Uma Ontologia de Domínio para Preservação de Privacidade em Dados Publicados pelo Governo Brasileiro

Alternative Title: A Domain Ontology for Privacy Preservation in Data Published by the Brazilian Government

Maria J. Queiroz
Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte, IFRN, Currais Novos - RN
jane.queiroz@ifrn.edu.br

Natasha C. Q. Lino
Universidade Federal da Paraíba, UFPB, Campus I
Av. dos Escoteiros, Mangabeira
João Pessoa – PB
natasha@ci.ufpb.br

Gustavo H. M. B. Motta
Universidade Federal da Paraíba, UFPB, Campus I
Av. dos Escoteiros, Mangabeira
João Pessoa – PB
gustavo@ci.ufpb.br

RESUMO

Apesar de considerar a transparência como regra e o sigilo como exceção, a Lei de Acesso à Informação (LAI) prevê a proteção aos dados pessoais dos cidadãos quando da publicação de dados relacionados ao setor público na Internet (Art. 31). Métodos sistemáticos de anonimização existentes podem ser aplicados para responder a esta necessidade. Assim, este artigo objetiva desenvolver uma ontologia de domínio para a área de preservação de privacidade em dados publicados, de forma a atender aos preceitos da LAI e às iniciativas do governo brasileiro, além de possibilitar a unificação semântica de termos da área de anonimização e a interoperabilidade entre ferramentas com esta finalidade.

Palavras-Chave

Dados governamentais; Ontologia de domínio; Preservação de privacidade.

ABSTRACT

While considering transparency as a rule and secrecy as an exception, the Access to Information Act (AIA) provides for the protection of citizens' personal data when the publication of data related to the public sector on the Internet (Art. 31). Existing systematic methods for anonymization can be applied to meet this need. Thus, this article aims to develop a domain ontology for privacy preservation area in published data in order to comply with the provisions of AIA and initiatives of the Brazilian government, besides enabling the semantic unification of terms of anonymisation area and interoperability between tools for this purpose.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – Representations (procedural and rule-based); K.4.1 [**Computers and Society**]: Public Policy Issues – Privacy; J.1 [**Computer Applications**]: Administrative Data Processing - Government, Law

General Terms

Security, Standardization, Legal Aspects.

Keywords

Government data; Domain ontology; Privacy preservation.

SBSI 2016, May 17–20, 2016, Florianópolis, Santa Catarina, Brazil.

1. INTRODUÇÃO

O governo brasileiro sancionou em 18 de novembro de 2011 a Lei nº 12.527, conhecida como Lei de Acesso à Informação (LAI) [1], e criou concomitantemente o Plano de Ação Brasileiro para transparência governamental. A lei torna obrigatória a publicação de dados relacionados ao setor público (como execução orçamentária, microdados obtidos a partir da realização de pesquisas, etc.) e normatiza as práticas de transparência no país, enquanto o Plano de Ação define as ações necessárias para a real promoção da transparência [2].

Dentre as ações realizadas em prol da abertura de dados estão: a elaboração de guias de orientação à publicação de dados governamentais abertos em sítios na Internet [3]; a criação do Portal Brasileiro de Dados Abertos [4] e das Seções de Acesso à Informação em portais de instituições públicas; a criação da INDA (Infraestrutura Nacional de Dados Abertos), composta por grupos de trabalho responsáveis pela criação das tecnologias de suporte à abertura de dados (como as ontologias disponíveis no Repositório de Vocabulários e Ontologias do Governo Eletrônico [5]) e pela publicação de documentos relacionados ao assunto, como o Guia de Abertura de Dados da INDA [6].

A seção 7 do guia mencionado anteriormente (intitulada Modelar os Dados Antes da Abertura) recomenda a anonimização de bases de dados que contenham informações pessoais antes de publicá-las, a fim de preservar a privacidade dos cidadãos [7]. Essa sugestão vai ao encontro do Art. 31 da LAI, que prescreve o respeito à "intimidade, vida privada, honra e imagem das pessoas, bem como às liberdades e garantias individuais" [1].

A anonimização visa ocultar a identidade ou dados confidenciais de cidadãos, de forma que, ainda que a informação seja divulgada e útil para análise por parte de receptores, a privacidade individual seja preservada [8]. Existem métodos sistemáticos e ferramentas específicas para esta finalidade que, no entanto, não interoperam entre si e possuem distinção semântica para termos idênticos. Além disso, os métodos e ferramentas existentes não têm sido aplicados às bases de dados brasileiras [9].

Em tal contexto, este artigo contribui com a criação de uma ontologia de domínio para representar métodos sistemáticos de anonimização, comumente disponíveis em ferramentas com tal finalidade. O objetivo é, em particular, colaborar para o Repositório de Vocabulários e Ontologias do Governo Eletrônico do governo brasileiro no que se refere à preservação de privacidade individual em bases publicadas na Internet, favorecendo, em geral, a interoperabilidade no uso de ferramentas de anonimização, bem como a promoção da uniformização semântica dos termos da área.

Para tanto, o presente artigo está organizado da seguinte forma: a seção 2 traz o método de pesquisa utilizado; a seção 3 apresenta as etapas realizadas para a criação da ontologia proposta; a seção 4 apresenta os trabalhos relacionados e por fim, a seção 5 traz a discussão e as considerações finais.

2. MÉTODO DE PESQUISA

Anteriormente à criação da ontologia, foi realizada uma revisão narrativa da literatura. Esse método é considerado como pesquisa qualitativa e tem como objetivo "descrever e discutir o desenvolvimento ou o 'estado da arte' de um determinado assunto, sob ponto de vista teórico ou contextual" [10]. Ainda de acordo com [10], esse método realiza uma "análise da literatura publicada em livros, artigos de revista impressas e/ou eletrônicas na interpretação e análise crítica pessoal do autor."

