Mapeando Dados Governamentais com uma Ontologia de Organizações

Lucas B. R. da Fonseca[1]
lfonseca@inf.ufes.br
Carlos L. B. Azevedo[1,2]
clbazevedo@inf.ufes.br
João Paulo A. Almeida[1]
jpalmeida@ieee.org

[1]*Núcleo de Estudos em Modelagem Conceitual e Ontologias (NEMO), UFES, Vitória, ES*
[2]*Services, CyberSecurity and Safety Research Group (SCS), University of Twente, Holanda*

Resumo
Em 18 de novembro de 2011 foi sancionada a Lei de Acesso à Informação (Lei nº 12.527/2011) que regula o acesso às informações mantidas pelo governo nas esferas federal, estadual e municipal. Com essa lei em vigor, um maior volume de dados de caráter público passou a ser disponibilizado na Internet. Apesar da maior disponibilidade, os dados fornecidos pelas várias organizações governamentais possuem formatos diferentes e advêm de fontes heterogêneas e não integradas, o que dificulta a utilização destes pelos cidadãos e sua apropriação para reuso em sistemas computacionais. Uma estratégia recente recomendada por órgãos de padronização como o W3C prevê o uso de ontologias e tecnologias semânticas para integrar e disponibilizar esses dados. Este artigo relata um estudo de caso na integração de dados governamentais no domínio de publicação de informações sobre organizações e estruturas organizacionais utilizando a W3C ORG Ontology. No estudo, foram recuperados dados publicados em seu formato corrente, realizado um mapeamento destes dados para a ontologia de referência da W3C e publicados os dados integrados em RDF em conformidade com a ontologia. O artigo apresenta a abordagem adotada, seus benefícios e discute as dificuldades encontradas assim como as lições aprendidas em cada uma das fases do processo.

Palavras-chave: Web Semântica, Dados Ligados, Ontologias, Integração de dados.

Abstract
On November 18 2011, Brazil passed its Freedom of Information Law (n. 12.527/2011) regulating access to government information at all levels (federal, state and municipal). With the enactment of this law, public government data has been increasingly made available on the Internet. Despite its availability, data published by government organizations often adopts ad hoc formats and originates from heterogeneous sources with little or no integration, creating barriers for data consumption. In order to address this challenge, several standards bodies such as W3C have proposed the use of ontologies and semantic technologies. This paper reports on case study on data integration in the domain of organizational structures using the W3C ORG Ontology. We discuss the approach that has been employed along with its potential benefits. Lessons learned throughout the integration process are discussed.

Key Words: Semantic Web, Linked Data, Ontologies, Data integration.

Introdução

Em 18 de novembro de 2011 foi sancionada a Lei de Acesso à Informação (Lei nº 12.527/2011) que regula o acesso às informações mantidas pelo governo nas esferas federal, estadual e municipal. Essa lei constituiu um avanço para a democratização da informação pública, com um grande volume de dados governamentais passando a ser disponibilizado na Internet. O objetivo é proporcionar ao cidadão maior visibilidade às ações de governo, melhor acesso aos serviços públicos e maior controle das contas públicas através de mecanismos de transparência.

Exemplos de iniciativas federais em resposta às demandas de acesso à informação pública incluem o Portal de Dados Abertos (disponível em http://dados.gov.br) e o Portal da Transparência (disponível em http://www.portaltransparencia.gov.br). Atualmente, os dados que são publicados por estes portais possuem formatos diferentes e advêm de fontes heterogêneas e não integradas, o que dificulta a utilização destes pelos cidadãos e sua apropriação para reuso em sistemas computacionais. A publicação dos dados é usualmente feita através de arquivos não estruturados (tais como planilhas e arquivos PDF), dificultando a leitura, integração e análise automatizadas.

Para lidar com as limitações das abordagens *ad hoc* para publicação de dados, o W3C (*World Wide Web Consortium*) vem recomendando um conjunto de boas práticas para publicar dados de forma estruturada e interligada (BIZER; HEATH; BERNERS-LEE, 2009). Esta abordagem preconiza o uso de vocabulários compartilhados que são representados através de *ontologias* formalizadas em linguagens de representação de conhecimento concebidas para a Web e objetivam o melhor compartilhamento, consumo e integração de dados.

Este artigo relata um estudo de caso empregando a abordagem de integração baseada em ontologias utilizando bases de dados governamentais. Foram mapeados os diversos dados em seus formatos correntes para uma ontologia de organizações padronizada pelo W3C (denominada *ORG Ontology* (REYNOLDS, 2014)). A *ORG Ontology* foi criada com o intuito de prover um modelo genérico para publicação de informações sobre organizações e estruturas organizacionais, incluindo organizações governamentais. A *ORG Ontology* permite descrever a estrutura de organizações, bem como as pessoas envolvidas, informações sobre a localização das organizações e seus históricos (como fusões e mudanças de nomes) (REYNOLDS, 2014).

Em especial, busca-se nesse trabalho: (i) avaliar as dificuldades existentes para integrar os dados abertos disponibilizados pelo governo brasileiro, incluindo prováveis dificuldades de interpretação desses dados, com o uso de uma ontologia recomendada pelo W3C; (ii) avaliar a abrangência e adequação da ontologia aos dados governamentais brasileiros e; (iii) considerar os benefícios e limitações do uso da abordagem em um caso real.

Este artigo está organizado da seguinte forma. A seção 1 aborda o referencial teórico relevante ao contexto deste trabalho. A seção 2 apresenta a abordagem para realização do mapeamento, desde da etapa conceitual até a sua implementação. A seção 3 apresenta uma aplicação que usa os resultados do mapeamento visando demonstrar o consumo dos dados em conformidade com a *ORG Ontology*. A seção 4 apresenta as discussões acerca do mapeamento, mostrando as dificuldades encontradas e os benefícios das tecnologia semânticas empregadas. Por fim, a última seção apresenta as conclusões do trabalho, limitações e propostas de trabalhos futuros.

1 Referencial Teórico

O mapeamento apresentado neste trabalho foi realizado com base em padrões do W3C, e envolve diretamente linguagens como RDF (*Resource Description Framework*) e SPARQL (*SPARQL Protocol and RDF Query Language*) dentro de contextos como a Web Semântica, Dados Ligados (*Linked Data*) e Ontologias. Essa seção apresenta estes elementos para contextualizar o trabalho.

