

UNIVERSIDADE FEDERAL DE SANTA CATARINA  
CENTRO TECNOLÓGICO  
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA  
CURSO DE SISTEMAS DE INFORMAÇÃO

Luiz Felipe Köhler

UMA PROPOSTA DE ONTOLOGIA DE PROVENIÊNCIA PARA  
PUBLICAÇÃO DE LINKED DATA

Florianópolis – SC  
2012

UNIVERSIDADE FEDERAL DE SANTA CATARINA  
CENTRO TECNOLÓGICO  
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA  
CURSO DE SISTEMAS DE INFORMAÇÃO

UMA PROPOSTA DE ONTOLOGIA DE PROVENIÊNCIA PARA  
PUBLICAÇÃO DE LINKED DATA

Trabalho de conclusão de curso  
apresentado como parte dos requisitos  
para obtenção do grau de Bacharel em  
Sistema de Informação.

AUTOR: LUIZ FELIPE KOHLER  
ORIENTADOR: PROF. DR. JOSÉ LEOMAR TODESCO  
CO-ORIENTADOR: PROF. DR. FERNANDO OSTUNI GAUTHIER  
CO-ORIENTADOR: ME. AIRTON ZANCANARO

Florianópolis - SC  
2012

Luiz Felipe Köhler

UMA PROPOSTA DE ONTOLOGIA DE PROVENIÊNCIA PARA  
PUBLICAÇÃO DE LINKED DATA

Trabalho de conclusão de curso  
apresentado como parte dos requisitos  
para obtenção do grau de Bacharel em  
Sistema de Informação.

BANCA EXAMINADORA

---

PROF. DR. JOSÉ LEOMAR TODESCO – UFSC

---

PROF. DR. FERNANDO OSTUNI GAUTHIER – UFSC

---

ME. AIRTON ZANCANARO - UFSC

## **AGRADECIMENTOS**

Inicialmente registro meus sinceros agradecimentos ao Professor Dr. José Leomar Todesco, por ter me orientado e permitido a realização deste trabalho.

Aos membros da banca, Prof. Dr. Fernando Ostuni Gauthier e Me. Airton Zancanaro, pelo apoio, críticas e sugestões.

À minha mãe, Lori Teresinha Köhler, por todo amor e carinho, pelas orações em meu favor, pela preocupação para que eu estivesse sempre no caminho correto.

Aos professores do curso de Sistemas de Informação/UFSC pelas excelentes aulas e pela ajuda sempre bem vinda, assim como todos os funcionários da universidade, que contribuíram em nossa trilha por esta etapa.

Aos colegas de turma, colegas de trabalho, amigos e familiares que de forma direta ou indireta, colaboraram.

## RESUMO

A Web Semântica é um novo passo no desenvolvimento da internet, marcado principalmente pela organização do conteúdo e pela interação inteligente do usuário com o material disponibilizado na rede. Em 2006 Tim Berners-Lee criou o conceito de *linked data*, o qual descreve um método de publicação de dados bem estruturados e interligados a outros dados, ou repositórios, de modo que possam ser consumidos por máquinas e humanos mais facilmente. Esse conceito tem ganhado força principalmente por ações de publicação de dados abertos e transparência governamental. Entretanto, a dificuldade em harmonizar informações extraídas de conjuntos de dados distintos, somados a questionável qualidade dos dados na Web, impulsiona um melhor controle no processo de publicação da informação. A proveniência de dados pode ser uma solução para esta problemática, pois visa o registro e a comunicação de todos os processos relacionados ao ciclo de vida dos dados. Para tratar esse problema, o presente trabalho tem como objetivo o levantamento dos estudos relacionados a este cenário e propor um modelo de proveniência para aplicação em dados abertos publicados no formato de *Linked Data*. Foram desenvolvidos dois conjuntos de ontologias para modelar o processo de Linked Data e aplicar práticas de proveniências sobre processos. Com base nestes modelos, partiu-se para a criação de instâncias da ontologia, e assim consultas de inferência foram realizadas demonstrando sua aplicabilidade.

**Palavras-chave:** web-semântica, linked data, proveniência, ontologia.

## ABSTRACT

The Semantic Web is a new step in the development of the Internet, mainly marked by the organization of the content and intelligent user interaction with the material available on the network. In 2006 Tim Berners-Lee created the concept of linked data, which describes a method of publishing data well structured and interlinked with other data, or repositories, so that they can be consumed by humans and machines more easily. This concept has gained strength mainly by shares publishing open data and government transparency. However, the difficulty in harmonizing information drawn from different data sets, plus the questionable quality of data on the Web drives better control the process of publishing information. The provenance of data can be a solution to this problem, since it seeks registration and communication of all processes related to the lifecycle of the data. To address this problem, this paper aims to survey the studies related to this scenario and propose a model of provenance for use in open data published in Linked Data format. We developed two sets of ontologies for modeling the process of Linked Data sources and apply practices on processes. Based on these models, we started with the creation of instances of ontology, and so inference queries were performed demonstrating its applicability.

**Keywords:** semantic web, linked data, provenance ontology.

## LISTA DE ILUSTRAÇÕES

Figura 2-7 Cenário de troca de informações na gestão de transplantes de órgãos.....	27
Figura 2-8 Tratamento de informações em um processamento estatístico .....	28
Figura 2-9 Diagrama para representação de derivação e proveniência de dados. Fonte: Autor, adaptado de .....	29
Figura 2-10 Representação de Vocabulário para Open Provenance .....	40
Figura 2-11 Representação das Estruturas Básicas e seus relacionamentos.....	41
Figura 2-1 Arquitetura de camadas da Websemântica .....	47
Figura 2-2 Processo de criação de Linked Data segundo Hyland .....	50
Figura 2-3 Processo de criação de Linked Data segundo Hausenblas... ..	50
Figura 2-4 Processo de criação de Linked Data segundo Villazón-terrazas .....	51
Figura 2-5 Processo de desenvolvimento da metodologia On-to-Knowledge .....	60
Figura 2-6 Diagrama de Representação do vocabulário DCAT .....	63
Figura 3-1 Listagem de Termos identificados para a Ontologia A. ....	67
Figura 3-2 Lista de Classes Ontologia A.....	68
Figura 3-3 Lista de Propriedades definidas para a ontologia A .....	69
Figura 3-4 Diagrama de classes e relacionamentos da Ontologia A .....	70
Figura 3-5 Lista de Termos da Ontologia B.....	72
Figura 3-6 Lista de Termos da Ontologia B (Continuação) .....	73
Figura 3-7 Lista de Classes Ontologia B.....	74
Figura 3-8 Lista de Classes Ontologia B(Continuação) .....	75
Figura 3-9 Figura 3 8 Lista de Classes Ontologia B(Continuação2).....	76
Figura 3-10 Lista de propriedades da ontologia B .....	77
Figura 3-11 Diagrama de classes e relacionamentos da Ontologia B ...	78
Figura 3-12 Criação de Instancias para o Projeto Linked Data Enem... ..	80
Figura 3-13 Criação das Instâncias para Etapa de Identificação de Dados .....	81

Figura 3-14 Criação das Instancias para Representação das Entidades de Dados gerados na etapa de Conversão .....	82
Figura 3-15 Consulta sobre Atividades que ocorreram no projeto Linked Data Enem.....	83
Figura 3-16 Consulta sobre as Entidades que Influenciaram o Dataset Linked Data Enem .....	84
Figura 3-17 Agentes (Organizações, Pessoas, Softwares) envolvidos no processo. ....	85



## **LISTA DE TABELAS**

Tabela 2-1 Resumo de Conceitos das Estruturas do Núcleo do modelo PROV-DM .....	42
---	----

## LISTA DE ABREVIATURAS E SIGLAS

**GRDDL** - *Gleaning Resource Descriptions from Dialects of Languages*

**HTML** - HyperText Markup Language

**JSON** - JavaScript Object Notation

**OWL** - Web Ontology Language

**RDF** – *Resource Description Framework* (Framework para Descrição de Recursos, em tradução livre), modelo de dados da Web de dados.

**RDF/XML** – Representação de um modelo de dados RDF pela linguagem XML.

**RDFa** - The Resource Description Framework in Attributes. Forma de nomear o conteúdo de modo a descrever um tipo específico de informação.

**R2RML** - RDB to RDF Mapping Language

**RDB2RDF** – Relational Database to RDF (Banco de dados relacional para RDF), termo que refere-se à transformação de dados em bancos de dados relacionais em DF.

**SGBD** – Sistema Gerenciador de Banco de Dados.

**SKOS** - SKOS é uma área de trabalho de desenvolvimento de especificações e normas para apoiar o uso de sistemas de organização do conhecimento (KOS), como tesouros, esquemas de classificação, listas de cabeçalhos de assunto e taxonomias no âmbito da Web Semântica.

**SPARQL** - *SPARQL Protocol and RDF Query Language*

**URI** – *Universal-Uniform Resource Identifier*

**Web** – World Wide Web (Rede de Alcance Mundial, em tradução livre) também conhecida como WWW.

**W3C** – *World Wide Web Consortium*, Consorcio responsável pela padronização da Web e de tecnologias para esta.

**WWW** – World Wide Web, rede de alcance mundial.

**XML** – *eXtensible Markup Language*, linguagem de marcação extensível utilizada para descrição e seriação de dados semiestruturados.

## SUMÁRIO

1.	INTRODUÇÃO .....	14
1.1.	Definição do problema.....	16
1.2.	Objetivos .....	17
1.2.1.	Objetivo Geral.....	17
1.2.2.	Objetivos Específicos.....	17
1.3.	Justificativa .....	17
1.4.	Método de Pesquisa.....	18
1.5.	Resultados Esperados.....	18
1.6.	Limitações do Trabalho.....	18
1.7.	Organização dos capítulos.....	18
2.	FUNDAMENTAÇÃO TEÓRICA .....	20
2.1.	Dados Abertos.....	21
2.1.1.	Lei de Acesso à Informação .....	21
2.1.2.	Transparência Ativa .....	22
2.2.	Proveniência.....	22
2.2.1.	Proveniência de Dados.....	23
2.2.2.	Uso de Proveniência.....	24
2.2.3.	Cenários de Aplicação.....	27
2.2.4.	Captura de Proveniência .....	28
2.2.5.	Rastreabilidade de Dados.....	30
2.2.6.	Captura e Rastreabilidade na Prática.....	30

2.2.7.	Modelos Conceituais para Proveniência .....	30
2.3.	Licenças para distribuição ou publicação de dados .....	43
2.3.1.	Direito Autoral.....	43
2.3.2.	Copyright .....	43
2.3.3.	Copyleft .....	43
2.3.4.	Creative Commons .....	44
2.4.	Web Semântica .....	45
2.4.1.	Arquitetura de camadas.....	46
2.5.	Linked Data.....	48
2.5.1.	Publicação de Dados em Linked Data .....	49
2.5.2.	Práticas para Publicação de Linked Data.....	50
2.6.	Ontologias.....	55
2.6.1.	Metodologias e Ferramentas para o Desenvolvimento de Ontologias.....	56
2.6.2.	Ontologias Disponíveis.....	62
2.7.	Considerações Finais .....	63
3.	<b>PROPOSTA DE UMA ONTOLOGIA DE PROVENIÊNCIA PARA PUBLICAÇÃO DE LINKED DATA .....</b>	<b>64</b>
3.1.	Procedimentos Metodológicos.....	64
3.2.	Revisão da Literatura .....	64
3.3.	Análise das metodologias de desenvolvimento de ontologias	64
3.4.	Ontologia sobre o processo de Linked Data .....	65
3.4.1.	Processo de Desenvolvimento .....	65
3.4.2.	Ontologia criada.....	70

3.5.	Ontologia sobre proveniência de dados publicados em Linked Data	71
3.5.1.	Processo de Desenvolvimento.....	71
3.5.2.	Ontologia criada.....	78
3.6.	Aplicar o modelo de proveniência proposto.....	79
3.6.1.	Cenário de uso.....	79
3.6.2.	Representação do projeto de publicação de dados do ENEM em Linked Data.....	79
3.6.4.	Consultas de Inferência.....	83
3.7.	Resultados.....	85
4.	CONCLUSÃO.....	87
4.1.	Considerações finais.....	87
4.3.	Trabalhos Futuros.....	88
	REFERÊNCIAS.....	89
	GLOSSÁRIO.....	92
	APÊNDICES.....	93
	APÊNDICE A: PERGUNTAS DE COMPETÊNCIA.....	93
	Apendice B: Código Fonte das Ontologias.....	95

# 1. INTRODUÇÃO

Na última década, com a ascensão da internet 2.0, o mundo presenciou e conheceu o poder e a força da colaboração. Tivemos um grande exemplo de sinergia, onde o coletivo é maior que a soma de suas partes.

Percebeu-se que ao disponibilizar uma informação, um dado, ou um conhecimento, outras pessoas teriam acesso a ele e poderiam acrescentar suas experiências e conhecimentos relacionados ao mesmo, dessa forma, melhorando-o e criando coisas maiores, pôde-se evoluir coletivamente.

Um bom exemplo disso é o portal colaborativo Wikipédia, no qual, pessoas do mundo todo (especialistas ou não), ajudam a descrever tudo e a todos, de tal forma que se crie a maior enciclopédia digital da história. Hoje, a Wikipédia em português possui 738.678 artigos válidos, e 1.010.711 usuários registrados mundialmente.

A internet passou a ser a maior fonte de conhecimento disponível, livros e documentos começaram a serem publicados digitalmente, vídeos e músicas divulgados em larga escala.

Percebendo-se desse poder coletivo, organizações de vários países viram neste cenário uma oportunidade; utilizar a internet, a colaboração mutua para ajudar a governar, a resolver problemas de pequena ou grande magnitude, explorar as necessidades e soluções de cada cidadão. Mas para que isto aconteça de maneira efetiva, estas organizações precisam de alguma forma, expor suas necessidades, seus problemas, seus dados na rede, de forma que possam ser consumidos, utilizados, reaproveitados.

Mas como executar isso? Em meio a tanto conteúdo, como procurar uma informação específica? Quais dados são corretos? Devido ao fato de o conteúdo da web ser pouco estruturado, utilizar estes dados de maneira inteligente torna-se um desafio.

Desta necessidade, surgiu o movimento da Web Semântica, idealizado por Tim-Burnes Lee, o mesmo que criou a internet. Este movimento visa dar significado e contexto aos dados publicados na internet, de forma que seja possível o uso de mecanismos de inferência sobre estes dados. Trata-se de um conjunto de regras que indicam um formato de publicação de dados na internet, de forma que uma informação ou dado explicita de maneira formal o contexto em que se encontra.

Num cenário como este, soluções para inúmeros problemas podem vir à tona, por exemplo: quando precisarmos fazer uma busca de um dado específico, seja possível delimitar de forma estruturada, os atributos relacionados a este dado, deixando a busca mais inteligente e objetiva.

Mais do que encontrar uma informação, ao realizarmos uma pesquisa em uma fonte de dados, como a Web, por exemplo, necessitamos extrair informações de acordo com aquele domínio de interesse. Tão importante quanto à informação adquirida, são os dados que dela fazem parte e como tais dados foram obtidos, manipulados e registrados. Desta forma, pode-se analisar estes dados de forma mais ampla, realizando associações, gerando conteúdo ou conhecimento novo.

Então não basta simplesmente encontrar um dado, é necessário que estejam disponíveis todas as referências e as informações relacionadas a este dado, e estes devem ser corretos, concretos e atualizados.

Como uma continuação da Web semântica, uma nova abordagem chamada *Linked Data foi publicada*, trata-se de uma proposta para publicação de dados na web que, seguindo os conceitos da *Web Semântica*, visa estabelecer relações e o referencialmento dos dados na web.

O principal benefício do Linked Data se dá no ganho de valor agregado de um dado na web. Quanto maior o número de vínculos (diretos e indiretos) este dado tiver, mais informações poderão ser obtidas através do mesmo e maior será a sua utilidade.

Infelizmente, a qualidade dos dados publicados na Web é questionável, uma vez que, na maioria dos casos, estes dados não são submetidos a um processo de controle em sua publicação. Perdem-se assim valiosas informações a respeito dos dados, a exemplo de sua validade, sua origem, em qual contexto ou sob qual objetivo este dado está sendo publicado.

Este tipo de informação é essencial para que se possibilite um relacionamento efetivo entre dados de diferentes bases de dados na web.

Para se alcançar este patamar, é importante que todas estas informações referentes ao ciclo de vida relacionado a um dado sejam registradas, ou seja, que a história dos fatores que envolvem o processamento e a geração desses dados seja capturada de forma a prover informações sobre a proveniência de um dado. Essas informações podem auxiliar em diferentes contextos, já que podem ficar à disposição e não apenas em forma de registros internos de controle.

Acredita-se que a proveniência permite aos usuários compartilhar, descobrir e reutilizar os dados, facilitando atividades colaborativas, reduzindo a possibilidade da repetição de erros, e promovendo a aprendizagem.

### **1.1. Definição do problema**

Hoje em dia, uma grande quantidade de dados é publicada na Web, e constantemente surgem novas aplicações que se utilizam desses dados de maneiras inovadoras. Um próximo desafio que deve ser tratado nestas aplicações é a avaliação da qualidade dos dados recuperados a partir da Web.

A qualidade dos dados e informações representam uma área importante e que está amadurecendo no campo de Gestão de Sistemas de Informação. As organizações buscam melhores dados e informações de qualidade. Esta busca é repleta de desafios, como descobrir as dificuldades que envolvem a definição, medição, análise e melhoria de qualidade de dados e informações.

Segundo (Marins, 2008), não é possível dar crédito a um dado, informação ou experimento/aplicação se não há como repeti-lo. Em ciência, assim como em muitos outros campos, as informações relativas às metodologias e mecanismos pelos quais os resultados são obtidos podem ser tão importantes quanto os resultados propriamente ditos.

Um estudo recente mostra que um dos principais fatores que influencia a confiança dos usuários de conteúdo da Web é a proveniência (Y. Gil, 2007). Assim, uma abordagem comum para os dados de avaliação da qualidade é a análise de informações de proveniência.

Por outro lado, a interoperabilidade também é característica essencial e necessária à utilização de dados publicados na Web ou mesmo em um sistema de informação de propósito isolado. A capacidade de interoperar está diretamente ligada à estratégia adotada para a concepção do respectivo modelo de dados e do formato em que é disponibilizado.

Para uma aplicação, seja ela Web ou não, ser bem sucedida, esta deve atender qualitativamente e quantitativamente a seus usuários. Contudo, sob um olhar estratégico, esta aplicação deve ser capaz de interoperar com outras aplicações e sistemas existentes e futuros.



Para isto, a adoção de formatos padronizados seria uma solução mais plausível do que o alinhamento à posteriori dos esquemas de dados. Neste sentido, o uso de abordagens de Web Semântica e *Linked Data* tornam-se adequadas ao problema em questão.

## **1.2. Objetivos**

Este trabalho tem como objetivo realizar pesquisas sobre proveniência de dados para uso de dados abertos.

### **1.2.1. Objetivo Geral**

Propor uma ontologia de proveniência para o registro do processo de publicação de dados abertos no formato de Linked Data.

### **1.2.2. Objetivos Específicos**

- Apresentar as principais características dos dados abertos, linked data, proveniência de dados e seus assuntos relacionados.
- Realizar levantamento das principais propostas de proveniência de dados.
- Criar uma ontologia de proveniência para aplicação na publicação de dados abertos no formato de Linked Data.
- Aplicar o modelo de proveniência proposto em um experimento.

## **1.3. Justificativa**

A relevância deste trabalho está em conhecer a relação e o grau de aplicabilidade de proveniência sobre dados abertos. Neste sentido o trabalho ora apresentado procura trazer um panorama fidedigno sobre os conceitos e práticas relacionados ao tema.

Para ajudar a lidar com os desafios relacionados à qualidade de dados e informações no contexto da Web, as organizações podem recorrer a um crescente corpo de pesquisas na área de Linked Data, qualidade dos dados e informações, e proveniência, contidos neste

trabalho, que reúne conteúdo de importantes pesquisadores e profissionais do campo.

#### **1.4. Método de Pesquisa**

- Coletar e analisar de maneira crítica conceitos e dados relacionados ao estudo;
- Desenvolver um modelo que atenda aos objetivos especificados;
- Aplicar o modelo em um cenário de uso.

#### **1.5. Resultados Esperados**

Após a conclusão deste trabalho, espera-se obter avanço na disponibilidade de modelos de proveniência de dados publicados no formato Linked Data.

#### **1.6. Limitações do Trabalho**

Este trabalho limita a aplicabilidade de sua pesquisa a ambientes de publicação de dados que seguem os princípios de Linked Data. Apesar de que seus estudos provavelmente possam ser adaptados para outras formas de publicação de dados, estas não serão consideradas.

#### **1.7. Organização dos capítulos**

Após a introdução, este primeiro capítulo, que contém a apresentação resumida do tema e os procedimentos metodológicos, o trabalho é dividido, por motivos práticos e didáticos, em outros capítulos e as considerações finais.

