas perspectivas por parte do cidadão que acessa o portal.

A arquitetura de integração proposta utiliza a estrutura de um *data warehouse* [KIMBALL; ROSS, 2011] que armazena os dados utilizados para alimentar o Portal da Transparência. Todas as fontes (oriundas dos diversos sistemas) são tratadas, integradas, deduplicadas[3] e correlacionadas durante os processos de carga desse *data warehouse*.

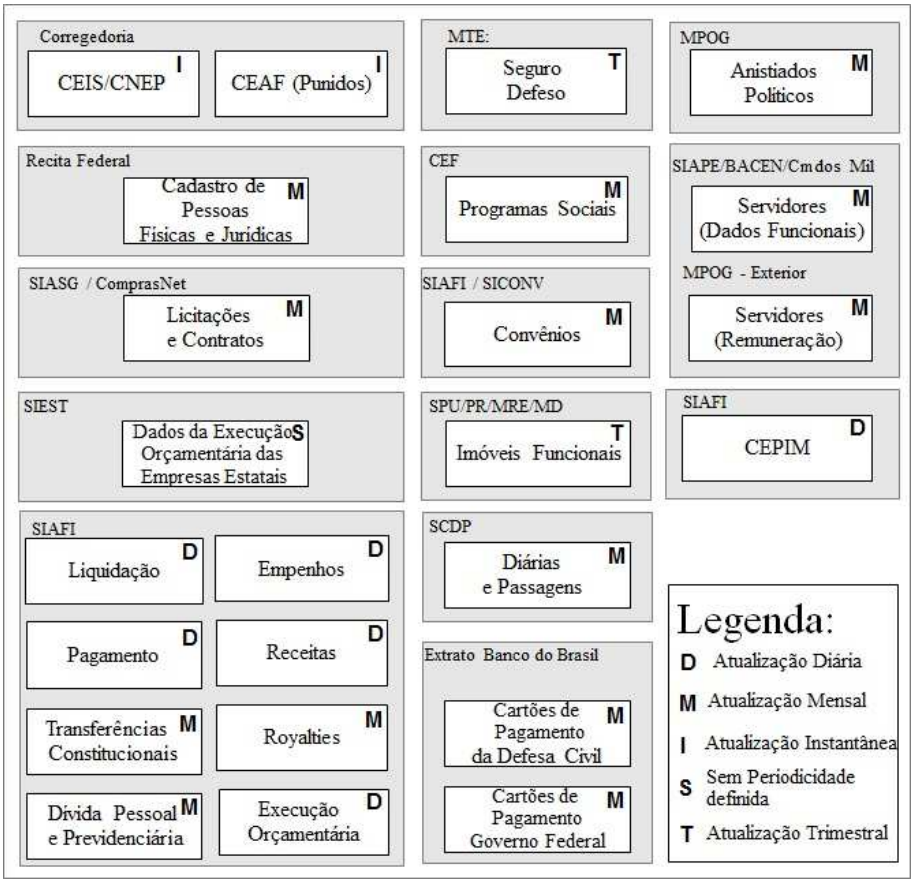

Figura 1: Mapa de assuntos do portal da transparência do governo federal brasileiro a serem integrados

[2] Convênios são acordos firmados por entidades públicas de qualquer espécie, ou entre estas e organizações particulares, para realização de objetivos de interesse comum dos partícipes [Meirelles 2001].

[3] Deduplicação refere-se à atividade de retirada de registros duplicados.

3.2. Arquitetura de integração dos dados do Portal da Transparência

O fato de os sistemas que proveem dados para o portal serem mantidos por entidades externas, cujo único compromisso é disponibilizar essas informações sem alterar suas rotinas já existentes, torna a estratégia de mapeamento *global-as-view* (GAV) [Lenzerini 2002] inviável, restando como alternativa a estratégia *local-as-view* (LAV), também definida em [Lenzerini 2002]. Outro fato que também impõe a opção pelo LAV é a necessidade de constantes inclusões de novas bases de dados ao Portal. O Portal da transparência é dinâmico e está sempre sendo demandado por novas bases de dados, que por questões políticas ou sociais devem entrar no ar. Como Halevy, Rajaraman e Ordille (2006) citam, uma das vantagens da utilização do LAV é a facilidade de inclusão de novas fontes de dados.

Sendo assim, a opção pela forma de mapeamento LAV se justifica pela própria configuração dos processos que operacionalizam a alimentação dos dados do portal. Para exemplificar tal situação, pode-se citar o caso dos dados oriundos do SIAFI[4]. Esse sistema controla toda a contabilidade do Governo Federal Brasileiro e é acessado por unidades de todo o Brasil. Dessa forma, por questões de segurança e desempenho, nenhum ator externo a esse sistema tem acesso a sua base de dados. Logo, todas as informações recebidas do SIAFI são oriundas de extrações específicas geradas por procedimentos controlados e mantidos pelos administradores do sistema, o que torna inviável a forma de mapeamento GAV, que exige acesso à base de dados original para a integração dos dados. Cabe ressaltar que situações análogas a essa ocorrem com todos os demais provedores de dados do portal.

Dentre as abordagens de integração citadas em [Ziegler e Dittrich 2007], a proposta é a utilização de uma área de armazenamento comum. Mais uma vez, a opção se dá devido às características dos dados recebidos pelo Portal da Transparência. Desta forma, a solução a ser empregada é a utilização de um *data warehouse* que consolida todas as informações recebidas em um ponto único de acesso. Esse *data warehouse* servirá como única fonte de informação para o Portal da Transparência, que atualmente lê dados de diversas fontes para alimentar suas páginas.

A Figura 2 demonstra a arquitetura da solução proposta, com os fluxos de dados numerados de 1 a 6, descritos a seguir.

No fluxo 1 são empregadas as estratégias de integração de dados citadas na seção 2. Os dados integrados são carregados em um repositório central denominado DW-CGU. As subseções 3.3 e 3.4 detalham as atividades que ocorrem nesse fluxo.