1.1 Dados Ligados e Ontologia

Berners-Lee (2006) define um conjunto de regras para a publicação de dados na Web, chamados de Dados Ligados (*Linked Data*). A intenção é que dados publicados seguindo esse conjunto de regras sejam unificados em único espaço global de dados. As regras propostas são: (1) Usar URI como nome para as coisas; (2) Usar URIs HTTP para que as pessoas possam procurar por estes nomes; (3) Fornecer informações úteis na recuperação de URIs, usando os padrões RDF e SPARQL; (4) Incluir links para outros URIs, para que seja possível descobrir mais informações.

Estas regras fornecem os princípios básicos para a publicação e conexão de dados através da Web. Para propiciar distinção semântica entre os dados publicados, possibilitando integração desses dados com dados de outras fontes, utilizam-se ontologias.

As ontologias mais comuns na Web possuem uma taxonomia e um conjunto de regras de inferência (BERNERS-LEE; HENDLER; LASSILA, 2001). Taxonomias definem classes de objetos e as relações entre elas. As regras de inferência permitem melhorar a qualidade da análise dos dados, através da descoberta de novas relações automaticamente sobre o conteúdo existente, além de detectar possíveis inconsistência neles.

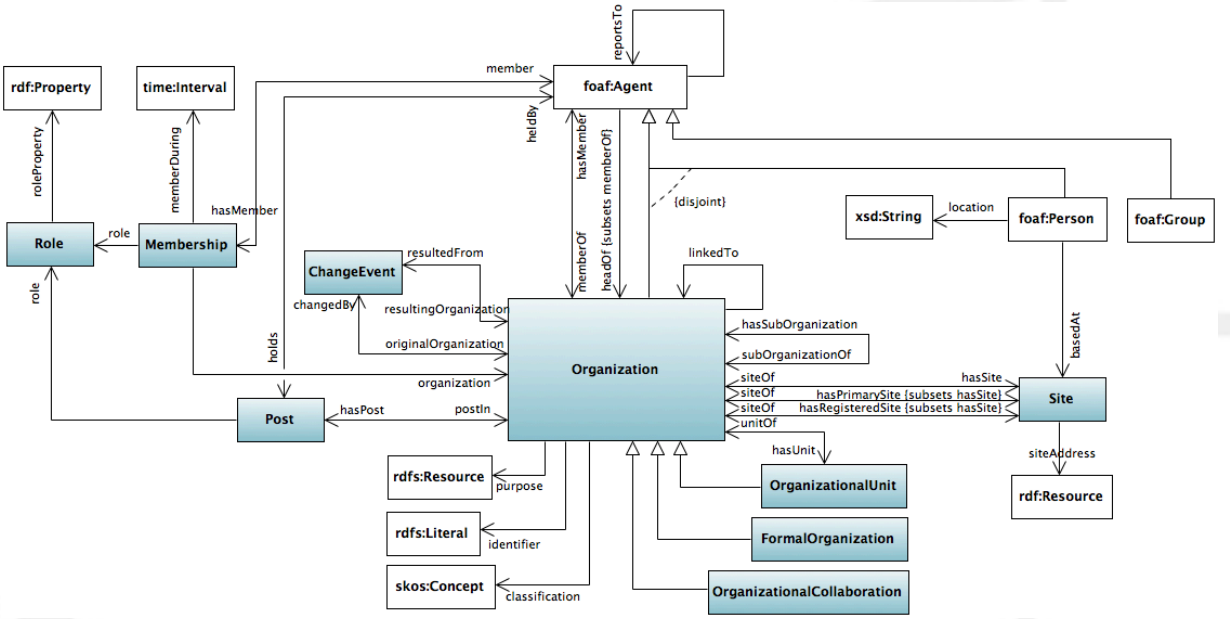
Para representar uma ontologia na Web, utiliza-se a Linguagem de Ontologia para Web (OWL - *Web Ontology Language*) que é uma extensão de RDF *Schema.* A linguagem OWL provê maior capacidade de descrição de classes e propriedades em relação ao RDF, incluindo a expressão de relações entre classes, restrições de cardinalidade, características de propriedades e classes enumeradas (MCGUINNESS; HARMELEN, 2004).

1.2 ORG Ontology

A *ORG Ontology* (*The Organizational Ontology*) é uma recomendação W3C, descrita em (REYNOLDS, 2014), projetada para permitir a publicação de informações sobre as organizações e estruturas organizacionais, incluindo organizações não-governamentais. A ontologia descreve a estrutura da organização, bem como as pessoas envolvidas nessa estrutura, informações sobre a localização da organização e o histórico da organização (como fusões e mudanças de nomes). A *ORG Ontology* possui um conjunto de conceitos bem estruturados e definidos, possibilitando que outros indivíduos entendam e os utilizem. Ela também reutiliza conceitos de outras ontologias usadas em larga escala (p. ex., FOAF, SKOS e *vCard*).

A *ORG Ontology* fornece um modelo genérico e reutilizável. Esse modelo pode ser estendido ou especializado para uso em situações particulares. A Figura 1 apresenta um diagrama (não normativo) da *ORG Ontology*.

Figura 1 – ORG Ontology



Fonte: (REYNOLDS, 2014)

Conforme a Figura 1, **Organização** (*Organization*) representa um conjunto de pessoas organizadas em uma comunidade ou outra estrutura social, comercial ou política. Uma **Organização** pode ser especializada em uma **Unidade Organizacional** (*OrganizationalUnit*), em uma **Organização Formal** (*FormalOrganization*) ou em uma **Colaboração Organizacional** (*OrganizationalCollaboration*). **Unidades Organizacionais** são entidades (como departamentos ou unidades de suporte) que pertencem a uma **Organização** e que não podem ser consideradas como uma entidade legal de direitos próprios. **Organizações Formais** são organizações reconhecidas mundialmente (como corporações, instituições de caridade, governo ou igreja) através de jurisdições legais, direitos e responsabilidades associadas. **Colaboração Organizacional** define um tipo de colaboração entre duas ou mais **Organizações** como um projeto. Ela atende aos critérios para ser uma organização na medida em que tem uma identidade e definição de propósito independente de seus membros em particular, mas não é nem uma entidade jurídica formalmente reconhecida nem uma unidade organizacional dentro de uma organização maior. Exemplos de **Colaboração Organizacional** são projetos conjuntos entre várias organizações ou consórcios de organizações, como os formados para licitações de grandes obras. **Adesão** (*Membership*) de uma **Organização** representa o **Papel** (*Role*) que o **Agente** (*Agent*) tem na **Organização** durante um **Intervalo** (*Interval*) de tempo. **Agente** detém de um **Cargo** (*Post*) e pode ser uma **Organização** (exceto a qual ele é membro), uma **Pessoa** (*Person*) ou um **Grupo** (*Group*). **Pessoa** e **Organização** têm **Endereço** (*Site*). **Organização** pode ser alterada por **Eventos de Mudança** (*ChangeEvent*) que representam eventos que resultaram em grandes alterações para uma **Organização** como uma fusão ou reestruturação completa. Além disso, a **Organização** possui um **identificador** (*identifier*), um **propósito** (*purpose*), uma **classificação** (*classification*), essa última podendo variar dentro de algum esquema de classificação, e um ou vários **Endereços** (REYNOLDS, 2014).

