2 Fundamentos

Este capítulo aborda os fundamentos principais aplicados neste trabalho.

2.1 Linked Data

Linked Data é um padrão de práticas a serem seguidas para a publicação e interligação de dados estruturados na Web. Seguindo esse padrão, eles precisam ser publicados com características específicas: machine-readable, semântica explicitamente definida, e a possibilidade de serem referenciados e referenciar conjuntos de dados externos.

Com o objetivo de distribuí-los em uma escala global de maneira estruturada, Berners-Lee (2009) definiu quatro regras conhecidas como “Princípios de Linked Data”, listadas a seguir.

1. Utilize Uniform Resource Identifiers (URIs) para nomear entidades;
2. Utilize URIs HTTP de forma que pessoas possam consultar esses nomes;
3. Quando alguém acessar uma URI, providencie informações úteis utilizando padrões (RDF, SPARQL);
4. Inclua ligações para outras URIs, a fim de permitir que outras entidades sejam descobertas.

A ideia do primeiro princípio é a criação de URIs não apenas para documentos, mas também para conceitos do mundo real, tanto concretos quanto abstratos (pessoas, sentimentos, cores, etc.).

O segundo princípio determina a utilização de URIs que sigam o protocolo HTTP para possibilitar a publicação de mais informações associadas a uma determinada entidade. Tais informações precisam ser acessíveis, e o protocolo HTTP é o principal protocolo para acesso a informações na Internet.

O terceiro princípio defende o uso de padrões, pois essa utilização permite que diferentes tipos de aplicações possam, de alguma forma, processar o conteúdo publicado.

Por fim, o quarto princípio pode ser considerado como o mais importante para o escopo do nosso trabalho. Nele, é incentivada a criação de ligações entre os dados publicados na Web. Tais ligações devem ser criadas não apenas entre documentos, mas principalmente entre entidades publicadas. Além disso, devem ser criadas não apenas ligações com dados proprietários, mas também com dados publicados por fontes de dados externas, possibilitando uma maior expansão de conhecimentos. Essas ligações entre entidades podem ser distribuídas nos três grupos abaixo (Heath & Bizer, 2011).

- Ligações de relacionamento;
- Ligações de vocabulário;
- Ligações de identidade.

As ligações de relacionamento representam algum tipo de relação entre as duas entidades interligadas. Por exemplo, a representação de uma relação de paternidade entre uma entidade pai e uma entidade filho.

As ligações de vocabulário interligam a entidade com o vocabulário utilizado para descrevê-la. Por exemplo, a característica de uma entidade e o termo de uma ontologia que descreve o significado de tal característica.

Por fim, as ligações de identidade são utilizadas para interligar entidades equivalentes, mas representados por diferentes URIs. Tais ligações permitem a existência de diferentes interpretações para um mesmo termo. Esse último grupo foi escolhido como escopo do nosso trabalho.

2.2 Web de Dados

Atualmente, muitas organizações estão alterando radicalmente a forma de compartilhar conhecimento, diminuindo as barreiras para publicação de seus dados na Web (Bizer, Heath, & Berners-Lee, 2009). Elas adotaram os princípios de Linked Data como forma de publicá-los, resultando em um espaço global e bem definido chamado Web de Dados: um grafo composto por bilhões de dados publicados de vários domínios e fontes de dados distintos.

Publicar dados na Web como Linked Data envolve os seguintes passos básicos:

- Atribuir URIs às entidades descritas pelo conjunto de dados e utilizar estas URIs no protocolo HTTP para representações RDF;
- Criar ligações RDF para outras fontes de dados na Web, permitindo que os clientes naveguem a Web de Dados como um todo, seguindo tais ligações;
- Fornecer metadados sobre os dados publicados, permitindo que os clientes possam acessar a qualidade desses dados, e escolher entre as diferentes formas de acessá-los.

Partindo destes passos, uma etapa importante para a publicação na Web é a escolha de URIs. As entidades devem possuir URIs únicas, mas podem existir diferentes URIs para descrever uma mesma identidade. Tratam-se de URI aliases.

Os URI aliases são importantes, pois permitem diferentes descrições para um mesmo objeto. Com isso, é possível a existência de diferentes opiniões expressas na Web. Os diferentes fornecedores de dados interligam URI aliases - quando as conhecem - através da propriedade owl:sameAs, propriedade utilizada para indicar que duas URIs diferentes representam uma mesma entidade.

Ou seja, além da escolha de URIs, ligações RDF precisam ser publicadas para permitir a descoberta de novas bases. Elas permitem a navegação entre diferentes fontes e acesso a informações relacionadas. Sendo assim, para fazer parte da Web de Dados, as fontes publicadoras devem criar ligações para entidades relacionadas que estejam em outras fontes. Devido ao grande número de entidades que normalmente uma fonte possui, é comum a geração desse tipo de ligação de maneira automática ou semiautomática.