O fluxo 2 carrega os dados do DW-CGU em um *Data Mart* específico para o portal (DM-Portal). O objetivo desse fluxo é aplicar todas as regras de sigilo necessárias para apresentação dos dados na web pois, conforme estabelecido pela Lei de acesso a informação [Brasil, Lei nº 12.527 2011], todas as informações protegidas por sigilo possuem prazos pré-determinados para a perda da classificação sigilosa. Sendo assim, como as regras de sigilo estão sendo aplicadas em outra camada, não há necessidade de reprocessamentos dos dados do repositório central após a expiração do prazo do sigilo. Para exemplificar a aplicação das regras de sigilo do DM-Portal, pode-se considerar a

[4] Sistema Integrado de Administração Financeira (SIAFI): Sistema Informatizado que processa e controla a execução orçamentária, financeira, patrimonial e contábil da União [Feijó 2006].

situação em que, por algum motivo de segurança pública, um determinado pagamento deva ser apresentado no portal com os campos favorecido e objeto da compra protegidos. Porém, independente dessas regras de privacidade, todo o processo de integração e carga dos dados do DW-CGU é executado sem nenhum tipo de filtro (visto que essas regras apenas dizem respeito à publicação dos dados na internet). As regras de sigilo só são aplicadas nos processos de carga do DM-Portal, que já carregam esses dados devidamente tratados, neste caso sem as informações do favorecido e do objeto da compra. Dessa forma, quando essa informação deixar de ser protegida pelo sigilo, basta uma recarga do DM-Portal, sem a utilização desses filtros de privacidade, eliminando-se assim a necessidade de recarga de todos os dados de pagamento do DW-CGU, e consequentemente todo o esforço de integração desse grande volume de dados. Outra vantagem em se manter uma camada própria para os dados a serem apresentados no Portal da Transparência é que novas visões de dados podem ser propostas sem a necessidade de se alterar a estrutura dos dados originais.

O fluxo 3 é utilizado para fazer a validação e homologação dos dados a serem carregados no portal. O volume de dados a ser carregado diariamente é muito grande (cerca de 45 mil registros diários) e propenso a erros. Dessa forma, faz-se necessária a existência de mecanismos que permitam que os gestores do portal tenham condições de validar e homologar os dados de forma rápida e sem a necessidade de conhecimentos específicos da área de banco de dados.

No fluxo 4 são coletadas informações de carga e tratamento de dados (hora de início e término, quantidades de registros, fontes utilizadas, erros reportados, etc.) geradas durante todo o processo. Tais informações (denominadas de dados de proveniência [Freire *et al.* 2008]) possibilitam o gerenciamento e a manutenção da rastreabilidade dos dados carregados, permitindo que seja possível a identificação da fonte de dados original utilizada para a carga de cada um dos registros, bem como o levantamento de estatísticas dos processos de carga.

Os fluxos 5 e 6 representam a migração dos dados para o ambiente de produção do Portal da Transparência e a utilização dos dados integrados por outras áreas da CGU, respectivamente.

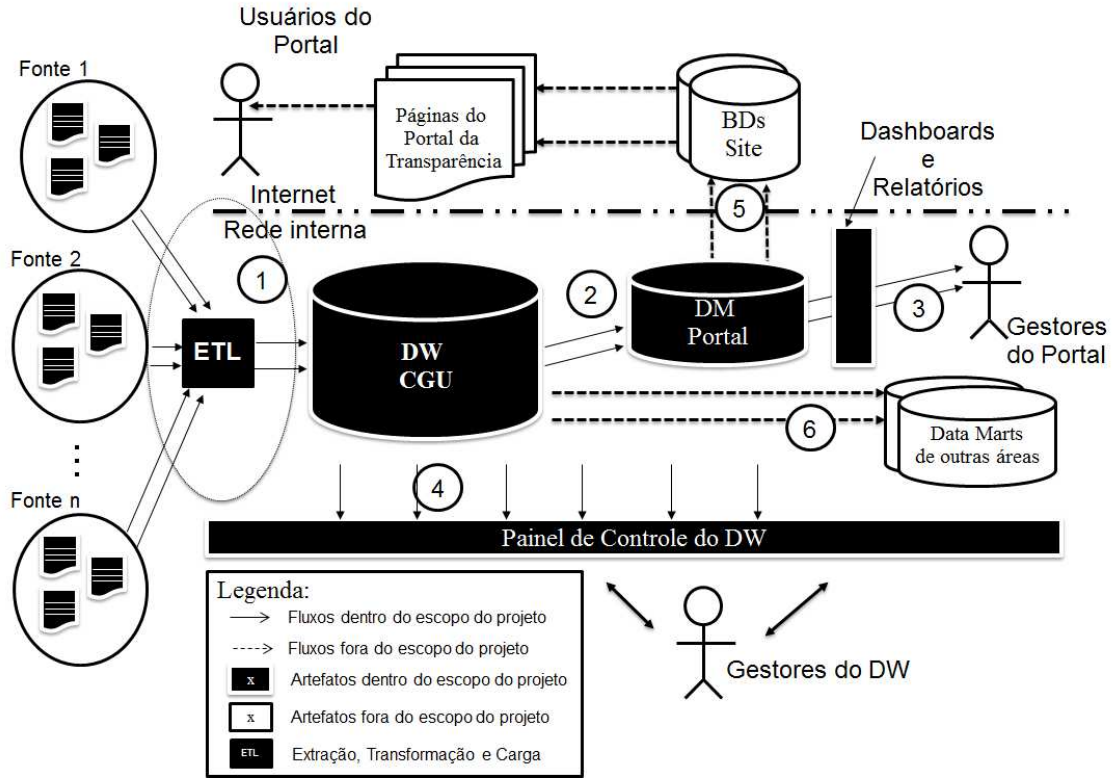

Figura 2: Arquitetura da solução de integração proposta

Pela análise da arquitetura apresentada na Figura 2, pode-se perceber claramente que a abordagem de integração utilizada é a área de armazenamento comum, e que o repositório utilizado para armazenar os dados integrados é o DW-CGU.

Liu et al. (2014) definem interoperabilidade pragmática como sendo a capacidade de agregação e otimização de diversos processos de negócios para alcançar as finalidades pretendidas dos diferentes sistemas de informação. Nesse sentido, o trabalho ora apresentado não aborda questões de interoperabilidade pragmática pelo fato de seu escopo ser a integração dos dados do Portal de Transparência, não tendo influência no resultado obtido por cada um dos sistemas que fornecem os dados a serem integrados.

3.3. Integração de esquemas dos dados do Portal da Transparência

Seguindo os quatro passos do processo de integração de dados sugeridos em [Batini *et al.* 1986] e [Mello 2002], a solução proposta pode ser descrita da seguinte forma:

- Pré Integração.

  Conforme citado na subseção 2.3, nessa fase define-se a estratégia de integração.

  Os dados apresentados no portal têm uma particularidade, todos os assuntos tratados possuem algum tipo de relacionamento com o assunto SIAFI, pois todos os pagamentos feitos pelo governo federal são executados por esse sistema.

Sendo assim, optou-se por começar o processo de integração pelos dados oriundos do sistema SIAFI.

Como citado na subseção anterior, a integração de dados é feita pela utilização de um *data warehouse*. Sendo assim, implementou-se primeiro as tabelas de dimensões e de fatos[5] que se originavam dos dados do SIAFI.

- Comparação de esquema

À medida que novos assuntos são incorporados, os conceitos que já estavam definidos no DW (através de dimensões) não são reimplementados, porém cada novo conceito dá origem a uma nova dimensão.

- Conformação de esquemas

Nas situações em que o conceito que já está implementado no esquema integrador possuir menos detalhes do que o que estava sendo recebido na nova base que está sendo integrada, faz-se a atualização do esquema de forma que o esquema mais detalhado passe a ser o utilizado por todos os outros assuntos.