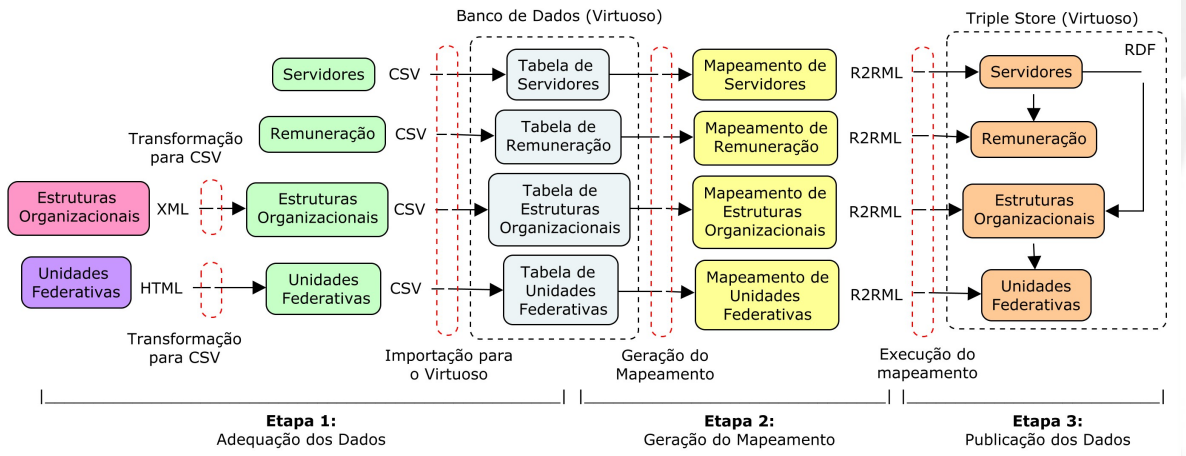
2 Mapeamento

Esta seção aborda as etapas para o mapeamento dos dados contidos nas bases governamentais selecionadas para dados estruturados, de acordo com a *ORG Ontology*. Para tal foi utilizada uma linguagem de mapeamento de dados relacionais para RDF, a linguagem R2RML (DAS; SUNDARA; CYGANIAK, 2012). A escolha das bases de dados baseou-se na relevância dos dados e em sua representatividade em relação a *ORG Ontology,* verificando se esses dados possuem interesse público, se podem ser mapeados aos termos do vocabulário da *ORG Ontology* e se, em conjunto, representam uma parte significativa de conceitos da ontologia. Foram escolhidos dados advindos de diversas fontes governamentais para verificar a integração existente entre os dados. Todas as bases de dados selecionadas pertencem a órgãos do governo federal. Com isso, busca-se verificar tanto a integração dos dados disponibilizados pelo governo quanto a abrangência da ontologia no escopo analisado.

Foram utilizadas as seguintes bases de dados: (i) **Servidores Civis (Servidores e Remuneração),** que apresenta dados em formato *Comma-Separated Values* (CSV) sobre cargo, função, situação funcional e remuneração dos servidores civis; (ii) **Estruturas Organizacionais,** que apresenta dados em formato XML contendo informações organizacionais do Poder Executivo Federal, (Administração Direta, Autarquias e Fundações), tais como: nomes, códigos e endereços de órgãos públicos e suas subdivisões administrativas e; (iii) **Catálogo de Unidades Federativas,** que apresenta dados em formato HTML contendo as 27 Unidades Federativas em acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE).

Para a realização do mapeamento, foram necessárias várias etapas para ter como resultado final um Grafo RDF que segue os princípios dos Dados Ligados e semântica de dados bem definidas. Essas etapas são ilustradas na Figura 2 e são descritas nas subseções a seguir.

Figura 2 – Etapas do Mapeamento



Fonte: Elaborado pelo autor

2.1 Etapa de Adequação dos Dados

Inicialmente, com a verificação de que as bases escolhidas retornam formatos diferentes, foi necessária a transformação de todas as bases de dados para um formato único. O formato CSV foi escolhido devido a maior quantidade dos dados publicados no escopo do trabalho já estarem neste formato, para que não fosse necessária alteração de formato dos dados

sobre Servidores e Remuneração. Os dados de Estruturas Organizacionais foram transformados de XML para CSV, com o auxílio da ferramenta *Google Refine*[1], que transforma dados XML em uma tabela, permitindo salvar no formato CSV. Como a quantidade de dados sobre Unidades Federativas era pequena, a transformação foi feita manualmente copiando os dados e inserindo-os em um arquivo CSV.

Em seguida, todos os dados foram importados para o banco de dados de triplas, convertendo os dados em CSV para tabelas lógicas através de uma opção presente no banco que permite realizar esse tipo de operação. O banco de dados de triplas escolhido foi o *Virtuoso*[2]. A escolha se baseou nos fatos de o banco permitir armazenar tanto os dados originais quanto os dados em RDF resultantes e, especialmente, possuir suporte a linguagem de mapeamento R2RML, utilizada no mapeamento dos dados com a ontologia.

2.2 Etapa de Geração do Mapeamento

Na etapa de geração do mapeamento, um mapeamento conceitual foi realizado de forma a classificar os elementos das bases de dados com os conceitos da *ORG Ontology*. A análise foi feita com base nas definições dos conceitos da ontologia e dos dados das bases de dados, ou seja, relacionando os elementos das bases com o conceito cuja definição fosse a mais apropriada. A Tabela 1 apresenta um fragmento do mapeamento conceitual, da base de dados de Servidores Civis com os conceitos da *ORG Ontology*.