Assim, no segundo capítulo, é apresentada a fundamentação teórica, que visa apresentar os conceitos e práticas relacionados ao tema. São abordadas as definições, programas, recursos, aplicações e benefícios da proveniência de dados e do uso de web semântica e Linked Data na atualidade.

O terceiro capítulo trata da proposta da pesquisa. Este capítulo apresenta os procedimentos metodológicos adotados, resultados obtidos a serem seguidos e cronograma de atividades a serem desenvolvidas.

O quarto capítulo trata do método e o desenvolvimento do experimento. Descreve os passos seguidos para a implementação do protótipo e avaliação do modelo proposto.

Em seguida, as considerações finais são elaboradas levando em conta todo o conteúdo deste trabalho, buscando cumprir os objetivos traçados e responder aos resultados esperados.

## 2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem como objetivo apresentar uma revisão de conceitos e práticas relacionados à Linked Data e Proveniência. O processo de desenvolvimento desta revisão baseou-se na realização de duas pesquisas à base de periódicos do Portal CAPES<sup>1</sup>. Para a execução destas pesquisas utilizou-se como estratégia de busca a utilização das palavras-chave em inglês “*linked data*”, para a primeira pesquisa, e “*provenance*” para a segunda. Em ambos os casos optou-se pela utilização de filtro de data de publicação para os últimos cinco anos. Como critério de seleção dos periódicos, os resultados obtidos foram ordenados por relevância e então uma análise e avaliação sobre os títulos e resumos dos primeiros 50 resultados foi realizada. Para cada periódico selecionado, houve também a realização de uma análise e avaliação sobre os periódicos citados em suas referências bibliográficas e trabalhos correlatos, no intuito de incluí-los no corpo de periódicos selecionados.

A seguir serão tratados os assuntos relacionados ao âmbito de Proveniência e *Linked Data*. Primeiramente serão abordados os conceitos de Dados Abertos, um movimento que impulsiona o uso de métodos padronizados na publicação de dados. Em seguida serão discutidos os domínios de proveniência, sua definição, seu uso e captura, e a utilização de modelos de aplicação universal e modelos de aplicação de proveniência para dados. Outro tópico pertinente se refere aos direitos relacionados à publicação de dados na Web. Neste sentido, serão tratadas as questões de direitos autorais, licenças de uso, dentre outras.

Posteriormente, serão apresentados os conceitos de Web Semântica, desde sua idealização, até a sua aplicabilidade em Linked Data propriamente. Em seguida, trataremos os aspectos de Dados Abertos, um movimento que impulsiona o uso de Linked Data e que é cada vez mais requisitado por organizações em diversos países. Será apresentado também um aprofundamento sobre ontologias, sobre os processos de seu desenvolvimento e algumas das principais ontologias disponíveis para re-úso e que podem ser relacionadas à Linked Data.

---

<sup>1</sup> Portal periodicos CAPES oferece acesso aos textos completos de artigos selecionados de mais de 21.500 revistas internacionais, nacionais e estrangeiras. <http://www.periodicos.capes.gov.br/>

## 2.1. Dados Abertos

A Open Definition<sup>2</sup> busca especificar o que exatamente representa a palavra “aberto” num contexto mais amplo, de Conhecimento Aberto, mas que contempla também os dados abertos. Para tal definição, conteúdos como musicas, fotos, filmes e livros; dados científicos, geográficos, históricos e outros; e informações governamentais e administrativas, são consideradas obras de conhecimento.

Segundo a Open Definition, “dado aberto é um dado que pode ser livremente utilizado, reutilizado e redistribuído por qualquer um”. Este dado deve seguir os seguintes princípios:

- **Disponibilidade e acesso:** o dado precisa estar disponível por inteiro e por um custo razoável de reprodução, preferencialmente por meio de download na Internet; também deve estar num formato conveniente e modificável.
- **Reuso e redistribuição:** o dado precisa ser fornecido em condições que permitam reutilização e redistribuição, incluindo o cruzamento com outros conjuntos de dados.
- **Participação universal:** todos podem usar, reutilizar e redistribuir, não havendo discriminação contra áreas de atuação, pessoas ou grupos (não são permitidas restrições como “não comercial”, que impedem o uso comercial, e restrições de uso para certos fins, como “somente educacional”).

(Bianco, 2011) sustenta que uma área que pode se beneficiar muito dessa construção coletiva de conhecimento, por exemplo, é a científica. Outra área que apresenta um grande movimento de publicação de dados abertos, provavelmente o mais notável, é a governamental. Possivelmente por esse ser um dos mais visíveis, e pelo seu impacto social.

### 2.1.1. Lei de Acesso à Informação

Em 16 de maio de 2012, é a data de início de vigência da Lei nº 12.527/2011, também conhecida como Lei de Acesso à Informação. A Lei permite que qualquer cidadão, sem necessidade de justificativa,

---

<sup>2</sup> Open Definition (<http://opendefinition.org/>)

solicite dados e informações a qualquer órgão ou entidade pública dos poderes Executivo, Legislativo e Judiciário, além do Ministério Público, nas esferas Federal, Estadual e Municipal. No executivo federal, os pedidos serão recepcionados eletronicamente, pelo sistema e-SIC, ou fisicamente pelo Serviço de Informação ao Cidadão, setor especificamente designado para essa finalidade em cada órgão ou entidade. Estes terão 20 dias, prorrogáveis por mais 10, para fornecer as informações solicitadas ou uma justificativa para o seu não fornecimento – nesse caso somente sendo admissíveis as justificativas específicas previstas na Lei. As organizações públicas tiveram 180 dias para se adaptarem à Lei.

### 2.1.2. Transparência Ativa

Além do fornecimento de informações sob demanda do cidadão, a Lei de Acesso à Informação também prevê que os órgãos e entidades devem se antecipar aos pedidos e publicar seus dados e informações na internet – a chamada transparência ativa. Exige, ainda, que os dados sejam publicados, inclusive em formatos abertos e não-proprietários. Essencialmente, demanda a publicação de dados abertos, embora não utilize este termo diretamente.

O governo federal se antecipou à vigência da Lei, disponibilizado, no dia 4 de maio, no Portal Brasileiro de Dados Abertos, com 78 conjuntos de dados e 850 recursos.

O Portal Brasileiro de Dados Abertos<sup>3</sup> foi construído colaborativamente entre pessoas e organizações interessadas da Sociedade Civil e servidores públicos, colocando em prática o conceito de Governo Aberto. A construção do Portal foi um dos compromissos cumpridos pelo Brasil na Parceria para Governo Aberto<sup>4</sup>.

## 2.2. Proveniência

O termo **proveniência** possui várias definições. No dicionário Aurélio, proveniência é (i) ato ou efeito de proceder. (ii) lugar de onde alguém ou algo procede. (iii) origem. Já no dicionário Michaelis, pode ser definida como (iv) lugar de onde alguma coisa provém. (v) fonte,

---

<sup>3</sup> <http://br.okfn.org/2012/05/10/novo-portal-dados-gov-br-feito-pela-sociedade/>

<sup>4</sup> Parceria para Governo Aberto. <http://www.opengovpartnership.org/>

origem, procedência. Dessas definições encontradas, as que mais se aproximam ao contexto envolvido na descrição da história de uma informação são as definições (ii) e (iv). O registro da história de uma informação é importante para que se possam ser conhecidos todos os passos em que um dado fez parte ou a partir dos quais um dado foi gerado e para se saber os motivos que levaram um dado a fazer parte de um processo.

Os aspectos fundamentais da proveniência não ficam restritos aos dados presentes em sistemas de computação. Pelo contrário, a proveniência já vem sendo utilizada e aplicada há tempos em situações do mundo real, em exemplos como:

(i) Belas artes e artefatos históricos para atestar a originalidade de uma obra ou mesmo evitar falsificações. A proveniência nesses casos assegura qualidade superior ao produto (Figura 5a);

(ii) Indústria de alimentos e no agronegócio brasileiro (AKABANE et al., 2010) para acompanhar os processos fabris da linha de produção e a qualidade final dos produtos alimentícios (Figura 5b e 5c, respectivamente)

(iii) Negócios em geral (CURBERA et al., 2008), na Ciência e até mesmo em publicações científicas (Figura 5d), para assegurar a reprodutibilidade e a veracidade do experimento e servir de evidência em disputas de patentes (SCIENCE, 2006, FREW et al., 2010).

Em todos esses casos a proveniência tem o potencial de transformar sistemas, tornando-os mais confiáveis e auditáveis.

### **2.2.1. Proveniência de Dados**

Informações de proveniência sobre um item de dados é a informação sobre a história do item, a partir da sua criação, incluindo informações sobre suas origens.

Em termos de sistemas computacionais, proveniência representa a ancestralidade de um objeto digital. Ela pode ser descrita de várias formas dependendo do domínio onde é aplicada (SIMMHAN et al., 2005, SRIVASTAVA, VELEGRAKIS, 2007). Atualmente, existem vários sinônimos para o termo —Proveniência (por exemplo, Data lineage, Data tracking, Data Pedigree, Data Provenance e Provenance Metadata (LING, ÖZSU, 2009).

Segundo (Bose, et al., 2005), os principais benefícios da proveniência para a qualidade de dados são:

- Comunica a qualidade de dados: confiabilidade, adequação, acurácia, atualidade, redundância;
- Melhora a interpretação do dado em função do reconhecimento da fonte;
- Contribui para a justificativa do uso de um determinado dado;
- Reduz a possibilidade de erros no juízo da precisão do dado;
- Permite que usuários não especialistas em dados entendam os passos do processamento;
- Permite identificar o processo utilizado para a condução da criação de dados científicos;
- Permite atualização de dados a partir de visões relacionais;
- Permite a modificação de schemas de visões relacionais;
- Possibilita o uso de fontes de dados históricas.

Tais benefícios revelam a aplicabilidade direta da proveniência de dados para obtenção de métricas de qualidade de dados.

### 2.2.2. Uso de Proveniência

Sistemas de proveniência podem ser construídos para suportar inúmeras formas de uso, Goble resume várias aplicações de informação de proveniência como:

- **Qualidade dos Dados:** A proveniência pode ser utilizada para estimar a qualidade e confiança dos dados com base na fonte e transformações de dados. Ele também pode fornecer declarações à prova em dados derivação.
- **Auditoria:** Proveniência pode ser usada para rastrear a auditoria de dados, determinar o uso de recursos, e detectar erros nos dados de geração.
- **Replicação:** informações de proveniência detalhada podem permitir a repetição e derivação de dados, ajudar manter suas características, e ser uma receita para a replicação.
- **Atribuição de Direitos:** Autenticidade pode ajudar a estabelecer direitos de autor e propriedade dos dados, permitir a sua citação, e determinar a responsabilidade em caso de dados errados.
- **Informativo:** A utilização genérica de proveniência é a consulta com base em metadados de linhagem para a



descoberta de dados. Ela também pode ser navegada para fornecer um contexto para interpretar os dados.

#### *2.2.2.1. Qualidade de Dados*

Proveniência sobre um conjunto de dados permite que o usuário avaliar sua qualidade para a sua aplicação. A qualidade dos dados de origem é importante, pois erros introduzidos por dados falhos tendem a aumentar a medida que se propagam a dados deles derivados [8].

O nível de detalhe incluído na proveniência determina a extensão em que a qualidade dos dados pode ser estimada. Metadados básicos sobre os dados, tais como as transformações aplicadas para criá-lo ou a fonte de dados de origem, pode auxiliar o usuário na criação da autenticidade dos dados e evitar fontes espúrias. Se o conhecimento semântico sobre a qualidade de dados é disponibilizada de forma autêntica, é possível avaliar automaticamente com base em métricas de qualidade e proporcionar um índice de qualidade usando técnicas de modelagem.

#### *2.2.2.2. Auditoria*

Proveniência pode servir como um meio para auditar os dados e o processo pelo qual foi produzido. Tal informação pode ser importante no estabelecimento de patentes sobre a descoberta de drogas ou para fins contábeis para empresas. Também pode ser usada para otimizar o processo de derivação, e para recolher estatísticas sobre o uso de recursos.

Proveniência sob a forma de uma linha de tempo de execução pode ajudar a verificar se qualquer exceção ocorreu na criação de dados. A utilização recorrente de proveniência deve voltar atrás e localizar os dados de origem ou processo que é a causa de erros encontrados nos dados obtidos e aplicar correções pertinentes.

#### *2.2.2.3. Replicação*

Informações de proveniência inclui os passos usados para derivar um conjunto de dados particular e pode ser pensado como um

receita para a criação que os dados. Se a proveniência contém detalhes suficientes sobre as operações, dados fontes, e parâmetros, pode ser possível repetir a derivação. Repetibilidade implica a disponibilidade de recursos semelhantes, como estava disponível quando o dado original foi criado. A derivação pode ser repetida para manter a atualidade de dados derivados, ou em casos de mudança da fonte ou se os módulos de processamento foram modificados. Tal uma reconstituição pode ser controlada para repetir apenas seções afetadas pela alteração de dados base ou do processo.

Pode ser possível e rentável para usar proveniência como meio de replicar os dados em vez de transportá-lo ou armazená-lo. Para que os dados derivados estejam fisicamente idênticos, várias dependências devem ser cumpridas, tais como o acesso aos mesmos dados de origem, processos e ambiente de processamento. Em alguns casos, como uma execução experimental, tal replicação de um byte-por-byte pode ser impossível, mas um produto de dado semanticamente equivalentes pode ser gerado. Tais propriedades podem ser estendidos para comparar dois conjuntos de dados apenas comparando sua linhagem, a reversão de mudanças, e engenharia reversa do dados.

#### *2.2.2.4. Atribuição de Direitos*

Informações de origem de um dado podem ajudar a determinar a posse desse. Os usuários podem procurar em uma árvore de derivação para analisar os criadores dos dados de origem e verificar a sua autoria. Do mesmo modo, criadores de propriedade intelectual podem olhar para baixo na cadeia de origem para ver quem está usando os dados que eles criaram. As citações são uma parte importante da publicação científicas e também pode ser usado como um meio de atribuição de responsabilidade em caso de erros no conjunto de dados.

#### *2.2.2.5. Informativo*

A utilização genérica de proveniência é como uma consulta a descrição de metadados os quais podem ser a base para a descoberta de conjuntos de dados, digamos pela procura com base em dados de origem ou um passo de processamento utilizado para gerá-los. Essas consultas podem localizar dados de interesse e evitar a duplicação de esforços se a derivação já foi realizada.

Anotações fornecidas junto com proveniência podem ajudar a interpretar os dados no contexto em que se pretendia, especialmente para os dados arquivados que são utilizados após um grande período de tempo depois que foram gerados. Isto ajuda a assimilar dados de forma inequívoca em um domínio do usuário. Proveniência também pode ser vista como uma árvore de derivação ou em outras formas gráficas, e agir como um ponto de partida para explorar outros metadados sobre os dados e processos.

### 2.2.3. Cenários de Aplicação

Para exemplificar casos em que proveniência se torna útil, Vazquez Salceda, Willmott 05-07, apresenta o cenário de instituições de saúde, quanto à gestão de transplante de órgãos:

Na Figura 2-1, são representadas as trocas de informações entre médico, doador, e laboratório para um caso de decisão pelo médico se um órgão para doação está apto ou não para transplante.

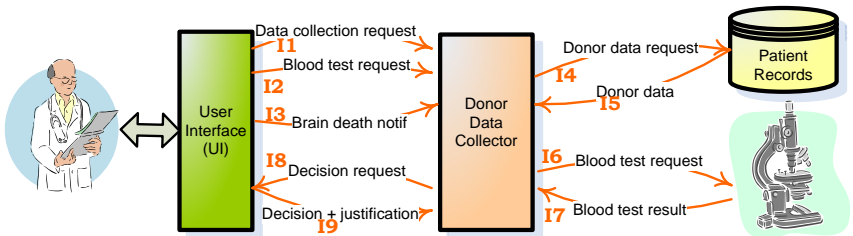


Figura 2-1 Cenário de troca de informações na gestão de transplantes de órgãos.

Neste cenário, podemos levantar as seguintes questões de proveniência:

1. Qual médico estava envolvido em uma decisão?
2. Por que um órgão para transplante foi rejeitado?
3. O órgão foi armazenado de acordo com as regras?

Outro exemplo é dado por Rocio Aldeco Perez, no qual uma Auditoria é realizada sobre o processamento de dados privados de uma organização.

Na Figura 2-2, são representados os passos pelos quais um conjunto de informações é tratado num processamento estatístico realizado por uma organização, neste caso, o cálculo da média de idade

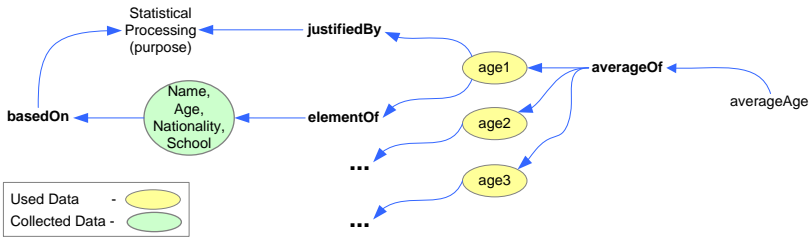


Figura 2-2 Tratamento de informações em um processamento estatístico de seus colaboradores.

Outras perguntas que podem ser identificadas quanto à proveniência deste cenário:

1. Os dados foram utilizados de forma compatível com o objetivo pelo qual foi capturado?
2. Os dados mais recentes foram utilizados no cálculo?
3. Os dados foram apagados após seu uso?

Para atender estas perguntas, alguns cuidados na captura de proveniência devem ser considerados, os quais serão abordados nas sessões seguintes.

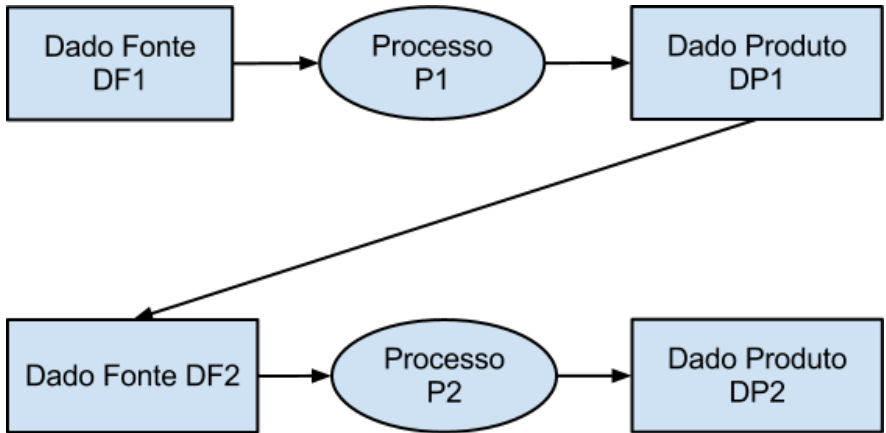
#### 2.2.4. Captura de Proveniência

Segundo (Freire, 2008), a proveniência pode ser capturada por meio de dois caminhos, prospectiva e retrospectiva. A forma prospectiva captura os passos que devem ser seguidos para a geração de um dado produto (por exemplo, os passos que devem ser seguidos para a execução de um conjunto de processos relacionados à publicação de um dado em Linked Data), permitindo desta forma o registro da especificação de tarefas computacionais (por exemplo, conjunto de processos, um script, etc).

A forma de captura de proveniência retrospectiva captura os passos executados por uma tarefa computacional assim como a informação sobre o ambiente utilizado para derivar um dado produto específico, ou seja, um log detalhado sobre a execução da tarefa (por

exemplo, os dados produzidos por um sistema computacional para busca por similaridade entre seqüências de DNA, além de todos os parâmetros envolvidos na execução desse sistema). Essas duas formas de captura de proveniência são independentes, ou seja, para a captura da proveniência prospectiva, a captura da proveniência retrospectiva não é estritamente necessária. O modelo conceitual de proveniência define como as informações de proveniência são representadas ou modeladas. Um modelo de proveniência pode representar informações de proveniência tanto prospectiva quanto retrospectiva. Além disso, esses modelos podem incluir anotações para prover mais semântica aos dados de proveniência.

A forma de captura da proveniência pode ser em maior ou menor nível de detalhe e pode ser classificada em dois níveis: "grão grosso" e "grão fino". A granularidade de proveniência chamada "grão grosso" está relacionada à proveniência de um conjunto de atividades, chamado workflow. Ela descreve a história da execução de um workflow e/ou da derivação de um conjunto de dados.



*Figura 2-3 Diagrama para representação de derivação e proveniência de dados. Fonte: Autor, adaptado de*

A Figura 2-3 mostra um exemplo de workflow onde os dados e seu processamento podem ser gravados e rastreados por meio da proveniência. Elipses representam os componentes do workflow, quadrados denotam dados.

### 2.2.5. Rastreabilidade de Dados

Se ao longo da cadeia produtiva há o registro de todas as atividades que aconteceram, é possível rastrear cada lote relacionado ao incidente. Se juntamente a essas atividades há também o registro de dados que descrevem a origem de cada lote, então rastrear os dados de um determinado lote é consultar dados que descrevem a origem de cada lote.

