PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Universidade Federal do Rio de Janeiro

DISSERTAÇÃO DE MESTRADO

ROGERS REICHE DE MENDONÇA

UMA ABORDAGEM PARA COLETA E PUBLICAÇÃO DE DADOS DE PROVENIÊNCIA NO CONTEXTO DE LINKED DATA

RIO DE JANEIRO
2013

Instituto de Matemática

Instituto Tércio Pacitti de Aplicações e Pesquisas Computacionais

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE MATEMÁTICA
INSTITUTO TÉRCIO PACITTI
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

ROGERS REICHE DE MENDONÇA

UMA ABORDAGEM PARA COLETA E PUBLICAÇÃO DE DADOS DE PROVENIÊNCIA NO CONTEXTO DE LINKED DATA

RIO DE JANEIRO

2013

ROGERS REICHE DE MENDONÇA

UMA ABORDAGEM PARA COLETA E PUBLICAÇÃO DE DADOS DE PROVENIÊNCIA NO CONTEXTO DE LINKED DATA

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Matemática, Instituto Tércio Pacitti, Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Orientadora: Prof ª. Dr ª. Maria Luiza Machado Campos

Rio de Janeiro

2013

M539 Mendonça, Rogers Reiche de.

Uma abordagem para coleta e publicação de dados de proveniência no contexto de Linked Data. / Rogers Reiche de Mendonça. -- 2013

143 f.: il.

Dissertação (Mestrado em Informática) – Universidade Federal do Rio de Janeiro, Instituto de Matemática, Instituto Tércio Pacitti, Programa de Pós-Graduação em Informática, Rio de Janeiro, 2013.

Orientadora: Maria Luiza Machado Campos

1. Proveniência. 2. Linked Data. 3. ETL. – Teses. I. Campos, Maria Luiza Machado (Orient.). II. Universidade Federal do Rio de Janeiro, Instituto de Matemática, Instituto Tércio Pacitti, Programa de Pós-Graduação em Informática. III. Título.

CDD:

ROGERS REICHE DE MENDONÇA

Uma abordagem para coleta e publicação de dados de proveniência no contexto de Linked Data

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Aprovada em: Rio de Janeiro, 17 de dezembro de 2013.

Prof[a]. Maria Luiza Machado Campos, Ph.D. – PPGI/UFRJ

Prof[a]. Adriana Santarosa Vivacqua, D.Sc. – PPGI/UFRJ

Prof[a]. Marta Lima de Queirós Mattoso, D.Sc. – COPPE/UFRJ

Prof. Sérgio Manuel Serra da Cruz, D.Sc. – ICE/UFRRJ

Agradecimentos

A Deus, pelo amor incondicional, e ao meu mestre Jesus Cristo, por estar sempre ao meu lado, renovando as minhas forças e me ensinando a perseverar na caminhada.

À minha amada esposa Monique, que acompanhou e testemunhou toda esta caminhada de perto. Meu agradecimento emocionado pelo carinho, pelo precioso apoio, pelas palavras de incentivo que mudaram as situações mais difíceis.

Aos meus queridos pais, Nério e Wilma, que me deram a vida e me ensinaram princípios e valores, e ao Rodrigo, meu estimado irmão.

À minha orientadora, professora Maria Luiza Machado Campos, pelos brilhantes conselhos, empenho, cuidado e apoio essenciais ao longo do desenvolvimento deste trabalho.

À Petrobras, por ter investido no meu interesse em cursar o mestrado e por todo o apoio. Aos meus líderes Martha Gomes, Marcos Sobral e Andrea Vallim, meu muito obrigado.

Aos meus amigos e padrinhos de casamento Jorge e Irene Fagim, por todo o incentivo e ajuda para enfrentar e vencer importantes desafios.

Aos demais familiares, amigos e colegas de trabalho que acompanharam todo o período de estudos do mestrado, pela torcida e por compreenderem minha ausência em diversas ocasiões.

Aos professores Adriana Vivacqua, Marta Mattoso e Sérgio Serra, pela gentileza de compor a banca examinadora e pelas ricas contribuições.

Aos professores Jonice Oliveira e Marcos Borges, pelas contribuições valiosas nos seminários de qualificação.

Aos companheiros Cristiano Expedito, Fabrício Firmino e Kelli Cordeiro, pelas contribuições durante o desenvolvimento deste trabalho. Aos demais colegas do Programa de Pós-Graduação em Informática (PPGI), pelo apoio e pelos momentos de descontração.

Ao colega Jonas De La Cerda e à professora Yoko, pela troca de experiências e pelo trabalho em conjunto para integração de nossas abordagens de proveniência.

Ao Aníbal e à Adriana, funcionários da Secretaria do PPGI, pela disponibilidade em ajudar em todos os momentos.

Meus agradecimentos especiais à Universidade Federal do Rio de Janeiro (UFRJ) por todo o suporte oferecido para apoiar o desenvolvimento desta dissertação.

Resumo

MENDONÇA, Rogers Reiche de. **Uma abordagem para coleta e publicação de dados de proveniência no contexto de Linked Data.** 2013. 143 f. Dissertação (Mestrado em Informática) - Programa de Pós-Graduação em Informática, Instituto de Matemática, Instituto Tércio Pacitti, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2013.

Nos últimos anos, um número crescente de provedores de dados vem adotando um conjunto de tecnologias e boas práticas da Web Semântica para publicar e interligar dados estruturados, seguindo os princípios de Linked Data. No entanto, sem subsídios para avaliação da qualidade e da confiabilidade dos dados publicados como Linked Data, o consumo e a exploração dos mesmos podem ser comprometidos. Com relação a este fato, o presente trabalho propõe a abordagem *ETL4LinkedProv*, uma abordagem de coleta e publicação de dados de proveniência para o processo de publicação de Linked Data realizado dentro dos limites de uma organização. A abordagem utiliza um agente de proveniência para atuar em um processo de publicação de Linked Data executado através de um *workflow* de ETL (Extração, Tranformação e Carga). Este agente, denominado Agente Coletor de Proveniência, coleta, interliga e armazena temporariamente os dados de proveniência prospectiva e retrospectiva, durante a execução do processo de publicação de Linked Data. Posteriormente, a proveniência coletada é também publicada no contexto de Linked Data, a fim de que os dados de domínio e seus respectivos dados de proveniência possam ser explorados conjuntamente, por meio de consultas SPARQL. Desta maneira, a abordagem oferece apoio para a avaliação da qualidade e da confiabilidade dos dados publicados no contexto de Linked Data. Um exemplo de aplicação envolvendo a publicação e integração de dados reais do CNPq e da RNP, organizações de incentivo à pesquisa no Brasil, foi realizado para evidenciar as contribuições da abordagem.

Palavras-chaves: Proveniência, Linked Data, ETL.

Abstract

MENDONÇA, Rogers Reiche de. **Uma abordagem para coleta e publicação de dados de proveniência no contexto de Linked Data.** 2013. 143 f. Dissertação (Mestrado em Informática) - Programa de Pós-Graduação em Informática, Instituto de Matemática, Instituto Tércio Pacitti, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2013.

In recent years, an increasing number of data providers has been adopting a set of Semantic Web technologies and best practices to publish and link structured data, following the Linked Data principles. However, without information to assess the quality and reliability of the data published as Linked Data, the consumption and exploration thereof may be compromised. Regarding this fact, this work proposes the *ETL4LinkedProv* approach, an approach for collecting and publishing provenance data to the Linked Data publishing process performed within the boundaries of an organization. The approach uses a provenance agent to act in the Linked Data publishing process executed through an ETL (Extract, Transform and Load) workflow. This agent, named Provenance Collector Agent, collects, links and temporarily stores prospective and retrospective provenance data during the execution of the Linked Data publishing process. Afterwards, the collected provenance is also published in the Linked Data context, so that the domain data and its respective provenance data can be explored jointly by means of SPARQL queries. Thus, the approach provides support for evaluating the quality and reliability of the data published in the context of Linked Data. An application example involving the publication and integration of real data from CNPq and RNP, research support organizations in Brazil, was held to highlight the contributions of the approach.

Keywords: Provenance, Linked Data, ETL.