Tabela 1 – Fragmento do mapeamento conceitual sobre as colunas da base de dados de Servidores Civis com os termos da ORG Ontology

Servidores	Conceito da Org Ontology
Servidor	Pessoa (*Person*)
Órgão de Lotação	Organização Formal (*FormalOrganization*)
Órgão de Exercício	Organização Formal (*FormalOrganization*)
Cargo	Cargo (*Post*)
Atividade	Cargo (*Post*)

Fonte: Elaborado pelo autor

Seguindo a Tabela 1, os elementos presentes na linha Servidor representam pessoas que mantém vínculos de trabalho com entidades governamentais. Esses elementos foram mapeados ao conceito de **Pessoa** (*Person*) da *ORG Ontogy*. Já os elementos presentes na segunda e terceira linha, Órgão de Lotação e Órgão de Exercício, apesar de possuírem significados diferentes em relação ao Servidor (o primeira representa onde o servidor está lotado e o segundo onde ele exerce suas atribuições), representam a entidade Órgão Governamental, e, por isso, são mapeados ao conceito **Organização Formal** (*Formal Organization*) da *ORG Ontology*. Caso semelhante a este, com mapeamento de dois elementos da base de dados a um único conceito ocorre no mapeamento sobre Cargo e Atividade (últimas duas linhas da Tabela 1), onde o primeiro representa um conjunto de atribuições inerentes ao agente público e o segundo um conjunto de atribuições inerentes ao exercício de funções especiais, como chefia e assessoramento. No entanto, esta distinção não existe na *ORG Ontology*. Dessa forma, o

[1] https://code.google.com/p/google-refine/

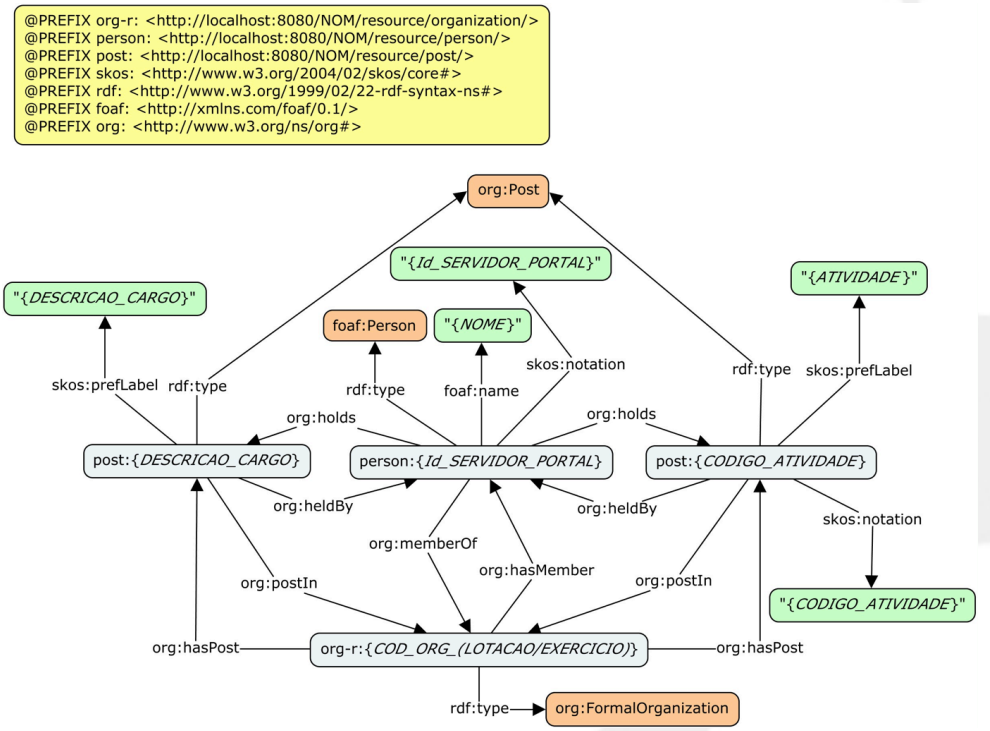
[2] Foi utilizado o banco de dados Virtuoso v 6.1.7. Disponível em: http://sourceforge.net/projects/virtuoso/files/virtuoso/6.1.7/

mapeamento foi feito ao conceito da ontologia mais adequado a ambos, no caso, ao conceito **Cargo** (*Post*).

Após a realização do mapeamento conceitual, foram criadas representações mais formais dos mapeamentos conceituais, de forma a se adequar a linguagem R2RML. Para representar o padrão de mapeamento aqui descrito, foi criada uma representação visual, em que um diagrama representa um *template* de um grafo que será gerado na saída do mapeamento. Esse *template* é parametrizado com variáveis que representam dados originários das bases de dados de entrada. São representados nesta notação os seguintes elementos: **Classes das Ontologias (laranja)**, que representam os termos dos vocabulários das ontologias contento o prefixo da ontologia mais o termo do vocabulário, denotados na forma <Prefixo>:<Termo do vocabulário>; **Entidades (azul)**, que representam um URI contendo o prefixo criado para a base de dados mais uma variável dada pelo valor da coluna mapeada, denotados na forma <Prefixo>:<{*Variável*}> e; **Literais (verde)**, que representam o valor da coluna mapeada, denotados na forma <"*Variável*">.

A Figura 3 apresenta o fragmento do mapeamento da relação entre servidores, cargos, órgãos de lotação e exercício.

Figura 3 – Fragmento referente ao mapeamento da relação entre servidores, cargos e organizações



Fonte: Elaborado pelo autor

Conforme a Figura 3, cada servidor é representado pelo seu identificador na base de dados (Id_SERVIDOR_PORTAL). Servidores são um tipo (*rdf:type*) de Pessoa (*foaf:Person*) e possuem um nome (*foaf:name*) representado pela coluna "NOME" e o identificador único (*skos:notation*) representado pela coluna "Id_SERVIDOR_PORTAL". Servidores têm (*org:holds*) cargo ou atividade e são membros (*org:memberOf*) de um Órgão de Lotação e Exercício. Como a base de dados não possui identificadores para cargos, seu URI e seu rótulo

preferencial (*skos:prefLabel*) são criados utilizando a coluna "DESCRICAO_CARGO". Já para atividades, temos que o URI utiliza o código (CODIGO_ATIVIDADE) e seu rótulo preferencial (*skos:prefLabel*) é representado pela coluna "ATIVIDADE". Além disso, Atividade possui o identificador único (*skos:notation*) representado pela coluna "COGIDO_ATIVIDADE". Ambos cargos e atividades são mapeados como um mesmo tipo (*rdf:type*) Cargo (*org:Post*) e são cargos (*org:postIn*) dentro de um órgão de lotação e exercício. Tanto Órgão de Lotação quanto de Exercício possuem membros (*org:hasMember*) e cargos e atividades (*org:hasPost*).