Dessa forma, a revisão narrativa da literatura realizada para a construção da ontologia proposta neste artigo iniciou-se pela leitura do *survey* presente em [8], seguida da busca artigos científicos complementares em repositórios *on line*, como o Google Scholar[1] e o IEEE Xplore[2], utilizando-se como palavras-chave o nome de modelos ou operações de anonimização.

Após a revisão narrativa da literatura, partiu-se para a criação da ontologia, empregando dois métodos combinados: o método 101 de criação de ontologias, desenvolvido por [11], e o processo de engenharia do conhecimento (do inglês, *Knowledge Engineering Process*), apresentado por [12]. Tanto o processo de engenharia do conhecimento quanto o método 101 consistem em um conjunto de 07 etapas, cada um e foram utilizados por serem considerados apropriados pelos autores deste artigo devido ao detalhamento dos passos necessários à elaboração de uma ontologia.

As etapas do processo de engenharia do conhecimento são: 1) identificação da tarefa; 2) seleção do conhecimento relevante; 3) decisão sobre o vocabulário a ser utilizado; 4) codificação do conhecimento geral sobre o domínio; 5) codificação de uma descrição da instância do problema específico; 6) criação de consultas ao procedimento de inferência e obtenção de respostas; 7) depuração da base de conhecimento.

As etapas do método 101 são: 1) determinação de domínio e abrangência da ontologia; 2) reutilização de ontologias existentes, quando possível; 3) enumeração de termos importantes; 4) definição das classes e hierarquia de classes; 5) definição das propriedades das classes; 6) definição das facetas dos slots (ou tipos de propriedades) e 7) criação de instâncias [11].

A seguinte integração das etapas dos métodos foi utilizada para elaborar a ontologia apresentada neste artigo, resultando em 04 etapas: Etapa 1) Definição do domínio e escopo da ontologia: como o próprio nome sugere, nesta etapa são definidos o domínio ao qual a ontologia deverá representar e os limites de sua abrangência, ou seja, quais questões ela deverá responder; Etapa 2) Seleção do conhecimento: compreende a aquisição e seleção dos conceitos relevantes para a construção da ontologia, realizada a partir da análise da literatura acadêmica ou entrevistas com especialistas da área de conhecimento a ser representada; Etapa 3) Criação de classes: como resultado da seleção do conhecimento relevante, esta etapa compreende a relação dos termos imprescindíveis para a área representada e que deverão compor a hierarquia de classes da ontologia; Etapa 4) Criação de propriedades: a partir da criação das classes, são extraídas de suas definições as propriedades de objetos e de tipos de dados possíveis.

3. ONTOLOGIA PARA PRESERVAÇÃO DE PRIVACIDADE

As seções a seguir descrevem em detalhes a execução das etapas elencadas anteriormente.

3.1 Etapa 01: Definição de domínio e escopo

Define o domínio do conhecimento a ser representado e as questões as quais tal representação deverá ser capaz de responder [12]. Para facilitar a realização dessa etapa, o método 101 propõe que algumas questões sejam respondidas [11]. Tais questões e suas respectivas respostas, considerando a ontologia proposta neste artigo, são apresentadas a seguir:

1) Qual o domínio que a ontologia irá cobrir?

R: O domínio da preservação de privacidade de dados, abrangendo as técnicas, modelos, operações, métricas e demais parâmetros clássicos relacionados ao tema.

2) Para que a ontologia será usada?

R: Para representar o domínio do conhecimento das técnicas e modelos clássicos para preservação de privacidade em dados publicados, além de promover a interoperabilidade entre soluções (como ferramentas, aplicações ou recomendações técnicas) usadas para este fim.

3) Para que tipos de perguntas a informação presente na ontologia deverá fornecer respostas?

R: As questões as quais a ontologia deverá responder, de acordo com um estudo inicial dos métodos de preservação de privacidade apresentados por [8], são as seguintes:

- Quais tipos de dados estão presentes em bases de dados para publicação?

[1] https://scholar.google.com.br/

[2] http://ieeexplore.ieee.org/Xplore/home.jsp

- Qual a possível classificação de cada atributo presente em uma base de dados?
- Que tipos de hierarquias de generalização podem ser aplicados a cada tipo de atributo?
- Quais modelos são mais adequados, de acordo com os tipos de identificadores presentes em uma base de dados?
- Quais as dependências ou restrições quando do uso conjunto de diferentes modelos?
- Quais tipos de operações de anonimização existentes e suas características?
- Quais métricas de utilidade podem ser aplicadas, considerando-se suas características em consonância com as características dos modelos e da base de dados utilizadas?

4) Quem vai usar e manter a ontologia?

R: Os responsáveis pela coleta e divulgação dos dados da administração pública poderão utilizar esta ontologia no processo de anonimização de dados, de forma sistemática, antes de publicá-los. Gestores dos sistemas de informação e técnicos de TI (Tecnologia da Informação) são alguns dos profissionais responsáveis por esta atividade.

Vale ressaltar que o escopo da ontologia não inclui a representação de algoritmos de anonimização porque, dentre as ferramentas de anonimização analisadas (UTD *Anonymization Toolbox* [13], Cornell *Anonymization Toolkit* [14] e ARX [15]), duas delas utilizam métodos de anonimização (Cornell *Anonymization Toolkit* e ARX), enquanto apenas uma emprega explicitamente algoritmos de anonimização (UTD *Anonymization Toolbox*). A escolha de tais ferramentas deveu-se ao fato de ambas permitirem a inserção de configurações de anonimização por meio de arquivos de texto adicionais, permitindo a análise destes para auxiliar na definição de classes e propriedades na ontologia.