Lista de Figuras

Lista de Quadros

Lista de Siglas

API	*Application Programming Interface*
BFO	*Basic Formal Ontology*
BI	*Business Intelligence*
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
DAG	*Directed Acyclic Graph*
DCMI	*Dublin Core Metadata Initiative*
DOLCE	*Descriptive Ontology for Linguistic and Cognitive Engineering*
DW	*Data Warehousing*
ETL	*Extract, Transform and Load*
FAP	Fundação de Amparo à Pesquisa
FINEP	Financiadora de Estudos e Projetos
GED	Gerenciamento Eletrônico de Documentos
GGG	*Giant Global Graph*
GPG	*GNU Privacy Guard*
HTML	*HyperText Markup Language*
IDT	*Interactive Data Transformation*
IRI	*Internationalized Resource Identifier*
JSON	*JavaScript Object Notation*
LDIF	*Linked Data Integration Framework*
LOD	*Linked Open Data*
LOV	*Linked Open Vocabularies*
MCTI	Ministério da Ciência, Tecnologia e Inovação
N3	*Notation 3*
OPM	*Open Provenance Model*
OPMV	*Open Provenance Model Vocabulary*
OPMO	*Open Provenance Model Ontology*
OPMW	*Open Provenance Model for Workflows*
OvO	*Open proVenance Ontology*
OWL	*Web Ontology Language*
PAV	*Provenance, Authoring and Versioning*

PGP	*Pretty Good Privacy*
PML	*Proof Markup Language*
PREMIS	*Preservation Metadata Implementation Strategies*
PROV-DM	*PROV Data Model*
PROV-O	*PROV Ontology*
QBE	*Query-By-Example*
RDF	*Resource Description Framework*
RDF-S	*RDF Schema*
REST	*Representational State Transfer*
RIF	*Rule Interchange Format*
RNP	Rede Nacional de Ensino e Pequisa
SGBD	Sistema de Gerenciamento de Banco de Dados
SGWfC	Sistema de Gerenciamento de *Workflow* Científico
SPARQL	*SPARQL Protocol and RDF Query Language*
SQL	*Structured Query Language*
SUMO	*Suggested Upper Merged Ontology*
SWP	*Semantic Web Publishing Vocabulary*
Turtle	*Terse RDF Triple Language*
URI	*Uniform Resource Identifier*
URL	*Uniform Resource Locator*
VoID	*Vocabulary of Interlinked Datasets*
WOT	*Web of Trust Schema*
WWW	*World Wide Web*
WYSIWYM	*What You See Is What You Mean*
W3C	*World Wide Web Consortium*
XML	*Extensible Markup Language*

Sumário

1 Introdução

1.1 Motivação

Proveniência é um conceito aplicado em diversos campos de estudo, entre eles, arquivologia (BEARMAN, LYTLE, 1985), museologia e arqueologia (MILLAR, 2002), agricultura e enologia (FABANI et al., 2009) e computação (MOREAU, 2010). As definições do termo proveniência (*provenance*) encontradas nos dicionários Oxford English Dictionary[1] e Merriam-Webster Online Dictionary[2] indicam-no como sendo todo tipo de informação que descreve a origem de um objeto e o processo de derivação do mesmo até determinado estado (MOREAU, 2010). Como exemplo no contexto das belas artes, são informações de proveniência de uma pintura como o quadro Diana e Acteon (Figura 1), o nome do pintor, o local e a data em que a pintura foi feita, o tipo de tinta e papel utilizados e os proprietários da obra de arte desde sua conclusão até o presente momento.

Figura 1. O quadro Diana e Acteon, do pintor italiano Ticiano, possui o registro integral de sua proveniência, incluindo a lista de proprietários desde que foi pintado para o rei Filipe II da Espanha, na década de 1550.

Na área da computação, um estudo realizado por GIL e ARTZ (2007) mostra que a proveniência é um dos principais fatores que influenciam a confiança dos usuários em um conteúdo da Web. Sendo assim, a análise dos dados de proveniência é fundamental para a

[1] http://www.oed.com

[2] http://www.merriam-webster.com/dictionary/provenance

avaliação da qualidade e da confiabilidade dos diferentes tipos de conteúdos disponibilizados na Web, como documentos não estruturados, imagens, vídeos e dados estruturados.

Nos últimos anos, um número crescente de provedores de dados vem adotando um conjunto de tecnologias e boas práticas da Web Semântica para publicar e interligar dados estruturados na Web, formando assim a Web de Dados. Os princípios determinados pela abordagem Linked Data[3] e adotados pela Web de Dados oferecem simplicidade e flexibilidade para que dados sejam representados, interligados e possam interoperar na Web. Estes dados são estruturados por meio de sentenças denominadas de triplas RDF (*Resource Description Framework*), que contêm identificadores no formato *Uniform Resource Indentifier* (URI) e descritores definidos por vocabulários (BIZER, HEATH, BERNERS-LEE, 2009).

A proliferação de uso e o acelerado crescimento da Web de Dados incentivaram o surgimento de uma nova geração de aplicações para consumir os dados publicados e explorar o potencial por eles oferecido (HEATH, BIZER, 2011). Por exemplo, encontramos iniciativas governamentais no Reino Unido (SHERIDAN, TENNISON, 2010), nos Estados Unidos (HENDLER, J. et al., 2012) e no Brasil (BREITMAN et al., 2012) que, fomentadas por acordos de transparência e colaboração, publicam seus dados abertos seguindo os princípios de Linked Data. Estas iniciativas, aliadas à propagação de clientes móveis e serviços de acesso público, trazem uma nova perspectiva de empoderamento e participação aos cidadãos (CORDEIRO, CAMPOS, M. L. M., BORGES, 2011).

No entanto, sem subsídios para a avaliação da qualidade e da confiabilidade dos dados publicados como Linked Data, o consumo e a exploração dos mesmos podem ser comprometidos. Assim sendo, a proveniência surge também no contexto de Linked Data com um papel de suporte importante para a avaliação da qualidade e da confiabilidade dos dados. A proveniência permite verificar se um dado publicado como Linked Data é oriundo de uma fonte oficial ou como um dado foi derivado a partir de outros dados, além de oferecer, por exemplo, indicadores para determinar o motivo de uma possível inconsistência entre dados procedentes de diferentes fontes. Devido a características como estas, John Sheridan, responsável pelo serviço de legislação do Arquivo Nacional do Reino Unido, declarou que a proveniência era a questão número um considerada ao publicar os dados governamentais britânicos como Linked Data no site data.gov.uk (SHERIDAN, 2010 apud OMITOLA et al., 2011).

[3] Este trabalho irá utilizar o termo original Linked Data, por ser um termo bem estabelecido para identificar os princípios definidos por BERNERS-LEE (2006).

No contexto de Linked Data, um conjunto de questões como "de quais fontes vieram os dados publicados?", "quando e por quem o processo de publicação foi projetado?", "quando e por quem o processo de publicação foi executado?", "houve alguma falha durante a execução do processo de publicação?", "quais as operações de preparação, modificação e integração foram aplicadas nos dados originais antes de publicá-los como triplas RDF?" podem ser respondidas por meio dos dados de proveniência. A motivação deste trabalho é que, fundamentadas em informações desta natureza, pessoas podem avaliar e máquinas podem inferir sobre a qualidade e a confiabilidade dos dados disponibilizados no contexto de Linked Data.

1.2 Caracterização do problema

As abordagens que atuam no contexto da proveniência em Linked Data têm se focado principalmente nos metadados sobre a criação dos dados ou sobre o acesso aos dados na Web (MARJIT, SHARMA, BISWAS, 2012). Entretanto, quando os dados primários são manipulados dentro dos limites de uma organização como, por exemplo, agências governamentais, empresas de negócio ou comunidades científicas, a proveniência relacionada ao respectivo processo de publicação dos dados como Linked Data apresenta um novo potencial a ser explorado.

Este processo peculiar de publicação de Linked Data, realizado dentro dos limites de uma organização, consiste normalmente da extração de dados de múltiplas fontes heterogêneas, seguida da execução, através do uso de diversas ferramentas, de uma gama de operações de preparação, modificação e integração dos dados antes de carregá-los em um banco de triplas RDF (CORDEIRO et al., 2011). Desta maneira, coletar e disponibilizar dados de proveniência relacionados a este processo de publicação de Linked Data e a suas respectivas operações de extração, preparação, modificação, integração e carga apresentam um subsídio adicional importante para a posterior análise da qualidade e da confiabilidade dos dados publicados.

Um estudo realizado por HARTIG (2009) dentro do foco principal de atuação das abordagens de proveniência no contexto de Linked Data, isto é, considerando a proveniência sobre a criação dos dados e sobre o acesso aos dados na Web, indicou uma carência de metadados de proveniência na Web de Dados. Este mesmo estudo concluiu que uma das principais razões para isto era a falta de ferramental para apoiar a coleta e a publicação dos

metadados de proveniência. Na análise do processo particular de publicação de Linked Data relatado no parágrafo anterior, verifica-se que este problema se mantém.