2.3 Etapa de Publicação dos Dados

Na etapa de publicação dos dados, foi realizada a codificação dos dados na linguagem de mapeamento R2RML com base no mapeamento conceitual e nos princípios dos Dados Ligados. A publicação, no entanto, é feita apenas localmente, com o intuito de mostrar, posteriormente, os resultados obtidos com a execução do mapeamento. A Listagem 1 apresenta o código referente a parte do mapeamento da base de dados de Servidores Civis na linguagem R2RML.

Listagem 1 – Código em R2RML referente a parte do mapeamento da base de dados de Servidores Civis

```
 DB.DBA.TTLP ('
@prefix rr: <http://www.w3.org/ns/r2rml#> .  @prefix org: <http://www.w3.org/ns/org#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> . @prefix skos: <http://www.w3.org/2004/02/skos/core#> .
<http://localhost:8080/NOM/resource#TriplesMapPerson>
  a rr:TriplesMap;
  rr:logicalTable
  [ rr:tableSchema "CSV"; rr:tableOwner "DBA"; rr:tableName "Agentes_csv" ];
  rr:subjectMap
  [ rr:template "http://localhost:8080/NOM/resource/person/{Id_SERVIDOR_PORTAL}";
    rr:class foaf:Person;
    rr:graph <http://localhost:8080/NOM/graph#>; ];
 rr:predicateObjectMap
  [ rr:predicate foaf:name;
    rr:objectMap [ rr:column "NOME" ]; ];
   rr:predicateObjectMap [
     rr:predicate org:memberOf;
     rr:objectMap [
     rr:parentTriplesMap <http://localhost:8080/NOM/resource#TriplesMapOrganization>;
        rr:joinCondition [ rr:child "COD_ORG_LOTACAO";
            rr:parent "COD_ORG_LOTACAO"; ]; ] ] .
', 'http://temp/person', 'http://temp/person');
```

Fonte: Elaborado pelo autor

No código, inicialmente são definidos os prefixos dos vocabulários RDF. O prefixo *rr* representa o vocabulário da linguagem R2RML. Em seguida, é definido um mapa de tripla como um tipo de mapa de triplas (*rr:TripleMap*). Após isso, é indicada a base de dados (*rr:logicalTable*) que será mapeada. Então, é definido o sujeito da tripla (*rr:subjectMap*) com um URI padrão (*rr:template*) usando um ou vários dados da base de dados, no caso, o identificador do servidor (Id_SERVIDOR_PORTAL). É definido então o tipo (*rr:class* que no resultado do mapeando vira um *rdf:type*) do URI e o grafo RDF (*rr:graph*) no qual será inserido as triplas RDF. Adiante, temos os mapeamentos de predicados e objetos (*rr:predicateObjectMap*) que serão ligados ao sujeito da tripla RDF. O predicado

(*rr:predicate*) será um URI de algum vocabulário. E o objeto pode ser um literal que é uma coluna (*rr:column*) da base de dados ou um outro URI de outro mapa de triplas. Neste ultimo caso, é feita uma junção (*rr:joinCondition*) com a outra base de dados através da coluna da base atual (*rr:child*) com a coluna da base do outro mapa de triplas (*rr:parent*). Ele consiste em ligar o sujeito do mapa atual com o sujeito do outro mapa (*rr:parentTripleMap*) quando os valores das duas colunas forem iguais. Por fim, o mapeamento é criado em um grafo temporário (http://temp/person).

Após a criação de todos os grafos temporários (aqui representado somente o de pessoa), ambos são inseridos em um grafo geral (http://temp/mix) através de uma inserção via SPARQL, conforme o código abaixo. O resultado, então, é inserido no grafo (representado no exemplo pelo grafo http://localhost:8080/NOM/graph#) indicado no mapeamento do sujeito. Por fim, é realizada uma inserção via SPARQL, conforme a Listagem 2, para transformar os dados do grafo, indicado no mapeamento do sujeito, para um grafo RDF representado por um URI pré-definido (no caso, http://localhost:8080/NOM/resource#).

Listagem 2 – Transformação dos dados para um grafo RDF com inserção via SPARQL

```
sparql insert in graph <http://temp/mix> { ?s ?p ?o }
from <http://temp/person> where { ?s ?p ?o };
exec ('sparql' || DB.DBA.R2RML_MAKE_QM_FROM_G ('http://temp/mix'));
sparql insert in graph <http://localhost:8080/NOM/resource#> { ?s ?p ?o } from
<http://localhost:8080/NOM/graph#> where { ?s ?p ?o };
```

Fonte: Elaborado pelo autor

A base de dados de Servidores Civis possuía cerca de 735 mil linhas e foram utilizadas 15 colunas. A base de dados de remuneração possuía cerca de 570 mil linhas e foram utilizadas 6 colunas. A base de Estruturas Organizacionais possuía um pouco mais de 70 mil linhas e foram utilizadas 19 colunas. O mapeamento gerou um total de cerca de 9,5 milhões de triplas.

3 Aplicação do Mapeamento e Uso dos Dados Mapeados

Seguindo os princípios dos Dados Ligados, uma aplicação local foi desenvolvida para demonstrar as potencialidades do mapeamento realizado e simular o uso dos dados publicados na Web. Nela é possível acessar os recursos através dos URIs que utilizam o protocolo HTTP, em um processo chamado *dereference*, que apresenta ao usuário algum conteúdo referente ao recurso acessado.

Na aplicação, o usuário consegue dereferenciar um URI qualquer e visualizar suas informações. A partir de um URI, o usuário pode navegar para os URIs relacionados ao URI original e acessar as suas informações, visualizando mais informações. A aplicação também permite que os termos das ontologias possam ser acessados e apresenta as suas definições.

A Figura 4 mostra um exemplo de um dereferenciamento do URI `http://localhost:8080/NOM/resource/person/1001212`, pertencente a servidora "Ana Cristina Ribeiro Alvim"[3], que trás os elementos conectados diretamente ao URI.

[3] A servidora específica foi escolhida aleatoriamente na base de dados. As informações dispostas neste trabalho foram recuperadas de dados publicados sob dados abertos governamentais, de acordo com a lei 12.527/2011 da República Federativa do Brasil.