Além disso, a partir do estudo da literatura existente, observou-se que os algoritmos de anonimização são compostos pela combinação de métodos de anonimização [13]. Assim, ainda que não adicionados em um primeiro momento, há a possibilidade de se estender a ontologia criada a fim de representar também os algoritmos de anonimização e introduzindo, dessa forma, novos relacionamentos entre os métodos de anonimização representados.

3.2 Etapa 02: Seleção do Conhecimento

Após definir o domínio e o escopo de conhecimento que a ontologia deverá representar, efetua-se a seleção do conhecimento relevante para a ontologia, conforme etapa 02 do processo de engenharia do conhecimento [12]. Como o método 101 sugere a reutilização de ontologias existentes (etapa 02) e a enumeração de termos importantes (etapa 03), tais etapas serão integradas à etapa 02 do processo de engenharia do conhecimento para compor o processo de aquisição e seleção do conhecimento, realizado neste trabalho. Entretanto, a reutilização não foi possível porque os domínios e objetivos das ontologias existentes identificadas e analisadas são distintos daqueles desta proposta.

Já a enumeração de termos (etapa 03 do método 101) foi realizada após a aquisição do conhecimento. Esse processo foi feito a partir de pesquisas em livros, artigos científicos, análise de ferramentas de anonimização e análise de bases de dados públicas brasileiras.

O livro produzido por [8] foi a fonte mais consultada e utilizada para o desenvolvimento desta pesquisa por se tratar de um *survey* (em português, pesquisa, estudo ou revisão), ou seja, uma revisão geral da literatura relevante na área de preservação de privacidade. Além deste livro, os artigos [16], [17], [18], dentre outros, também foram consultados.

Foram utilizadas ainda, no processo de aquisição de conhecimento, informações obtidas a partir da análise da ferramenta de anonimização ARX [15] e de bases de dados públicas brasileiras, como a base de dados do Censo da Educação Superior Brasileira, realizado e divulgado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [19].

Além do estudo de livros e artigos científicos, outra forma eficaz de aquisição de conhecimento é a realização de entrevistas com especialistas da área que se pretende representar. No entanto, como na literatura estudada não foram identificados autores brasileiros na área de preservação de privacidade, este meio não foi empregado neste trabalho.

3.3 Etapa 03: Criação de classes

Como resultado do processo de aquisição de conhecimento na área de preservação de privacidade em dados publicados, a etapa 03 compreende a seleção dos termos adequados para a criação da hierarquia de classes e de suas definições, a partir da integração entre a etapa 03 do processo de engenharia do conhecimento (decisão sobre o vocabulário a ser utilizado) e a etapa 04 do método 101 (definição das classes e hierarquia de classes).

Para a nomeação das classes, optou-se pela língua inglesa por ser uma língua estrangeira universalmente utilizada e porque as fontes de pesquisa consultadas também estavam nesta língua. Além disso, o uso desta língua facilita o compartilhamento e/ou reutilização da ontologia por pesquisadores e estudiosos.

Para facilitar seu uso por servidores públicos e pesquisadores brasileiros, pretende-se disponibilizar uma versão desta ontologia em português no Repositório de Vocabulários e Ontologias do Governo Eletrônico. Por isso, as classes nomeadas em inglês e descritas a seguir já apresentam sua tradução em português.

3.3.1 Anonymization Method

A classe *Anonymization Method* (em português, Método de Anonimização) – Figura 1 – é a classe raiz do domínio da preservação de privacidade em dados publicados.

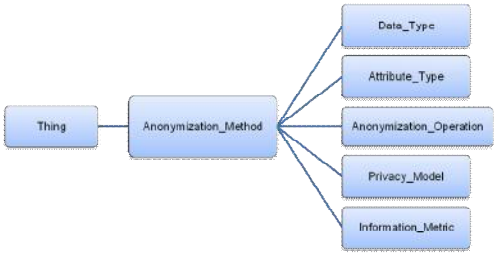

Figura 1. Classe *Anonymization Method* e suas subclasses.

Ela representa os métodos utilizados para a anonimização de bases de dados, sendo composta por um conjunto de modelos, técnicas e procedimentos cujo objetivo é evitar a re-identificação

de registros de um indivíduo (re-identificação individual) e, ainda assim, possibilitar a publicação de uma base de dados útil ao propósito ao qual se destina [8].

As demais classes do domínio da preservação de privacidade em dados publicados derivam da classe *Anonymization Method*, conforme apresentado na Figura 1: *Data Type, Attribute Type*, *Anonymization Operation, Privacy Model* e *Information Metric*. Cada uma dessas classes derivadas possui suas próprias subclasses, que serão apresentadas nas seções a seguir.

3.3.2 Data Type

A classe *Data Type* (em português, tipo de dado) estabelece o formato em que o dado é apresentado [15]. Para a definição das seis subclasses derivadas de *Data Type*, efetuou-se uma análise dos formatos dos dados presentes na base de dados do Censo da Educação Superior Brasileira de 2013 [19] e dos tipos de dados suportados pelas ferramentas ARX [15], UTD *Anonymization Toolbox* [13] e Cornell *Anonymization Toolkit* [14].

O Quadro 1 apresenta as descrições para cada formato de dado.

Quadro 1. Descrição dos formatos de dado possíveis.

Classe	*Definição*
String	Sequência de caracteres.
Ordered String (String Ordenada)	String com escala ordinal.
Decimal	Número com componente fractional.
Integer (Inteiro)	Número sem componente fractional.
Date (Data)	Data, incluindo dia, mês e ano.
Time (Tempo)	Marcação de tempo. Pode incluir dia, mês, ano, hora, minuto e segundo.