- Junção e reestruturação

Nessa fase são analisados os critérios de qualidade definidos em [Batini *et al.* 1986] e [Dang e Feldmann 2010]: Completude, Minimalidade e Entendibilidade. Com relação à completude, é verificado se há alguma informação que é recebida do sistema origem e não está sendo representada no repositório integrador. Quanto ao critério de minimalidade, verifica-se se há situações em que o mesmo conceito está sendo representado mais de uma vez. Porém, por questões de desempenho, algumas informações redundantes são mantidas no repositório central, mas essa redundância é devidamente controlada. Apesar da entendibilidade ser um critério subjetivo, toda a integração é feita com o intuito de se prover um entendimento mais fácil do dado. Inclusive, são inseridas informações não presentes nos sistemas de origem, a fim de tornar o entendimento mais fácil. Por exemplo, os empenhos do SIAFI têm um atributo denominado Ação, que descreve a ação de governo referente àquele gasto, porém essa ação é apresentada em termos muito técnicos. Então, durante o processo de integração incorpora-se mais um atributo denominado linguagem cidadã (que não existe nos sistemas originais) a fim de tornar o dado mais claro. Dessa forma, a ação de governo 8442, denominada Programa de Transferência de Renda Diretamente às famílias em condições de Pobreza e Extrema Pobreza [lei nº 10.836, de 2004], é também representada pela linguagem cidadã nomeada por “Programa Bolsa Família”.

3.4. Conflitos de Integração dos dados do Portal da Transparência

Dado o tamanho e diversidade do processo de integração proposto, muitos dos conflitos de integração referenciados em [Batini *et al.* 1986] foram encontrados. Nessa subseção serão demonstrados alguns exemplos desses conflitos, bem como os tratamentos aplicados.

3.4.1 Conflito de nome nos dados governamentais

[5] Tabelas de dimensões e de fatos são utilizadas na modelagem dimensional, que é uma técnica de modelagem de bancos de dados para responder consultas em um data warehouse. Tabelas de dimensão se ligam a uma tabela fato central formando um modelo estrela.

- Homonímia

Um caso de homonímia acontece com o conceito de órgão representado nos sistemas SIAFI e SIAPE[6]. Ambos os sistemas têm uma tabela chamada órgão cuja chave principal é um código numérico de 5 dígitos.

Apesar das relações apresentarem estruturas semelhantes e de alguns órgãos possuírem correspondências diretas entre os conceitos representados nos dois sistemas, existem situações em que um mesmo órgão do SIAFI está sendo representado por mais de uma tupla na tabela de órgão do SIAPE e vice-versa, conforme pode ser observado na Figura 3.

Cod Órgão Siafi	Cod Órgão Siape	Nome Orgão
22000	13000	MINIST. DA AGRICUL.,PECUARIA E ABASTECIMENTO
22802	13000	INSTITUTO NACIONAL DE METEOROLOGIA/MAPA
22803	13000	SECRETARIA NAC. DE DEFESA AGROPECUARIA/MAPA
22804	13000	SUBSECRETARIA DE PLANEJ.,ORCAM.E ADM./MAPA

(a) Mesmo Órgão no SIAPE sendo representado por Órgãos diferentes no SIAFI

Cod Órgão Siafi	Cod Órgão Siape	Nome Orgão
20101	11000	PRESIDENCIA DA REPUBLICA
20101	20101	PRESIDENCIA DA REPUBLICA

(b) Mesmo Órgão no SIAFI sendo representado por Órgãos diferentes no SIAPE

Cod Órgão Siafi	Cod Órgão Siape	Nome Orgão
22211	22200	COMPANHIA NACIONAL DE ABASTECIMENTO

(c) Órgão no SIAFI sendo representado por outro código no SIAPE

Figura 3: Comparação Código Órgão (SIAFI X SIAPE)

Analisando a situação, verificou-se que, embora houvesse algumas semelhanças, os conceitos representados entre os dois sistemas eram diferentes. Essa é uma diferença sutil, e há a necessidade de uma interpretação semântica para identificá-la. A Figura 3 demonstra algumas ocorrências que evidenciam que, apesar das semelhanças existentes entre os conceitos de órgão representado nos dois sistemas, essas informações não podem ser integradas.

Também foi identificado que os demais sistemas, diferentes do SIAPE e SIAFI, quando faziam referência ao conceito de órgão, utilizavam o conceito adotado pelo sistema SIAFI.

Como solução, optou-se pela utilização de duas estruturas distintas no esquema integrador: uma para representar os dados da tabela de órgão do SIAFI, que recebeu o nome de DimOrgao, e outra para armazenar os dados oriundos do sistema SIAPE, denominada DimOrgaoSiape.

[6] Sistema Integrado de Administração de Recursos Humanos (SIAPE): sistema de abrangência nacional destinado à gestão do pessoal civil do Poder Executivo Federal [Feijó 2006].

Dessa forma, os registros que fazem referência ao conceito de órgão e são oriundos do sistema SIAPE, são ligados à tabela DimOrgaoSiape. Já os registros que fazem referência a órgão e são oriundos de outros sistemas (diferentes do SIAPE) são associados à tabela DimOrgao no esquema integrador.

- Sinonímia

A sinonímia ocorre quando se tem mais de um nome referenciando um mesmo conceito. Esse tipo de conflito ocorre com o conceito de Pessoa Jurídica. Na base de dados dos Cartões de Pagamento do Governo Federal usa-se o termo Estabelecimento para designar esse conceito. Já no Cadastro de Entidades Privadas Sem Fins Lucrativos Impedidas (CEPIM)[7] utiliza-se o nome Entidade privada. No Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS)[8] utiliza-se o termo Empresa e em outras bases também são encontrados diferentes nomes para esse mesmo conceito.

Como solução optou-se por utilizar um nome padronizado. O sistema integrador passou a utilizar uma tabela única para esse conceito, cujo nome escolhido foi Pessoa Jurídica (DimPessoaJuridica).

3.4.2 Conflito de estrutura nos dados governamentais

- Conflito de tipo

Um caso de conflito de tipo ocorre com o conceito de convênio. Os convênios firmados pelo poder executivo federal devem ser registrados no sistema SICONV[9]. Porém, esses convênios são pagos através de ordens bancárias[10] emitidas no sistema SIAFI. O convênio é referenciado apenas como um atributo da tabela de ordens bancárias do SIAFI, ou seja, no sistema SIAFI, o convênio é representado apenas pelo seu número, não sendo considerado nenhum outro atributo.

Já no sistema SICONV, o convênio tem uma série de outros atributos, tais como: data de celebração, vigência, valor inicial, objeto, etc. O SICONV trata um convênio como uma entidade. Essas diferenças trouxeram dificuldades na geração do esquema de conciliação.