Para facilitar a geração dessas ligações entre diferentes bancos de dados publicados, uma abordagem é utilizar nomes conhecidos como identificadores dos dados (por exemplo, utilizar o International Standard Book Number (ISBN) quando o domínio de banco for de publicações). Quando é utilizado um esquema de nomes conhecidos, ligações implícitas entre entidades de diferentes fontes podem se tornar explícitas com as ligações geradas.

Há casos em que as fontes de dados não utilizam essa nomenclatura compartilhada. Nestes casos, é comum as ligações RDF serem geradas baseadas na similaridade das entidades dos diferentes bancos de dados. Existem vários

frameworks para identificação de entidades similares, ou seja, possíveis ligações. Na maioria, por exemplo, é possível especificar que tipo de ligações RDF serão criadas (através de uma linguagem declarativa); que combinação de similaridade de métricas é utilizada para comparar entidades; e como valores de similaridade para propriedades específicas são agregadas em um valor mais geral. Tais frameworks contribuem para uma Web de Dados cada vez mais conectada.

2.3 OWL e ligações owl:sameAs

OWL, uma sigla para Web Ontology Language, é desenvolvida como uma extensão do vocabulário RDF, acrescentando novas propriedades que permitem expressar mais características dos dados publicados na Web (por exemplo, cardinalidades e equivalências) (Bechhofer, et al., 2004).

Dentre essas novas propriedades, destacamos a de equivalência. Na Web, não é possível determinar que duas URIs distintas correspondem a entidades diferentes ou equivalentes, a menos que essa característica esteja publicada. Para publicar esse tipo de característica, OWL fornece duas propriedades sobre a identidade das entidades:

- owl:differentFrom
- owl:sameAs

A propriedade owl:differentFrom indica que duas URIs referem-se a entidades diferentes.

Em contrapartida, a propriedade owl:sameAs, utilizada no escopo deste trabalho, permite representar que duas entidades são, de alguma forma, equivalentes. Ou seja, uma declaração sameAs entre duas URIs indica que ambas referem-se a uma mesma entidade e possuem uma mesma identidade.

A Figura 1 contém um exemplo do uso desse tipo de declaração para afirmar que duas URIs correspondem a uma mesma pessoa. No exemplo, as URIs “Luiz_Inacio_Lula_da_Silva” e “Presidente_Lula” descrevem uma mesma entidade. Ou seja, suas propriedades podem ser agrupadas.

```
<rdf:Description rdf:about="#Luiz_Inacio_Lula_da_Silva">
  <owl:sameAs rdf:resource="#Presidente_Lula"/>
</rdf:Description>
```

Figura 1: Exemplo de tripla RDF contendo a propriedade owl:sameAs

A partir do aumento do número de fontes de dados publicadas na Web como Linked Data, a propriedade owl:sameAs tem sido cada vez mais utilizada para conectar dados equivalentes e publicados por diferentes fontes (Halpin & Hayes, 2010).

```
<owl:Class rdf:ID="FootballTeam">
  <owl:sameAs rdf:resource="http://sports.org/US#SoccerTeam"/>
</owl:Class>
```

Figura 2: Exemplo de uso alternativo da propriedade owl:sameAs

Essa propriedade possui ainda outra aplicação. Em OWL Full¹, podemos utilizá-la para definir igualdade de classes quando dois conceitos possuem o mesmo significado. No exemplo apresentado na Figura 2, a ligação owl:sameAs entre as classes “FootballTeam” e “SoccerTeam” define que ambas denotam o mesmo conceito, apesar de serem de ontologias diferentes.

2.4 Uma crítica ao uso da propriedade owl:sameAs

A utilização de ligações owl:sameAs é imprescindível para a Web de Dados. No entanto, existe uma suspeita de mau uso dessas ligações, uma vez que podem determinar que duas entidades são “de alguma forma” equivalentes, permitindo diferentes interpretações. Halpin & Hayes (2010) delineiam quatro interpretações alternativas para a equivalência entre entidades:

- **Igual, mas referencialmente opaca:** são equivalentes, mas nem todas as propriedades de uma são aceitas pela outra, não podendo se substituir;
- **Igual, mas em contexto diferente:** são equivalentes e todas as propriedades de uma são aceitas pela outra, mas são de contextos diferentes e não podem ser reutilizadas em outros contextos;

¹ OWL atualmente possui três sublinguagens, em ordem de expressividade: OWL Lite, OWL DL e OWL Full. Elas foram projetadas para uso por comunidades específicas de programadores e usuários.

- **Representação:** não são equivalentes, sendo uma a representação da outra. Por exemplo, uma ligação entre uma pessoa e seu Web site;
- **Muito semelhante:** não são equivalentes, sendo apenas parecidas.

Obviamente, a questão de como expressar as relações de identidade entre dados é mais complexa do que apenas a aplicação de ligações de equivalência. Ao mesmo tempo, uma abordagem mais detalhada que cubra essas quatro possíveis interpretações seria um passo útil para a evolução da Web de Dados.