3.3.3 Attribute Type

A classe *Attribute Type* (em português, tipo de atributo) representa as possíveis classificações dos tipos de informação presentes em uma base de dados, considerando o grau de sensibilidade dessas informações [8], segundo a literatura da área.

Existem quatro tipos de atributos, conforme ilustrado na Figura 2. Dessa forma, cada coluna em uma tabela pode ser classificada como *Explicit Identifier, Quasi Identifier, Non-Sensitive Attribute* ou *Sensitive Attribute*. As definições de cada tipo de atributo, de acordo com [8], são apresentadas a seguir.

- *Explicit Identifier* (em português, Identificador Explícito): identifica explicitamente um indivíduo. Por exemplo: nome completo ou CPF (Cadastro de Pessoa Física);
- *Quasi Identifier* (em português, Quase Identificador): possui potencial para a identificação individual a partir da combinação de conjuntos de quase identificadores (também referenciados como QIDs). Por exemplo: um conjunto de dados composto por data de nascimento, sexo e etnia;
- *Sensitive Attribute* (em português, Atributo Sensível): possui informação confidencial. Por exemplo: doença;
- *Non-Sensitive Attribute* (em português, Atributo Não-sensível): representa informações que não sejam classificadas em um dos tipos de atributos apresentados anteriormente.

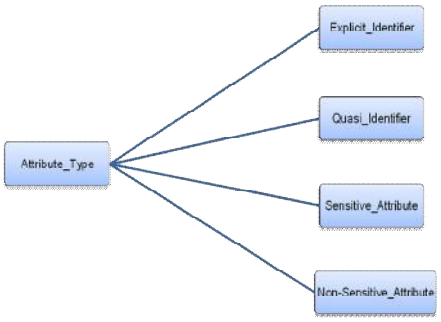

Figura 2. Classe *Attribute Type* e suas subclasses.

Alguns autores utilizam uma nomenclatura diferente, como é o caso de [20] que utiliza *Identifier, Confidential* e *Non-Confidential* para se referir a *Explicit Identifier, Sensitive Attribute* e *Non-Sensitive Attribute*, respectivamente. Por se tratarem de denominações pouco aplicadas na literatura pesquisada, elas foram adicionadas como sinônimos na ontologia.

3.3.4 Anonymization Operation

A classe *Anonymization Operation* (em português, operação de anonimização) representa as operações utilizadas para tornar os registros indistinguíveis quanto aos valores de quase identificadores, dificultando a re-identificação individual [8]. As principais operações, agrupadas sob a superclasse *Anonymization Operation*, estão representadas como subclasses na Figura 4.

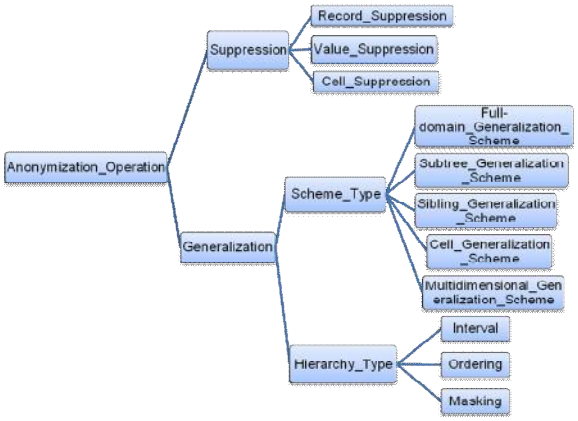

Figura 3. Classe *Anonymization Operation* e suas subclasses.

A classe *Suppression* representa as operações de supressão que consistem na substituição de valores (ou parte deles) por caracteres especiais. Existem três tipos de supressão: de registro (*Record Suppression*), de valor (*Value Suppression*) ou de célula (*Cell Suppression*). O primeiro substitui um registro (linha em uma tabela) por caracteres especiais; o segundo substitui os valores de um atributo (ou seja, uma coluna em uma tabela) por caracteres especiais e o terceiro substitui alguns valores em células por caracteres especiais [8].

A classe *Generalization* representa operações de generalização que substituem valores específicos por valores gerais. A Figura 3 apresenta duas subclasses derivadas da classe *Generalization*:

Hierarchy Type (em português, Tipo de Hierarquia) e *Scheme Type* (em português, Tipo de Esquema) [8].

A classe *Scheme Type* estabelece os tipos de generalizações válidas e possíveis de serem aplicadas a um domínio de atributo. Domínio de atributo é um conjunto de valores possíveis para um atributo [18]. Existem cinco tipos de esquema, conforme ilustrado na Figura 3, apresentados a seguir.

- *Full-domain Generalization Scheme* (em português, Esquema de generalização de domínio completo): generaliza todos os valores para um mesmo nível do domínio de atributo [8];
- *Subtree Generalization Scheme* (em português, Esquema de generalização de sub-árvore): os valores de um nó podem ser generalizados para um nó não-folha, não sendo obrigatório que isto aconteça para os demais nós [8];
- *Sibling Generalization Scheme* (em português, Esquema de generalização de irmãos): valores semelhantes podem ser generalizados, sem a obrigatoriedade de que todos o sejam [8];
- *Cell Generalization Scheme* (em português, Esquema de generalização de célula): algumas instâncias de um nó podem ser generalizadas, sem a obrigatoriedade de que todas sejam [8];
- *Multidimensional Generalization Scheme* (em português, Esquema de generalização Multidimensional): considera mais que um tipo de atributo e suas árvores taxonômicas no momento da generalização [8].