Interpretamos *rastreabilidade* como a capacidade de realizar consultas para rastrear a origem (*trace*) e rastrear o destino (*track*) (Dorp, 2004). Na Figura 3, ao seguir as setas a partir de um dado específico, DP2 por exemplo, é possível rastrear os dados de todos os processos relacionados.

Portanto, rastrear dados é realizar consultas à proveniência. Essas consultas não são uma exclusividade da área industrial e estão presentes também em outras áreas. Por exemplo, na área de gerência de configuração de software (Staa, 2003), especificamente no versionamento de software, localizar o rastro para a origem identifica possíveis causas de um componente defeituoso.

De forma complementar, identificar o rastro para o destino viabiliza a análise de impacto de componentes construídos a partir de um componente potencialmente defeituoso. Assim, aqui também é essencial consultar a proveniência sobre as diferentes versões de componentes existentes.

### 2.2.6. Captura e Rastreabilidade na Prática

Computacionalmente, capturar e rastrear dados levantam duas questões a serem analisadas.

1. Computadores são bons em produzir resultados rapidamente;
2. Computadores são ruins para explicar suas ações passadas.

Neste sentido, diversos estudos vêm sendo realizados buscando identificar os princípios para se alcançar proveniência de dados em sistemas computacionais. A seguir são descritos resumidamente os principais modelos utilizados como subsídios para tal.

### 2.2.7. Modelos Conceituais para Proveniência

Um modelo é uma representação do que existe no intelecto humano como solução para um determinado problema. Como cada ser

humano tem uma forma diferente de interpretar o mundo, existem diferentes modelos para o mesmo tipo de problema (Sowa, 1999).

#### *2.2.7.1. Modelo W7*

Modelos conceituais convencionais não fornecem explicitamente um mecanismo simples de capturar a semântica de proveniência de dados, e ainda não está claro como informações de proveniência podem estar relacionadas com os dados do aplicativo no nível conceitual.

Em resposta a este problema, (RAM, 2007) propôs um modelo de proveniência genérico chamado de modelo W7 para capturar a semântica de dados proveniência.

O modelo W7 representa diferentes componentes de proveniência e suas relações com o outro. Conceituamos como proveniência a combinação de sete elementos interligados, incluindo "o que", "quando", "onde", "Como", "quem", "que" e "porque". Cada um destes componentes pode ser utilizado para rastrear eventos que afetam os dados durante a sua vida útil.

#### **O quê (What)**

O bloco básico de construção do modelo W7 é o elemento "o que". Sua semântica é definida como: uma seqüência de eventos que afetam um objeto de dados durante o seu tempo de vida.

O "que" consiste em três subconjuntos, incluindo:

IL\_EVENT: um conjunto de eventos do ciclo de vida das informações que capturam a criação, transformação, uso, e exclusão de dados durante o seu ciclo de vida.

IR\_EVENT refere-se a direitos de propriedade intelectual relacionados a eventos que desencadeiam a atribuição dos direitos de propriedade intelectual, incluindo a propriedade, direitos autorais e patentes aos dados.

AR\_EVENT captura eventos de arquivamento destinadas para preservar o dado e disponibilizá-lo mais tarde para uma comunidade designado. Estes três subconjuntos são disjuntos, isto é, cada um destes sub-grupos pode ter ainda mais eventos classificados.

#### **Quando (When)**

QUANDO representa um conjunto de marcas de tempo  $\{t_1, t_2, \dots, t_n\}$  associado com vários eventos proveniência. Enquanto alguns eventos podem ser instantâneos, outros podem ocorrer durante um

intervalo de tempo. Assim, podemos especificar dois subconjuntos disjuntos de QUANDO incluindo *Instant* e *TIME\_PERIOD*. *INSTANT* é um conjunto de instantes. Cada caso é um ponto na linha do tempo. *TIME\_PERIOD* é um conjunto de períodos de tempo. Um período de tempo refere-se ao tempo entre dois instantes com um começo e um fim.

### Onde (Where)

O elemento "onde" no modelo W7 capta locais de evento. ONDE denota um conjunto de localizações onde acontecem vários eventos. As formas mais comuns de representar locais são física e geográfica. Locais físicos especificam a posição de lugares ou pontos baseado em um sistema de coordenadas global, enquanto localizações geográficas significam uma área ou limite regido por uma lei comum e são normalmente organizados hierarquicamente.

Além da localização física e geográfica, é introduzido o conceito de localização de transação, que liga um objeto de dados para a sua localização em um servidor ou banco de dados. Este conceito é importante uma vez que os dados podem viajar entre fontes de informação devido a operações de armazenamento ou transferência. A transação de localização pode muitas vezes ser representado por uma URI, e pode ser digitado a origem e destino.

### Como (How)

"Como" documenta ações que levam à ocorrência de um evento. Bunge postula em (Bunge, 1977), que a história de uma coisa evolui se está sob a ação de um outro. Uma ação é vista como um sistema de "obras", em que os agentes trabalham em determinados objetos, a fim de se obter um resultado desejado. "Ações" são causas de evento, e os eventos são trazidos à existência como resultado de ações executadas por agentes. Informações sobre as ações normalmente inclui:

- *Pré-condições* que se referem às condições que devem conter antes da promulgação de uma ação.
- *Métodos* que fornecem descrições detalhadas sobre o que foi feito e capturam vários parâmetros de ação.
- *Entradas* que se referem a objetos de dados que são manipulados através da publicação de uma ação. Uma ação pode, assim, ser vista principalmente como um processo de transformação de um conjunto de entradas em saídas.
- *Recursos* que se referem a itens de apoio disponíveis de realizar várias ações, por exemplo, armas e veículos são



muitas vezes os recursos utilizados em atividades de guerra.

“Como” é definido como uma tupla (AÇÃO, pré-requisitos, MÉTODOS, ENTRADAS, RECURSOS P, M, I, R), onde:

- AÇÃO = conjunto de ações, e os conceitos mencionados anteriormente, como Pré-condições, métodos, insumos e recursos também são definidos como conjuntos.
- P: AÇÃO → PRE-CONDIÇÕES é uma função que mapeia uma ação à sua condição; M: Ação → MÉTODOS mapeia uma ação ao seu método, I: AÇÃO → ENTRADAS mapeia uma ação à sua entrada, e R: AÇÃO → RECURSOS associa uma ação com o recurso utilizado na mesma.

Seguindo (Konolige, et al., 1993), podemos classificar as ações em primitivas e complexas. Consequentemente, especificamos que a ação consiste em dois subgrupos PRIMITIVE\_ACTION e COMPLEX\_ACTION.

Uma ação é considerada primitiva se nenhuma decomposição irá revelar qualquer informação adicional que é de interesse. Ações complexas, por outro lado, podem ser arbitrariamente atividades que podem ser decompostos em ações primitivas que ocorrem sequencialmente ou simultaneamente.

Além disso, estudos anteriores, tais como (Frew, et al., 2001) se focam em capturar os procedimentos utilizados para processamento de dados, descrevendo o fluxo de trabalho de uma experiência. Assim, definimos uma relação "Depends\_on" que captura o fluxo de controle de ações primitivas dentro de uma ação complexa, como uma seqüência de concorrência, etc.

Definimos uma ação complexa  $c = (P, \text{DEPENDS\_ON})$ , onde:

- $P = \{p_1, p_2, \dots, p_k\}$  é um conjunto de ações primitivas que constituem a ação complexa  $c$ .
- $\text{DEPENDS\_ON} = \{D_1, D_2, \dots, D_k\}$  é um conjunto de relações. Cada relação  $d$  é um par ordenado  $(p_i, p_j)$ , onde  $p_i, p_j \in P$ .

### Quem (Who)

"Quem" refere-se a agentes envolvidos nos eventos. Os principais conceitos associados "quem" são de um agente e um papel. Um agente é "uma entidade intencional", que é que tem algum propósito de ideias  $d$  e que orienta suas ações (Koubarakis, et al., 2002). Neste caso é utilizado o termo "agente" em vez de pessoa para a generalidade,

de modo que ele pode ser usado para se referir a indivíduos, organizações, tão bem como agentes artificiais. Um papel é definido como "um conjunto coerente de atividades a ser atribuído a um agente como uma responsabilidade funcional" (Curtis, et al., 1992). Cada agente assume um papel certo para fazer alguma contribuição para a ação, o que leva a um evento. Por exemplo, um agente federal pode desempenhar o papel de supervisor na criação do roteiro de uma correspondência suspeita.

Definimos "Como" uma tripla (AGENTE PAPEL, RL), onde:

- AGENTE = {a1, a2, ..., an} é um conjunto de agentes que estão envolvidos em vários eventos.
- PAPEL = {r1, r2, ..., rn} é um conjunto de funções que agentes estão autorizados a assumir.
- RL: AGENTE → PAPEL é uma função que associa um agente com o papel que desempenhou em um evento particular.

"Como" inclui três subgrupos, ou seja, um conjunto de indivíduos *INDIVIDUAL*, um conjunto de organizações *ORGANIZATION*, e um conjunto de agentes artificiais *ARTIFICIAL\_AGENT*. Muitas vezes precisamos capturar a posição e filiação de um agente individual. Quando um agente individual participa de sua filiação, ele já não é inteiramente livre para escolher seus objetivos e ações.

Em vez disso, ele realiza algumas atividades de acordo com a sua posição. A posição, a qual é chamada papel organizacional em [15], representa um conjunto de responsabilidades de um indivíduo em sua filiação.

Como resultado, especificamos uma função PA: *INDIVIDUAL* → *POSICÃO FILIAÇÃO*, que mapeia um agente individual de sua posição e afiliação.

### Qual (Which)

O elemento "Qual" descreve quais os dispositivos são usados na criação, análise e transformação do dado. Os dispositivos podem ser distinguidos em instrumentos (por exemplo, equipamentos e hardware) e aplicações. Quando um evento envolve um dispositivo, algum nível de detalhe sobre o dispositivo em que está alojado deve ser capturado. Além disso, algumas ações são especificamente suportadas ou oferecidas por determinados dispositivos, por meio de que as

características e capacidades que os dispositivos podem desempenhar um papel importante na descrição do comportamento da ação.

A informação relacionada com um dispositivo é logicamente dividida em três classes dependendo do tipo de informação que fornecem: a descrição do dispositivo, a função e configurações. A descrição do dispositivo contém as informações básicas relacionada com um dispositivo, tal como o seu nome, versão, vendedor, etc. A função de um dispositivo pode ser especificado em termos das variáveis do dispositivo em si, por exemplo, a função de uma bateria pode ser especificada como de fornecimento de uma tensão eléctrica medido em volts. Mais frequentemente, um dispositivo é composto de peças ou componentes, e sua função é expressa em termos das variáveis dos seus componentes. Como um exemplo, um computador pode ter um processador de 2,0 GHz e memória de 256 MB. Diferentes das propriedades funcionais que raramente mudam ao longo da vida de um dispositivo, suas configurações contém informações voláteis referentes ao dispositivo, como nível atual de uso da CPU e nível de energia restante na do computador. As configurações de um dispositivo geralmente variam entre aplicativos, e especifica o o desempenho dos componentes de um dispositivo durante um evento.

Definindo “Qual” uma tupla (dispositivo, as configurações, DESCRIÇÃO, FUNÇÃO, S, D, F), em que:

- **DISPOSITIVO** = {D1, D2, ..., dn} é um conjunto de dispositivos utilizados em vários eventos. É constituído por dois disjuntos subconjuntos aparelho e da aplicação.
- **CONFIGURAÇÕES** denota um conjunto de configurações que um dispositivo pode assumir.
- **S:** **DISPOSITIVO** → **CONFIGURAÇÕES**, **D:** **DISPOSITIVO** → descrição e **F** **DISPOSITIVO** → **FUNÇÃO**, representam mapeamentos a partir de um dispositivo para suas configurações, descrições e funções.

### Por quê (Why)

Definimos **POR QUE** como um conjunto de lógica de decisão {y1, y2, ..., yn} associado a vários eventos proveniência.

O esquema para representar "por que" é baseado em grande parte no Modelo de Crença-Desejo-Intenção, que identifica crenças, desejos e intenções como fatores importantes que afetam a tomada de decisão. Crenças representam o conhecimento do mundo, os desejos são os objetivos atribuídos ao agente, e intenções são compromissos de um

agente para alcançar objetivos particulares. Aqui, juntamos desejos e intenções em metas. Como resultado, nós especificamos dois subconjuntos de POR QUE, ou seja, crença e OBJETIVO. O primeiro representa um conjunto de crenças e este último um conjunto de metas.

Uma forma natural de responder "por quês" é rastreá-los aos objetivos. Por exemplo, o por quê de um marco ser estabelecido em uma ação anti-terrorista pode estar relacionado com o objetivo de que a ação seja concluída a tempo. A representação explícita de metas é importante porque nos permite estudar um evento específico a partir de um certo ponto de vista. Também é definido um relacionamento "is\_reduced\_to" para capturar a estrutura de objetivo  $\rightarrow$  sub-objetivo. Um objetivo pode ter diversos objetivos pais, pois pode ocorrer em várias reduções. Nós definimos  $IS\_REDUCED\_TO = \{S1, S2, \dots, sn\}$  como um conjunto de relações que representam estruturas objetivo  $\rightarrow$  sub-objetivo. Cada relação é de um par ordenado  $(g_i, g_j)$ , onde  $g_i, g_j \in OBJETIVO$ . Além disso, cada relação  $s \in IS\_REDUCED\_TO$  não deve ser simétrica. Assim, especificar uma restrição  $s$  quanto  $s \in S\_REDUCED\_TO$  e  $s-1 \in S\_REDUCED\_TO \Rightarrow s = s-1$ .

Outro conceito importante associado com o "por que" é o conceito da crença. Os agentes têm uma visão subjetiva do mundo, onde eles formam suas crenças. Diferente dos objetivos que um agente pretende cumprir através de uma ação, crenças referem-se ao que um agente acredita antes da ação, e eles formam o fundo sobre o qual um agente pode escolher agir de uma maneira particular [18]. Ainda classificamos crenças em suposições e hipóteses.

### Considerações sobre o Modelo 7W

O modelo 7W é um modelo genérico de proveniência de dados que se destina a ser facilmente adaptável para representar um domínio ou requisitos específicos de um aplicação com proveniência através de modelagem conceitual.

O objetivo principal deste modelo de proveniência é apoiar a avaliação das qualidades de dados, tais como confiabilidade, precisão, e oportunidade.

Há diferentes áreas que se beneficiariam de um modelo deste tipo. Entre elas estão:

- Engenharia de Software, na demanda por rastreabilidade de requisitos, por exemplo, identificando a proveniência de artefatos que não possuam requisitos;

- e-Science, na necessidade de repetição de experimentos, por exemplo, com a verificação dos métodos e valores utilizados e obtidos durante o processamento de um fluxo de experimentos;
- Industrial, na imposição de conformidade através de legislações vigentes, por exemplo, com a capacidade de consulta a proveniência dos eventos de uma cadeia produtiva atendendo a normas governamentais para o setor de produção de alimentos;
- Gestão de Conteúdo Digital, na auditoria e histórico de *logs*, por exemplo, na identificação do editor e outros dados de proveniência que descrevam o contexto das alterações de um documento digital;
- Preservação Digital, na preservação de documentos arquivados a longo prazo, por exemplo, na identificação dos eventos de arquivamento para garantir e creditar a autenticidade da fonte produtora que repassa esses documentos para arquivamento;
- Gestão de Projetos, na análise de tarefas realizadas contra as planejadas, por exemplo, através de consultas a proveniência de eventos considerados particularmente importantes (*milestones*) e rastro dos dados relativos às tarefas que contribuíram para o seu alcance.

#### 2.2.7.2. Proveniência Aberta

O Modelo de Proveniência Aberto é um modelo de proveniência proposto por Luc Moreau, que se destina a atender aos seguintes requisitos:

(1) Permitir que a informação de proveniência ser trocados entre os sistemas, por meio de uma camada de compatibilidade com base em um modelo compartilhado de proveniência.

(2) Permitir aos desenvolvedores construir e compartilhar ferramentas que operam sobre tal modelo de proveniência.

(3) Para de proveniência ne num precisa, tecnologia maneira agnóstica.

(4) Apoiar uma representação digital de proveniência para qualquer coisa, tanto dados produzidos pelos sistemas informáticos ou não.

(5) Permitir que coexistam vários níveis de descrição de dados.

(6) Unir um conjunto de regras que identificam as inferências válidas que podem ser feitas na representação de proveniência.

## Definições

O modelo de Proveniência Aberta toma como base três entidades primárias, que seguem:

Definição 1 (Artefato) peça Imutável de Estado, que pode ter um corpo físico em um objecto físico ou de uma representação digital de um sistema de computador.

Definição 2 (Processo) Ação ou série de ações realizadas em ou causados por artefatos, resultando em artefactos e novos.

Definição 3 (agente) entidade Contextual agindo como um catalisador de um processo, permitindo, facilitando, controle, afetando a sua execução.

Como relações de dependência o modelo define:

Definição 4 (relação causal) A relação de causalidade é representada por um arco e denota a presença de uma dependência causal entre a fonte do arco (o efeito) e o destino do arco (a causa). Cinco relações causais são reconhecidas: o processo utilizado por um artefato, um artefato foi gerado por um processo, um processo foi desencadeado por um processo, um artefato foi derivado de um artefato, e um processo foi controlado por um agente.

Definição 5 (Artefato usado por um processo) em um gráfico, conectando um processo a um artefato por uma aresta destina-se a indicar que o processo exigiu a disponibilidade do artefato para completar sua execução. Quando vários artefatos são ligados a um mesmo processo pelo uso de múltiplas arestas, todas elas foram necessárias para o processo se concretizar.

Definição 6 (artefatos gerados por Processos) em um gráfico, a ligação de um artefato a um processo por uma borda "wasGeneratedBy" pretende indicar que o processo era necessário para iniciar a sua execução para o artefacto a ser gerado. Quando vários artefatos estão ligados a um mesmo processo por arestas múltiplas "wasGeneratedBy", o processo tinha de ter começado, para que todos eles sejam gerados.

Definição 7 (processo desencadeado pelo processo) Uma ligação de um processo P2 de um processo de P1 por uma ligação " was

triggered by" indica que o início do processo P1 foi necessário para P2 ser capaz de completar.

Definição 8 (Artefato Derivado do Artefato) Uma aresta "was derived from " entre dois artefatos A1 e A2 indica que o artefato A1 pode ter sido usado por um processo derivado de A2.

Definição 9 (processo controlado por agente) A afirmação de uma aresta "was controlled by " entre um processo P e um agente Ag indica que o início e o final do processo P foi controlado pelo agente Ag.

Um "papal" é usado em uma anotação, wasGeneratedBy e wasControlledBy.

Definição 10 (Papal) Um papel designa um artefato ou função de agente em um processo. Um papel é usado para a diferenciação entre os vários usos, geração, ou as relações de controle.

1. Um processo pode usar (gerar) mais de um artefato. Cada relação "wasGeneratedBy" utilizada pode ser distinguida por um papel único em relação a tal processo. Por exemplo, um processo pode utilizar vários arquivos, a leitura a partir de parâmetros de um lado, e dados de leitura de outro. As relações utilizadas seriam rotuladas com papéis distintos.

2. Um artefato pode ser usado por mais do que um processo, possivelmente para fins diferentes. Neste caso, as relações usadas podem ser distinguidas das funções associadas com as relações usadas. Por exemplo, um dicionário pode ser utilizado por um processo de procurar a ortografia de "proveniência", (papel = "procurar proveniência"), enquanto outro processo usa o mesmo dicionário para manter aberta a porta (papel = "peso para porta").

3. Um agente pode controlar mais de um processo. Neste caso, os diferentes processos podem distinguir-se pela função associada com a relação "wasControlledBy". Por exemplo, um jardineiro pode controlar o processo de escavação (papel = "cavar o leito"), bem como plantar uma roseira (papel = "plantar") e regar o arbusto (papel = "irrigação").

4. Um processo pode ser controlado por mais de um agente. Neste caso, cada agente pode ter uma função de controle distinta, o que pode ser distinguido por funções associadas com as relações "wasControlledBy". Por exemplo, embarcar no trem pode ser controlado pelo agente vendedor de bilhetes (papel = "vender bilhete"), pelo agente porteiro (papel = "tomar bilhete ") e o agente mordomo (papel = "guia para a acento").

Um exemplo que ilustra todos os conceitos e algumas das dependências causais é exibido na Figura 2-4.

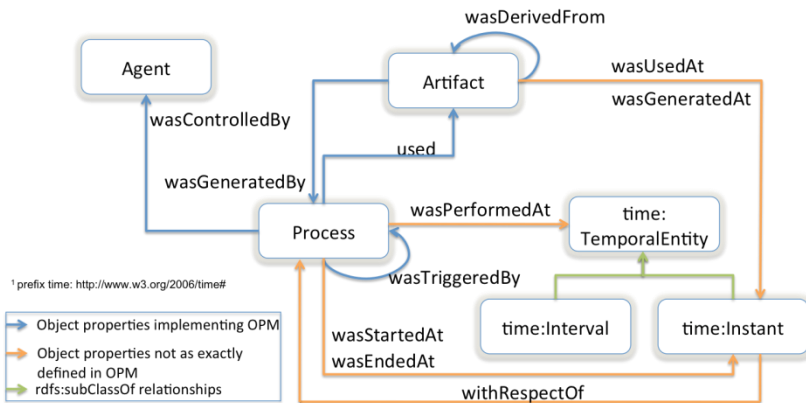


Figura 2-4 Representação de Vocabulário para Open Provenance

### 2.2.7.3. Modelo PROV-Data Model

Neste modelo, proveniência é definida como um registro que descreve pessoas, instituições, entidades, e atividades que envolvem produção, influencia, ou disponibilizam partes de dados ou algo.