Além disso, com relação à recuperação efetiva dos dados de proveniência publicados no contexto de Linked Data, surge ainda a necessidade de que estes estejam vinculados aos dados aos quais se referem, sendo com eles publicados (HARTIG, ZHAO, 2010), para que, assim, tanto a partir de um determinado dado na Web seja possível recuperar seus dados de proveniência, quanto a partir de um determinado dado de proveniência seja possível recuperar os dados por ele referenciados. Neste sentido, também configura-se como um problema, a dificuldade de interligar os dados de proveniência com os demais dados publicados pelo processo de publicação de Linked Data.

Ainda sobre esta dificuldade de interligação entre os dados e sua proveniência no contexto de Linked Data, cabe fazer um adendo. Uma das facetas de categorização dos dados de proveniência classifica-os como dados de proveniência prospectiva e dados de proveniência retrospectiva (FREIRE et al., 2008; CRUZ, CAMPOS, M. L. M., MATTOSO, 2009; LIM et al., 2010). Os dados de proveniência prospectiva indicam a proveniência relacionada à definição e composição do processo, enquanto que os dados de proveniência retrospectiva indicam a proveniência relacionada à execução do processo. De maneira análoga à dificuldade de interligar os dados de proveniência com os demais dados publicados como Linked Data, também configura-se como um problema, a dificuldade de interligar os dados de proveniência retrospectiva com seus respectivos dados de proveniência prospectiva.

Feita esta exposição introdutória, segue abaixo a caracterização do problema analisado por este trabalho de pesquisa:

"A carência de ferramental para coletar e publicar os dados de proveniência relacionados com o processo de publicação de Linked Data realizado dentro dos limites de uma organização e a dificuldade de interligar, permitindo uma exploração conjunta, (i) a proveniência da composição (prospectiva) com a proveniência da execução (retrospectiva) do processo e (ii) os dados publicados com seus respectivos dados de proveniência."

1.3 Objetivo

O objetivo deste trabalho de pesquisa é propor uma abordagem de coleta e publicação de proveniência para o processo particular de publicação de Linked Data realizado dentro dos limites de uma organização. A abordagem, denominada *ETL4LinkedProv*, utiliza um agente de proveniência para atuar no processo de publicação de Linked Data executado através de um *workflow* de Extração, Transformação e Carga (ETL - *Extract, Transform and Load*). Este agente coleta, interliga e armazena temporariamente os dados de proveniência prospectiva e retrospectiva, durante a execução do processo de publicação de Linked Data. Posteriormente, os dados de proveniência coletados são também triplificados e publicados no contexto de Linked Data. A hipótese analisada é se a utilização da abordagem *ETL4LinkedProv* supre a carência de ferramental para coleta e publicação dos dados de proveniência no contexto de Linked Data e resolve a dificuldade de interligação entre os dados publicados e a respectiva proveniência, a fim de que possam ser explorados conjuntamente através de consultas *SPARQL Protocol and RDF Query Language* (SPARQL).

A abordagem *ETL4LinkedProv* foi concebida de maneira independente do uso de uma ferramenta de ETL específica. Neste trabalho, a ferramenta Pentaho Data Integration, também conhecida como Kettle, ferramenta de *workflow* ETL da plataforma de Inteligência de Negócio (BI - *Business Intelligence*) Pentaho, foi utilizada para implementar um protótipo da abordagem. O motivo da escolha do Kettle deveu-se às suas características de ser *open source*, possuir uma ampla comunidade de usuários e ser extensível através de uma *Application Programming Interface* (API) Java. No entanto, outras ferramentas de *workflow* ETL podem ser utilizadas para implementar a abordagem proposta por este trabalho.

Por fim, na etapa de validação, a abordagem *ETL4LinkedProv* foi aplicada em um cenário envolvendo dados reais relacionados às organizações Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Rede Nacional de Ensino e Pequisa (RNP). Além dos dados destas organizações, os respectivos dados de proveniência foram coletados e publicados como Linked Data, com posterior execução de consultas SPARQL e discussão dos resultados obtidos. Os resultados evidenciaram as contribuições da abordagem para possibilitar diferentes tipos de exploração conjunta entre os dados de domínio e os dados de proveniência, oferecendo assim subsídios para a avaliação da qualidade e da confiabilidade dos dados de domínio publicados. Adicionalmente, o impacto causado pela coleta e publicação dos dados de proveniência com nível de granulosidade fina foi apresentado a partir de uma

análise quantitativa dos tempos de execução dos *workflows* ETL utilizados no cenário de aplicação e dos números de triplas RDF publicadas.

1.4 Organização do trabalho

Este trabalho é dividido em outros cinco capítulos. O capítulo 2 apresenta o referencial teórico e alguns trabalhos relacionados sobre Web Semântica e Linked Data, com exemplos aplicados em dados da empresa Petrobras publicados como Linked Data no *dataset* da DBPedia. O capítulo 3 apresenta o referencial teórico e alguns trabalhos relacionados sobre proveniência e propõe uma taxonomia para classificação da proveniência no contexto de Linked Data. O capítulo 4 apresenta a abordagem *ETL4LinkedProv* proposta por este trabalho para coleta e publicação de dados de proveniência no contexto de Linked Data, assim como a implementação de um protótipo da abordagem utilizando a ferramenta de *workflow* ETL Pentaho Data Integration (Kettle). O capítulo 5 apresenta um exemplo de aplicação da abordagem *ETL4LinkedProv* em um cenário envolvendo dados reais do CNPq e da RNP, publicados pelo protótipo implementado, e evidencia as contribuições da abordagem por meio da análise dos resultados obtidos por consultas SPARQL de exploração conjunta entre os dados do CNPq, os dados da RNP e seus respectivos dados de proveniência. O capítulo 5 apresenta ainda uma análise quantitativa dos tempos de execução dos *workflows* ETL utilizados no exemplo de aplicação e dos números de triplas RDF publicadas. Por fim, o capítulo 6 destaca as principais contribuições deste trabalho, as limitações identificadas e os trabalhos futuros.

2 Web Semântica e Linked Data

2.1 A evolução dos dados na Web

No final da década de 80, a *World Wide Web* (WWW) proposta por BERNERS-LEE (1989) surgiu como uma Web de Documentos, cujas unidades primárias eram documentos escritos com a linguagem de marcação *HyperText Markup Language* (HTML), para disponibilizar informações na Internet por meio de hipertexto. Esta Web de Documentos é análoga a um sistema de arquivos (*filesystem*) global de documentos interligados e foi projetada para consumo humano, por meio de navegadores Web.

O processo de interpretação das informações da Web de Documentos é feito pelos usuários a partir da semântica implícita ao conteúdo dos documentos, ou seja, os documentos não apresentam uma semântica explícita de maneira que também possam ser acessados e processados por computadores em qualquer parte da Internet. Além disso, a relação entre dois documentos interligados também é implícita, pois o HTML não apresenta a expressividade necessária para permitir a conexão, através de ligações (*links*) tipadas, entre os dados presentes nos documentos (BIZER, HEATH, BERNERS-LEE, 2008).

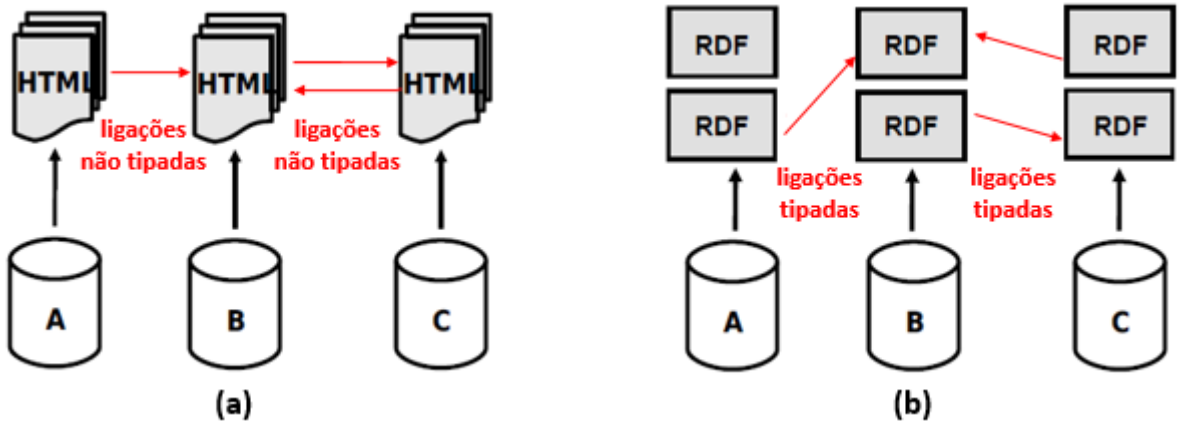

Figura 2. A evolução dos dados na Web: da (a) Web de Documentos para (b) Web de Dados. Adaptado de BIZER, HEATH e BERNERS-LEE (2008).