Um esquema pode aplicar diferentes tipos de hierarquia na generalização de valores em uma tabela. Hierarquia de generalização é a forma de substituição de valores específicos por valores gerais, empregada conforme o tipo de dado (seção 3.3.2) a ser substituído [15]. Existem três tipos de hierarquia de generalização:

- *Interval* (em português, Intervalo): substitui um valor por um intervalo ao qual o valor substituído pertence. Utilizado para dados do tipo numérico (inteiro, decimal, data e tempo) [15];
- *Ordering* (em português, Ordenação): substitui grupos de valores específicos por valores gerais, utilizando funções de agregação, como valores delimitadores, prefixo comum, intervalo, conjunto de valores ou conjunto de prefixos. Utilizado para qualquer tipo de dado, sendo recomendado para dados do tipo string ou string ordenada [15];
- *Masking* (em português, Mascaramento): substitui partes da informação ou a informação por completo por caracteres especiais. Utilizado para qualquer tipo de dado [15].

Existem outras operações de anonimização além da supressão e da generalização (como anatomização e permutação, por exemplo), porém o foco deste artigo é apenas nestas duas por serem as mais utilizadas de acordo com [8]. Posteriormente, serão adicionadas outras operações de anonimização à ontologia.

3.3.5 Privacy Model

A classe *Privacy Model* (em português, modelo de privacidade) representa modelos utilizados para anonimização de dados que agem de forma a prevenir que atacantes obtenham informações sensíveis sobre uma vítima a partir de uma base de dados publicada na Internet [8]. Essa classe é apresentada na Figura 4.

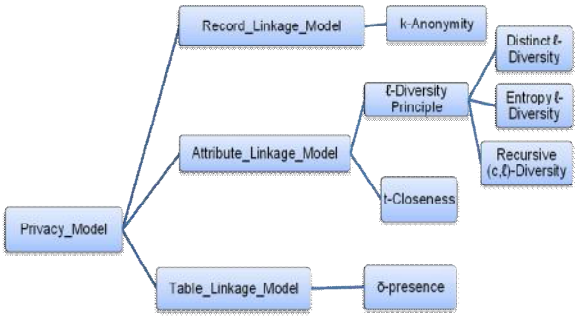

Figura 4. Classe *Privacy Model* e suas subclasses.

Os autores [14] e [15] empregam o termo *Privacy Criteria* (em português, critério de privacidade) em lugar de *Privacy Model*. Por isso, *Privacy Criteria* foi adicionado como sinônimo para *Privacy Model* na ontologia.

Existem três categorias principais de modelos de privacidade, dependendo do tipo de ataque que evitam: modelos contra ataques por vinculação de registro, de atributo ou de tabela [8]. Essas categorias e os modelos clássicos pertencentes a cada uma delas são apresentados na Figura 4. Existem outras categorias e modelos, que serão adicionados posteriormente na ontologia, mas que por questões de limite de espaço não foram incluídos aqui.

A classe *Record Linkage Model* (em português, modelo de vinculação de registro) representa modelos que visam evitar a vinculação de um indivíduo a um registro (linha) em uma tabela, através da correspondência entre seus valores de QIDs e os valores presentes na tabela analisada [8]. Um exemplo de modelo desta categoria é o *k-Anonymity*.

A classe *k-Anonymity* representa o modelo de mesmo nome, o qual define que um grupo deve ter, no mínimo, ***k*** registros idênticos com respeito aos valores de QID. Esse modelo possui parâmetro (representado pela letra ***k*** e cujo valor deve ser maior ou igual a dois) e utiliza operações de anonimização, principalmente generalizações e supressões [16].

A classe *Attribute Linkage Model* (em português, modelo de vinculação de atributo) representa modelos que visam evitar a vinculação de um indivíduo a um atributo sensível em uma tabela, através da realização de inferências [8].

Modelos clássicos contra a vinculação de atributos e baseados no Princípio da ℓ-Diversidade foram agrupados sob a classe *ℓ-Diversity Principle*. Esse princípio estabelece que deva haver ℓ valores "bem-representados" para atributos sensíveis em um grupo de registros [17]. Por "bem-representados" compreende-se que, em cada grupo, deve haver pelo menos ***ℓ*** valores diferentes para atributos sensíveis. Três subclasses representam os modelos baseados neste princípio:

- *Distinct ℓ-Diversity*: aplica estritamente o que estabelece o Princípio da ℓ-Diversidade.
- *Entropy ℓ-Diversity*: adiciona entropia (imprevisibilidade) ao grupo de valores sensíveis anonimizado.

- *Recursive (c,ℓ)-Diversity*: controla a entropia do modelo anterior, fazendo com que valores frequentes não apareçam com tanta frequência e valores raros não apareçam tão raramente. Possui ainda um parâmetro adicional, representado pela letra ***c*** (constante que permite controlar a frequência de valores em grupos criados a partir da execução do modelo).

Os modelos apresentados anteriormente possuem um parâmetro em comum, representado por $\boldsymbol{\ell}$ e utilizam operações de supressão. Além disso, quando utilizados em conjunto com o modelo *k-Anonymity*, o valor de $\boldsymbol{\ell}$ deve ser menor ou igual a $\boldsymbol{k}$.

Existem outros modelos contra a vinculação de atributos e que não são baseados no Princípio da ℓ-Diversidade, como o modelo *t-Closeness* apresentado na Figura 4. Esse modelo objetiva assegurar que a distribuição de um atributo sensível em cada grupo seja próxima à sua distribuição na tabela inteira. Para isso, o modelo *t-Closeness* utiliza uma função denominada *Earth Mover Distance* (EMD ou, em português, Distância da Retro-Escavadeira) para medir a proximidade entre as distribuições e a proximidade definida através da variável ***t***, configurada pelo responsável pela publicação da base de dados [8].