O Prov Data Model, PROV-DM, foi especificado pelo grupo W3C, World Wide Web Consortium, órgão regulador de padrões e especificações técnicas para a WEB. Trata-se de um modelo voltado para o ambiente WEB, com o intuito de permitir que domínios e representações de aplicações específicas de proveniência sejam traduzidos em um modelo intercambiável entre diferentes sistemas.

O modelo foi projetado de forma modular e é estruturado de acordo com seis componentes, afim de prover cobertura para as várias facetas de proveniência:

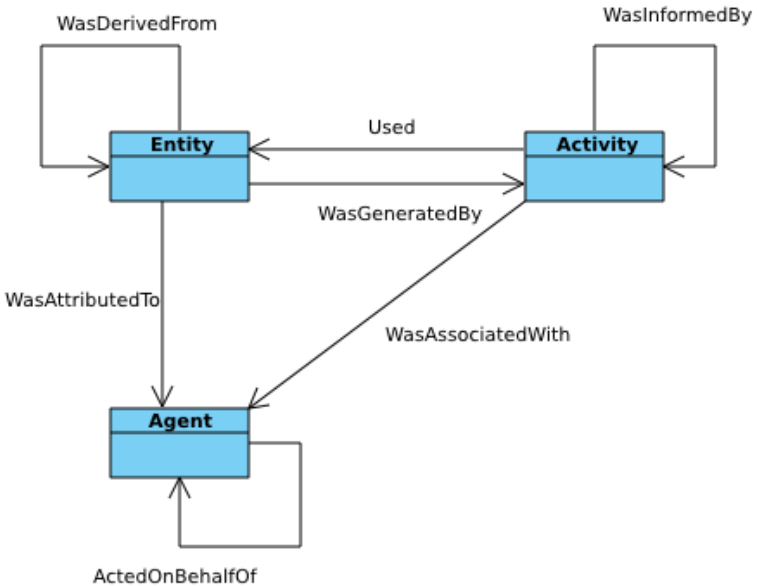
- Componente 1: entidades e atividades, e o horário em que cada um foi criado, usado ou finalizado;
- Componente 2: derivações de entidades de outras entidades;
- Componente 3: agentes que assumem a responsabilidade de entidades que foram geradas ou atividades que aconteceram;
- Componente 4: pacotes, uma maneira de apoiar a proveniência da proveniência;
- Componente 5: Propriedades que relacionam entidades equivalentes, que se referem a mesma coisa;



- Componente 6: Coleções formando uma estrutura lógica para os seus membros.

### Estruturas básicas

Em geral, proveniência descreve o uso e produção de entidades por atividades, que podem ser influenciados de várias maneiras pelos agentes. Estes tipos de estruturas básicas e suas relações são ilustrados pelo diagrama UML da Figura 2-5.



*Figura 2-5 Representação das Estruturas Básicas e seus relacionamentos*

Estas estruturas básicas compõem o núcleo do modelo PROV-DM. A seguir são introduzidos os conceitos de cada uma destas estruturas, que estão apresentados de forma resumida na Tabela 2-1.

Na primeira coluna apresentam-se os conceitos, a segunda coluna indica se o conceito é mapeado como um tipo ou de uma relação, enquanto a terceira coluna contém o nome correspondente no modelo, tal como aparece na Figura 3.

Nomes de relações usam a forma verbal no passado para expressar o que aconteceu no passado, ao contrário do que pode ou vai acontecer. No núcleo deste modelo, todas as relações entre entidades se dão de forma binária, ou seja, relacionam somente uma Entidade A à uma entidade B ou ela própria.

<b>Conceitos</b>	<b>Tipos ou Relações PROV-DM</b>	<b>Nome</b>
Entidade	Tipos	Entity
Atividade		Activity
Agente		Agent
Geração	Relações	WasGeneratedBy
Uso		Used
Comunicação		WasInformedBy
Derivação		WasDerivedFrom
Atribuição		WasAttributedTo
Associação		WasAssociatedWith
Delegação		ActedOnBehalfOf

*Tabela 2-1 Resumo de Conceitos das Estruturas do Núcleo do modelo PROV-DM*

### **Entidades e Atividades**

Neste modelo, coisas no qual queremos descrever proveniência são chamadas entidades e têm alguns aspectos fixos. O termo “coisas” abrange uma ampla diversidade de noções, incluindo objetos digitais, como um arquivo ou página web, as coisas físicas, como uma montanha, um edifício, um livro impresso, ou um carro, bem como conceitos abstratos e ideias.

Segundo a W3C:

*“An entity is a physical, digital, conceptual, or other kind of thing with some fixed aspects; entities may be real or imaginary.”*

Do Inglês:

Uma entidade é uma coisa física, digital, conceitual, ou de outro tipo, com alguns aspectos fixos; entidades podem ser reais ou imaginárias.

## **2.3. Licenças para distribuição ou publicação de dados**

### **2.3.1. Direito Autoral**

Direito autoral, direitos autorais ou direitos de autor são as denominações empregadas em referência ao rol de direitos aos autores de suas obras intelectuais que pode ser literárias, artísticas ou científicas. Neste rol encontram-se dispostos direitos de diferentes naturezas. A doutrina jurídica clássica coube por dividir estes direitos entre os chamados direitos morais que são os direitos de natureza pessoal e os direitos patrimoniais (direitos de natureza patrimonial)

### **2.3.2. Copyright**

Direitos do Autor não são necessariamente o mesmo que copyright em inglês. O sistema anglo-saxão do copyright difere do de direito de autor. Os nomes respectivos já nos dão conta da diferença: de um lado, tem-se um direito à cópia, copyright ou direito de reprodução, do outro, um direito de autor; neste, o foco está na pessoa do direito, o autor; naquele, no objeto do direito (a obra) e na prerrogativa patrimonial de se poder copiar.

Deve perceber as diferenças entre o direito autoral de origem romano-germânica, com base no sistema continental europeu do chamado Sistema romano-germânico e o sistema anglo-americano do copyright baseado na Common Law, havendo por característica diferencial, o fato de que o Direito Autoral tem por escopo fundamental a proteção do criador e ao contrário o copyright protege a obra em si, ou seja o produto, dando ênfase à vertente econômica, à exploração patrimonial das obras através do direito de reprodução. No efetramento do direito de reprodução, o titular dos direitos autorais poderá colocar à disposição do público a obra, na forma, local e pelo tempo que desejar, a título oneroso ou gratuito.

### **2.3.3. Copyleft**

Copyleft é uma forma de usar a legislação de proteção dos direitos autorais com o objetivo de retirar barreiras à utilização, difusão e modificação de uma obra criativa devido à aplicação clássica das normas de propriedade intelectual, exigindo que as mesmas liberdades sejam preservadas em versões modificadas. Ele difere assim do domínio público, que não apresenta tais exigências; enquanto o domínio público

permite qualquer utilização de uma obra, o copyleft, tem, via de regra, a única exigência de se poder copiar e distribuir não comercialmente uma obra. O copyleft também não proíbe a venda da obra pelo autor, mas implica a liberdade de qualquer pessoa fazer a distribuição não comercial da obra.

O copyleft denomina genericamente uma ampla variedade de licenças que permitem, de diferentes modos, liberdades em relação a uma obra intelectual. Seu nome se origina do trocadilho com o termo "copyright"; literalmente, copyleft pode ser traduzido como "esquerdo de cópia" ou "permitida a cópia".

Richard Stallman foi um dos responsáveis pela popularização inicial do termo copyleft, ao associá-lo, em 1988, à licença GPL. De acordo com Stallman, o termo foi-lhe sugerido pelo artista e programador Don Hopkins, que incluiu a expressão "Copyleft - all rights reversed." numa carta que lhe enviou. A frase é um trocadilho com expressão "Copyright - all rights reserved." usada para afirmar os direitos de autor.

Uma obra, seja de software ou outros trabalhos livres, sob uma licença copyleft requer que suas modificações, ou extensões do mesmo, sejam livres, passando adiante a liberdade de copiá-lo e modificá-lo novamente.

Uma das razões mais fortes para os autores e criadores aplicarem copyleft aos seus trabalhos é porque desse modo esperam criar as condições mais favoráveis para que mais pessoas se sintam livres para contribuir com melhoramentos e alterações a essa obra, num processo continuado.

#### **2.3.4. Creative Commons**

Creative Commons é uma organização não governamental sem fins lucrativos localizada em São Francisco, Califórnia, nos Estados Unidos, voltada a expandir a quantidade de obras criativas disponíveis, através de suas licenças que permitem a cópia e compartilhamento com menos restrições que o tradicional todos direitos reservados. Para esse fim, a organização criou diversas licenças, conhecidas como licenças Creative Commons.

A organização foi fundada em 2001 por Larry Lessig, Hal Abelson, e Eric Eldred (Boyle, 2010) com apoio do Centro de Domínio Público. O primeiro conjunto de licenças copyright foram lançadas em dezembro de 2002 (Creative Commons, 2009). Creative Commons é governado por um conselho de diretores e um conselho técnico. Joi Ito é

atualmente o coordenador do conselho e CEO (Creative Commons, 2010). Creative Commons tem sido abraçada por muitos criadores de conteúdo, pois permite controle sobre a maneira como sua propriedade intelectual será compartilhada. Alguns criticam a ideia acusando-a de não ser suficientemente abrangente.

As licenças Creative Commons foram idealizadas para permitir a padronização de declarações de vontade no tocante ao licenciamento e distribuição de conteúdos culturais em geral (textos, músicas, imagens, filmes e outros), de modo a facilitar seu compartilhamento e recombinação, sob a égide de uma filosofia copyleft.

As licenças criadas pela organização permitem que detentores de copyright (isto é, autores de conteúdos ou detentores de direitos sobre estes) possam abdicar em favor do público de alguns dos seus direitos inerentes às suas criações, ainda que retenham outros desses direitos. Isso pode ser operacionalizado por meio de um sortimento de módulos-padrão de licenças, que resultam em licenças prontas para serem agregadas aos conteúdos que se deseja licenciar.

Os módulos oferecidos podem resultar em licenças que vão desde uma abdicação quase total, pelo licenciante, dos seus direitos patrimoniais, até opções mais restritivas, que vedam a possibilidade de criação de obras derivadas ou o uso comercial dos materiais licenciados.

## 2.4. Web Semântica

A *Web* tem sido um dos principais meios de comunicação em escala global. Ao mesmo tempo, ela se comporta como um grande depósito de dados, contendo documentos, imagens, vídeos e diversos outros tipos de mídias digitais. No entanto, o uso de informações neste depósito tão volumoso é uma tarefa complicada. Os documentos na *web* não agregam valor semântico em seus conteúdos, o que impossibilita a pesquisa e o referenciamento preciso dessas informações.

A *Web Semântica*, idéia de Tim Berners-Lee, James Hendler e Ora Lassila publicada em 2001 visa dar contexto ao conteúdo na *web*, ou seja, atribuir um significado (sentido) aos dados publicados na Web de modo que seja perceptível tanto pelo humano como pelo computador. Tecnicamente, a Web semântica é uma extensão da Web atual, e o termo passou a ser também conhecido como *Web de dados*, ou *Web 3.0*.

Tim Berners-Lee acreditava que, com os dados invisíveis publicados numa camada escondida das páginas, poderíamos criar uma

estrutura que daria semântica aos conteúdos da Web, permitindo o reconhecimento de contexto e significado por parte de sistemas computacionais. Isso então poderia ser usado para enriquecer pesquisas, possibilitando resultados mais relevantes; permitir que sistemas de informação interagissem, extraíndo ou fornecendo dados entre si, aumentando a interoperabilidade de dados.

(Bianco, 2011) defende que ao longo da última década surgiram várias iniciativas para se implementar a Web Semântica, a maioria baseada em linguagens ou padrões de marcação adicionados ao código das páginas HTML. Dentre estes, os Microformatos foi uma das iniciativas mais amplamente adotadas. Os microformatos consistem em um conjunto de especificações criadas para descrever determinados objetos do mundo real ou conceitos abstratos. Existem classes para descrever agendas, pessoas, notícias, referências geográficas, entre outros. Essas anotações são feitas no código HTML.

Bernes-Lee, no entanto, diz (em tradução livre) que “a Web Semântica não é apenas colocar dados na web. É acima de tudo fazer links, pois assim pessoas e máquinas podem explorar a Web (rede) de dados”. Faz-se assim um paralelo com a Web de documentos hipertextos, onde os mesmos são ligados por hyperlinks. Essa ligação é importante pois permite que a partir de um documento, ou dado, possa-se encontrar outros documentos, ou dados, relacionados, permitindo a navegação, exploração e enriquecimento de informação.

Hoje, a W3C (World Wide Web Consortium) declara que:

*"... to make the Web of Data a reality, it is important to have the huge amount of data on the Web available in a standard format, reachable and manageable by Semantic Web tools."*

Do inglês:

*"Para tornar a web de dados uma realidade, é importante ter a enorme quantidade de dados na Web disponível em um formato padrão, alcançável e gerenciável através de ferramentas da Web Semântica."*

A ideia trata do uso de uma série de tecnologias livres, como RDF, OWL, SKOS, SPARQL, dentre outras; que, em conjunto, possibilitam um ambiente no qual é possível consultar os dados, fazer cruzamentos, buscas mais eficientes e inteligentes, tirar conclusões usando vocabulários e etc.

#### 2.4.1. Arquitetura de camadas

O W3C propõe uma arquitetura em camadas, que indica os passos a serem tomados para concretizar a implementação da Web Semântica. Uma camada é construída sobre uma inferior e não depende da camada superior, cada camada tende a ser mais especializada e complexa do que as que estão dispostas em níveis inferiores. Existe também a possibilidade de desenvolvê-las separadamente e integrá-las posteriormente. A Figura 2-6, apresenta as camadas da arquitetura de um modelo proposto para web semântica.

No nível mais baixo, composto por Unicode e URI, temos respectivamente um padrão internacional de caracteres (Unicode) e um modo de identificar unicamente um recurso URI (Uniform Resource Identifier).

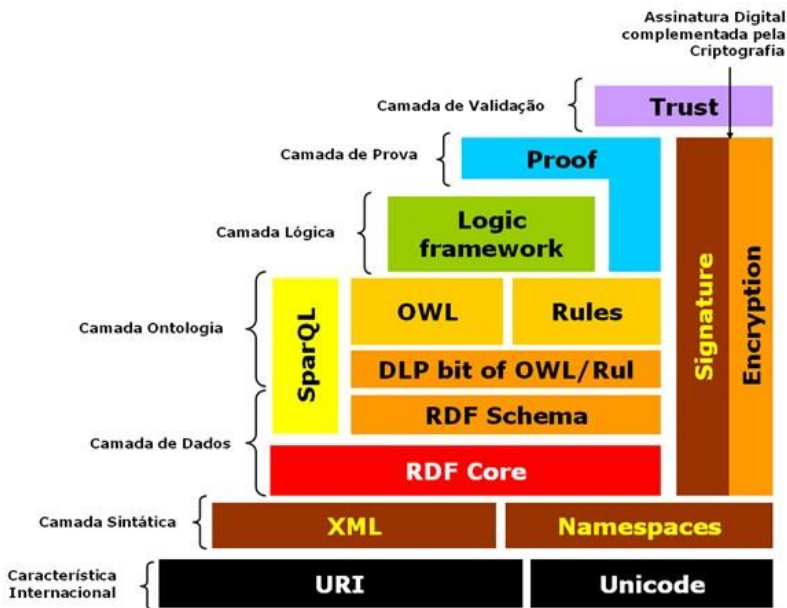


Figura 2-6 Arquitetura de camadas da Websemântica

Na camada XML, composta por Namespaces (NS) e xmlschema, temos a estruturação dos dados com um vocabulário definido pelo usuário. Na camada RDF + rdfschema, é criada uma estrutura de metadados, que são organizadas como triplas (sujeito,

predicado, objeto). Neste vocabulário torna-se possível associar uma propriedade a um recurso através da sua URI.

Este vocabulário é enriquecido e pode ser expandido pela camada *Ontology vocabulary*, que estende o repertório de conceitos e relações semânticas envolvidas. Nas camadas *Logic*, *Proof* e *Trust* estão em desenvolvimento ainda. A camada lógica expressa as regras de tratamento de informações definidas nos níveis inferiores, permitindo ao agente inferir sobre as estruturas de dados. A camada *Proof* verifica a consistência da informação acessível na *Web Semântica*, utilizando uma lógica previamente definida em níveis inferiores, aplicando essas regras para gerar novos conhecimentos. O grau de confiança da inferência é definido na camada *Trust*. Já a camada *Digital Signature* é introduzida nas outras camadas para garantir segurança da informação via criptografia e assinatura digital (Ramalho, et al., 2007).

## 2.5. Linked Data

Em 2005, Tim Berners-Lee publicou uma nova abordagem chamada *Linked Data*, trata-se de uma proposta para publicação de dados na *web* que, seguindo os conceitos da *Web Semântica*, visa estabelecer relações e o referenciamento dos dados na *web*.

Esta proposta se dá num conjunto de melhores práticas para publicar e conectar esses dados, os quais serão abordados a seguir.

Para se alcançar *Linked Data*, os dados devem estar disponíveis em um formato comum, com identificadores que possibilitem o referenciamento dos mesmos, e os dados relacionados devem ser organizadas através de links, os quais indiquem sua origem.

Tim Berners-Lee apresentou como princípios de *Linked Data* os seguintes termos (em tradução livre):

- Use URIs para identificar as coisas.
- Use HTTP URIs para que estas coisas possam ser referenciadas por pessoas e agentes (computacionais).
- Fornecer informações úteis sobre a coisa, quando o seu URI é referenciado, usando formatos padrão como RDF / XML.
- Incluir links para outros dados, relacionados com URIs nos dados expostos para melhorar a descoberta de outras informações relacionadas na *web*.

*“Eu me refiro aos passos acima como regras [referindo-se aos princípios do linked data], mas elas*



*são expectativas de comportamento. Quebrá-las não destrói nada, mas perde-se uma oportunidade de construir dados interconectados. Este, por sua vez, limita as formas que pode depois ser reutilizado de forma inesperada. É esse uso inesperado da informação que é o valor adicionado pela web.”* (BERNERS-LEE, 2006) em tradução livre.

O principal benefício em relação ao Linked Data se dá no ganho de valor agregado de um dado na web. Quanto maior o número de vínculos (diretos e indiretos) este dado tiver, mais informações poderão ser obtidas através do mesmo e maior será a sua utilidade.

### **2.5.1. Publicação de Dados em Linked Data**

Publicar é conceber, criar, capturar, transformar, disseminar, arquivar, procurar e recuperar informação e conhecimento acadêmico e profissional (Wills, 1996).

Em relação à publicação eletrônica, esta pode ser descrita como “o uso de meios eletrônicos de comunicação para tornar a informação disponível ao público” ou como “uma publicação online que está organizada como uma revista científica impressa tradicional, tanto pode ser uma versão online de uma revista científica, como uma que só tenha uma existência online” (Arms, 2000).

De forma prática, para a publicação de dados seguindo os princípios de Linked Data, é necessária a disponibilização dos mesmos em um formato comum e aberto. Como sugere (Berners-Lee, 2006) citado anteriormente, o padrão RDF/XML por exemplo, atende de forma adequada.

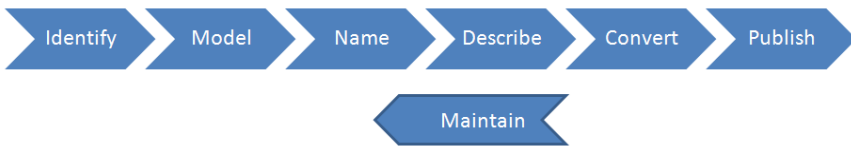
Para isto, pode-se utilizar dois métodos: o primeiro, converter os dados através de ferramentas e disponibilizá-los diretamente em seus fontes, ou então, utilizar uma camada de conversão *on-the-fly*, no qual um mecanismo acessa as bases de dados existentes (relacional, XML, HTML, etc.) e faz a conversão sob demanda em tempo de execução, e então disponibiliza o que podemos chamar de RDFs virtuais (criados somente para aquela requisição).

Também é interessante a configuração de *end-points* (pontos de acesso), os quais possibilitam a realização de consultas aos dados de forma mais conveniente. Estas consultas são realizadas através de uma

linguagem orientada a dados, no caso de dados em RDF, a linguagem SPARQL seria uma possibilidade.

### 2.5.2. Práticas para Publicação de Linked Data

(Hyland, et al., 2012). propôs um processo de criação de Linked Data que consiste nas seguintes etapas: (1) Identificar, (2) modelo (3), Nome, (4) Descreva, (5) Converter (6) Publicar, e (7) mantain.