Em 2001, para suprir a necessidade de interligação entre os dados da Web de Documentos e também suprir a carência de semântica explícita, a fim de que os dados pudessem ser acessados e processados por agentes computacionais, BERNERS-LEE, HENDLER e LASSILA (2001) propuseram uma extensão para a Web existente até então. A visão proposta por eles, denominada Web Semântica, consistia na criação de uma Web de Dados, onde os

dados seriam disponibilizados e interligados na Web, de maneira que pudessem ser usados por computadores não somente com propósitos de apresentação, mas para automação, integração e reutilização entre as aplicações.

2.2 Pilha da Web Semântica (*Semantic Web Stack*)

Para atingir a visão da Web Semântica, BERNERS-LEE (2003) propôs a implementação de um conjunto de tecnologias e boa práticas, representadas através de uma pilha de camadas denominada Pilha da Web Semântica (*Semantic Web Stack*). A Pilha da Web Semântica foi aprimorada nos anos seguintes pelo *World Wide Web Consortium* (W3C)[4] e pode ser ilustrada conforme a Figura 3 abaixo (BERNERS-LEE et al., 2006).

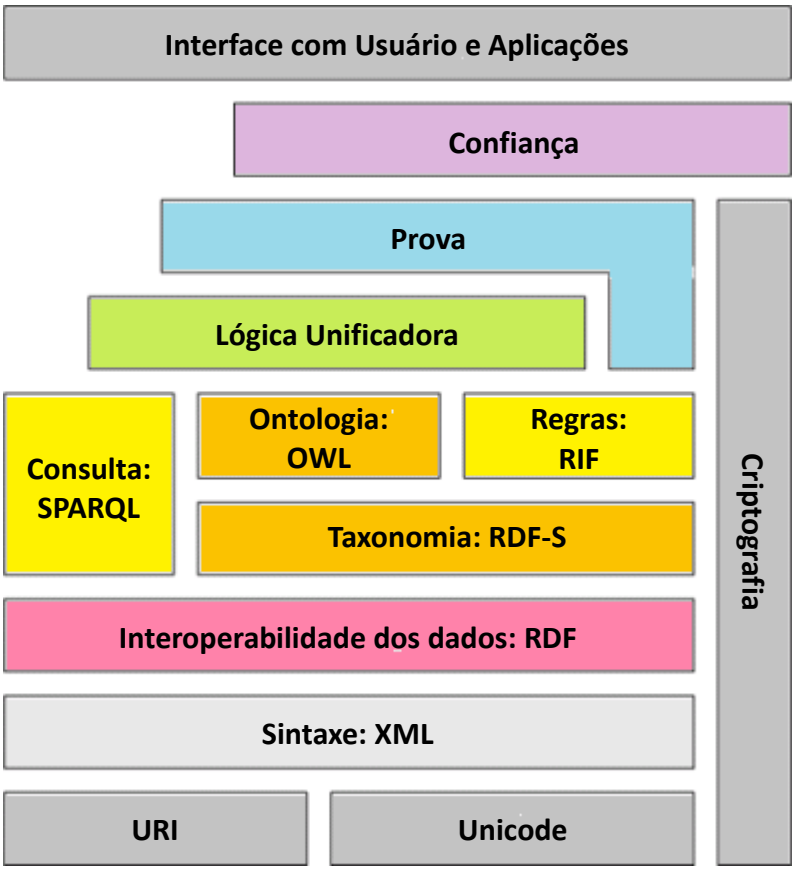

Figura 3. Pilha da Web Semântica (*Semantic Web Stack*). Adaptado de BERNERS-LEE et al. (2006).

As tecnologias e boas práticas de cada camada da Pilha da Web Semântica serão abordadas nas próximas subseções, seguindo um sentido de baixo para cima. As capacidades proporcionadas pelas camadas inferiores da Pilha da Web Semântica são utilizadas e exploradas pelas camadas superiores.

[4] http://www.w3.org

2.2.1 URI e Unicode

Na base da Pilha da Web Semântica, encontram-se a camada URI, responsável pela identificação dos dados, e a camada Unicode, responsável pela representação do texto em diferentes idiomas. *Internationalized Resource Identifier* (IRI) (DUERST, SUIGNARD, 2005), generalização de *Uniform Resource Identifier* (URI) (BERNERS-LEE, FIELDING, MASINTER, 2005) com emprego da codificação de caracteres Unicode[5], tornou-se um padrão do W3C para identificar unicamente os dados na Web, também denominados de recursos.

Uma URI é uma sequência de caracteres semelhante a um *Uniform Resource Locator* (URL), que é utilizada, por exemplo, para acessar um documento HTML através de um navegador Web. Entretanto, uma URI, assim como uma IRI, não necessariamente precisa especificar a localização do recurso identificado na rede e o mecanismo para recuperá-lo (MEALLING, DENENBERG, 2002).

http://dbpedia.org/resource/Petrobras é um exemplo de IRI que identifica a empresa Petrobras[6] no contexto da DBPedia[7].

2.2.2 XML

A próxima camada está relacionada à linguagem *Extensible Markup Language* (XML) (BRAY et al., 2008), linguagem de marcação que provê um conjunto de regras de sintaxe para criação de documentos compostos de dados semiestruturados (DACONTA, OBRST, SMITH, 2003). O uso de XML *Namespaces* (BRAY et al., 2009), um mecanismo simples para a criação de nomes globais exclusivos para elementos e atributos da linguagem de marcação XML a fim de evitar conflitos entre vocabulários, tornou-se também uma recomendação do W3C.

Como será visto na próxima subseção, além da linguagem XML, o W3C recomenda uma notação alternativa para conceber a sintaxe dos dados, denominada *Notation 3* (N3) (BERNERS-LEE, CONNOLLY, 2011). Além do formato N3, cabe ressaltar também que, apesar de ainda não ser uma recomendação do W3C, existem propostas em andamento para utilização de *JavaScript Object Notation* (JSON) (CROCKFORD, 2006) como alternativa à linguagem XML para a representação sintática dos dados (DAVIS, STEINER, 2011).

---

[5] http://www.unicode.org/standard/standard.html

[6] http://www.petrobras.com.br

[7] http://dbpedia.org/About

2.2.3 RDF

Acima da camada XML, encontra-se a camada RDF. *Resource Description Framework* (RDF) (KLYNE, CARROLL, 2004) é o modelo de dados recomendado pelo W3C para representar informações sobre os recursos na Web. Estas informações são descritas através de triplas, sentenças compostas por três partes: sujeito, predicado e objeto. O sujeito é uma IRI que identifica um recurso, o predicado é uma IRI que identifica uma propriedade do sujeito e o objeto pode ser uma IRI para um recurso ou um valor literal, como um número, um texto ou uma data. Os valores literais podem ser planos (*plain literal*), um texto combinado com uma marcação opcional de idioma, ou tipados (*typed literal*), um texto combinado com uma URI indicando o tipo do literal. Um recurso também pode ser tipado, usando uma tripla com o recurso a ser tipado no sujeito, a propriedade *http://www.w3.org/1999/02/22-rdf-syntax-ns#type* no predicado e a IRI do recurso que identifica o tipo no objeto.

O grafo ilustrado na Figura 4 representa a tripla RDF que descreve a informação "o nome da empresa Petrobras é Petróleo Brasileiro S.A.". O recurso Petrobras representa o sujeito, é identificado pela IRI *http://dbpedia.org/resource/Petrobras* e é ilustrado pela elipse à esquerda; a propriedade nome representa o predicado, é identificada pela IRI *http://dbpedia.org/property/name* e é ilustrada pela aresta direcionada do sujeito para o objeto e o valor literal "Petróleo Brasileiro S.A." representa o objeto e é ilustrado pelo retângulo à direita.

Figura 4. Tripla RDF que descreve a informação "o nome da empresa Petrobras é Petróleo Brasileiro S.A.".

Outras triplas RDF podem ser incluídas neste grafo, representando outras propriedades da empresa Petrobras e também interligando-a a outros recursos. O grafo da Figura 5 ilustra quatro triplas RDF, que descrevem as propriedades nome, ano de fundação e produto do recurso Petrobras; sendo o objeto da propriedade produto, o recurso petróleo, identificado pela IRI *http://dbpedia.org/resource/Petroleum*. Este último recurso possui a propriedade resumo, identificada pela IRI *http://dbpedia.org/ontology/abstract*.