A classe *Table Linkage Model* (em português, modelo de vinculação de tabela) representa modelos que visam evitar a vinculação de um indivíduo a uma tabela publicada, possibilitando a obtenção de informações confidenciais sobre ele [8]. A classe *δ-presence* representa um modelo desta categoria. Tal modelo requer que sejam definidos limites máximos e mínimos (representados por $\boldsymbol{\delta_{min}}$ e $\boldsymbol{\delta_{max}}$) para as probabilidades de um atacante inferir a presença de um indivíduo em uma tabela.

3.3.6 Information Metric

A classe *Information Metric* (em português, métrica de informação) representa as métricas aplicadas para medir a utilidade da informação após a anonimização de uma base de dados. Existem três categorias de métricas de informação, representadas pelas classes presentes na Figura 5.

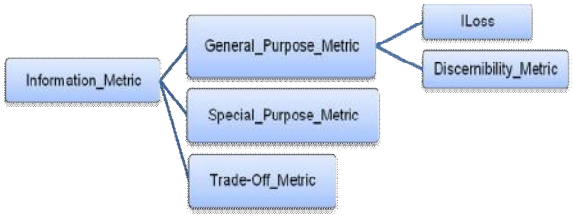

Figura 5. Classe *Information Metric* e suas subclasses.

A classe *General Purpose Metric* (em português, métrica de propósito geral) representa as métricas utilizadas quando a finalidade do receptor dos dados (aquele que analisa os dados) não é conhecida. Duas métricas de propósito geral são apresentadas na Figura 5 e representadas pelas classes ILoss e a *Discernibility Metric*.

A métrica ILoss mede a perda de informação devido a generalizações ou supressões, permitindo a atribuição de pesos aos quase identificadores para delimitar a quantidade máxima de perda tolerada, conforme a importância do atributo definida pelo publicador de dados [21].

A *Discernibility Metric* (em português, métrica da discernibilidade) mede a perda de informação, cobrando uma penalidade a cada registro idêntico quanto aos valores de quase identificadores encontrado.

A classe *Special Purpose Metric* (em português, métrica de propósito especial) representa as métricas utilizadas quando a finalidade do receptor de dados é conhecida [8]. Por fim, a classe *Trade-Off Metric* (em português, métrica de equilíbrio) representa as métricas utilizadas para aferir o equilíbrio entre a utilidade e a preservação de privacidade em bases de dados publicadas [8].

As métricas categorizadas como de equilíbrio ou de propósito específico e outras métricas de propósito geral não foram representadas aqui por limitações de espaço, porém serão adicionadas à ontologia posteriormente.

3.4 Etapa 04: Criação de propriedades

Para que a representação do domínio seja uma ontologia, é necessária a criação de relacionamentos entre os conceitos, ou seja, a criação de propriedades (correspondente à etapa 05 do método 101). Para isso, mais uma vez, serão utilizados termos e conhecimentos extraídos da literatura da área de preservação de privacidade em dados publicados.

Existem dois tipos de propriedades: as de objeto e as de tipos de dados. O primeiro tipo descreve relacionamentos entre classes (ou instâncias) e o segundo descreve características sobre uma classe (ou instância). As propriedades de objeto identificadas para esta ontologia são apresentadas a seguir:

operates_on (*Anonymization_Operation*, *Attribute_Type*): ou em português, opera_sobre (Operação_de_Anonimização, Tipo_de_Atributo). Estabelece que uma determinada operação de anonimização opera sobre um tipo de atributo específico, por exemplo:

- Value_Suppression ***operates_on*** *Explicit_Identifier*
- *Generalization* ***operates_on*** *Quasi_Identifer*
- *Generalization* ***operates_on*** *Sensitive Attribute*

generalize (*Hierarchy_Type*, *Data_Type*): ou em português, generaliza (Tipo_de_Hierarquia, Tipo de Dado). Define o(s) tipo(s) de hierarquia(s) de generalização adequado(s) a certos tipos de dados, por exemplo:

- *Interval* ***generalize*** *Date, Time, Integer and Decimal*
- *Ordering* ***generalize*** *Date, Time, Integer, Decimal, String and Ordered String*
- *Masking* ***generalize*** *Date, Time, Integer, Decimal, String and Ordered String*

Conforme dito anteriormente, o tipo de hierarquia *Ordering* é mais indicado para dados do tipo *String* e *Ordered String*. Essa especificação será indicada por meio de uma propriedade de anotação na ontologia.

employ (*Privacy_Model*, *Anonymization_Operation*): ou em português, emprega (Modelo_de_Privacidade, Operação_de_Anonimização). Define que um modelo de privacidade emprega uma ou mais operações de anonimização para a trans-

formação de uma base de dados original em uma base anonimizada, por exemplo:

- *k-Anonymity* ***employ*** *Generalization and Suppression*
- *ℓ-Diversity Principle* ***employ*** *Suppression*
- *t-Closeness* ***employ*** *Generalization*

apply_to (*Privacy_Model*, *Attribute_Type*): ou em português, aplica-se_a (Modelo_de_Privacidade, Tipo_de_Atributo). Define os modelos de privacidade adequados a cada tipo de atributo, por exemplo:

- *k-Anonymity* ***apply_to*** *Quasi_Identifier*
- *ℓ-Diversity Principle* ***apply_to*** *Sensitive_Attribute*
- *t-Closeness* ***apply_to*** *Sensitive_Attribute*

measures_loss (*Information_Metric*, *Anonymization_Operation*): ou em português, mede_perda (Métrica_de_Informação, Operação_de_Anonimização). Indica quais operações de anonimização são analisadas para o cálculo da perda de informação, também chamada perda de utilidade da informação. Por exemplo:

- *ILoss* ***measures_loss*** *Generalization and Suppression*
- *Discernibility_Metric* ***measures_loss*** *Generalization and Suppression*

A propriedade de tipo de dado extraída a partir da literatura foi ***has_parameter*** (*Privacy_Model*, *parameter-list*), ou em português, possui_parametro (Modelo_de_Privacidade, lista-de-parâmetros). Ela indica os parâmetros que modelos de privacidade possuem, por exemplo:

- *t-Closeness* ***has_parameter*** *parameter-list*
- *k-Anonymity* ***has_parameter*** *parameter-list*
- *ℓ-Diversity Principle* ***has_parameter*** *parameter-list*
- *Recursive (c, ℓ)-Diverse* ***has_parameter*** *parameter-list*

4. TRABALHOS RELACIONADOS

Foram encontrados três trabalhos que utilizam ontologias para representar e/ou melhorar métodos de anonimização.

O artigo [22] apresenta a ontologia denominada *Personal Health Information Ontology* (em português, Ontologia de Informação de Saúde Pessoal) que representa informações pessoais de atores da área da saúde (como médicos, enfermeiros, recepcionistas) e considera a lei canadense de proteção de dados pessoais PIPEDA (*Personal Information Protection and Electronic Documents Act*, ou em português, Lei de Proteção de Informações Pessoais e Documentos Eletrônicos). Foi desenvolvida e integrada ao modelo *k-anonymity* objetivando solucionar as fragilidades deste último. Também foi criado um *software* para indicar informações que devem ser suprimidas ou generalizadas para preservar a privacidade individual [22].

A pesquisa [23] apresenta a *Ontology Generalization Hierarchy* (em português, Ontologia para Hierarquia de Generalização) que realiza, a partir de inferências, a mínima generalização possível para cada par de quase-identificadores, até atingir a última coluna de uma tabela, respeitando-se o parâmetro ***k*** configurado para a aplicação do modelo *k-Anonymity*. Esse método não utiliza supressão e por isso melhora a utilidade dos dados, resultando em uma tabela minimamente generalizada (MGT ou, em inglês, *Minimal Generalized Table*) [23].

A tese de doutorado [20] descreve o desenvolvimento de um *framework* geral que usa tesauros e ontologias existentes (baseadas em *WordNet* e SNOMED CT) e métodos novos ou melhorados de anonimização e controle de descoberta estatística (do inglês, *Statistical Disclosure Control* ou SDC) para anonimizar atributos categóricos, considerando a similaridade semântica e a distribuição de valores ao longo da base de dados. Originalmente, métodos SDC eram utilizados apenas com atributos numéricos, sendo estendidos por [20] para serem utilizados também em atributos categóricos. Os resultados obtidos comprovaram que o uso de semântica melhora significativamente a utilidade e qualidade dos dados anonimizados.

5. DISCUSSÃO E CONCLUSÃO

Embora os objetivos dos autores pesquisados na seção 4 sejam distintos dos objetivos da ontologia proposta neste artigo, o domínio é o mesmo e, portanto, foram realizadas buscas na Internet a fim de obter versões das ontologias descritas em [20], [21] e [22] para uma possível reutilização, porém tais versões não foram encontradas. No caso de [22], por exemplo, tentou-se contato com os autores por e-mail[3], mas não houve resposta até o momento.

A partir de análise realizada no Repositório de Vocabulários e Ontologias do Governo Eletrônico [5], verificou-se que há uma ascensão do uso de ontologias para representar estruturas organizacionais e funcionamento de órgãos públicos brasileiros, como é o caso do Modelo Ontológico da Classificação das Despesas do Orçamento Federal Brasileiro – que representa conceitos da área orçamentária e automatiza o tratamento de dados do orçamento público federal [24] – e do Esboço de Modelagem Conceitual para Estruturas Organizacionais Governamentais Brasileiras e o SIORG (Sistema de Informações Organizacionais do Governo Federal), que é um esboço de uma ontologia para representar as estruturas organizacionais governamentais com o objetivo de "possibilitar a interoperabilidade com outras esferas e poderes de governo" [25].

Como a ontologia apresentada neste artigo possui domínio e escopo diferentes das ontologias analisadas em [5] – visto que se propõe a representar conceitos da área de preservação de privacidade e não estruturas organizacionais de instituições públicas–, a reutilização das ontologias presentes no Repositório de Vocabulários e Ontologias do Governo Eletrônico se tornou inviável.

Ainda assim, pretende-se – após finalização e validação da ontologia a partir de sua instanciação em uma base de dados real –, disponibilizar a ontologia apresentada neste artigo para o Repositório de Vocabulários e Ontologias do Governo Eletrônico, a fim possibilitar seu uso pela administração pública em todas as esferas de governo (Federal, Estadual e Municipal), objetivando a aplicação padronizada de métodos sistemáticos de anonimização às bases de dados governamentais, conforme recomendações da

[3] **K-Anonymity Privacy Protection Using Ontology**. Mensagem enviada por jane.mjq@gmail.com em 14 de mar. de 2016.

LAI e do Guia de Abertura de Dados, em prol da transparência e da preservação da privacidade individual.

6. REFERÊNCIAS

[1] Brasil. Lei nº 12.527, de 18 de novembro de 2011. *Diário Oficial [da] República Federativa do Brasil*, Brasília, DF, 18 nov. 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm>. Acesso em: 09 out. 2015.