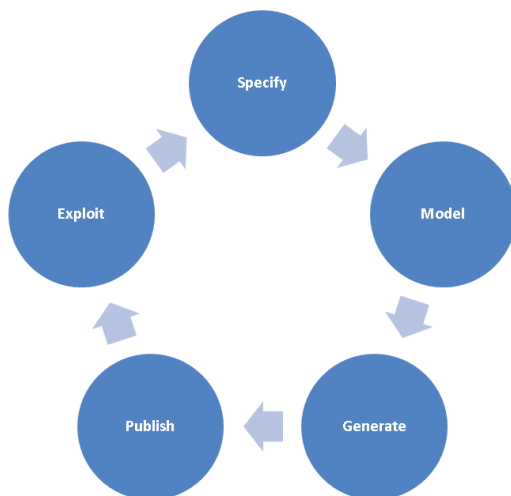
*Figura 2-7 Processo de criação de Linked Data segundo Hyland*

Seguindo a mesma fonte, para Hausenblas, um ciclo de vida para linked Data que consiste nas seguintes etapas: (1) de conscientização de dados, (2), modelagem (3), edição (4) descoberta, integração (5), e os casos (6) de uso.



*Figura 2-8 Processo de criação de Linked Data segundo Hausenblas*

Enquanto que Villazón-terrazas alegaram que o processo de publicação de Linked Data Governamental deve ter um ciclo de vida, da mesma maneira da Engenharia de Software, em que cada projeto de desenvolvimento tem um ciclo de vida. De acordo com nossa experiência este processo tem um modelo interativo de ciclo de vida incrementais, que é baseado na melhoria contínua e extensão do Governo Linked Data resultou da realização de várias iterações.



*Figura 2-9 Processo de criação de Linked Data segundo Villazón-terrazas*

Com base nestas três propostas, a W3C publicou uma análise, sob revisão dos autores citados (Hyland, et al., 2012), definindo os seguintes processos como itens base para o ciclo de vida de Linked Data.

**IDENTIFICAR:** O primeiro passo é identificar os conjuntos de dados que podem ser publicados.

**MODELAR:** Modelar um esboço dos principais objetos de dados, descrevendo a forma como eles são relacionados uns com os outros. Desnormalizando os dados quando necessário. Não considerando, neste momento, as necessidades imediatas de qualquer aplicação e modelo de dados.

**NOMEAR:** Usar HTTP URIs como nomes para os objetos. Tendo em conta uma estratégia de nomeação URI.

**VOCABULÁRIOS PADRONIZADOS:** Descrever os objetos com vocabulários padronizados sempre que possível.

**REPRESENTAÇÃO CONSISTENTE:** Fornecer representações de um recurso consistente e previsível.

**DESCRIÇÕES:** Definir descrições legíveis para humanos e para máquinas dos dados linkados.

**CONVERTER:** Converter os dados de origem em uma representação de dados linkada, também chamado de serialização RDF, incluindo Turtle, Notation-3 (N3), N-Triples, XHTML integrado com RDFa, e RDF / XML.

**ESPECIFICAR LICENÇA:** Especificar uma licença adequada.

**ANUNCIAR:** Hospedar os Dados Abertos Linkados na Web pública e anunciá-los.

**CONTRATO SOCIAL:** A manutenção do serviço e dos dados é crítica. Se os dados que estão publicados na Web forem movidos ou removidos, pode-se quebrar aplicativos de terceiros ou mashups. Assim, um contrato social deve garantir que o conjunto de dados que uma organização publica permaneça disponível onde estiver anunciada.

#### *2.5.2.1. URI*

A própria especificação do Identificador Uniforme de Recurso (URI) não define ou limita o que é um recurso, ou seja, serve para identificar tudo desde objetos tangíveis como pessoas, frutas, carros, etc. e também conceitos abstratos como amor, guerra, possui, contém, etc. Consiste de uma sequência de caracteres que segue uma regra de sintaxe, a qual é essencialmente um nome de conjunto como “HTTP”, “FTP”, “mailto”, “URN”, entre outros, seguido do caractere dois pontos e por fim a parte específica do conjunto.

Esta sintaxe consiste em uma sequência hierárquica de componentes: sistema, autoridade, caminho, consulta e fragmento. Um tipo muito comum de URI é o Uniform Resource Locator (URL) que podemos definir como um localizador uniforme de recursos. Este tipo de URI não apenas identifica, mas localiza este recurso permitindo que o computador o encontre na Web, por exemplo, <http://www.inf.ufsc.br/graduacao/> onde “HTTP” é o nome do conjunto, “www.inf.ufsc.br” é a autoridade, e “/graduacao/” é o caminho apontando para o documento.

Tecnicamente URL e URN funcionam como identificadores de recursos, no entanto, muitos conjuntos não podem ser categorizados como somente um ou outro, por que todas as URIs podem ser tratados como nomes, e alguns conjuntos integram aspectos de ambas ou de nenhuma das categorias.

No contexto linked data, URIs são usadas para identificar objetos e conceitos, permitindo que eles sejam referenciados para

obtenção de informações a seu respeito. Assim, a referência de uma URI resulta em uma descrição RDF do recurso identificado. Por exemplo, a URI “<http://www.w3.org/People/Berners-Lee/card#i>” identifica o pesquisador Tim Bernes-Lee.

#### 2.5.2.2. URI Aliases

Em um ambiente tão amplo como a Web, pode acontecer de links falarem de um mesmo recurso mas de informações diferentes, por exemplo uma cidade ou pessoa famosa, como eles não tem ligação existem dois URIs para identifica-los. Por exemplo, o DBPedia<sup>5</sup> usa a URI <http://dbpedia.org/resource/Berlin> para identificar a cidade de Berlin, enquanto a Geonames<sup>6</sup> utiliza a URI <http://sws.geonames.org/2950159/>. Como as duas URIs referenciam o mesmo objeto elas são chamadas de aliases URIs. É algo muito comum na Web. Através de owl: sameAs esses links são ligados, ou seja, entrando nas informações da cidade de Berlin pelo Geonames é possível obter outras informações vindas do DBPedia (Berners-Lee, 2010).

#### 2.5.2.3. Escolhendo as URIs

Ao publicar linked data, é muito importante dedicar algum esforço para escolher bons URIs para seus recursos.

- Usar HTTP URIs para tudo. O esquema de <http://> é o esquema de URI só que é amplamente apoiado em ferramentas atuais e infra-estrutura. Todos os outros regimes exigem um esforço extra para serviços web resolver, lidar com registradores, identificadores, e assim por diante. Os argumentos em favor do uso de HTTP são discutidos em vários lugares, por exemplo, nomes e endereços de Norman Walsh e URNs, Namespaces e Registros (rascunho) pela TAG W3C.
- Definir os URIs em um namespace HTTP sob controle próprio, onde realmente se pode torná-los referenciáveis. Não defini-los em alguém do namespace. Manter a implementação cruft fora dos URIs. Considerar estes dois exemplos:

---

<sup>5</sup> *DBpedia* é um projeto cujo objetivo é extrair conteúdo estruturado das informações da Wikipédia. Essa informação estruturada é então disponibilizada na Web. (<http://dbpedia.org>)

<sup>6</sup> Geonames é um banco de dados geográficos disponíveis e acessíveis através de vários serviços da Web, sob uma Licença Creative Commons. (<http://www.geonames.org/>)

- a. <http://dbpedia.org/resource/Berlin>
- b. <http://www4.wiwiss.fu-berlin.de:2020/demos/dbpedia/cgi-bin/resources.php?id=Berlin>
  - Tentar manter as URIs estáveis e persistentes. Mudar as URIs depois vai quebrar todos os links já estabelecidos, por isso é aconselhável dedicar alguma reflexão extra para eles, numa fase inicial.
  - Os URIs que podem ser escolhidos são limitadas pelo ambiente técnico. Se o servidor é chamado `demo.serverpool.wiwiss.example.org` e recebendo outro nome de domínio não é uma opção, então o URI terá que começar com `http://demo.serverpool.wiwiss.example.org/`. Se não pode executar o servidor na porta 80, então o URI pode ter que começar com `http://demo.serverpool.example.org:2020/`. Se possível se deve limpar os URIs, adicionando algumas regras URI reescrita para a configuração do servidor web.
  - Que muitas vezes acabam com três URIs relacionados a um recurso de informações não-simples:
    - 1.um identificador para o recurso;
    - 2.um identificador para um recurso de informações relacionadas adequadamente para navegadores HTML (com uma representação de página web);
    - 3.um identificador para um recurso de informações relacionadas adequado para navegadores RDF (com uma representação RDF / XML).

Aqui estão várias possibilidades para a escolha destes URIs relacionados:

1. <http://dbpedia.org/resource/Berlin>
2. <http://dbpedia.org/page/Berlin>
3. <http://dbpedia.org/data/Berlin>

Ou:

1. <http://id.dbpedia.org/Berlin>
2. <http://pages.dbpedia.org/Berlin>
3. <http://data.dbpedia.org/Berlin>

Ou:

1. <http://dbpedia.org/Berlin>
2. <http://dbpedia.org/Berlin.html>
3. <http://dbpedia.org/Berlin.rdf>

- Muitas vezes será preciso usar algum tipo de chave primária dentro do seu URIs, para se certificar de que cada

um é único. Se puder, deve ser usada uma chave que é significativa dentro do seu domínio. Por exemplo, quando se trata de livros, fazendo a parte número ISBN da URI é melhor do que usar a chave primária de uma tabela do banco de dados internos. Isso também faz mineração de equivalência para derivar links RDF mais fácil.

## 2.6. Ontologias

Na computação existem várias definições versando sobre ontologias, uma delas foi inicialmente proposta por Gruber em 1993 — Ontologia é uma especificação formal e explícita de uma abstração, uma visão simplificada de um domínio de conhecimento.

(Gruber, 1995) ainda completa - Uma ontologia modela uma parte do “mundo”, ou seja, um domínio de conhecimento, definindo um vocabulário comum.

No entanto, (Guarino, 1998) posteriormente discute a definição proposta por Gruber sob a luz de diversos trabalhos disponíveis na literatura. O autor propõe uma definição mais abrangente: — Uma ontologia é uma descrição parcial e explícita de uma conceituação. Para ele, o grau de especificação de uma conceituação depende do propósito desejado para uma ontologia. É introduzida a ideia de classificações, hierarquias e da existência de regras que regulam a combinação entre os termos e as relações. As relações entre os termos são criadas por especialistas, e os usuários formulam consultas usando os conceitos especificados. Uma ontologia define assim uma linguagem (conjunto de conceitos) que será utilizada para formular consultas.

(Noy, et al., 2001) complementa que tal modelo pode ser utilizado tanto por humanos quanto por agentes de software, a fim de estabelecer um entendimento comum sobre os conceitos e relacionamento desse domínio de conhecimento.

Diversos são os benefícios apresentados na literatura para a utilização de ontologias (Noy, et al., 2001), alguns deles são relacionados ao: compartilhamento, reuso, estruturação da informação, interoperabilidade e confiabilidade.

(Beduschi, et al., 2012) afirma que a utilização de ontologias para descrição semântica de um determinado vocabulário permite um entendimento amplo das características e propriedades das classes pertencentes a um domínio, assim como seus relacionamentos. O uso de componentes de inteligência artificial está cada vez mais presente em

aplicações da área de Engenharia de Software, em particular, nas atividades que envolvem uma grande manipulação de informações para tomada de decisão.

Portanto em computação ontologia pode ser definida como um modelo de dados que representam um conjunto de conceitos dentro de um domínio e os seus relacionamentos, ou seja, a definição formal das relações entre termos e conceitos (Gilchrist, 2003). A ontologia é utilizada como uma forma de representação do conhecimento sobre o mundo ou alguma parte dele. Geralmente são descritos os seguintes conceitos:

- **Indivíduos:** os objetos.
- **Classes:** coleções, conjuntos ou tipos de objetos.
- **Atributos:** características, propriedades ou parâmetros que os objetos podem ter e compartilhar.
- **Relacionamentos:** as formas como os objetos podem se relacionar com outros objetos.

As ontologias podem ser classificadas em diferentes tipos, de acordo com seu grau de generalidade, e podem ter as seguintes descrições (Gómez-Pérez, et al., 1999):

- **Ontologias de representação** definem as primitivas de representação, como frames, axiomas, atributos e outros, de forma declarativa. Essa ideia abstrai os formalismos de representação. (Chandrasekaran, et al., 1999)
- **Ontologias gerais** (ou de topo) trazem definições abstratas necessárias para a compreensão de aspectos do mundo, como tempo, processos, papéis, espaço, seres, coisas, etc.
- **Ontologias centrais** (core ontologies) ou genéricas de domínio definem os ramos de estudo de uma área ou conceitos mais genéricos e abstratos desta área.
- **Ontologias de domínio** tratam de um domínio mais específico de uma área genérica de conhecimento, como direito tributário, microbiologia, etc.
- **Ontologias de aplicação** procuram solucionar um problema específico de um domínio, como identificar doenças do coração, a partir de uma ontologia de domínio de cardiologia. Normalmente, ela referencia termos de uma ontologia de domínio.

### **2.6.1. Metodologias e Ferramentas para o Desenvolvimento de Ontologias**



Segundo (Gasevic, et al., 2006), para o desenvolvimento de ontologias é necessário um esforço considerável de engenharia, disciplina e rigor; onde princípios de projeto, atividades e processos de desenvolvimento, tecnologias de suporte e metodologias sistêmicas devem ser empregados. Neste sentido, surge a Engenharia de Ontologias preocupando-se com o conjunto de atividades, o processo de desenvolvimento de ontologias, o ciclo de vida de ontologias, os métodos e metodologias para desenvolver ontologias e as ferramentas e linguagens de suporte à construção de ontologias (Gómez-Pérez, et al., 2004).

Segundo Pinto e Martins (2004) e Ye et al (2007), a terminologia de Engenharia de Ontologias é baseada na Engenharia de Software. Por conseguinte, no processo de desenvolvimento de ontologias, usualmente, são aceitas as atividades de especificação, conceitualização, formalização, implementação e manutenção. A cada uma destas atividades existem tarefas a serem executadas, como seguem:

- Especificação: identificar o propósito e o escopo da ontologia. O propósito responde a questão “por que a ontologia é construída?”, enquanto o escopo responde a questão “quais são as intenções de uso e usuários da ontologia?”
- Conceitualização: descrever, em modelo conceitual, a ontologia a ser construída, de acordo com as especificações encontradas no estágio anterior. Cabe ressaltar que o modelo conceitual de uma ontologia pode ser construído mediante o emprego de ferramentas formais e informais. Tal modelo consiste em conceitos do domínio, as relações entre os conceitos e as propriedades dos conceitos.
- Formalização: transformar a descrição conceitual em um modelo formal. Nesta fase, conceitos são definidos através de axiomas que restringem as possíveis interpretações de seu significado e também organizados hierarquicamente através de relações estruturais, tais como “é-um” ou “parte-de”.
- Implementação: implementar a ontologia formalizada em uma linguagem de representação do conhecimento. Para isso, um pré-requisito é a escolha da linguagem de representação adequada.
- Manutenção: atualizar e corrigir a ontologia desenvolvida, de acordo com o surgimento de novos requisitos.

Além disso, Pinto e Martins (2004) também pontuam outras atividades devem ser executadas durante o ciclo de vida de uma ontologia, sendo elas:

- **Aquisição do conhecimento:** adquirir conhecimento sobre um domínio por meio de técnicas de eliciação do conhecimento com especialistas de domínio ou recorrer à bibliografia relevante. Várias técnicas podem ser utilizadas, como brainstorming, entrevistas, questionários, análise de texto e técnicas indutivas.

- **Avaliação:** julgar tecnicamente a qualidade da ontologia por meio da:

- o **Avaliação técnica:** julgar a ontologia e a documentação diante um frame de referência. Há duas tarefas envolvidas:

- **verificação**, a qual garante a correção da ontologia de acordo com o entendimento aceito sobre o domínio em fontes de conhecimento especializadas; e

- **validação**, a qual garante que a ontologia corresponde a sua suposta finalidade, de acordo com os documentos de especificação de requisitos.

- o **Avaliação dos usuários:** julgar a ontologia do ponto de vista do usuário, em relação a sua usabilidade e utilidade; e do ponto de vista da (re)utilização em outras aplicações conforme a sua documentação.

- **Documentação:** relatar o que, como e por que foi feito. Uma documentação associada com os termos presentes na ontologia é importante, não somente para melhorar a clareza da ontologia, mas também para facilitar a manutenção, uso e reuso.

Ressalta-se que o conjunto de atividades anteriormente enumeradas pode não ser contemplado totalmente em uma metodologia para desenvolvimento de ontologias. Para Corcho et al (2003), existem metodologias que são empregadas em tarefas específicas na Engenharia de Ontologias. Corroborando, para Fernandez-López e Gómez-Pérez (2002) em cada metodologia proposta existem atividades que deixam de estar compreendidas. Por isso, segundo os autores e para Sure e Studer (2003) e Brusa et al (2008), uma combinação de metodologias se torna pertinente no processo de desenvolvimento de ontologias.

### *2.6.1.1. Ontokem*

O ontoKEM foi concebido e desenvolvido no Laboratório de Engenharia do Conhecimento (LEC) do Programa de Engenharia e Gestão do Conhecimento (EGC) da Universidade Federal de Santa

Catarina. O desenvolvimento contou com a participação dos professores José Leomar Todesco e Fernando Álvaro Ostuni Gauthier, dos doutorandos do EGC Sandro Rautenberg e Rafael Speroni e dos graduandos Polite Lottin e Cleiton E. J. Duarte. Seu primeiro uso se deu na construção de uma ontologia de domínio para um projeto projeto P&D “Sistemas de Conhecimento para Gestão da Rede de Distribuição de Média Tensão”. Esta ferramenta tem apoiado em projetos de pesquisa e em atividades de ensino de graduação e pós-graduação.

Para se construir ontologias, várias metodologias foram propostas. Contudo, Fernandez-López & Gómez-Peréz (2002) comenta que não há metodologia completamente madura para o propósito de construção de ontologias. Em cada metodologia existem atividades que deixam de estar compreendidas. Segundo os autores, uma combinação de metodologias se torna interessante mediante um processo de construção de ontologias. Este é o pilar de sustentação do ontoKEM como ferramenta de EC para construção de ontologias, baseado-se nas metodologias 101 (NOY & MCGUINNESS, 2008), On-to-Knowledge (FENSEL & HERMELEN, 2008) e METHONTOLOGY (Gómez-Pérez, et al., 2004) onde se enumera-se como contribuições:

- Metodologia 101: metodologia que prega a construção de ontologias num processo iterativo de sete passos (determinar o escopo da ontologia, considerar o reuso, listar termos, definir classes, definir propriedades, definir restrições e criar instâncias). Este processo é compreendido no ontoKEM.

- On-to-Knowledge: metodologia que incute as questões de competência como modo simples e direto para determinar o escopo de uma ontologia e permite identificar conceitos, propriedades, relações e instâncias. O ontoKEM faz uso deste instrumento tanto na compreensão da aplicabilidade da ontologia, quanto na disposição do artefato de documentação pregado.

- METHONTOLOGY: metodologia que formaliza a construção de ontologias através de uma rica gama de artefatos de documentação (documentos-texto e quadros). Estes artefatos são usados como modelos de documentos no ontoKEM.

Resumindo e definindo, o ontoKEM é uma ferramenta para EC que suporta um processo de construção e documentação de ontologias, baseado no processo de desenvolvimento da metodologia 101 (NOY & MCGUINNESS, 2008) e nos artefatos documentais das metodologias para ontologias On-to-Knowledge (FENSEL & HERMELEN, 2008) e METHONTOLOGY (GOMÉZ-PERÉZ et al, 2004).

### 2.6.1.2. Methontology

METHONTOLOGY está entre as metodologias de engenharia de ontologias mais abrangentes, pois é um processo para a construção de ontologias a partir do zero, reutilização de ontologias de terceiros em sua forma original, ou por um processo de re-engenharia. A estrutura permite a construção de ontologias ao nível de conhecimento, ou seja, o nível conceitual, em oposição ao nível de implementação. A estrutura consiste de: identificação do processo de desenvolvimento de ontologias, com a identificação das principais atividades, tais como: configuração de avaliação, gestão e execução conceitualização, integração, um ciclo de vida com base na evolução dos protótipos, e a própria metodologia especifica os passos para a realização as atividades, as técnicas utilizadas, os resultados e sua avaliação. Eles descrevem muito detalhadamente o processo de construir uma ontologia para sistemas baseados em ontologias centralizados.