Figura 5. Grafo RDF que representa outras propriedades da empresa Petrobras e interliga-a ao produto petróleo.

Para representar as triplas RDF em um formato que possa ser armazenado em um banco de triplas (*triplestore*), existem duas recomendações do W3C em uso: o formato XML (BECKETT, 2004) e o formato N3 (BERNERS-LEE, CONNOLLY, 2011). *Terse RDF Triple Language* (Turtle) (BECKETT, BERNERS-LEE, 2011) e NTriple (GRANT, BECKETT, 2004) são subconjuntos do formato N3, também utilizados para serializar triplas RDF. As serializações do grafo ilustrado na Figura 5, nos formatos Turtle e RDF/XML respectivamente, são exibidas abaixo:

```
1  @prefix dbpedia: <http://dbpedia.org/resource/> .
2  @prefix dbpprop: <http://dbpedia.org/property/> .
3  @prefix dbpedia-owl: <http://dbpedia.org/ontology/> .
4
5  dbpedia:Petrobras
6        dbpprop:name "Petróleo Brasileiro S.A." ;
7        dbpedia-owl:foundingYear 1953 ;
8        dbpedia-owl:product dbpedia:Petroleum .
9
10 dbpedia:Petroleum
11       dbpedia-owl:abstract "Petróleo (...) é uma substância oleosa,
   inflamável, (...) Além de gerar a gasolina (...), vários produtos
   são derivados do petróleo como, a parafina, GLP, (...)" .
```

```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <rdf:RDF
3        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4        xmlns:dbpprop="http://dbpedia.org/property/"
5        xmlns:dbpedia-owl="http://dbpedia.org/ontology/">
6
7  <rdf:Description rdf:about="http://dbpedia.org/resource/Petrobras">
8        <dbpprop:name>Petróleo Brasileiro S.A.</dbpprop:name>
9        <dbpedia-owl:foundingYear>1953</dbpedia-owl:foundingYear>
10       <dbpedia-owl:product
   rdf:resource="http://dbpedia.org/resource/Petroleum" />
11 </rdf:Description>
12
13 <rdf:Description rdf:about="http://dbpedia.org/resource/Petroleum">
14       <dbpedia-owl:abstract>Petróleo (...) é uma substância oleosa,
   inflamável, (...) Além de gerar a gasolina (...), vários produtos
   são derivados do petróleo como, a parafina, GLP, (...)</dbpedia-
   owl:abstract>
15 </rdf:Description>
16
17 </rdf:RDF>
```

Na descrição do grafo RDF com o formato RDF/XML, cada marcação *rdf:Description* representa um recurso, com as respectivas propriedades definidas através de submarcações. XML *Namespaces* são utilizados para definir prefixos para as marcações que definem as propriedades dos recursos. Estes prefixos determinam a IRI do vocabulário utilizado para representar a propriedade. De maneira análoga, os prefixos também são representados no formato N3, através do uso da diretiva *@prefix*. Neste tipo de formato, as triplas RDF são representadas através de uma sentença simples com o formato "<*sujeito*> <*predicado*> <*objeto*> .". O ponto e vírgula pode ser empregado após o objeto para descrever outra propriedade de um mesmo sujeito e a vírgula pode ser empregada após o objeto para descrever outro objeto de um mesmo predicado.

2.2.4 RDF-S, OWL, SPARQL e RIF

Acima da camada RDF na Pilha da Web Semântica, encontram-se as camadas RDF-S, OWL, SPARQL e RIF. Estas quatro camadas também apresentam tecnologias padronizadas pelo W3C.

RDF Schema (RDF-S) (BRICKLEY, DAN, GUHA, 2004) é o padrão do W3C que provê um vocabulário básico para definição de uma taxonomia para as informações descritas em RDF. Como visto na subseção anterior, a função do RDF é descrever informações sobre

recursos na Web através de triplas, entretanto, o RDF não fornece expressividade para definir a estrutura seguida na descrição das informações. Com esta finalidade, o RDF-S estende o RDF para permitir a definição de classes (*rdfs:Class*) e suas propriedades, assim como uma hierarquia entre as classes (*rdfs:subClassOf*) e entre as propriedades (*rdfs:subPropertyOf*), o domínio de classes aceitas pelo sujeito de uma propriedade (*rdfs:domain*) e o domínio de classes ou tipos de dados aceitos pelo objeto de uma propriedade (*rdfs:range*). A Figura 6 ilustra o emprego das propriedades *rdfs:subClassOf* e *rdfs:domain* a partir da tripla RDF que relaciona Petrobras e petróleo.

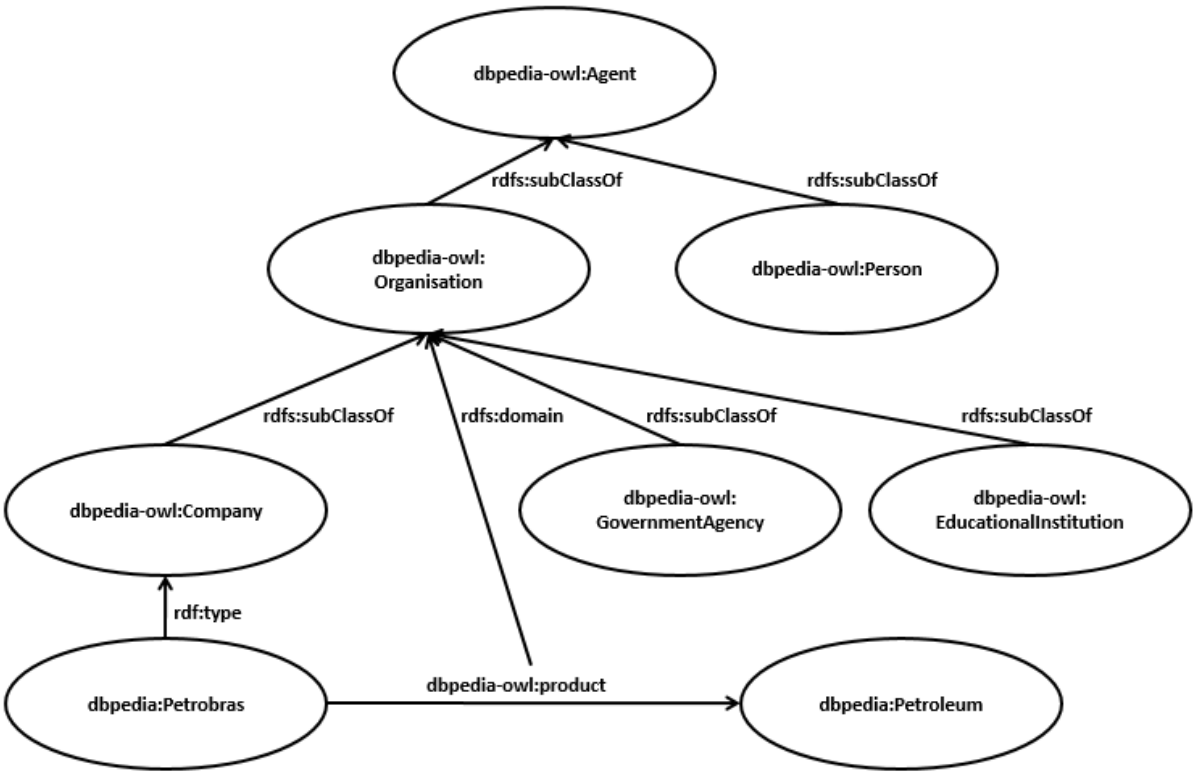

Figura 6. Grafo RDF que ilustra o emprego das propriedades *rdfs:subClassOf* e *rdfs:domain* a partir da tripla RDF que relaciona Petrobras e petróleo.

A linguagem *Web Ontology Language* (OWL) (MCGUINNESS, VAN HARMELEN, 2004) estende o RDF-S, adicionando um vocabulário para permitir a descrição das classes e propriedades com maior enriquecimento semântico. Entre as principais funcionalidades oferecidas, a linguagem OWL permite descrever (i) relações de equivalência e disjunção entre as classes e suas instâncias; (ii) operações de interseção, união e complemento, análogas às da teoria de conjuntos, sobre as classes; (iii) a cardinalidade das propriedades de uma classe e (iv) características de simetria, reflexividade e transitividade entre as propriedades. Assim sendo, a OWL, que oferece três sublinguagens (OWL *Lite*, OWL DL e OWL *Full*) com expressividade crescente projetadas para uso de comunidades específicas de usuários e tipos de aplicação,

tornou-se o padrão do W3C para descrição de ontologias. Ontologia é um termo inicialmente ligado à esfera filosófica como sendo a ciência que estuda em quais tipos e estruturas se classificam os objetos, as propriedades, os eventos, os processos e as relações existentes no mundo (SMITH, WELTY, 2001), mas posteriormente o termo foi herdado pela área de Computação, especialmente pelas comunidades ligadas à Inteligência Artificial e Sistemas de Informação, exercendo o papel de possibilitar "uma descrição parcial e explícita de uma conceituação" (GUARINO, 1998).