Para solucionar esse problema, passou-se a dotar a tabela mais detalhada como referência (oriunda do sistema SICONV). Como a estratégia de integração adotada foi a criação de um *data warehouse*, optou-se pela utilização de uma dimensão denominada

---

[7] Cadastro de Entidades Privadas Sem Fins Lucrativos Impedidas (CEPIM): relação de entidades privadas sem fins lucrativos que estão impedidas de celebrar convênios, contratos de repasse ou termos de parceria com a Administração Pública Federal [Controladoria Geral da União 2004].

[8] Cadastro Nacional de Empresas Inidôneas e Suspensas (CEIS): relação das empresas que sofreram sanções que impliquem restrição ao direito de participar em licitações ou de celebrar contratos com a Administração Pública, nos três Poderes e em todas as esferas federativas [Controladoria Geral da União 2004].

[9] Sistema de Gestão de Convênios (SICONV): desenvolvido para permitir o registro de contratos de execução firmados pelo órgão que celebra o convênio, com valores superiores a R$ 450.000,00, e para atender a determinações de dispositivos legais (Parágrafo 2º do Artigo 116 da Lei nº 8.666/93 e Artigo 2º da Lei nº 9.452/97) [Controladoria Geral da União 2004].

[10] Ordem Bancária – OB: principal documento de saque à conta única, cumprindo funções assemelhadas às dos cheques de movimentação das contas correntes bancárias [Feijó 2006].

DimConvenio com os registros oriundos do SICONV. Essa dimensão passou a ser utilizada como referência para ambos os sistemas.

Porém, como a atualização dos dados do SIAFI é diária e a dos dados do SICONV é mensal, em algumas situações, convênios referenciados pelo SIAFI ainda não estão carregados na dimensão de integração de convênios. Quando ocorre esse tipo de situação, adota-se a estratégia de se inserir no esquema de convênios um novo registro apenas com o número do convênio (oriundo do sistema SIAFI), e quando é feita a atualização das informações com dados do SICONV, atualiza-se esses convênios com os dados complementares.

- Conflito de chave

Um exemplo de conflito de chave ocorre para o conceito de município. Por exemplo, a tabela do sistema SIAFI referência o município beneficiado com um determinado recurso com um código de 4 dígitos, uma vez que, o sistema SIAFI utiliza uma representação própria de municípios e essa relação possui como chave principal um código próprio, adotado apenas por esse sistema. Já o sistema que trata de programas sociais, como por exemplo o Bolsa Família, referência o município do beneficiário do programa como um código de 7 dígitos, uma vez que é utilizado o código de município adotado pelo IBGE.

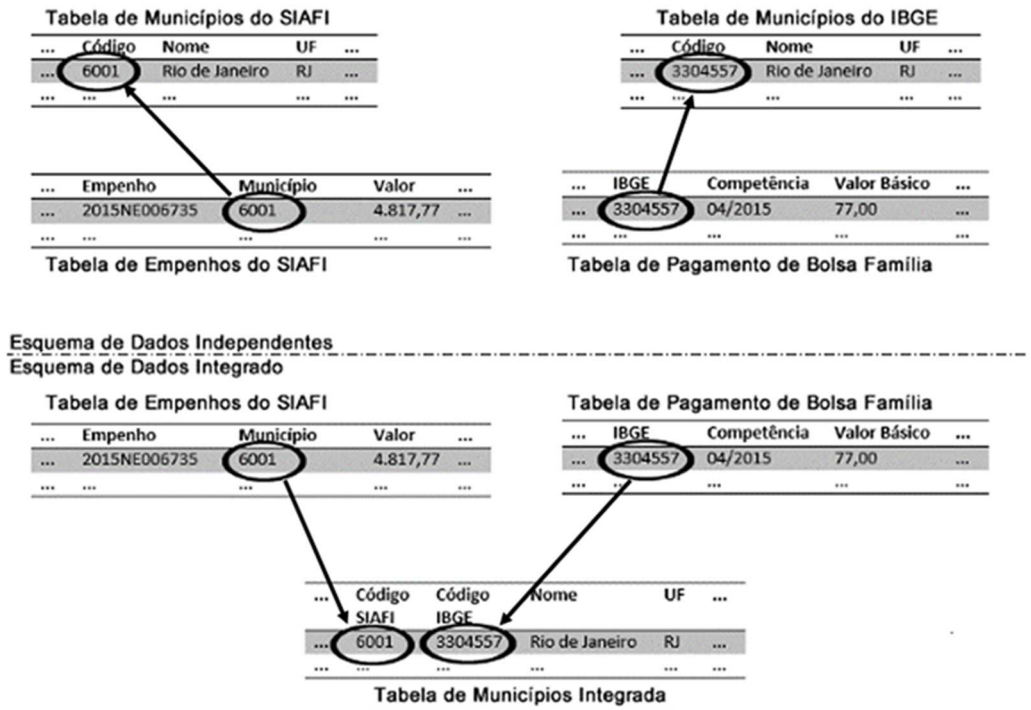

Figura 3: Tratamento do conflito de tipo

Esse tipo de conflito foi tratado com a inserção dos dois códigos de município (do SIAFI e do IBGE) na tabela do esquema integrador, conforme pode ser observado na Figura 3. Dessa forma, quando algum registro oriundo do SIAFI faz referência a um município, utiliza-se o código do SIAFI, porém, quando a referência é feita por algum registro oriundo do sistema de programas sociais, utiliza-se o código do IBGE.

3.5. Integração semântica dos dados do Portal da Transparência

Como Ziegler e Dittrich (2007) afirmam, a questão semântica é a parte mais difícil da tarefa de integração, pois ela engloba percepções mais profundas do que simplesmente a configuração dos dados.

Como visto na subseção 2.5, a integração semântica pressupõe um perfeito entendimento das regras de negócio e do contexto em que os dados estão inseridos.

No caso dos dados do Portal da Transparência, como eles englobam praticamente todas as áreas do Poder Executivo Federal, é necessário se entender o funcionamento de toda a Administração Pública, bem como conhecer os relacionamentos, explícitos e implícitos, existentes entre cada um dos assuntos.

Cabe ressaltar que, na maioria das vezes, tais relacionamentos não são tratados pelos sistemas estruturantes utilizados pelo governo.

Para transpor essa dificuldade, fazem-se necessários esforços que envolvam pessoas com conhecimentos das diversas áreas. Sendo assim, foram criados grupos específicos que ficaram responsáveis por estudar determinadas áreas e propor formas de relacionar os dados presentes nos diversos sistemas.

Um exemplo de estudo para a avaliação semântica dos dados ocorreu com o caso do SIAFI. Atualmente, recebe-se dados de pagamento oriundos do SIAFI de diferentes periodicidades, alguns arquivos de dados com atualização diária e outros mensal. Os dados que chegam diariamente representam o pagamento na forma mais granular possível, enquanto que os dados mensais trazem algumas visões consolidadas específicas.