### 2.6.1.3. On-to-knowledge (OTK)

On-to-Knowledge é uma metodologia de desenvolvimento de ontologias fruto da cooperação de várias entidades europeias (Fensel, et al., 2008), tendo como intuito desenvolver ontologias para serem empregadas em Sistemas de Gestão do Conhecimento. Conforme mostra a Figura 2-5, esta metodologia é dividida em cinco fases (Sure, et al., 2003), sendo elas:

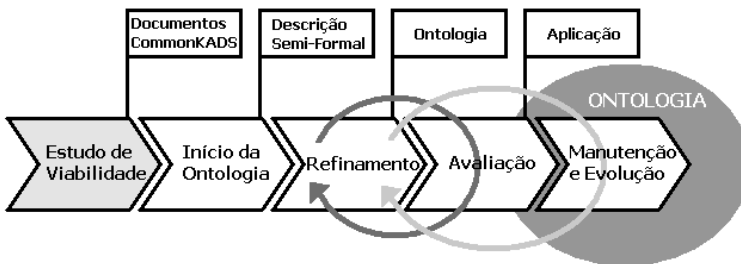


Figura 2-10 Processo de desenvolvimento da metodologia On-to-Knowledge

1. Estudo de viabilidade: é uma fase anterior ao desenvolvimento de ontologias. Amparada pela

metodologia de Engenharia do Conhecimento CommonKADS (Schreiber, 2002), o estudo de viabilidade destina-se a identificar problemas e oportunidades de uma organização, objetivando mapear a real necessidade do desenvolvimento de uma ontologia.

2. Início da ontologia: na metodologia, o desenvolvimento de uma ontologia inicia nesta fase. Fazendo uma analogia ao processo de software, aqui se objetiva produzir documentos de especificação de requisitos, definindo o domínio e objetivos da ontologia, utilizando padrões de projeto, identificando as fontes de conhecimento, definindo atores e cenários, enumerando questões de competência, definindo o ambiente de desenvolvimento da ontologia, entre outros.
3. Refinamento: o objetivo desta fase é desenvolver uma ontologia a ser utilizada em um Sistema de Gestão do Conhecimento, de acordo com os documentos produzidos nas fases anteriores. Para tanto, engenheiros do conhecimento se valem de técnicas de elicitación do conhecimento ao interagir com os especialistas de domínio, modificando e estendendo a ontologia em desenvolvimento em direção de uma versão estável.
4. Avaliação: o objetivo desta fase é a aferição da completude e precisão da ontologia mediante a documentação gerada durante o desenvolvimento da ontologia e um frame de referência, o qual pode corresponder às questões de competência enumeradas na fase “início da ontologia”.
5. Manutenção e Evolução: esta é uma fase de responsabilidade da organização. É importante ter ciência dos atores responsáveis pela manutenção da ontologia e das regras para sua manutenção.

A característica principal da On-to-Knowledge é sua preocupação com as fases iniciais do estudo de viabilidade e início da ontologia. Neste sentido, corroborando o guia *Ontology Development 101*, é clara a necessidade de definir o domínio e o escopo da ontologia, sobretudo, na utilização de questões de *competência para tal delineamento*.

#### 2.6.1.4. Ferramenta protege

Desenvolvido pelo departamento de informática médica da Universidade de Stanford, a ferramenta Protegé é utilizada para construir ontologias de domínio, personalizar formulários de entrada de dados, inserir e editar dados, possibilitando então, a criação de bases de conhecimento guiadas por uma ontologia.

Possui cinco áreas de visualização (views) que funcionam como módulos de navegação e edição de classes, atributos, formulários, instâncias e pesquisas na base de conhecimento, propiciando a entrada de dados e a recuperação das informações.

## **2.6.2. Ontologias Disponíveis**

A seguir serão apresentadas ontologias amplamente difundidas e utilizadas com propósito de compartilhamento e interoperabilidade de dados entre sistemas computacionais.

### *2.6.2.1. SKOS*

A Simple Knowledge Organization System (SKOS) é um modelo de dados comum para compartilhamento e ligação de sistemas de organização do conhecimento através de um documento Web.

Segundo a W3C, a SKOS fornece um modo padrão para representar os sistemas de conhecimento da organização usando o Resource Description Framework (RDF). Que codifica esta informação em RDF permite que seja passado entre aplicações de computador de uma maneira interoperável.

### *2.6.2.2. DCAT*

Segundo a W3C, DCAT é um vocabulário RDF projetado para facilitar a interoperabilidade entre catálogos de dados publicados na web. Seu objetivo principal é a expressão de catálogos do governo de dados, como data.gov ou data.gov.uk, em RDF. Ele está sendo produzido pelo W3C eGovernment Interest Group.

Fazendo uso do DCAT para descrever conjuntos de dados em catálogos de dados, pode-se aumentar a descoberta e permitir que aplicativos consumam facilmente metadados de vários catálogos. Além disso, permite a publicação descentralizada de catálogos e facilita a pesquisa a datasets federados. A agregação de metadados DCAT pode servir como um registro para facilitar a preservação digital.

A seguir, a Figura 2-11 apresenta as classes e propriedades propostas pelo vocabulário DCAT.

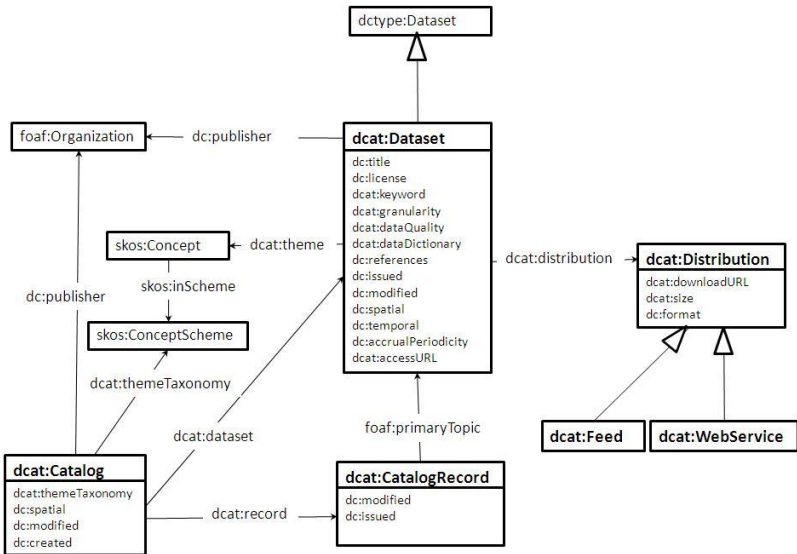


Figura 2-11 Diagrama de Representação do vocabulário DCAT

## 2.7. Considerações Finais

Ao estudar os princípios da Web semântica, Linked Data, ontologia, proveniência e demais conceitos envolvidos, temos a possibilidade de um melhor entendimento da proposta das técnicas de Linked Data, assim como da importância do uso da proveniência para assegurar a qualidade de dados publicados na Web.

A utilização de Linked open data vem sendo adotada por várias organizações, sem se preocupar com os tipos de plataformas e armazenamentos. Muitos projetos vêm surgindo para diversos fins e inclusive há incentivos governamentais em diversos países para a disponibilização de informações no modelo de Dados Abertos.

Nota-se uma crescente busca pela melhoria na qualidade dos dados publicados na Web, e que modelos de proveniência tem por objetivo apoiar diretamente esta questão.

### **3. PROPOSTA DE UMA ONTOLOGIA DE PROVENIÊNCIA PARA PUBLICAÇÃO DE LINKED DATA**

#### **3.1. Procedimentos Metodológicos**

- Revisão da Literatura
- Analisar metodologias de desenvolvimento de ontologia
- Escolher uma metodologia de uso
- Construir a ontologia sobre o processo de Linked Data.
- Construir a ontologia sobre proveniência de Linked Data.
- Aplicar a ontologia de proveniência de Linked em um cenário de uso.

#### **3.2. Revisão da Literatura**

Através dos estudos realizados e apresentados no capítulo anterior, foi possível tomar conhecimento de uma série de técnicas para a publicação de dados abertos no formato Linked Data; assim como processos para desenvolvimento de ontologias e formas de aplicação de proveniência e seus modelos de uso.

Pode-se observar a existência de modelos já bem encaminhados como o vocabulário DCAT, e o modelo de proveniência PROV-DM, os quais poderão ser reutilizados para o desenvolvimento da ontologia proposta.

No entanto, faz-se necessária a modelagem do processo de Linked Data, para assim, possibilitar a aplicação das práticas de proveniência de dados e por fim a criação de uma ontologia de proveniência de Linked Data.

Em vista disto, foi construída uma primeira ontologia com objetivo de descrever o processo Linked Data, com base nas técnicas apresentadas neste trabalho, para que, a posteriori, se desse a construção de uma segunda ontologia aplicando proveniência sobre a ontologia citada primeiramente, buscando assim, os objetivos almejados neste trabalho.

#### **3.3. Análise das metodologias de desenvolvimento de ontologias**



Para a construção das ontologias descritas, é importante que o processo de desenvolvimento considere o reuso de ontologias disponíveis, uma vez que já existe a disposição ontologias que descrevem objetos necessários para a ontologia objetivo deste trabalho.

Caso do vocabulário DCAT, que descreve catálogos e datasets, itens resultantes do processo Linked Data. Assim como a ontologia PROV-DM, que já define um amplo conjunto de conceitos relativos à proveniência de dados.

A Metodologia 101 prega a construção de ontologias num processo iterativo de sete passos (determinar o escopo da ontologia, considerar o reuso, listar termos, definir classes, definir propriedades, definir restrições e criar instâncias).

Seguindo esta metodologia, deu-se início a construção da ontologia de publicação de dados em Linked Data.

### **3.4. Ontologia sobre o processo de Linked Data**

Para facilitar o entendimento, denominarei esta primeira ontologia, sobre o processo de publicação de Linked Data, como Ontologia A.

#### **3.4.1. Processo de Desenvolvimento**

##### *3.4.1.1. Definição do Escopo*

No intuito de auxiliar na delimitação do escopo desta ontologia, fez-se uso do processo de especificação proposto pela metodologia *ontokem*. Deste modo, criou-se um conjunto de perguntas de competência visando levantar os conceitos relacionados e aplicá-los.

Foram tomados como base os conceitos indicados pelas propostas de práticas citadas no item 2.5.2 deste trabalho. O resultado desta etapa encontra-se disponível no Apêndice A.

##### *3.4.1.2. Reutilização de ontologias disponíveis:*

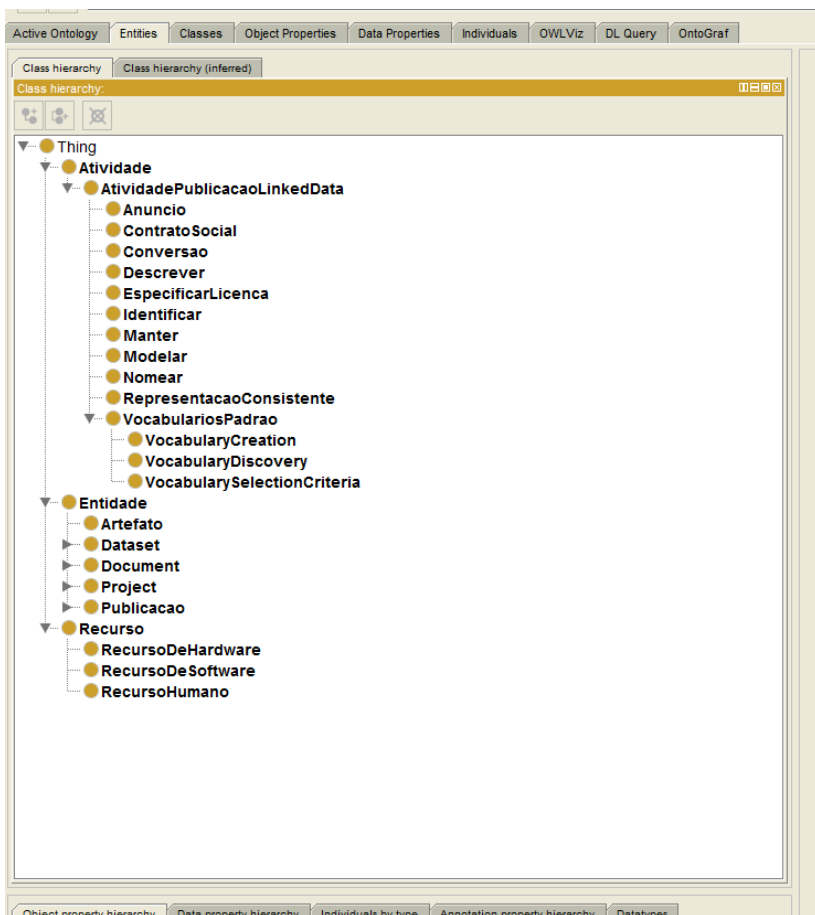
Nesta etapa, foi realizada uma análise das ontologias disponíveis no intento de responder à pergunta “Que recursos poderiam ajudar nesta solução?”; A publicação de Linked Data gera como resultado um *dataset* disponibilizado na Web. A ontologia DCAT

fornece suprimentos para modelagem dos mesmos; Outra necessidade refere-se aos conceitos envolvidos nos procedimentos de criação, relacionamento de dados em um ambiente Linked Data. Para tal, a SKOS fornece a modelagem para conceituar as relações e ligações entre sistemas de organização de conhecimento através da semântica da Web.

#### *3.4.1.3. Listar Termos*

Neste ponto, sobrevieram as etapas de refinamento da ontologia e da importação de ontologias disponíveis. No intuito de auxiliar tais objetivos, a ferramenta de desenvolvimento Protégé foi selecionada com adequada para tais necessidades.

A seguir serão apresentadas imagens do desenvolvimento da ontologia e o andamento das etapas segundo as telas da ferramenta.



*Figura 3-1 Listagem de Termos identificados para a Ontologia A.*

Na Figura 3-1 é apresentada a listagem de termos (entidades) identificados para a Ontologia A, referente ao processo de publicação de Linked Data. Os termos foram registrados diretamente na Interface da ferramenta Protégé.

### 3.4.1.4. Definir Classes

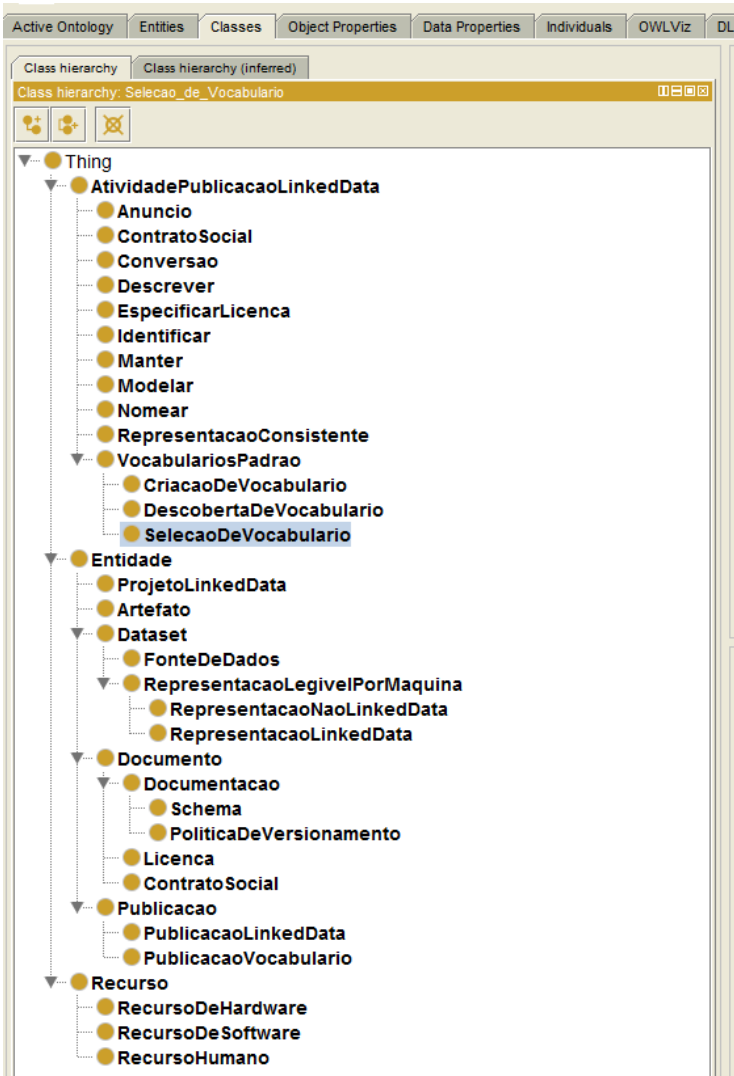


Figura 3-2 Lista de Classes Ontologia A

A Figura 3-2 descreve as classes definidas para Ontologia A, com base nos modelos apresentados no capítulo 2.

### 3.4.1.5. Definir Propriedades

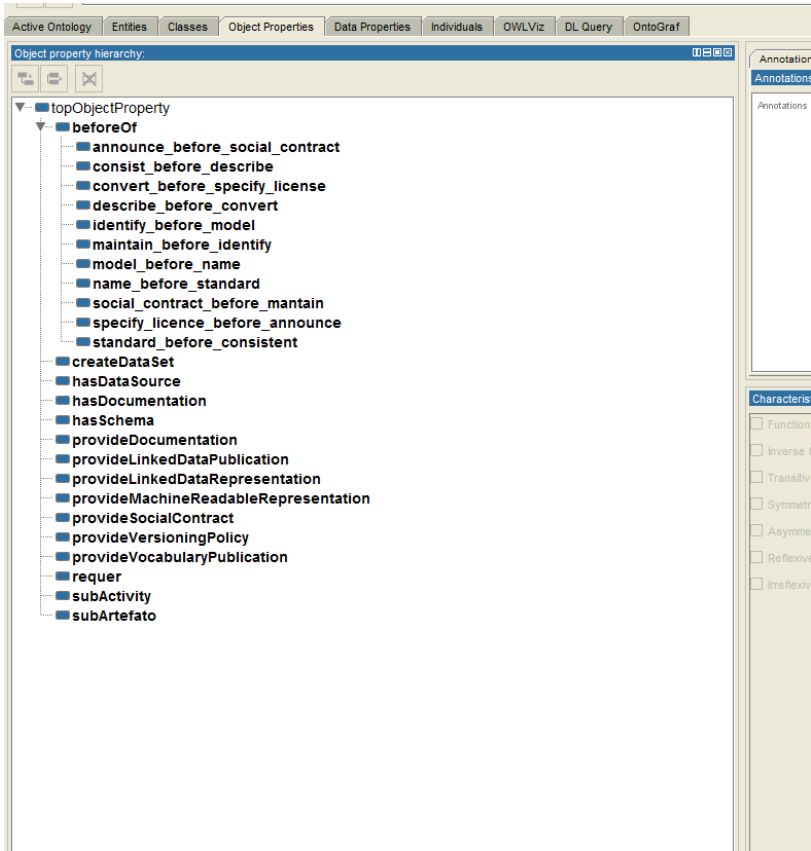


Figura 3-3 Lista de Propriedades definidas para a ontologia A

A figura 3-3 relaciona a lista de propriedades da Ontologia A. Cada propriedade representa um tipo de relacionamento entre as classes e define quais possibilidades de interação entre as mesmas.

### 3.4.2. Ontologia criada

#### a) Grafo Final da Ontologia:



Figura 3-4 Diagrama de classes e relacionamentos da Ontologia A

Na figura 3-4 é apresentado um diagrama em formato de grafo referente ao resultado final da ontologia, neste diagrama pode-se observar a hierarquia das classes e entender o fluxo no processo de publicação de *Linked Data*.

### **3.5. Ontologia sobre proveniência de dados publicados em Linked Data**

#### **3.5.1. Processo de Desenvolvimento**

Seguindo o mesmo processo de desenvolvimento utilizado na Ontologia A, deu-se início ao desenvolvimento da ontologia sobre proveniência de dados publicados em Linked Data.

##### *3.5.1.1. Definição do Escopo*

Esta ontologia se define no mapeamento de proveniência sobre o processo de Linked Data, processo este modelado na Ontologia A.

Cabe neste processo identificar as entidades da ontologia A e relacioná-las aos métodos de registro de proveniência descritos no item 2.2 deste trabalho.

##### *1. Reutilização de ontologias disponíveis:*

Dentre as ontologias disponíveis para reutilização, optou-se pela PROV-DM por atender um maior número de propriedades relacionadas ao registro de informações de proveniência sobre processos e atividades, sejam estas, humanas ou computacionais.