Para recuperar e manipular dados descritos em RDF, incluindo sentenças que envolvem RDF-S e OWL, o W3C padronizou a linguagem *SPARQL Protocol and RDF Query Language* (SPARQL) (CLARK, GRANT, TORRES, 2008). Assim como dados armazenados em banco de dados relacionais podem ser consultados e manipulados com a linguagem *Structured Query Language* (SQL), dados armazenados na forma de triplas podem ser consultados e manipulados com a linguagem SPARQL, por meio de SPARQL *Endpoints*, serviços *Representational State Transfer* (REST) que implementam o protocolo SPARQL.

A linguagem SPARQL especifica quatro formas de consultas com diferentes tipos de propósitos para recuperar os dados. A consulta SELECT extrai dados RDF e retorna os resultados estruturados em formato de tabela; a consulta CONSTRUCT possibilita a construção de novos grafos RDF a partir dos dados extraídos; a consulta ASK provê uma resposta booleana (verdadeiro ou falso) para perguntas que verificam se um padrão de consulta possui resultados e a consulta DESCRIBE retorna um grafo RDF contendo informações sobre determinado recurso identificado por uma IRI. Em cada forma de consulta, um bloco WHERE deve ser usado para restrição do conjunto de resultados, sendo a restrição opcional para consultas do tipo ASK e DESCRIBE. O Quadro 1 exibe o resultado da execução da consulta SPARQL SELECT descrita a seguir, no SPARQL *Endpoint* da DBPedia[8]:

[8] http://dbpedia.org/sparql

```
1  PREFIX dbpedia: <http://dbpedia.org/resource/>
2  PREFIX dbpprop: <http://dbpedia.org/property/>
3  PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
4
5  SELECT ?name ?foundingYear ?product
6  WHERE {
7        dbpedia:Petrobras
8              dbpprop:name ?name ;
9              dbpedia-owl:foundingYear ?foundingYear ;
10             dbpedia-owl:product ?product .
11 }
```

Quadro 1. Nome, ano de fundação e produtos da empresa Petrobras retornados pela consulta SPARQL SELECT no SPARQL *Endpoint* da DBPedia.

name	**foundingYear**	**product**
"Petróleo Brasileiro S.A."@en	"1953-01-01T00:00:00+02:00" ^^<http://www.w3.org/2001/XML Schema#gYear>	http://dbpedia.org/resource/Fertilizer
"Petróleo Brasileiro S.A."@en	"1953-01-01T00:00:00+02:00" ^^<http://www.w3.org/2001/XML Schema#gYear>	http://dbpedia.org/resource/Natural_gas
"Petróleo Brasileiro S.A."@en	"1953-01-01T00:00:00+02:00" ^^<http://www.w3.org/2001/XML Schema#gYear>	http://dbpedia.org/resource/Petroleum
"Petróleo Brasileiro S.A."@en	"1953-01-01T00:00:00+02:00" ^^<http://www.w3.org/2001/XML Schema#gYear>	http://dbpedia.org/resource/Petroleum_product
"Petróleo Brasileiro S.A."@en	"1953-01-01T00:00:00+02:00" ^^<http://www.w3.org/2001/XML Schema#gYear>	http://dbpedia.org/resource/Biofuel
"Petróleo Brasileiro S.A."@en	"1953-01-01T00:00:00+02:00" ^^<http://www.w3.org/2001/XML Schema#gYear>	http://dbpedia.org/resource/Lubricant
"Petróleo Brasileiro S.A."@en	"1953-01-01T00:00:00+02:00" ^^<http://www.w3.org/2001/XML Schema#gYear>	http://dbpedia.org/resource/Petrochemical

A versão 1.1 da linguagem SPARQL (ARANDA et al., 2013) é a versão corrente recomendada pelo W3C. Esta versão estende os recursos de consulta da versão 1.0 do SPARQL, oferecendo novas funcionalidades como os operadores de caminhos de propriedade (*property paths*) e as funções IRI e BOUND, utilizados nas consultas SPARQL apresentadas no capítulo 5. O SPARQL 1.1 provê ainda suporte para operações de atualização (INSERT DATA, DELETE DATA, DELETE/INSERT, LOAD e CLEAR) e gerenciamento (CREATE, DROP, COPY, MOVE e ADD) em um banco de triplas.

A última camada a ser abordada nesta subseção está relacionada com o padrão do W3C denominado *Rule Interchange Format* (RIF) (BOLEY et al., 2013). Dada a existência de uma grande variedade de sistemas de regras em uso com características sintáticas e semânticas distintas, RIF visa ser um padrão para propiciar interoperabilidade de regras de inferência na Web Semântica (KIFER, 2008). Para isso, RIF ofecere três dialetos, que são linguagens extensíveis com sintaxe e semântica rigorosamente definidas: *RIF Core Dialect*, *RIF Basic Logic Dialect* e *RIF Production Rule Dialect*. Estes dialetos são compatíveis com XML, RDF, RDF-S e OWL (BRUIJN, WELTY, 2013). O padrão associado a esta camada semântica não chegou a ser explorado neste trabalho.

2.2.5 Criptografia, Lógica Unificadora, Prova, Confiança e Interface com Usuário e Aplicações

As demais camadas da Pilha da Web Semântica, ou seja, as camadas de Criptografia (*Cryptrography*), Lógica Unificadora (*Unifying Logic*), Prova (*Proof*), Confiança (*Trust*) e Interface com Usuário e Aplicações (*User Interface and Applications*) contêm tecnologias ainda não padronizadas ou contêm somente indicações de boas práticas a serem implementadas.

A camada de Criptografia está relacionada com a atividade de criptografar as triplas RDF, por meio de assinatura digital, para proteger os dados da Web Semântica e assegurar que eles procedem de fontes seguras (KUMAR, S. et al., 2010). A camada de Lógica Unificadora está relacionada com a atividade de inferir novas informações e verificar a consistência dos dados, através do processamento, por agentes computacionais, das ontologias e das regras de inferência proporcionadas pelas camadas inferiores da Pilha da Web Semântica (HU, BOLEY, 2010). A camada de Prova se destina a justificar as informações obtidas pelos agentes computacionais, durante o consumo e o processamento das informações da Web Semântica (SIZOV, 2007). A camada de Confiança visa endereçar as questões relacionadas à confiabilidade das informações consumidas pelos agentes computacionais (ARTZ, GIL, 2007).

De acordo com BIZER e OLDAKOWSKI (2004), as sentenças da Web Semântica só devem ser consideradas como fatos depois de se evidenciar a confiabilidade das mesmas. Por fim, a camada de Interface com Usuário e Aplicações permite a interação dos usuários com os dados da Web Semântica, através das aplicações (JANIK, SCHERP, STAAB, 2011).

Conforme indicado nos artigos de SIZOV (2007) e ARTZ e GIL (2007), a proveniência possui um papel importante de suporte para as atividades relacionadas às camadas de Prova e Confiança da Pilha da Web Semântica. Essa importância da proveniência para o suporte da avaliação da qualidade e da confiabilidade dos dados da Web Semântica será abordada no capítulo 3.

2.3 Linked (Open) Data

Baseado nas tecnologias e boas práticas da Web Semântica, BERNERS-LEE (2006) estabeleceu quatro regras, listadas abaixo, para publicar e interligar dados estruturados na Web:

1. Usar URIs como nomes para os itens.
2. Usar URIs HTTP para que as pessoas possam consultar esses nomes.
3. Quando alguém consultar uma URI, prover informação RDF útil.
4. Incluir sentenças RDF interligando URIs, a fim de permitir que itens relacionados possam ser descobertos.

Estas regras constituíram-se como os princípios fundamentais do termo Linked Data. Segundo BIZER, HEATH e BERNERS-LEE (2009), "Linked Data refere-se a dados publicados na Web, de tal forma que sejam legíveis por máquina, seus significados sejam explicitamente definidos, estejam interligados a outros conjuntos de dados externos e possam ser interligados de outros conjuntos de dados externos". Assim, a adoção dos princípios de Linked Data forneceu meios para concretizar a visão da Web Semântica e levou à criação da Web de Dados, um Grafo Gigante Global (GGG - *Giant Global Graph*) (BERNERS-LEE, 2007) contendo bilhões de triplas RDF que representam e interligam dados provenientes de diversos domínios.