Figura 4 – Exemplo de URI Dereferenciado

Property	Value
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://xmlns.com/foaf/0.1/Person
http://www.w3.org/2004/02/skos/core#notation	1001212
http://xmlns.com/foaf/0.1/name	ANA CRISTINA RIBEIRO ALVIM
http://www.w3.org/ns/org#memberOf	http://localhost:8080/NOM/resource/organization/26411
http://www.w3.org/ns/org#memberOf	http://localhost:8080/NOM/resource/organizational-unit/CAMPUS%20BCENA%20-%20DIRETORIA%20ADMIN%20E%20PLANEJA
http://www.w3.org/ns/org#holds	http://localhost:8080/NOM/resource/post/CONTADOR
http://www.w3.org/ns/org#remuneration	http://localhost:8080/NOM/resource/remuneration/brl/2013/11/person/1001212
http://www.w3.org/ns/org#remuneration	http://localhost:8080/NOM/resource/remuneration/usd/2013/11/person/1001212

Fonte: Elaborado pelo autor

As representações descritas acima permitem que usuários da aplicação visualizem a informação requerida de forma mais fácil em relação ao formato original, além de permitir uma melhor compreensão das informações, dado os termos estão mapeados a uma ontologia de referência.

Consultas específicas podem ser realizadas através da linguagem SPARQL, sendo possível obter o resultado de consultas em diversos formatos (RDF, HTML, CSV, etc.). O formato de acesso também permite que os elementos sejam usados por aplicações externas. O Banco de Dados de Triplas disponibiliza um SPARQL *Endpoint* que pode ser acessado pelo usuário. Ele possibilita a realização de consultas de forma programática (através de código) e de forma visual interativa (com consultas através de grafos visuais). O SPARQL *Endpoint* possibilita ao usuário realizar consultas com inferência, ou seja, consultas que realizam raciocínio automático através de regras definidas para descobrir novas relações, i.e., relações não anteriormente mostradas entre os elementos do grafo.

Para exemplificar o uso de SPARQL com inferência, a Listagem 3 apresenta uma consulta que visa encontrar o total de membros da "Fundação Nacional de Saúde". Na primeira linha, temos definido o grafo que possui as regras de inferência usadas na consulta, no caso, o grafo da *ORG Ontology* (representado pelo URI http://www.w3.org/ns/org#). A *ORG Ontology* define em seu modelo diversas regras (como relações inversas e transitividade) que permitem tirar conclusões automáticas sobre os dados.

Listagem 3 – Exemplo de consulta SPARQL com inferência

```
DEFINE input:inference <http://www.w3.org/ns/org#>
PREFIX org: <http://www.w3.org/ns/org#>
SELECT COUNT(?person) WHERE {
   ?org org:hasMember ?person .
   ?org skos:prefLabel "FUNDACAO NACIONAL DE SAUDE" . }
```

Fonte: Elaborado pelo autor

Ao realizar a consulta, temos como resposta que a "Fundação Nacional de Saúde" possui um total 13254 servidores. Sem o uso de inferências, não seria possível encontrar este valor, dado que existe apenas elementos relacionados pelo predicado *memberOf* que são encontrados automaticamente através da relação inversa com o predicado *hasMember* definido na *ORG Ontology*.

4 Discussões

A realização do mapeamento possibilitou melhorias ao uso e compreensão dos dados. Entretanto, algumas dificuldades para se chegar a essa situação foram encontradas, devido a problemas existentes nos dados recebidos e em sua interpretação. As subseções a seguir apresentam os problemas encontrados para a correta realização do mapeamento e o uso dos dados, assim como os benefícios advindos do uso dos dados mapeados e integrados.

4.1 Problema da Falta de Identificação Única

Um problema encontrado foi a falta de identificação única para as entidades. Esse problema faz com que seja difícil a correta identificação, relacionamento e separação dos dados. Isso faz com que dados iguais, porém apresentados diferentemente em sua forma sintática são classificados como dados diferentes. Da mesma forma, dados diferentes, porém apresentados com a forma sintática igual, podem ser classificados como o mesmo elemento.

Nesse trabalho, como não foi encontrado um identificador único para os dados, para remediar esse problema, foram utilizados identificadores sintáticos para realizar a identificação única dos elementos. Além disto, o uso de maiúsculas e minúsculas, assim como caracteres especiais também dificultou a identificação única dos elementos. Para tal, uma transformação dos dados com a retirada de caracteres especiais e com a modificação das bases para caracteres em maiúsculos também foi efetuada. Como exemplo, a Tabela 2 apresenta alguns cargos da base de dados de Servidores Civis que aparentemente são os mesmos, mas, por possuírem distinções sintáticas e não possuírem identificação única, são mapeados como entidades diferentes.

Tabela 2 - Exemplo de Cargos semelhantes sem identificadores da base de dados de Servidores Civis

CARGO
AGENTE DE TELEC E ELETRICIDADE
AGENTE DE TELEC ELETRICIDADE
AGENTE DE TELECOMUNI E ELETRICIDADE
AGENTE DE TELECOMUNIC E ELETRICIDADE

Fonte: Elaborado pelo autor

Esse problema afeta, por exemplo, a precisão de relatórios e estatísticas que poderiam ser gerados sobre esses dados. A falta de identificação única afeta também, em alguns casos, o mapeamento de determinadas entidades distintas, mas que possuem a mesma descrição e/ou nome descritivo. Para exemplificar essa situação, a Figura 5 apresenta o caso onde o nome "Coordenação de Planejamento" é usado para diversas organizações. Provavelmente trata-se de diferentes unidades organizacionais, dado o fato de pertencerem a órgãos diferentes. Apesar disto, não é possível saber ao certo se trata-se de uma coincidência de descrições ou de fato de entidades diferentes.

O uso de identificadores únicos resolveria os problemas mencionados nessa seção. Eles unificariam entidades semanticamente iguais independentemente de suas descrições e permitiriam a distinção de entidades com as mesmas descrições.

Figura 5 – Exemplo de problema da falta de identificadores no mapeamento de unidades organizacionais

Property	Value
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.w3.org/ns/org#OrganizationalUnit
http://www.w3.org/2004/02/skos/core#prefLabel	COORDENACAO DE PLANEJAMENTO
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/36205
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/25000
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/40112
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/45203
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/20224
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/28000
http://www.w3.org/ns/org#unitOf	http://localhost:8080/NOM/resource/organization/26442

Fonte: Elaborado pelo autor

4.2 Problema de Falta de Integração entre Diferentes Bases de Dados

Outro problema encontrado refere-se à falta de integração entre as diferentes bases de dados do governo. Esse problema faz com que seja difícil a correta identificação, separação e relacionamento dos dados entre as diversas bases. Entidades, quando possuem identificação em uma base específica, não possuem a identificação compartilhada com as demais bases do governo, de forma que não é possível precisar com certeza a identificação única e correta de elementos referenciados nas várias bases.