## 2. Listar Termos

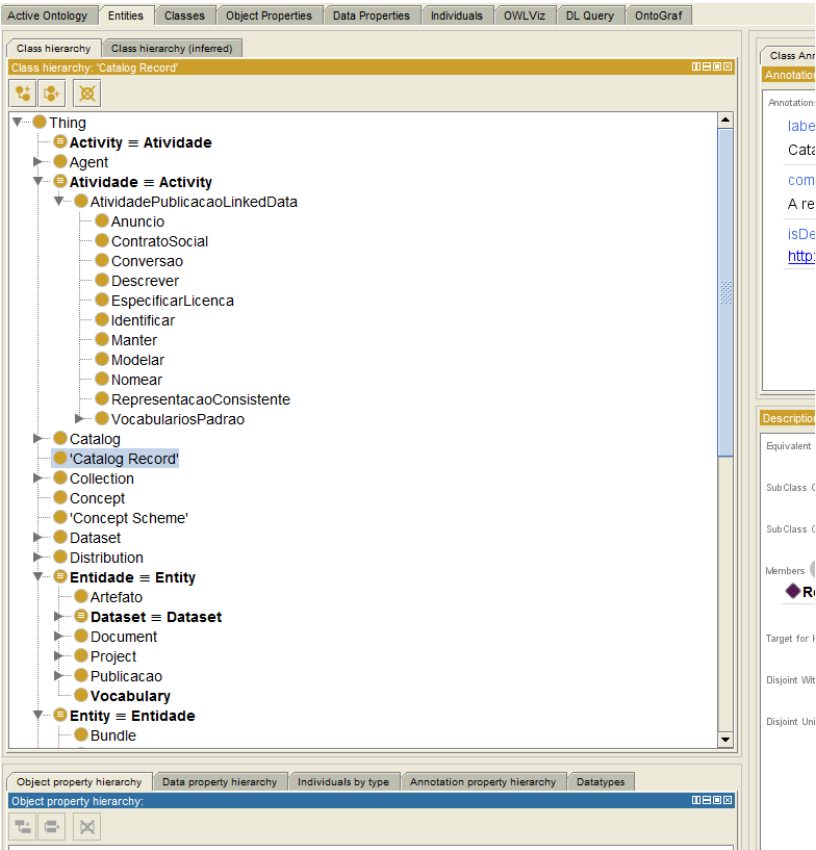


Figura 3-5 Lista de Termos da Ontologia B.

A Figura 3-5 apresenta a lista de termos da Ontologia B, referente ao registro de dados e proveniência sobre os processos e atividades da ontologia A. A continuação da listagem pode ser observada na Figura 3-6.



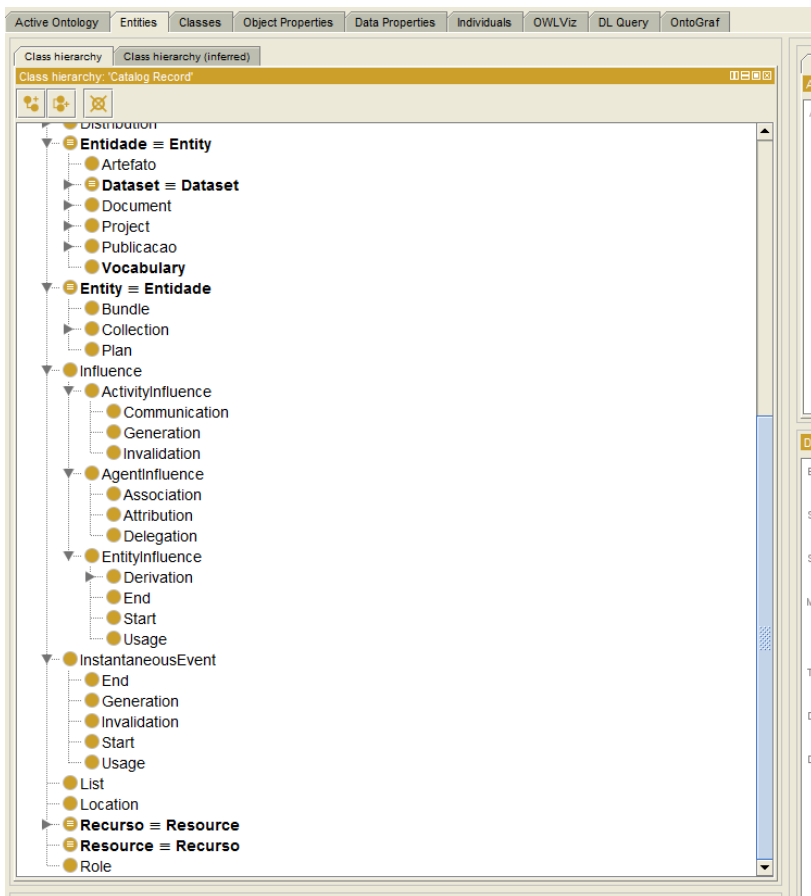


Figura 3-6 Lista de Termos da Ontologia B (Continuação)

### 3. Definir Classes

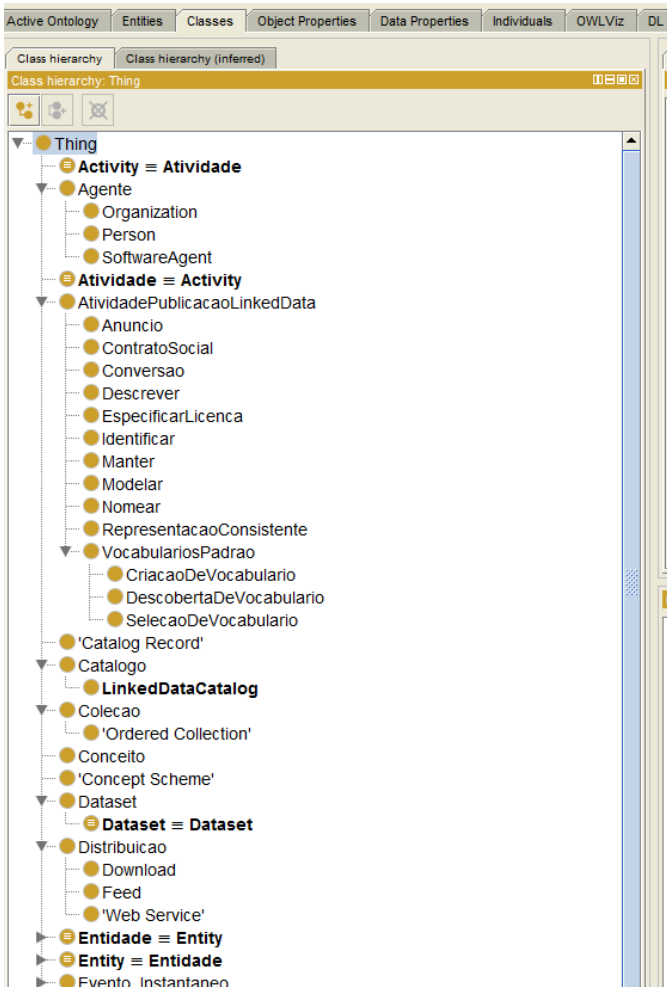


Figura 3-7 Lista de Classes Ontologia B

A Figura 3-7 relaciona as Classes definidas para a Ontologia B, neste ponto a ontologia A está incorporada ao processo para relacionamento das entidades, classes e propriedades.

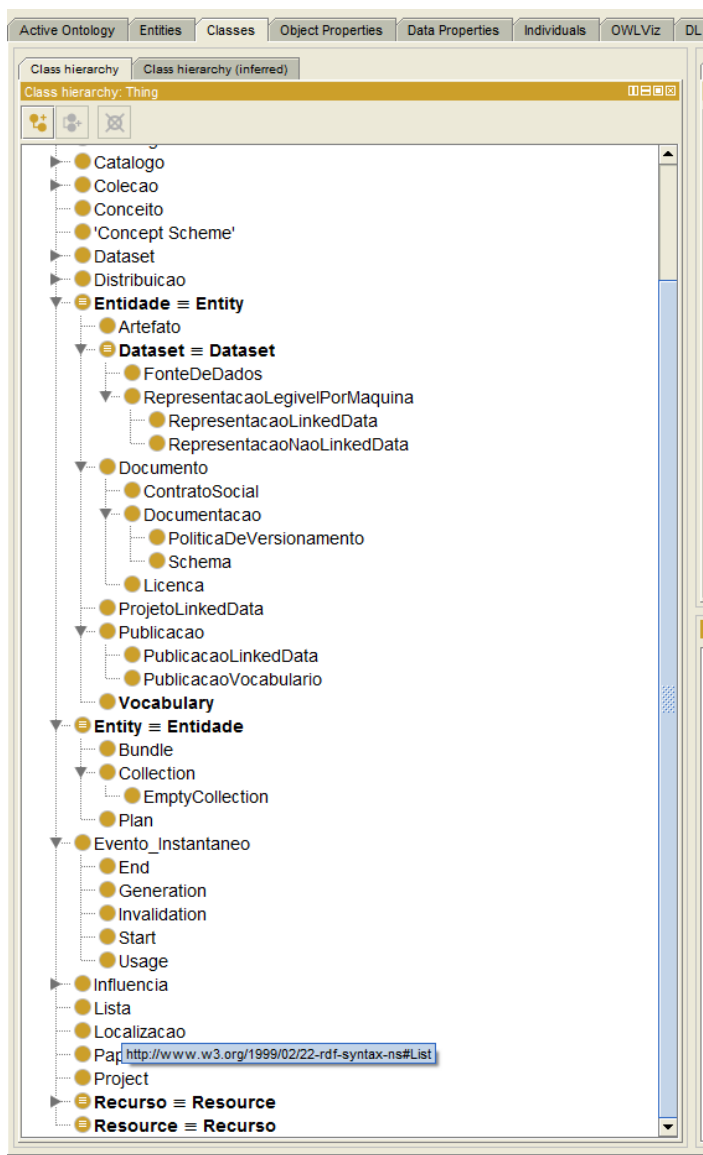


Figura 3-8 Lista de Classes Ontologia B(Continuação)

A Figura 3-8 e a Figura 3-9 apresentam a continuação referente às classes definidas para a Ontologia B (Figura 3-7).

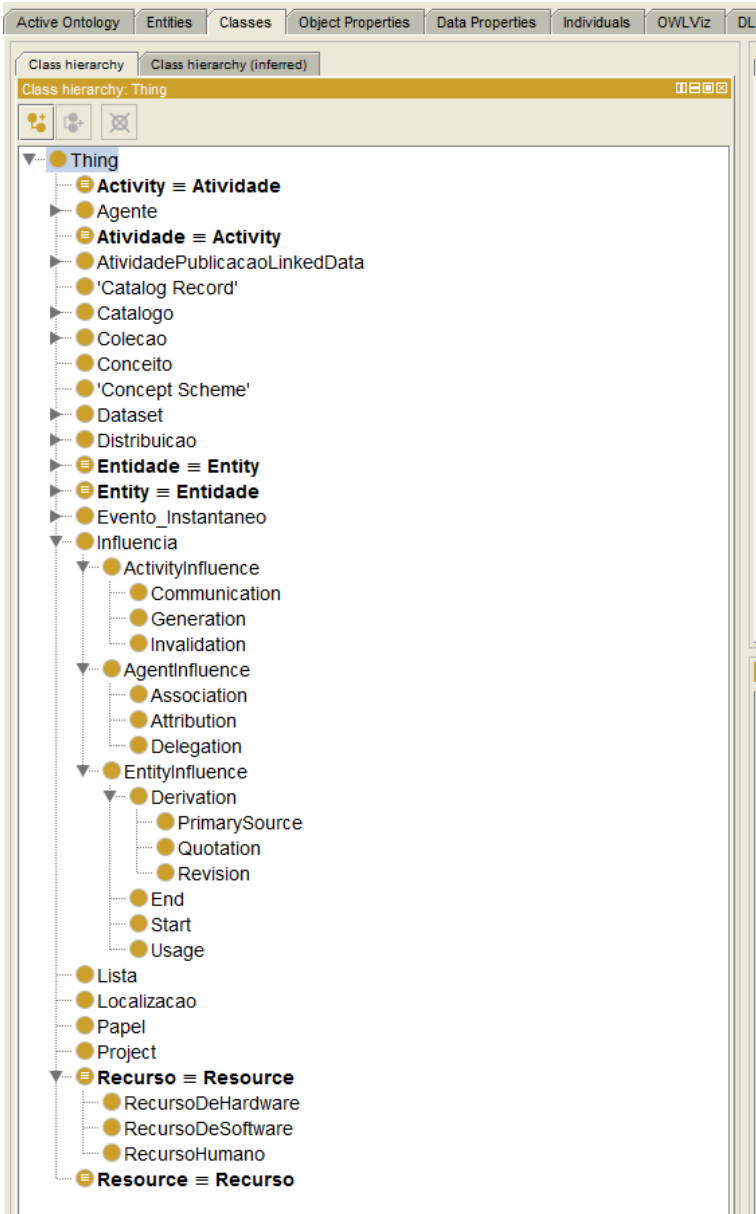


Figura 3-9 Figura 3 8 Lista de Classes Ontologia B(Continuação2)

#### 4. Definir Propriedades

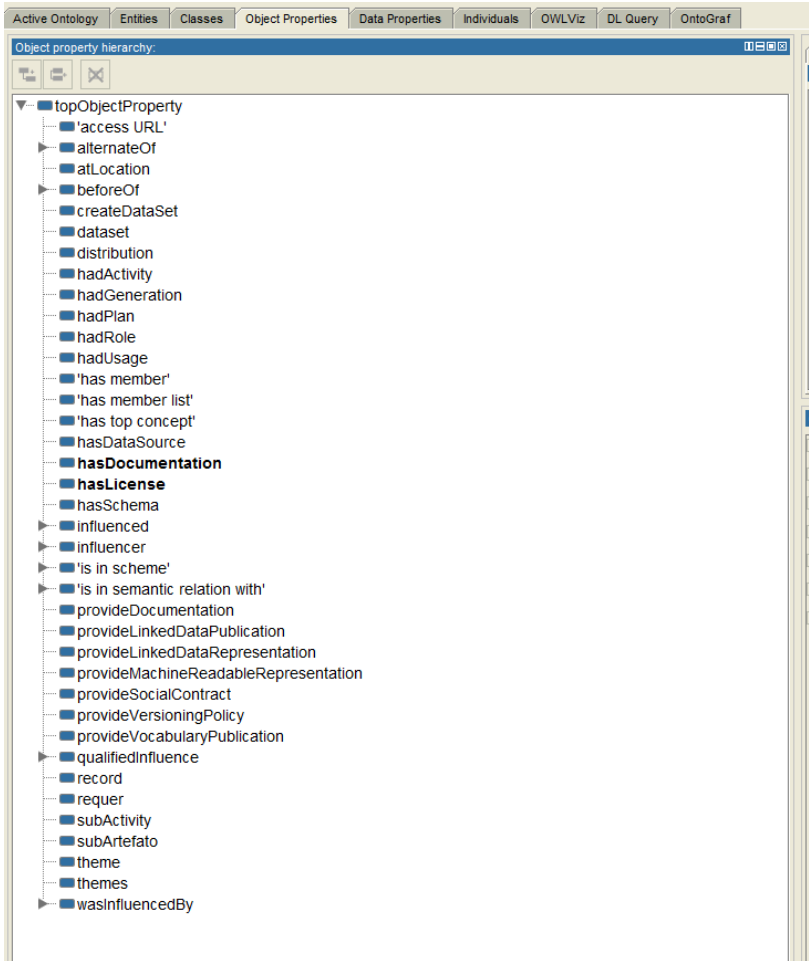


Figura 3-10 Lista de propriedades da ontologia B

A figura 3-10 apresenta as propriedades relacionadas a ontologia B, neste passo podemos observar os relacionamentos entre e processos da ontologia A com agentes de proveniência, como por exemplo, as atividades “*hadActivity*” ou “*hadRole*”.

### 3.5.2. Ontologia criada

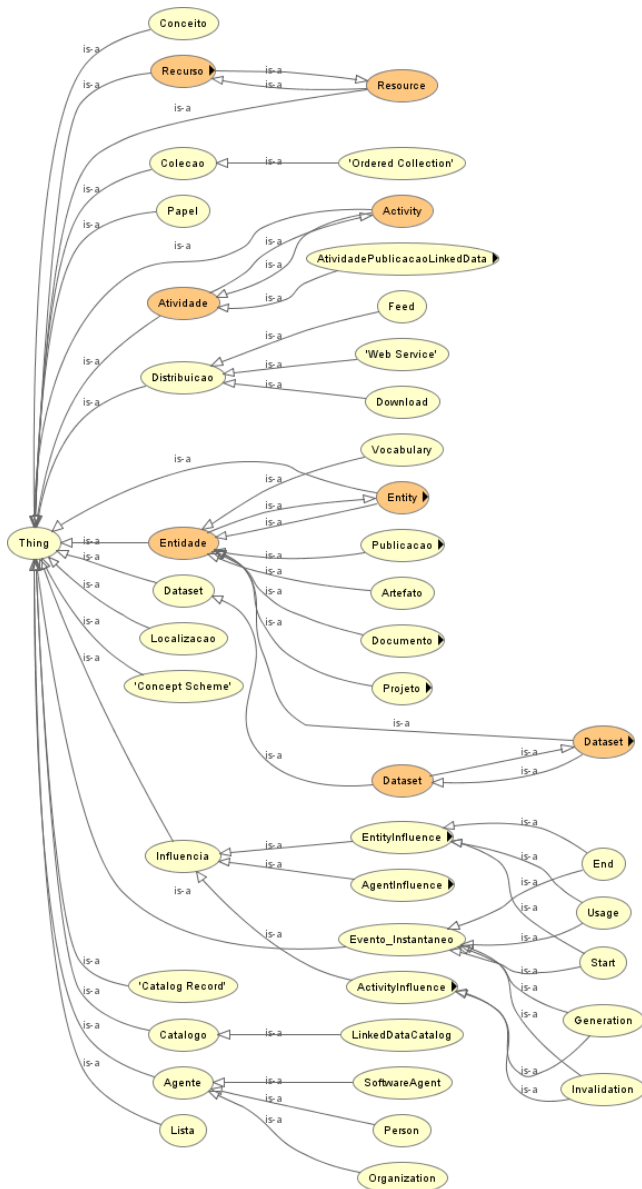


Figura 3-11 Diagrama de classes e relacionamentos da Ontologia B

A figura 3-11 apresenta um diagrama em formato de grafo referente ao resultado final da ontologia B, neste diagrama pode-se observar o relacionamento das classes da ontologia A com as novas classes referentes ao preceito de proveniência.

### **3.6. Aplicar o modelo de proveniência proposto**

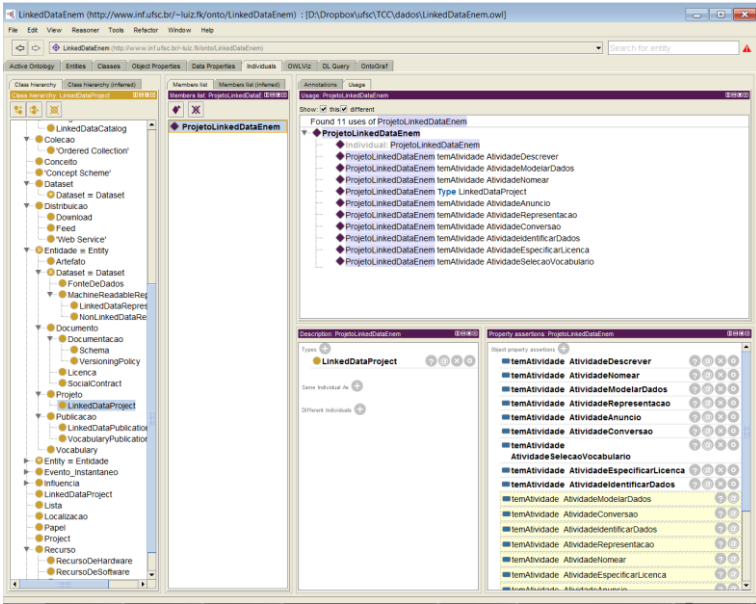
Para a aplicação prática da ontologia de proveniência sobre Linked Data, um *case* de publicação Linked Data foi utilizado como modelo para a criação de instâncias da ontologia, a fim de realizar consultas de inferência.

#### **3.6.1. Cenário de uso**

BEDUSCHI N. B., CABRAL S. P; (2012) publicou um trabalho relatando a publicação de dados do Exame Nacional do Ensino Médio (ENEM) em formato Linked Data. Neste trabalho, foram apresentadas ferramentas utilizadas, a preparação dos dados e o processo de publicação. Utilizaremos a seguir as informações documentadas neste trabalho, em caráter de exemplo.

#### **3.6.2. Representação do projeto de publicação de dados do ENEM em Linked Data**

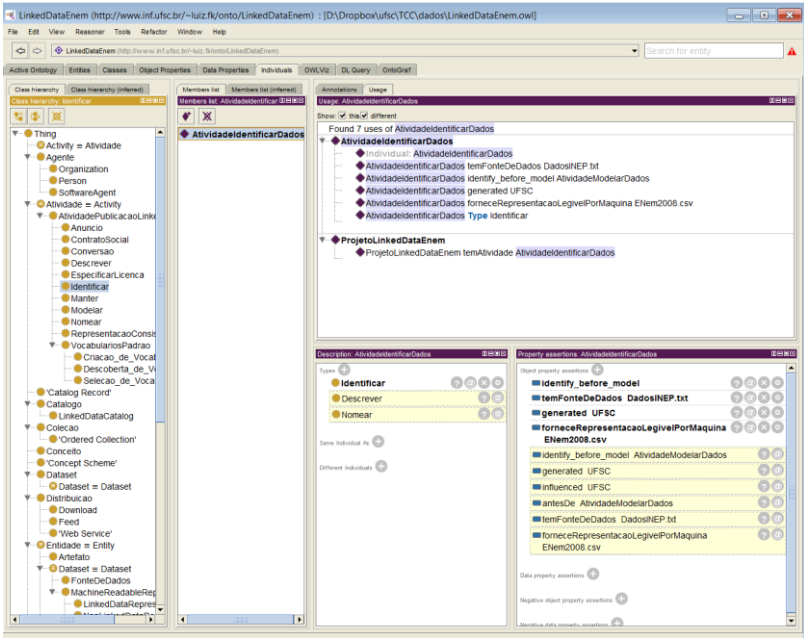
O processo foi representado através da ferramenta Protégé, seguindo o modelo sugerido pela ontologia proposta neste trabalho. Foram criadas instâncias para as etapas do projeto, os agentes envolvidos (organizações, pessoas, agentes de software), as entidades utilizadas e geradas durante o processo (arquivos fonte, arquivos de transformação, conversão, etc), registros de documentação, dentre outros. A seguir serão apresentadas imagens referentes à criação das instâncias.