O conceito Linked Data pode ser especializado para *Linked Open Data* (LOD) (BAUER, KALTENBÖCK, 2011), quando está vinculado com o conceito de Dados Abertos, ou seja, "dados que podem ser usados, reutilizados e redistribuídos livremente, sujeitos, no

máximo, à exigência de atribuição e compartilhamento pela mesma licença"[9]. O projeto Linking Open Data[10], um esforço comunitário fundado em 2007 e suportado pelo W3C, tornou-se a iniciativa mais visível de LOD. O objetivo desta iniciativa consiste em identificar conjuntos de dados disponibilizados com licença aberta, para convertê-los e publicá-los na Web de acordo com os princípios de Linked Data (HEATH, BIZER, 2011). O diagrama da Figura 7 exibe os conjuntos de dados publicados pelo projeto Linking Open Data até setembro de 2011. Cada nó do diagrama representa um conjunto de dados distinto e as arestas representam as interligações entre os conjuntos de dados. As diferentes cores dos nós indicam os diferentes domínios, aos quais os conjuntos de dados pertencem.

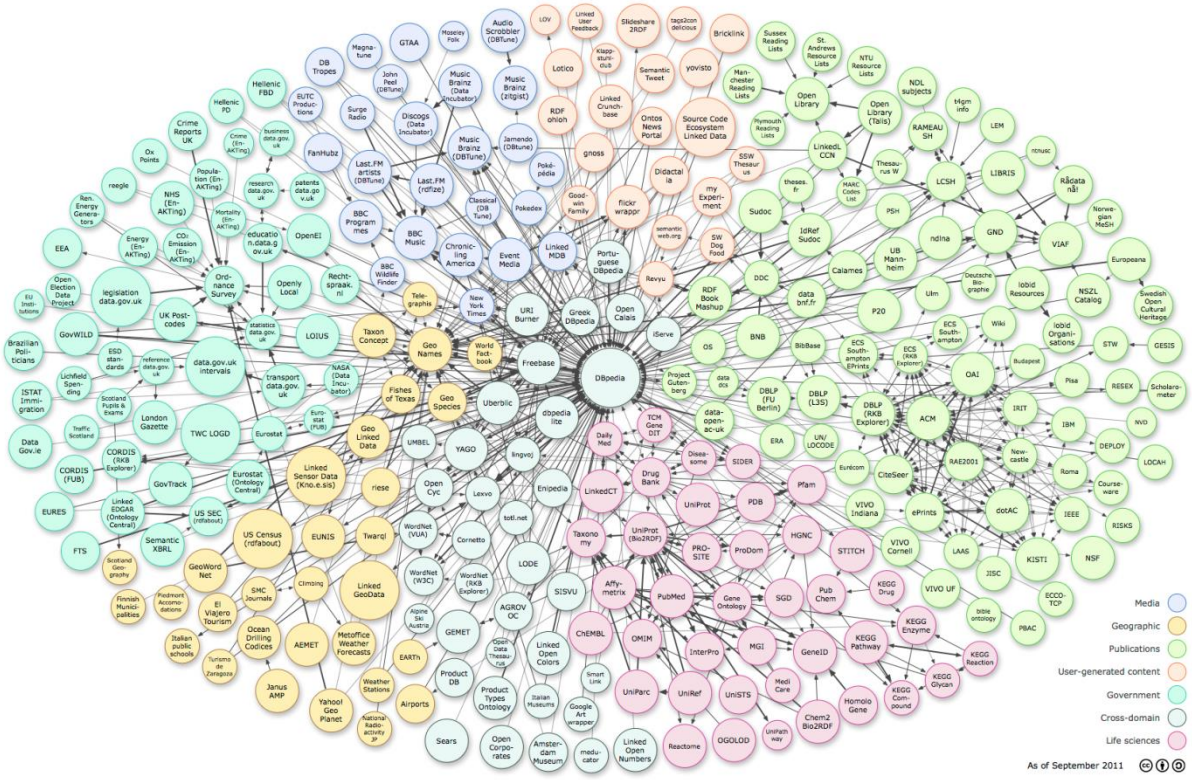

Figura 7. Visão geral dos conjuntos de dados publicados pelo projeto Linking Open Data até setembro de 2011. CYGANIAK e JENTZSCH (2011).

Antes de concluir esta seção, cabe ressaltar que o conjunto de dados DBPedia, que disponibiliza dados estruturados provenientes da Wikipedia[11], configura-se como o núcleo da Web de Dados publicada pelo projeto Linking Open Data (AUER et al., 2007). Informações

[9] http://opendefinition.org

[10] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[11] http://www.wikipedia.org

desse conjunto de dados sobre a empresa Petrobras foram utilizadas nos exemplos das seções 2.2.1, 2.2.3 e 2.2.4.

2.4 Publicação de Linked Data

2.4.1 A abordagem do projeto LOD2

Desde o estabelecimento dos princípios de Linked Data, arquiteturas, *frameworks* e ciclos de vida foram constituídos para gerenciar as atividades desafiadoras relacionadas com o processo de publicação e interligação de dados no contexto de Linked Data. Um exemplo relevante é a arquitetura definida pelo projeto europeu LOD2[12], denominada como Pilha do LOD2 (*LOD2 Stack*) (AUER et al., 2012).

A Pilha do LOD2 consiste de uma arquitetura extensível formada por um conjunto de ferramentas distribuídas e integradas para apoiar todas as etapas do ciclo de vida definido pelo projeto LOD2. D2R Server (BIZER, CYGANIAK, 2006), Openlink Virtuoso (ERLING, MIKHAILOV, 2007), OntoWiki (HEINO et al., 2009), Silk (VOLZ et al., 2009) e Linked Data Integration Framework (LDIF) (SCHULTZ et al., 2011) são exemplos de ferramentas definidas para apoiar as atividades da Pilha do LOD2. D2R Server é utilizado para extrair e triplificar dados provenientes de banco de dados relacionais; Openlink Virtuoso é o banco de triplas utilizado para armazenamento dos dados; OntoWiki é a ferramenta Wiki cujos recursos são utilizados para definição de autoria do conteúdo semântico; Silk é o *framework* utilizado para descoberta e criação de interligações entre diferentes fontes de dados e LDIF é o *framework* utilizado para integração de dados provenientes de diferentes conjuntos de dados no contexto de Linked Data, cujo controle da proveniência dos dados contribui para a análise da qualidade dos mesmos.

O ciclo de vida do LOD2 é composto pelas 8 etapas ilustradas na Figura 8, cujas atividades são descritas abaixo:

1. Extração: emprego de tecnologia para acessar, extrair e triplificar dados de fontes heterogêneas, com o intuito de publicá-los na Web de Dados;
2. Armazenamento / Recuperação: emprego de tecnologia para armazenamento das triplas RDF com otimização da escalabilidade, das consultas, do *cache* das junções (*joins*) e do processamento dos grafos;

[12] http://lod2.eu

3. Autoria / Revisão: emprego de tecnologia para facilitar a autoria de bases de conhecimento com riqueza semântica, através do paradigma "O que você vê é o que você quer dizer" (WYSIWYM - *What You See Is What You Mean*), e emprego de técnicas voltadas para autoria e revisão no contexto de redes sociais distribuídas e colaboração semântica;
4. Interligação / Fusão: emprego de tecnologia para criação e manutenção de ligações entre os dados de maneira automática ou semiautomática, a fim de estabelecer coerência e facilitar a integração dos dados;
5. Classificação / Enriquecimento: emprego de tecnologia para integração dos dados publicados com ontologias para possibilitar, entre outras atividades, a integração, a fusão e a busca dos dados;
6. Análise de Qualidade: emprego de tecnologia para suportar a análise da qualidade dos dados, através de características como proveniência, contexto, cobertura ou estrutura;
7. Evolução / Reparo: emprego de tecnologia para detectar problemas nos conjuntos de dados publicados, devido às mudanças ou atualizações nos dados, vocabulários e ontologias, e automaticamente sugerir estratégias de reparo;
8. Busca / Navegação / Exploração: emprego de tecnologia para buscar, navegar, explorar e visualizar dados espaciais, temporais e estatísticos, publicados segundo os princípios de Linked Data, a fim de torná-los acessíveis para os usuários.