Nesse trabalho a identificação única se tornou especialmente complexa devido à falta de identificadores únicos verificada em dados recuperados de uma mesma base, conforme descrito na seção 4.1. A abordagem utilizada para remediar esses problemas foi a de classificar os dados de acordo com as descrições textuais. Todos os dados foram transformados para terem caracteres em maiúsculos e foram retirados os caracteres especiais de sua forma sintática para diminuir a probabilidade de separações incorretas. Reitera-se porém que o trabalho pode ter classificado o mesmo elemento, porém apresentado diferentemente em sua forma sintática nas diversas bases como dados diferentes. Da mesma forma, dados diferentes, porém apresentados com as características sintáticas iguais podem ter sido incorretamente classificados como o mesmo dado. Reitera-se a complexidade dado que a representação em cada base pode apresentar distinções nos caracteres utilizados para um mesmo dado, como abreviações e siglas.

A Tabela 3, onde "Ministério da Fazenda" possui diversos códigos advindos das diversas bases de dados, exemplifica o problema da falta de identificação única apresentado.

Tabela 3 – Diferentes identificadores para Ministério da Fazenda nas bases de dados

NOME	CÓDIGO	CÓDIGO PAI
MINISTERIO DA FAZENDA	1929	26
MINISTERIO DA FAZENDA	118699	1929
MINISTERIO DA FAZENDA	118700	1929
MINISTERIO DA FAZENDA	118701	1929
MINISTERIO DA FAZENDA	118702	1929
MINISTERIO DA FAZENDA	118703	1929

MINISTERIO DA FAZENDA	118704	1929
MINISTERIO DA FAZENDA	118708	1929

Fonte: Elaborado pelo autor

Continuando no exemplo da Tabela 3, nesse caso, no mapeamento automático os órgãos descritos acima são mapeados como "Ministério da Fazenda". Entretanto, ao se realizar uma navegação manual nos dados abertos do governo (disponibilizador dos dados), é possível verificar, a exceção do primeiro elemento, que os outros órgãos descritos são órgãos vinculados ao Ministério da Fazenda e não o próprio Ministério da Fazenda.

Para a solução do problema acima descrito, a abordagem preferencial seria que os sistemas governamentais fossem integrados, de forma que elementos semanticamente iguais tivessem um mesmo identificador. Isso poderia ser realizado tanto com um código identificador quanto com uma identificação sintática única.

A integração dos dados possibilitaria um compartilhamento maior e interpretação melhor das informações, com o cruzamento de dados advindos de bases diferentes.

4.3 Problema da Falta de Expressividade na ORG Ontology

A *ORG Ontology* foi utilizada como ontologia de referência e para adicionar semântica aos termos mapeados nesse trabalho. A *ORG Ontology* abrange o escopo de organizações, sendo uma recomendação W3C. Entretanto, nem todos os conceitos necessários no escopo do problema encontraram mapeamento único na *ORG Ontology*. Essa limitação gerou sobrecarga semântica em alguns conceitos mapeados (GUIZZARDI, 2005). Por exemplo, Cargo e Atividade, apesar de possuírem significados distintos, onde o primeiro representa um conjunto de atribuições inerentes ao agente público e o segundo um conjunto de atribuições inerentes ao exercício de funções especiais, como chefia e assessoramento, foram mapeados ao conceito **Cargo** (*Post*) da *ORG Ontology*.

A *ORG Ontology* permite ser estendida ou especializada para uso em situações não previstas no padrão, o que resolveria os mapeamentos de conceitos diferentes para uma conceituação igual. Embora isso mitigue o problema da falta de expressividade semântica, afeta o benefício descrito na seção 4.4.4. Como o escopo do trabalho era realizar o estudo de caso com tecnologias existentes, não foi realizada a extensão.

4.4 Benefícios do Mapeamento

O mapeamento dos dados seguindo as etapas descritas na seção 2 traz benefícios ao uso dos dados, tanto neste trabalho como em outros. Esses benefícios incluem a: (i) melhor integração dos dados e qualidade de consultas; (ii) possibilidade de inferência; (iii) semântica bem definida e; (iv) utilização de conceitos conhecidos. As subseções a seguir descrevem esses benefícios.

4.4.1 Integração dos dados e qualidade de Consultas.

A realização do mapeamento apresentou melhorias à utilização dos dados. A integração é o principal deles. Com a integração, os dados passaram a ser unificados. A informação se torna mais qualificada dado o maior relacionamento e agregação entre os dados. Também aumenta o relacionamento entre informações diversas e se torna mais fácil o cruzamento de

informações, aumentando o poder de análise e melhorando a tomada de decisões baseada nas informações recuperadas.

Isso advém da maior quantidade de dados possíveis de serem utilizados para uma mesma consulta, o que abre a possibilidade de realizar consultas que antes não eram possíveis.

4.4.2 Possibilidade de Inferência nos dados.

O raciocínio automático através de regras de inferência permite que agentes de software entendam e tirem conclusões lógicas sobre os dados existentes. É possível descobrir relações entre dados que se conectam, diretamente ou indiretamente, assim como entre dados que não se conectam, permitindo buscar conexões que não são simples de serem observadas. Por exemplo, no trabalho efetuado é possível através de inferência saber o gasto com pessoal em uma determinada organização, o que não era diretamente possível com a forma original dos dados. Também é possível utilizar inferências para verificar os dados, descobrindo possíveis inconsistências entre eles. Como exemplo, se o usuário possuir dados de orçamentos dos órgãos poderia saber a proporção dos gasto do pessoal em relação ao orçamento do órgão, assim como verificar sua coerência em caso de proporções notadamente inconsistentes.

4.4.3 Utilização de semântica bem definida.

O mapeamento dos dados para uma ontologia permite que a semântica dos dados seja provida pelos conceitos aos quais eles se mapeiam na ontologia. O uso de ontologias com termos bem definidos faz com que os usuários dos dados compreendam sobre o que os dados tratam e aumenta a possibilidade de interpretação correta dos mesmos. Esse mapeamento também afasta a possibilidade de problemas como a Falsa Concordância (GUARINO, 1998), em que os usuários troquem informações sobre elementos diferentes acreditando se tratar do mesmo elemento e não identifiquem essa divergência, levando a interpretação e, consequente, tomada de decisões incorreta. Nesse trabalho, o mapeamento dos dados foi realizado para a ontologia *ORG Ontology*.

4.4.4. Utilização de conceitos amplamente utilizados

Outro benefício do mapeamento é ter os dados mapeados para termos de ontologias que são amplamente utilizadas. O uso de ontologias bastante utilizadas diminui o tempo necessário para aprendizado sobre o que as informações tratam e aumentam o seu potencial de uso, dado que vários desenvolvedores e usuários de ontologias já conhecem os seus significados. Isso também aumenta a probabilidade dos dados do trabalho serem usados e cruzados com outros dados, tanto em aplicações de terceiros como em inclusão de novos dados por outros usuários.