*Figura 3-12 Criação de Instancias para o Projeto Linked Data Enem*

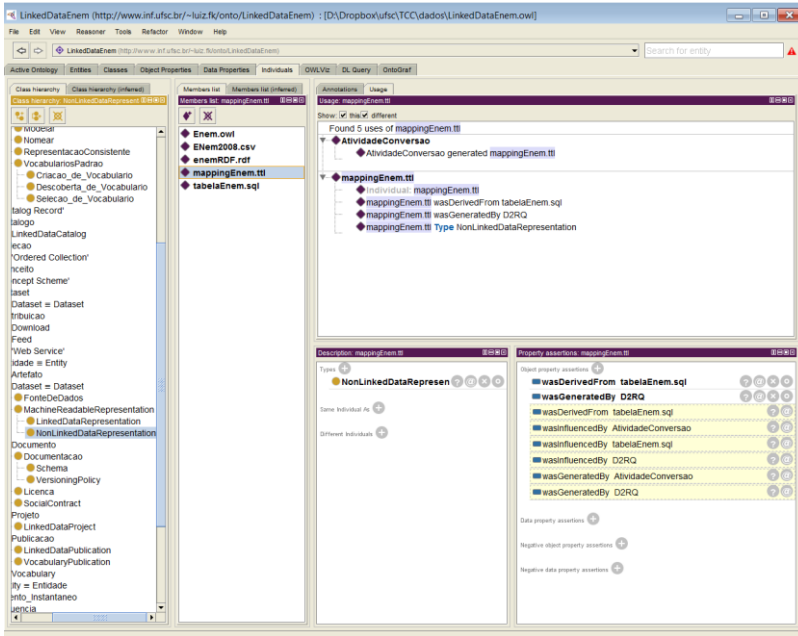
A Figura 3-12 demonstra a tela de criação de Instâncias na ontologia, cada instancia representa uma entidade (atividade ou recurso) utilizada pelo projeto.





*Figura 3-13 Criação das Instâncias para Etapa de Identificação de Dados*

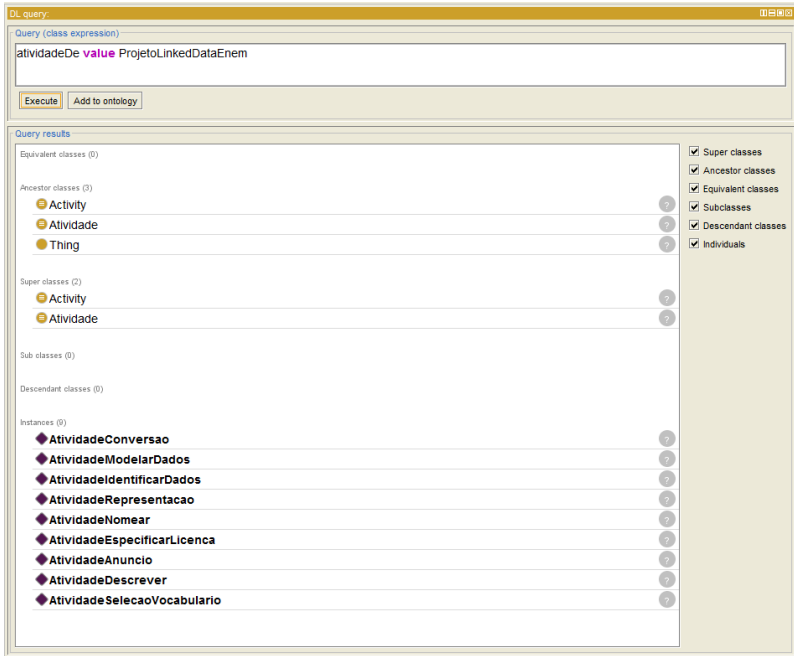
A Figura 3-13 exemplifica a de criação de Instâncias na ontologia para a etapa de Identificação de Dados.



*Figura 3-14 Criação das Instâncias para Representação das Entidades de Dados gerados na etapa de Conversão*

A Figura 3-14 exemplifica a de criação de Instâncias na ontologia para a etapa de Conversão de Dados.

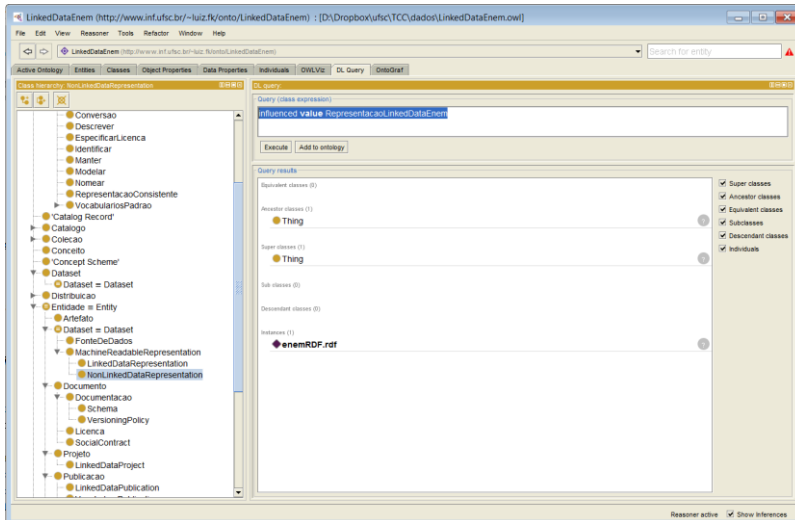
### 3.6.4. Consultas de Inferência



*Figura 3-15 Consulta sobre Atividades que ocorreram no projeto Linked Data Enem*

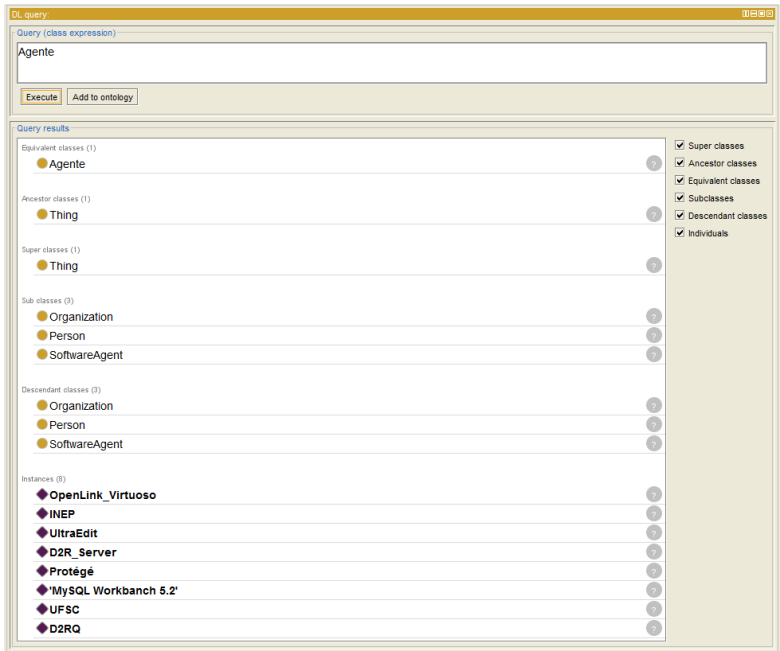
Na figura 3-15, é exemplificada uma consulta sobre as Atividades que ocorreram no projeto Linked Data Enem.

A consulta sobre as Atividades realizadas no projeto é realizada diretamente em linguagem SPARQL, para fins de experimentação a própria ferramenta Protegé pode ser utilizada em conjunto com o mecanismo de inferência para realizar consultas sobre a ontologia.



*Figura 3-16 Consulta sobre as Entidades que Influenciaram o Dataset Linked Data Enem*

Na figura 3-16, outro exemplo é dado em relação a consultas sobre a ontologia, no caso, buscando o arquivo gerado pelo processo de publicação, que representa o *dataset* do projeto



*Figura 3-17 Agentes (Organizações, Pessoas, Softwares) envolvidos no processo.*

Na Figura 3-17 é exemplificado a listagem de todos os agentes envolvidos no processo, são recuperados registros das ferramentas utilizadas e das entidades organizacionais relacionadas.

### 3.7. Resultados

Partindo inicialmente da especificação de boas práticas para a publicação de Linked Data, foi realizado um processo de construção de uma ontologia através da metodologia 101, visando modelar o processo de publicação de Linked Data.

Com esta ontologia foi possível analisar as atividades executadas em Linked Data e assim partir para um estudo sobre a aplicabilidade de proveniência sobre o processo de publicação. Como resultado deste estudo, foi realizado um novo processo de construção de uma ontologia para proveniência de dados publicados em Linked Data.

Para a aplicação prática desta, um case de publicação Linked Data foi utilizado como modelo para a criação de instâncias da ontologia, a fim de realizar consultas de inferência.

## 4. CONCLUSÃO

### 4.1. Considerações finais

A questionável qualidade dos dados na Web nos obriga a procurar e definir processos de controle no processo de publicação da informação. A Web Semântica trás princípios que podem auxiliar esta necessidade. No entanto, a dificuldade em harmonizar informações extraídas de conjuntos de dados distintos ainda é se faz presente, tornando a prática de *Linked Data* complexa e custosa e de resultado também questionável.

A proveniência de dados pode ser uma solução para esta problemática, pois visa o registro e a comunicação de todos os processos relacionados ao ciclo de vida dos dados.

O presente trabalho apresentou um levantamento dos estudos relacionados a este cenário e propôs um modelo de proveniência para aplicação em dados abertos publicados no formato de *Linked Data*.

A literatura aponta para várias formas para se publicar dados em *Linked Data*, a dificuldade em garantir a proveniência para estes aspectos, se dá em definir de forma concreta um modelo geral para a publicação. Estes modelos ainda devem amadurecer e assim ferramentas e metodologias de publicação entrar em um formato padrão.

O desenvolvimento de uma ontologia ajuda neste amadurecimento, pois busca a definição dos conceitos e procedimentos envolvidos no processo, harmonizando o entendimento sobre o seu uso.

No trabalho foram desenvolvidos dois conjuntos de ontologias para modelar o processo de *Linked Data* e aplicar práticas de proveniência sobre processos. Com base nestes modelos, partiu-se para a criação de instâncias da ontologia, e assim consultas de inferência foram realizadas demonstrando sua aplicabilidade.

### 4.3. Trabalhos Futuros

A aplicação da ontologia para proveniência de dados publicados em Linked Data pode ser aplicado em qualquer área de domínio. A título de trabalho futuro caberia à aplicação do modelo apresentado em um ambiente dedicado ou real de Linked Data, para obtenção de resultados mais abrangentes e aprofundados. O refinamento e a evolução das ontologias geradas são importantes, entretanto não se encontravam no escopo deste trabalho.

O consumo de dados em Linked Data é facilmente realizado através de aplicações *mashups*. Também seria interessante a realização de experimentos de consumo aos dados de proveniência por um mashup, proporcionando interação direta para usuários, construção de indicadores, gráficos, dentre outros.

Por fim, poderiam ser realizados experimentos com maior número de indivíduos, para verificar possíveis relações entre o tempo de armazenamento e o ciclo de vida dos dados, bem como avaliar a qualidade dos dados frente ao uso de inferências na ontologia.



## REFERÊNCIAS

- ARMS, William. 2000. **Digital libraries**. [Online] 2000.  
<http://www.cs.cornell.edu/wya/DigLib/MS1999/index.html>.
- BEDUSCHI, Nitay Batista e CABRAL, Samuel Pierri. 2012. **Disponibilização de Dados do Enem no formato Linked Data**. 2012.
- BERNERS-LEE, T, HENDLER, J. e LASSILA, O. 2001. **The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities**. *Scientific American*. 2001, Vol. 5.
- BERNERS-LEE, Tim. 2006. **Linked Data - Design Issues**. 2006.
- BIANCO, R. 2011. **Disponibilização de Dados Abertos Utilizando Linked Data: Uma avaliação teórico-prática**. 2011.
- BOSE, R. e Frew, J. 2005. **Lineage retrieval for scientific data processing: a survey**. *ACM Comput. Surv.* 37, 2005.
- BOYLE, James. 2010. **The Public Domain**. [Online] 2010.  
<http://yupnet.org/boyle/archives/169#4>.
- BUNGE, M. 1977. **Treatise on Basic Philosophy. The Furniture of the World**. 1977, Vol. 3.
- CHANDRASEKARAN, B, JOSEPHSON, John R e BENJAMINS, V Richard. 1999. **What Are Ontologies, and Why Do We Need Them?** *IEEE Intelligent Systems*. 1999, Vol. 14.
- CREATIVE COMMONS. 2009. *History of Creative Commons*. [Online] 2009.
- . 2010. **Board of Directors**. [Online] 2010.  
<http://creativecommons.org/board>.
- CURTIS, B., KELLNER, M. e OVER, J. 1992. **Process modeling**. *Communication of ACM*. 1992, Vol. 35.
- FENSEL, D. e Hermelen, F. van. 2008. **On-To-Knowledge: Content-Driven Knowledge Management Tools through Evolving Ontologies**. 2008.
- FREIRE, J., Koop, D., Santos, E., Silva, C.T.. 2008. **Provenance for Computational Tasks: A Survey**. *Computing in Science and Engineering*. 2008, Vol. 10.
- FREW, J. e BOSE, R. 2001. **Earth System Science Workbench: A Data Management Infrastructure for Earth Science Products**. 2001.

- GASEVIC, D., et al. 2006. *Model Driven Architecture and Ontology*. Heidelberg: Springer. 2006.
- GILCHRIST, Alan. 2003. **Thesauri, taxonomies and ontologies – an etymological note**. *Journal of Documentation*. 2003, Vol. 59.
- GÓMEZ-PÉREZ, A., Corcho, O. e Fernández-Lopez, M. 2004. **Ontological Engineering: with examples from the areas of knowledge management, e-commerce and the semantic web**. Heidelberg: Springer. 2004.
- GÓMEZ-PÉREZ, Asunción e Rojas-Amaya, Ma Dolores. 1999. **Ontological Reengineering for Reuse**. *Lecture Notes in Computer Science*. 1999, Vol. 1621, pp. 139-156.
- GRUBER, T R. 1995. **Toward principles for the design of ontologies used for knowledge sharing**. *International Journal of Human and Computer Studies*. 1995, Vol. 43.
- GUARINO, N. 1998. **Formal Ontology and Information System**. 1998.
- HYLAND, B, Villazón-Terrazas, B e Hausenblas, M. 2012. Best practices for publishing linked data. [Online] 2012. <https://dvc.w3.org/hg/gld/raw-file/default/bp/index.html>.
- KONOLIGE, K. e Pollack, M. E. 1993. **A Representationalist Theory of Intention**. 1993.
- KONRATH, Mathias, et al. 2012. **SchemEX — Efficient construction of a data catalogue by stream-based indexing of linked data**. *Web Semantics: Science, Services and Agents on the World Wide Web*. November 2012, Vol. 16, pp. 52-58. <http://www.sciencedirect.com/science/article/pii/S1570826812000716>.
- KOUBARAKIS, M. e PLEXOUSAKIS, D. 2002. **A formal framework for business process modeling and design**. *Information Systems*. 2002, Vol. 27.
- MARINS, A. 2008. **Modelos Conceituais para Proveniência**. 2008.
- NOY, Natalya F e McGuinness, Deborah L. 2001. **Ontology Development 101: A Guide to Creating Your First Ontology**. 2001.
- RAMALHO, Rogério Aparecido Sá, VIDOTTI, Silvana Aparecida Borsetti Gregorio e FUJITA, Mariângela Spotti Lopes. 2007. **Web**

- semântica: uma investigação sob o olhar da Ciência da Informação.** *DataGramaZero*. 6, 2007, Vol. 8.
- SCHREIBER, G. 2002. **Knowledge engineering and management: the. *Mit Press***. 2002.
- SOWA, J. 1999. **Conceptual graphs: draft proposed American National Standard.** *Lecture notes in artificial intelligence*. 1999, Vol. 1640.
- SURE, Y. e Studer, R. 2003. **A Methodology for ontology-based Knowledge Management.** *Towards the semantic web: ontology-driven knowledge management*. 2003.
- WILLS, Gordon. 1996. **Embracing electronic publishing.** *Electronic Networking Application and Policy*. 4, 1996, Vol. 6.
- Y. Gil, D. Artz. 2007. **Towards Content Trust of Web Resources.** *Journal of Web Semantics*. MeZ de 2007, Vol. 5, 4, pp. 227-239.

## GLOSSÁRIO

**Metadados:** Metadados são dados sobre os dados, ou seja, são informações que possibilitam organizar, classificar, relacionar e inferir novos dados sobre o conjunto de dados. A quantidade e a qualidade dos metadados de um conjunto de dados podem determinar a utilidade daquele conjunto de dados. Em outras palavras, mais e melhores metadados agregam mais valor ao conjunto de dados, além de melhorar sua classificação e a busca sobre ele.

**Framework:** arcabouço, em tradução livre, em desenvolvimento de software, e uma abstração que une códigos comuns entre vários projetos de software provendo uma funcionalidade genérica.

## APÊNDICES

### APÊNDICE A: PERGUNTAS DE COMPETÊNCIA

**1) Como acessar Linked Data? R. Pode-se acessar diretamente um End-Point de consulta ou através de uma aplicação Mashup.**

**Termos Sugeridos:** End-Point; LinkedData; Mashup

**Relações Sugeridas:** acessar

**2) Como alcançar proveniência de dados? R. Garantindo a rastreabilidade dos fatos/eventos/acontecimentos do ciclo de vida dos dados envolvidos.**

**Termos Sugeridos:** CicloDeVida; Dado; Evento

**Relações Sugeridas:**

**3) Como estruturar um RDF? R. O RDF deve ser baseado em triplas, seguindo a gramática: subject, predicate, object; e em cada item deve ser aplicado um Identificador Uniforme de Recursos (URI)**

**Termos Sugeridos:** Object; Predicate; RDF; Subject; Tripla; URI

**Relações Sugeridas:** estruturar; identificar

**4) Como executar uma consulta? R. Através da linguagem SPARQL (Query Language for RDF)**

**Termos Sugeridos:** SPARQL

**Relações Sugeridas:** consultar

**5) Como Mashups são utilizados? R. Mashups são utilizados por consumidores interessados nos serviços específicos oferecidos pelo mesmo.**

**Termos Sugeridos:** Consumidor; Mashup

**Relações Sugeridas:** consomeDe

**6) Como persistir dados em RDF? R. Um RDF pode ser serializado em diferentes formatos, é preferível a utilização de formatos padrão e abertos, como RDF/XML, RDFa, N3, Turtle, JSON**

**Termos Sugeridos:** Formato; JSON; N3; RDF; RDF/XML; RDFa; Turtle

**Relações Sugeridas:** serializar

**7) Como publicar Linked Data? R. Disponibilizando na WEB seus dados em modelo/fonte de dados RDF.**

**Termos Sugeridos:** FonteDeDados; LinkedData; RDF; WEB

**Relações Sugeridas:** publicar

**8) Como referenciar dados em outras fontes de dados? R.**

**Termos Sugeridos:**

**Relações Sugeridas:**

**9) O que são Mashups? R. Um mashup é um site personalizado ou uma aplicação web que consulta e usa conteúdo de mais de uma fonte para criar um novo serviço completo.**

**Termos Sugeridos:** FonteDeDados; Mashup; Serviço

**Relações Sugeridas:** consultar; criar

**10) Quais eventos do ciclo de Vida dos Dados? R. Criação, Publicação, Acesso, Morte/Invalidez**

**Termos Sugeridos:** Acesso; Criacao; Dado

**Relações Sugeridas:**

**11) Que informações são necessárias para a rastreabilidade dos eventos de ciclo de vida de um dado? R. Para cada processo/evento no ciclo de vida de um dado, identificar: Dado:"o que","onde"; Agente:"Quem"; Processo: "quando", "Como", "por que" Artefato:**

**Termos Sugeridos:** Agente; Artefato; Dado; LicençaDeUso; Processo

**Relações Sugeridas:** derivadoDe; geradoPor; iniciadoPor; licenciadoPor; realizadoEm; utilizadoPor

**12) Quem pode acessar dados em Linked Data?**

**Termos Sugeridos:** Aplicacao; Organizacao; Pessoa

**Relações Sugeridas:** acessadoPor

## **Apendice B: Código Fonte das Ontologias**

O Código Fonte das Ontologias se encontra anexado em CD.