[2] Open Government Partnership (Brasil). *2º Plano de Ação Brasileiro*. Versão traduzida. 2013. Disponível em: <http://edemocracia.camara.gov.br/documents/980199/980230/2%C2%BA%20Plano+de+A%C3%A7%C3%A3o/>. Acesso em: 07. jun. 2015.

[3] Controladoria-Geral da União. *Guias e orientações*. Disponível em: <http://www.acessoainformacao.gov.br/lai-para-sic/sic-apoio-orientacoes/guias-e-orientacoes>. Acesso em: 03 jul. 2015.

[4] Secretaria de Logística e Tecnologia da Informação. *Portal Brasileiro de Dados Abertos*. Disponível em: <http://dados.gov.br/>. Acesso em: 25 jan. 2016.

[5] *Repositório de Vocabulários e Ontologias do Governo Eletrônico*. Disponível em: <http://vocab.e.gov.br/>. Acesso em: 25 jan. 2016.

[6] Ministério do Planejamento, Orçamento e Gestão. *Conheça o Programa de Governo Eletrônico Brasileiro*. Disponível em: <http://www.governoeletronico.gov.br/o-gov.br>. Acesso em: 02 jul. 2015.

[7] Governo Eletrônico. *0067 - Consulta Pública para Validação das Orientações e sugestão de novas ideias ao Guia de Abertura de Dados da Infraestrutura Nacional de Dados Abertos - INDA*. 2012. Disponível em: <https://www.consultas.governoeletronico.gov.br/ConsultasPublicas/consultas.do?acao=exibir&id=93>. Acesso em: 09 jun. 2015.

[8] Fung, B. C. M. et al. *Introduction to privacy-preserving data publishing: concepts and techniques*. CRC Press, 2010.

[9] Queiroz, M. J.; Motta, G. H. M. B. Privacidade e Transparência no Setor Público: Um Estudo de Caso da Publicação de Microdados do INEP. In: *XV Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais* – SBSeg, 2015. Florianópolis. Anais... Porto Alegre: Sociedade Brasileira de Computação, 2015. Resumos Estendidos, p. 362-365. Disponível em: <http://sbseg2015.univali.br/anais/>. ISSN: 2176-0063.

[10] Rother, Edna Terezinha. Systematic literature review X narrative review. *Acta Paulista de Enfermagem*, v. 20, n. 2, p. v-vi, 2007.

[11] Noy, N.; Mcguinness, D. *Ontology development 101*: A guide to creating your first ontology. Development, v. 32, p. 1–25, 2001.

[12] Russell, S.; Norvig, P. *Artificial Intelligence*: A Modern Approach. 3rd ed. New Jersey: Prentice Hall, 2002, 1132 p.

[13] University of Texas AT Dallas. *UT Dallas Anonymization Toolbox*. Disponível em: <http://cs.utdallas.edu/dspl/cgi-bin/toolbox/anonManual.pdf>. Acesso em: 02 ago. 2015.

[14] Cornell University. *CAT*: The Cornell Anonymization Toolkit Introduction. 2011. Disponível em: <http://sourceforge.net/projects/anony-toolkit/files/Documents/cat-mannual-1.0.PDF/download>. Acesso em: 2 ago. 2015.

[15] Prasser, F.; Kohlmayer, F.; Lautenschlaeger, R.; Kuhn, K. A. ARX – A Comprehensive Tool for Anonymizing Biomedical Data. In: *Proceedings of the AMIA 2014 Annual Symposium*, November 2014, Washington D.C., USA.

[16] Sweeney, L. K-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, v. 10, n. 05, p. 557–570, 2002.

[17] Machanavajjhala, A.; Gehrke, J.; Kifer, D.; Venkitasubramaniam, M. ℓ-Diversity: Privacy beyond k-anonymity. In: *Proceedings – InternationalConference on Data Engineering*, p. 1-12, 2006.

[18] Li, T. L. T.; Li, N. L. N. Optimal k-Anonymity with Flexible Generalization Schemes through Bottom-up Searching. In: *Sixth IEEE International Conference on Data Mining - Workshops* (ICDMW'06), 2006.

[19] Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Microdados para download*. 2011. Disponível em: <http://portal.inep.gov.br/basica-levantamentos-acessar>. Acesso em: 27 fev. 2015.

[20] Lluís, S. M. *Ontology Based Semantic Anonymisation of Microdata*. 2013. Tese de Doutorado. Universitat Rovira i Virgili.

[21] Fletcher, S.; Islam, M. Z. Measuring Information Quality for Privacy Preserving Data Mining. *International Journal of Computer Theory and Engineering*, v. 7, n. 1, p. 21-28, 2015.

[22] Omran, E.; Bokma, A.; Abu-Almaati, S. A K-anonymity Based Semantic Model For Protecting Personal Information and Privacy. In: *Advance Computing Conference*, 2009. IACC 2009. IEEE International. IEEE, 2009. p. 1443-1447.

[23] Talouki, M. A.; Nematbakhsh, M.; Baraani, A. K-anonymity privacy protection using ontology. In: *Computer Conference*, 2009. CSICC 2009. 14th International CSI. IEEE, 2009. p. 682-685.

[24] Secretaria de Orçamento Federal. *Modelo ontológico da Classificação das Despesas do Orçamento Federal Brasileiro*. 2013. Disponível em: <http://vocab.e.gov.br/2013/09/loa>. Acesso em: 11 jan. 2016.

[25] Batista, A. H. et. al. *Esboço de Modelagem Conceitual para Estruturas Organizacionais Governamentais Brasileiras e o SIORG*. 2011. Disponível em: <http://vocab.e.gov.br/2011/09/org#ref-1>. Acesso em: 11 jan. 2016.