Com o objetivo de manter a qualidade da minimalidade, definidas em [Dang e Feldmann 2010] e [Batini *et al.* 1986], foram montados grupos de estudo responsáveis por estudar o assunto em questão, de forma aprofundada, a fim de se obter as informações que são recebidas mensalmente a partir dos dados que são recebidos diariamente.

Os grupos de estudo, responsáveis pelos diversos temas, consultaram especialistas de outros órgãos, a legislação e literatura pertinentes e os próprios dados em si, a fim de encontrarem regras que depois de homologadas seriam repassadas para a equipe de modelagem do DW-CGU.

Após os estudos, algumas das extrações mensais foram suprimidas e passaram a ser obtidas pela aplicação de regras específicas nos dados diários, como por exemplo a visão de pagamentos por programa de governo. No entanto, outras informações, como por exemplo o pagamento do Fundo de Participação dos Municípios (FMP)[11], não puderam ser suprimidas.

No caso do pagamento do FPM, o grupo de estudo levantou que esse tipo de pagamento era feito por uma ordem bancária (OB) única para o Banco do Brasil, e que essa ordem bancária, com o valor global destinado a todos os municípios, que era

[11] O Fundo de Participação dos Municípios é uma transferência constitucional (CF, Art. 159, I, b), da União para os Estados e o Distrito Federal, composto de 22,5% da arrecadação do Imposto de Renda (IR) e do Imposto sobre Produtos Industrializados (IPI).

recebida nas extrações diárias. Posteriormente, o Banco do Brasil, de acordo com percentuais calculados pelo Tribunal de Contas União[12], fazia a distribuição dos percentuais corretos para os diversos municípios, repartindo uma parcela do montante da OB com o valor global para cada um dos municípios beneficiados. Ou seja, se fossem usados apenas os dados diários, o detalhamento de quanto cada município recebeu seria perdido. Esse exemplo, ilustra a necessidade de um aprofundamento na semântica dos dados para se fazer o trabalho de integração e evidencia que as questões tecnológicas não são os únicos dificultadores no desafio de integrar bases de diferentes fontes.

Adicionalmente, esse grupo interdisciplinar se reúne periodicamente para tratar de questões relacionadas aos dados do portal da transparência. Dentre tais questões abordam-se regras de negócio referentes aos diversos assuntos e a possibilidade de inserção de novas fontes de dados ao projeto, sendo que cada nova inserção pressupõe a sua integração com os dados que já fazem parte do conjunto de dados do Portal da Transparência, e consequentemente o estudo da semântica envolvida nesses diferentes conjuntos de dados.

4. Resultados

A arquitetura proposta já controla o processo de integração de uma parte dos dados tratados pelo Portal da Transparência. Essa primeira versão da implementação contempla os dados dos assuntos de pessoas físicas e Jurídicas, os extratos de cartões de pagamento do Banco do Brasil, as informações referentes a servidores públicos e imóveis funcionais, além de todos os arquivos oriundos do SIAFI. Dessa forma, de acordo com a figura 1, ainda restam ser implementados os assuntos de licitações e contratos, diárias e passagens, anistiados políticos, seguro defeso, além dos dados oque vêm da corregedoria e de empresas estatais. Sendo assim, toda a arquitetura e modelagem dos dados que atualmente fazem parte do escopo do projeto já estão prontas, porém os processos de carga e integração de alguns assuntos ainda estão sendo desenvolvidos. No entanto, como a implementação e implantação do projeto ocorre por fases iterativas e incrementais, alguns resultados práticos já podem ser colhidos e outros resultados potenciais podem ser projetados.

O foco desse artigo é a arquitetura e metodologias empregadas no processo de integração dos dados do Portal da Transparência do Governo Federal, e não nos dados em si, visto que, conforme citado anteriormente, o portal é dinâmico e, constantemente, novas bases de dados são inseridas nesse escopo. Essa seção apresenta alguns resultados obtidos a partir dos dados integrados pela aplicação da arquitetura proposta, a fim de demonstrar a utilidade do trabalho desenvolvido.

O principal benefício atingido com a aplicação da solução proposta foi a possibilidade de se oferecer uma visão integrada dos diversos sistemas coorporativos do Governo Federal brasileiro.

A proposta apresentada continua focada nos diversos assuntos que são tratados de forma individualizada por cada um dos sistemas coorporativos, porém, consegue

[12] Essa regra está prevista no artigo 161 da Constituição Federal de 1988 e no artigo 92 do Código Tributário Nacional (Lei 1572 de 25/10/1966).

integrar esses assuntos através de informações comuns a diferentes sistemas. Sendo assim, cada um dos assuntos tratados (retângulos brancos da Figura 1) são modelados como uma tabela "Fato", e todas as informações que dizem respeito a esses assuntos compõem as dimensões do modelo. As dimensões se relacionam às suas respectivas tabelas fato, sendo que as dimensões comuns a mais de um assunto se relacionam a mais de uma tabela fato e são responsáveis por integrar esses assuntos.

Serão apresentados alguns recortes do modelo de dados obtido a fim de ilustrar alguns resultados possíveis, visto que o total potencial desse banco de dados integrado depende da criatividade do usuário que estiver consultando-o, uma vez que as possibilidades de cruzamento e recuperação de informações são muitas.

A Figura 4 apresenta uma modelagem simplificada de quatro assuntos tratados no DW-CGU. O modelo não apresenta todas as dimensões, nem os atributos das tabelas por uma questão de legibilidade. Sendo assim, apenas algumas dimensões são inseridas a fim de demonstrar uma possível aplicabilidade da solução proposta.