Nesse trabalho as ontologias como FOAF (*Friend of a Friend*), SKOS (*Simple Knowledge Organization System*), *vCard* e *GoodRelation* são exemplos de ontologias que possuem seus termos utilizados em larga escala, inclusive por empresas como Google, Wikipedia e IBM.

Conclusão

Dados governamentais são frequentemente publicados de forma não estruturada ou em formatos não padronizados. Frequentemente, diferentes bases de dados utilizam formatos distintos e há pouca ou nenhuma integração entre elas. Neste trabalho foi realizado um estudo de caso utilizando bases de dados governamentais e mapeando os diversos dados para uma ontologia de organizações, denominada *ORG Ontology*, com a publicação de dados integrados baseados nos princípios de Dados Ligados e semanticamente definidos pelos termos de uma ontologia.

Foi realizado um mapeamento conceitual de bases de dados governamentais para a *ORG Ontology*, utilizada como ontologia de referência. Com o uso da linguagem R2RML, o mapeamento conceitual foi codificado e executado, gerando um grafo com mais de 9,4 milhões de triplas, interligando todas as bases de dados entre si e com os conceitos da ontologia. Finalmente, uma aplicação local foi desenvolvida para demonstrar as potencialidades do mapeamento realizado e simular o uso dos dados publicados na Web. Nela é possível realizar consultas específicas com base nos conceitos da ontologia ou em dados individuais (com resultados em diversos formatos), inclusive acessando-os individualmente por meio de seus URIs ou fazendo inferências sobre eles. Ademais, também é possível realizar inferências entre os dados utilizando relacionamentos originalmente somente providos entre os conceitos da ontologia para a qual os dados foram mapeados.

O trabalho mostrou diversos benefícios do uso de uma abordagem com a utilização de dados integrados e semanticamente definidos, conforme descritos na seção 4.4. Em especial, ressalta-se: (i) a integração dos dados, possibilitando maior qualidade na resposta às consultas e maior quantidade de consultas possíveis; (ii) a possibilidade de inferência nos dados, descobrindo-se relações não explicitamente presente nos dados isolados; (iii) a utilização de semântica bem definida, via o mapeamento para uma ontologia de referência, o que aumenta a compreensão dos dados, assim como gera a possibilidade de uso e inclusão de dados por terceiros sem inconsistências (assumindo que se respeite a ontologia) e; (iv) a utilização de conceitos amplamente utilizados pela comunidade diminuindo a curva de aprendizado para desenvolvimento de aplicações por terceiros, facilitando o consumo e análise.

O mapeamento realizado também objetivou a verificação na prática da ocorrência de problemas estruturais ou semânticos, assim como objetivou a verificação dos benefícios para o uso de dados abertos governamentais. Durante a abordagem, foi necessário lidar com problemas, em especial: (i) a falta de identificação única para os elementos recuperados das bases de dados, tanto dentro de uma mesma base de dados quanto entre diferentes bases de dados; (ii) a falta de integração entre bases de dados governamentais, forçando o consumidor a recorrer a suposições para a integração os dados; (iii) a falta de descrição da semântica dos dados publicados, sendo o usuário dos dados responsável por supor a sua semântica com base em rótulos, e; (iv) à falta de expressividade da recomendação W3C *ORG Ontology* para o mapeamento de todos os conceitos assumidamente presentes nas bases de dados, levando a sobrecarga semântica de alguns conceitos.

Os problemas encontrados levam a menor confiabilidade das informações. A precisão do grafo gerado pelo mapeamento foi comprometida devido às limitações nas bases de dados, reduzindo a possibilidade de uso das mesmas pela população. Limita-se a possibilidade de tomada de decisões e o correto acompanhamento das políticas públicas utilizando esses dados, uma motivação para a publicação da lei e da disponibilização dos dados em questão.

No trabalho, foram propostas e aplicadas soluções para remediar imediatamente os problemas identificados, de forma a possibilitar a conclusão do trabalho, porém também foram propostas soluções permanentes, a serem tomadas no âmbito governamental para evitar os problemas identificados. Em especial, propõe-se no trabalho a integração das bases de dados governamentais e o seu mapeamento para ontologias de referência, para prover semântica e ajudar a compreensão dos mesmos.

Como perspectivas futuras, maior esforço é necessário para aumentar o escopo do mapeamento de dados para incluir outros elementos, além de servidores e organizações (como gastos e convênios, por exemplo). Isto permitiria gerar uma "nuvem" maior de informações governamentais estruturadas. Essa nuvem de informações possibilitaria um compartilhamento e entendimento maior dos dados governamentais. Também viabilizaria que o acesso às informações se torne cada vez maior através de aplicações externas ligadas aos dados governamentais. Para o governo de um país que prega transparência da informação, ter os dados publicados de forma como a proposta neste trabalho possibilitaria à população extrair a informação desejada e analisá-la com maior facilidade.

Agradecimentos

Este trabalho contou com o apoio do W3C Brasil, da CAPES, da FAPES (projeto de no. 59971509/12) e do CNPq (projetos de nos. 310634/2011-3 e 485368/2013-7).

Referências

BERNERS-LEE, T. Linked Data. **Design Issues**, 2006.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific american**, v. 284, n. 5, p. 28–37, 2001.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data - The Story So Far. **International journal on semantic web and information systems**, v. 5, n. 3, p. 1–22, 2009.

DAS, S.; SUNDARA, S.; CYGANIAK, R. **R2RML: RDB to RDF Mapping Language**. Disponível em: <http://www.w3.org/TR/r2rml/>. Acesso em: 29 abr. 2014.

GUARINO, N. **Formal ontology in information systems: Proceedings of the first international conference (FOIS'98), June 6-8, Trento, Italy**. Amsterdam: IOS Press, 1998. p. 3–15

GUIZZARDI, G. **Ontological foundations for structural conceptual models**. The Netherlands: CTIT, Centre for Telematics and Information Technology, 2005.

MCGUINNESS, D. L.; HARMELEN, F. VAN. **OWL Web Ontology Language Overview**. Disponível em: <http://www.w3.org/TR/owl-features/>. Acesso em: 7 maio. 2014.

REYNOLDS, D. **The Organization Ontology**. Disponível em: <http://www.w3.org/TR/vocab-org/>. Acesso em: 29 abr. 2014.