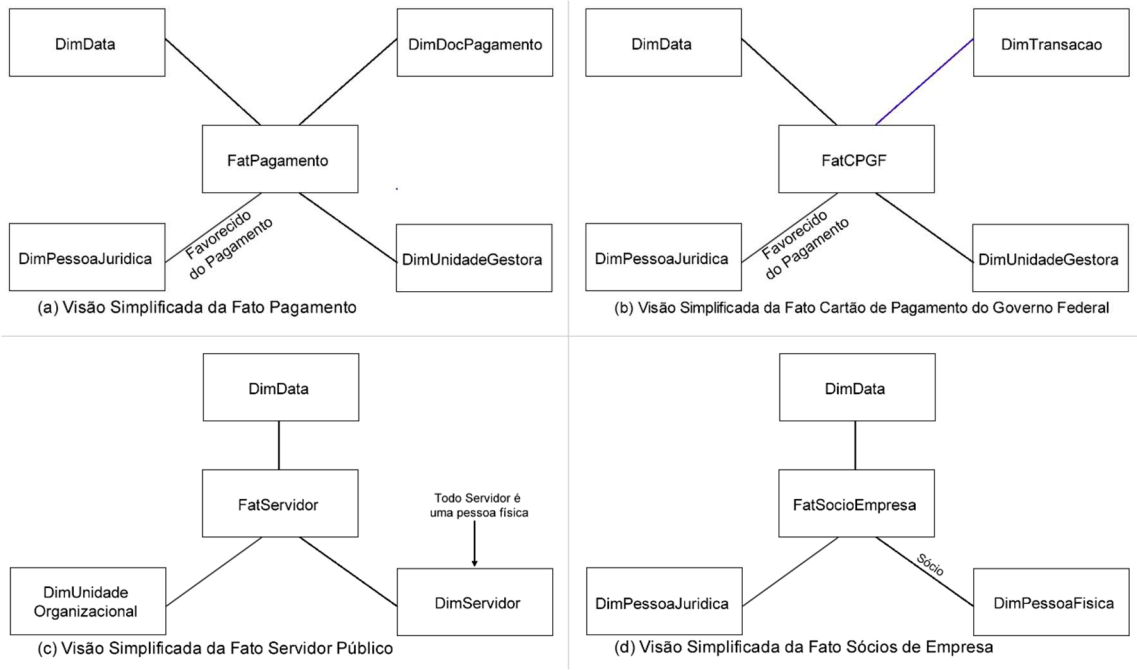

Figura 4: Modelagem Simplificada de 4 Tabelas Fato do DW-CGU

Através dos relacionamentos entre o assunto Pagamento (oriundo do SIAFI) e Cartões de Pagamento do Governo Federal – CPGF (oriundo de extratos gerados pelo Banco do Brasil), foi possível descobrir que uma mesma empresa recebeu, no ano de 2013, pagamentos de uma mesma unidade gestora pelos dois sistemas, pelo SIAFI (através do processo de execução do orçamento anual da União) e pelo CPGF (por compras diretas via Cartão de Crédito Coorporativo). Esse levantamento foi possível devido ao vínculo criado pelas dimensões Pessoa Jurídica (DimPessoaJuridica) e Unidade Gestora (DimUnidadeGestora), que agrega as informações oriundas desses dois sistemas (partes a e b da Figura 4). A Tabela 1 apresenta os valores obtidos.

Tabela 1: Valores pagos por uma unidade gestora a uma mesma empresa pelo SIAFI e pelo CPGF no ano de 2013

Valor Oriundo do SIAFI (2013)	Valor Oriundo do CPGF (2013)
R$ 2.461.393,84	R$ 7.243,03

Pelos relacionamentos com o assunto Pessoa Jurídica, oriundo da base da Receita Federal (parte c da Figura 4), também foi possível verificar os sócios dessa empresa, e com a utilização dos dados oriundos do SIAPE, através do relacionamento com a DimPessoaFisica (parte d da Figura 4), levantou-se que 12 dentre os 15 sócios da referida empresa eram servidores lotados em uma unidade organizacional vinculada ao mesmo órgão relacionado à unidade gestora que executou os pagamentos. As identificações da empresa e do órgão em questão foram omitidas pelo fato de tais questões estarem fora do escopo deste trabalho. O exemplo ora apresentado foi feito com o único intuito de demonstrar um dos possíveis resultados que podem ser obtidos a partir da integração das bases de dados que compõem os diferentes sistemas coorporativos do governo.

5. Validação

Esta seção descreve o processo de validação da arquitetura proposta. A validação utiliza-se de estratégias tanto qualitativas quanto quantitativas.

É importante ressaltar que o processo de validação da solução proposta confunde-se com o próprio processo de validação dos dados a serem apresentados no Portal da Transparência, pois os erros identificados podem ter sido causados por falhas na implementação das regras de negócio do integrador ou por falhas nos próprios dados recebidos.

A validação se dá pelo auxílio dos fluxos de dados 3 (coleta de informações dos processos de carga) e 4 (fluxo para validação de especialistas do negócio), apresentados na Figura 2 da seção 3, sendo que o fluxo de dados 3 auxilia na validação qualitativa, enquanto que o fluxo de dados 4 é utilizado na validação quantitativa.

A validação de cada um dos assuntos tratados começa a partir do momento em que a primeira versão do processo de carga do assunto em questão fica pronta e estende-se indefinidamente, uma vez que também é responsável por garantir a qualidade dos dados recebidos, e esses sempre estão sujeitos a erros. Observou-se que, à medida que os processos de carga vão sendo amadurecidos, a quantidade de falhas diminui consideravelmente.

Esse processo de validação pode ser dividido em 2 momentos distintos: um que ocorre durante a implantação da solução propostas para o assunto em questão, e outro, que ocorre periodicamente durante os processos de carga. Essa periodicidade varia de acordo com os assuntos, uma vez que alguns assuntos são carregados diariamente, outros semanalmente, e outros mensalmente.

5.1. Validação na Implantação

A validação realizada durante o processo de implantação dos assuntos é composta de 2 fases: na primeira fase é realizada uma validação quantitativa, enquanto que na segunda há uma validação qualitativa dos dados carregados.

A validação quantitativa consiste da verificação de percentuais de ocorrência de valores atípicos. Por exemplo, quando um registro a ser inserido em uma tabela fato não tem um valor correspondente em uma das dimensões em que essa tabela fato está ligada, o campo que faz essa ligação com a dimensão em questão (chave estrangeira[13]) recebe um dos seguintes valores: -3, -2, ou -1 (que correspondem às descrições "inconsistente", "não se aplica" e "sem informação", respectivamente), de acordo com sua situação.

Por exemplo, o campo skEmpenho da tabela Fato do assunto pagamento (FatPagamento) é responsável por fazer a ligação entre o empenho referenciado por um determinado pagamento com os dados desse empenho na tabela da dimensão de empenho (DimEmpenho). Caso, durante o processo de carga, o empenho referenciado no pagamento não seja um empenho válido e consequentemente não esteja presente na DimEmpenho, o campo skEmpenho recebe o valor -3 (inconsistente), caso o pagamento em questão não faça referência a um empenho (não tenha o campo de empenho preenchido), o campo skEmpenho recebe o valor -1 (sem informação).

Durante o processo de validação quantitativa da tabela FatEmpenho (que trata dos empenhos), verificou-se que o campo "skFuncionalProgramatica", que faz a ligação da tabela fatEmpenho com a DimFuncionalProgramatica[14], apresentava uma percentagem de 50,78% dos seus registros com o valor -3 (inconsistente), o que caracteriza uma atipicidade.

Outra inconsistência levantada, foi o fato de existirem pagamentos inseridos na tabela fatPagamento, cujo valor referente a chave "skEmpenho" (que faz a ligação com a dimEmpenho) era -3 (inconsistente). Apesar de apenas 0,01% dos registros apresentarem esse valor, o que não seria estatisticamente representativo, isso caracteriza uma incoerência, pois, de acordo com as regras de contabilidade pública, não é possível que um pagamento não esteja associado a um empenho.

A validação estatística dá origem a um relatório que transcreve essas e outras inconsistências encontradas. Esse relatório é passado para a equipe de desenvolvimento do DW-CGU, que é responsável por verificar e corrigir as inconsistências, além de gerar uma nova carga dos dados do assunto em questão.

Com relação aos 2 exemplos citados acima, foi verificado que havia um erro na regra da ligação entre o empenho e sua respectiva classificação funcional programática e, após a correção desse erro, o índice de valores inconsistentes caiu de 50,78% para 0,001%. Também foi verificado que havia a necessidade de uma carga inicial dos empenhos antigos (anteriores a data inicial de carga do DW, que é 2013), pois esses empenhos pagos se referiam a empenhos de restos a pagar[15] e por isso não constavam nas extrações recebidas regularmente. Após essa recarga, todos os pagamentos passaram a ter um empenho correspondente na tabela DimEmpenho. Já a validação qualitativa se

---

[13] Chave estrangeira: campo que estabelece o relacionamento entre duas tabelas. Esse campo impõe a ligação entre os dados de duas tabelas, de forma que um nenhum valor possa ser inserido na coluna de chave estrangeira, sem que haja um valor correspondente na tabela referenciada.

[14] Classificação Funcional Programática: Agrupamento das ações do governo nas grandes áreas de sua atuação, para fins de planejamento, programação e orçamentação.

[15] Restos a Pagar: Despesas empenhadas, mas não pagas, até o final do exercício financeiro.

dá através do auxílio de uma ferramenta OLAP[16]. Essa ferramenta permite que pessoas sem conhecimentos específicos da área de banco de dados consigam interagir com as informações carregadas no *data warehouse*.

A validação qualitativa é executada pelo Grupo Operacional do Portal da Transparência. Essa equipe é formada por 7 especialistas com formações diversas e com conhecimento das diversas áreas da Administração Pública tratadas no Portal da Transparência.

Nessa fase qualitativa da análise dos dados, os especialistas comparam os valores obtidos diretamente pelos processos de carga com as mesmas informações presentes nos sistemas coorporativos, a fim de verificar se há alguma divergência nessas informações. Nessa fase, os especialistas também fazem agregações dos dados para verificar se os resultados obtidos pela consolidação desses dados fazem sentido, utilizando para isso o conhecimento das regras de negócio de cada um dos assuntos tratados. Por exemplo, de acordo com a legislação brasileira, o processo da execução da despesa pública é composto de 3 fases: empenho, liquidação e pagamento. Essas fases devem ser sequenciais e nenhuma fase pode ocorrer sem que a fase anterior já tenha acontecido. Sendo assim, durante a agregação dos dados, se algum especialista do negócio constatar que o valor pago referente ao orçamento de um determinado ano é superior ao valor empenhado ou liquidado, isso quer dizer que algo de errado está acontecendo com esse conjunto de dados.

5.2. Validação durante a Operação

O processo de validação que ocorre durante a operação do sistema utiliza-se de informações estatísticas levantadas durante o próprio processo de carga e é realizado de forma automatizada.

Esse processo faz uso de algumas métricas que são calculadas durante a carga dos dados, e compara esses valores com a média dos dados históricos dos demais processos de carga. Alguns dos exemplos das métricas utilizadas como fonte de comparação no processo de carga do SIAFI são listadas abaixo:

- Quantidade de empenhos carregados;
- Quantidade de liquidações carregadas;
- Quantidade de pagamentos carregados;
- Tempo médio de carga.

Essa comparação com os dados históricos permitiu a detecção de uma grande redução dos registros de liquidações a partir de uma determinada data, possibilitando assim a identificação de um problema relacionado às liquidações. Após alguns estudos, verificou-se que esse problema estava ocorrendo em razão da alteração de algumas regras de negócio que não tinham sido replicadas no processo de extração dessas informações no sistema fornecedor desses dados.

[16] OLAP - On-Line Analytical Processing: as Ferramentas de OLAP fornecem uma interface com o usuário que permite a analise e visualização dos dados corporativos de forma rápida, consistente e interativa.

5.3. Rastreabilidade dos dados

Outra atividade de extrema importância durante a validação é a manutenção da rastreabilidade de todos os registros inseridos no repositório central de integração. Porque, tão importante quanto identificar um determinado erro é saber todo o caminho que essa informação defeituosa percorreu até ser carregada no seu destino final, pois só assim será possível se identificar em qual parte do processo esse erro foi ocasionado.

Existem centenas de processos de carga e modelo de arquivos de dados utilizados durante todo o processo de integração, sendo que um mesmo modelo pode dar origem a vários arquivos diferentes (por exemplo, cada um dos modelos de arquivos oriundos do SIAFI, que tem atualização diária dá origem a um arquivo distinto a cada dia), logo, para saber quais foram os processos de carga e arquivos utilizados na carga de um determinado registro, é necessário algum mecanismo de rastreabilidade.

A solução proposta implementa um mecanismo de controle, realizado pelo fluxo 4 da Figura 2, que prevê a manutenção dessa rastreabilidade. Os registros carregados no repositório central recebem um identificador que permite a verificação de quais processos foram utilizados no tratamento dessas informações e de quais fontes de dados foram utilizadas para a sua geração. Dessa forma, no momento da identificação de um registro falho, também se identifica todos os demais processos de carga e arquivos fontes envolvidos nessa operação defeituosa, bem como a data e hora do processamento. Esse mecanismo é especialmente útil para os casos de validação durante a operação, pois nessa situação é muito mais difícil de se descobrir a fonte do erro.

6. Conclusão

Esse trabalho apresentou uma solução de integração para os dados do Portal da Transparência do Governo Federal brasileiro. Essa solução, além de controlar todo o processo de carga dos dados, também implementa os tratamentos necessários à perfeita harmonização dos diferentes conjuntos de dados que alimentam o Portal da Transparência do Governo Federal.

Dessa forma, a principal contribuição desse artigo é a proposta de uma arquitetura capaz de gerenciar todo o processo de integração dos dados de um grande portal coorporativo. Essa arquitetura gerencia questões ligadas à integração em si, assim como os processos de carga e validação dos dados, mantendo todo o histórico de rastreabilidade dos dados carregado em um repositório de controle, o que permite a identificação dos processos de carga e arquivos fonte utilizados no carregamento de um registro defeituoso, facilitando assim a manutenibilidade dos dados.

Durante o desenvolvimento dos trabalhos ficaram evidentes as dificuldades de se unificar dados heterogêneos. Os conflitos de esquema e as diferenças semânticas apresentadas em [Batini *et al.* 1986] e [Ziegler e Dittrich 2007], respectivamente, se constituem os principais desafios da tarefa de proporcionar uma visão unificada de dados oriundos de diferentes sistemas de informação.

Concluiu-se que um projeto de integração de dados envolve muitas outras questões que vão além dos desafios tecnológicos, devendo-se considerar a interpretação semântica e o entendimento dos dados e a padronização das regras de negócio envolvidas nos diversos processos de trabalho; razão essa que impõe o comprometimento de todas as áreas de negócio envolvidas no projeto de unificação dos

dados. Esse fato pôde ser evidenciado pelo caso do pagamento do Fundo de Participação dos Municípios, citado na subseção 3.5, que trata da integração semântica dos dados do Portal da Transparência.

Em trabalhos futuros, pretende-se utilizar esses dados integrados para a aplicação de técnicas de mineração de dados a fim de adquirir novos conhecimentos sobre as atividades governamentais, gerando-se assim, novas informações a serem apresentadas no Portal da Transparência.

Referências

Agner, L. (2008). Governo eletrônico e transparência do Estado. Revista Webinsider, Brasília,

Alves Costa, T. and Salgado, C. B. (2005). O gerenciador de consultas de um sistema de integração de dados. Recife: UFPE, 2005. Dissertação (Mestrado em Informática) – Programa de Pós-graduação em Ciência da Computação, Universidade Federal d Pernambuco.

Araújo, R. M., Cappelli, C. A. and Leite, J. (2010). A importância de um Modelo de Estágios para avaliar Transparência. Revista TCMRJ, setembro, n. 45, p. 97.

Arfaoui, N. and Akaichi, J. (2015). Automating schema integration technique case study: generating data warehouse schema from data mart schemas. Beyond Databases, Architectures and Structures. Springer. p. 200–209.

Barbosa, A. C. (2001). Middleware para integração de dados heterogêneos baseado em composição de frameworks. Tese de Doutorado, PUC-Rio, Brazil.

Batini, C., Lenzerini, M. and Navathe, S. B. (1986). A comparative analysis of methodologies for database schema integration. ACM computing surveys (CSUR), v. 18, n. 4, p. 323–364.

Bellström, P. (2006). Bridging the gap between comparison and conforming the views in view integration. In 10th East-European Conference on Advances in Databases and Information Systems.

Beluzo, J. (2015). IDEO – Integrador de dados da Execução Orçamentária Brasileira: Um estudo de caso da integração de dados das receitas e despesas nas Esferas Federal, Estadual – Governo de São Paulo, e Municipal – Municípios do Estado de São Paulo [Dissertação]. São Paulo: Universidade de São Paulo,

Bernstein, P. A. and Melnik, S. (2004). Meta data management. Data Engineering, 2004. Proceedings. 20th International Conference on, 2004, pp. 875-875.

BRASIL ([S.d.]). Lei Complementa no 131 de 27 de maio de 2009. Disponibilização em tempo real de Informações. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 28 mai. 2009.

BRASIL ([S.d.]). Lei no 10.836, de 9 janeiro de 2004. Lei 10.836. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 12 jan. 2004

BRASIL ([S.d.]). Lei no 12.527, de 18 de novembro de 2011. Lei de Acesso à Informação. Diário Oficial [da] República Federativa do Brasil, Brasília, DF, 18 nov. 2011.

Controladoria Geral da União (2004). Portal da Transparência nos Recursos Públicos Federais. http://transparencia.gov.br/, [accessed on Apr 18].

Dang, L. H. and Feldmann, D.-I. M. (2010). A Guideline for the Conduction of Data Integration for Heterogeneous Information Systems. Ph. D. thesis. Technische Universität Dresden - Institute of Systems Architecture.

Doan, A., Halevy, A. and Ives, Z. (2012). Principles of data integration. Elsevier.

Feijó (2006). Curso de SIAFI: uma abordagem prática da execução orçamentária e financeira. Brasília, Editora Gestão Pública

Freire, J., Koop, D., Santos, E., and Silva, C. T. (2008). Provenance for computational tasks: a survey. Computing in Science and Engineering 10 (3), p. 11–21.

Halevy, A., Rajaraman, A. and Ordille, J. (2006). Data integration: the teenage years. In Proceedings of the 32nd international conference on Very large data bases. VLDB Endowment.

Hull, R. (1997). Managing semantic heterogeneity in databases: a theoretical prospective. In Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. ACM.

Jardim, J. M. (2004). A construção do e-gov no Brasil: configurações político-informacionais. Encontro Nacional da Ciência da Informação, v. 5.

Kimball, R. and Ross, M. (2011). The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons.

Lee, M. L. and Ling, T. W. (1995). Resolving structural conflicts in the integration of entity-relationship schemas. OOER'95: Object-Oriented and Entity-Relationship Modeling. Springer. p. 424–433.

Lenzerini, M. (2002). Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM.

Liu, S., Li, W. and Liu, K. (2014). Pragmatic oriented data interoperability for smart healthcare information systems. In Cluster, Cloud and Grid Computing (CCGrid), 2014 14th IEEE/ACM International Symposium on. . IEEE.

Meirelles, H. L. (2001). Direito administrativo brasileiro. Atualizada por Eurico de Andrade Azevedo, Délcio Balestero Aleixo e José Emmanuel Burle Filho. São Paulo: Malheiros, p. 08–2005.

Mello, R. dos S. (2002). Uma abordagem bottom-up para a integração semântica de esquemas XML. Porto Alegre: UFRGS, 2002. Tese (Doutorado em Informática) – Programa de Pós-Graduação em Computação, Universidade Federal do Rio Grande do Sul.

Naiman, C. F. and Ouksel, A. M. (1995). A classification of semantic conflicts in heterogeneous database systems. Journal of Organizational Computing and Electronic Commerce, v. 5, n. 2, p. 167–193.

Nazario, D. C., Silva, P. F. Da and Rover, A. J. (2012). Avaliação da qualidade da informação disponibilizada no Portal da Transparência do Governo Federal. Revista Democracia Digital e Governo Eletrônico, n. 6.

Ouksel, A. M. and Sheth, A. (1999). Semantic interoperability in global information systems. ACM Sigmod Record, v. 28, n. 1, p. 5–12.

Prado, O. and Loureiro, M. R. G. (2008). Governo eletrônico e transparência: avaliação da publicização das contas públicas das capitais brasileiras. Revista Alcance, v. 13, n. 3, p. 355–372.

Prado, O., Ribeiro, M. M. and Diniz, E. (2012). Governo eletrônico e transparência: olhar crítico sobre os portais do governo federal brasileiro. Estado, sociedade e interações digitais: expectativas democráticas, p. 13–39.

Rommel, Carvalho, De Paiva, E., Da Rocha, H. and Mendes, G. (2013). Methodology for Creating the Brazilian Government Reference Price Database. X Encontro Nacional de Inteligência Artificial e Computacional.

Rommel, Carvalho, De Paiva, E., Da Rocha, H. and Mendes, G. (2014). Using Clustering and Text Mining to Create a Reference Price Database. Learning and NonLinear Models, v. 12, p. 38–52.

Ziegler, P. and Dittrich, K. R. (2004). Three Decades of Data Intecration—all Problems Solved? Building the Information Society. Springer. p. 3–12.

Ziegler, P. and Dittrich, K. R. (2007). Data integration—problems, approaches, and perspectives. Conceptual Modelling in Information Systems Engineering. Springer. p. 39–58.