De acordo com o projeto LOD2, essas etapas não devem ser tratadas de maneira isolada, mas sempre de maneira integrada, ou seja, as atividades de determinada etapa devem estar alinhadas e trazer insumos para as atividades realizadas pelas demais etapas do ciclo de vida. Por exemplo, o enriquecimento semântico realizado na etapa Classificação / Enriquecimento deve colaborar para a detecção de inconsistências e problemas de modelagem realizada na etapa Evolução / Reparo, que por sua vez deve trazer benefícios para as atividades realizadas na etapa de Interligação / Fusão. Da mesma maneira, os dados de proveniência utilizados na etapa de Análise de Qualidade devem proceder do registro de evolução das bases de conhecimento através do ciclo de vida. Este registro é realizado pelas ferramentas que apóiam cada etapa do ciclo de vida (AUER et al., 2012).

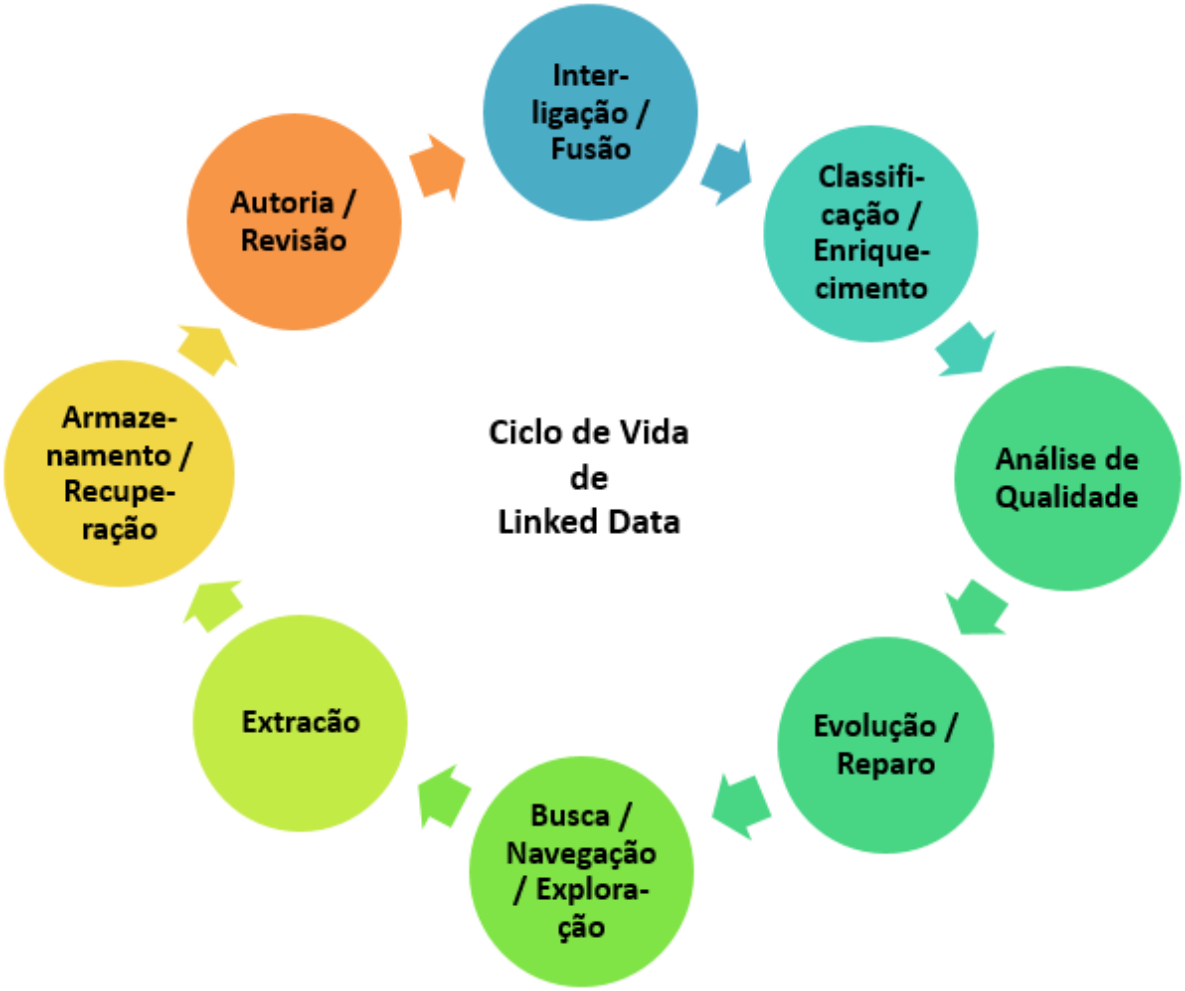

Figura 8. Etapas do ciclo de vida de Linked Data suportado pela Pilha do LOD2. Adaptado de AUER et al. (2012).

2.4.2 Publicação de Linked Data dentro das organizações

Com relação à etapa de Extração do ciclo de vida do projeto LOD2, vimos que ela remete à atividade de extração dos dados a partir de fontes heterogêneas e à posterior transformação dos dados extraídos em triplas RDF. No entanto, complexas atividades de limpeza, consolidação, agregação e integração dos dados necessitam, frequentemente, ser executadas entre a extração e a triplificação, especialmente no contexto de organizações como agências governamentais, empresas públicas e privadas e instituições de pesquisa.

Neste cenário, embora pareça contradizer a abordagem liberal das iniciativas de LOD, a publicação de Linked Data configura-se como mais um processo que necessita ser monitorado e gerenciado pela organização. Este processo é ilustrado na Figura 9. Na verdade, a figura representa dois processos complementares. Além do processo relacionado com a etapa de extração do ciclo de vida de Linked Data, denominado Processo de Preparação e Transformação

dos Dados, há também o processo relacionado com a etapa de interligação, denominado Processo de Interligação dos Dados.

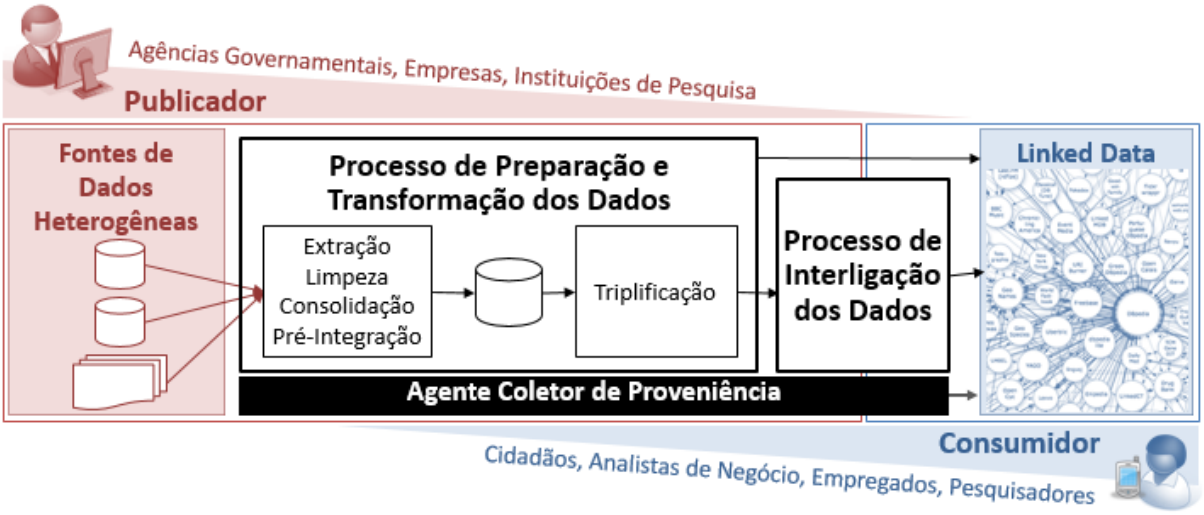

Figura 9. Visão geral do processo de publicação de Linked Data realizado dentro das organizações.

O Processo de Preparação e Transformação dos Dados traz a necessidade de estender a etapa de extração para contemplar os passos de preparação e pré-integração que ocorrem antes da triplificação. O uso de um *workflow* cuidadosamente planejado para controlar, programar e executar a sequência de atividades do Processo de Preparação e Transformação dos Dados torna-se uma importante estratégia de governança dos dados dentro das organizações. Além disso, a sequência de atividades do Processo de Preparação e Transformação dos Dados pode ser naturalmente mapeada por um processo de Extração, Transformação e Carga (ETL - Extract, Transform and Load), de maneira análoga ao que ocorre nos sistemas de *Data Warehousing* (DW) (CORDEIRO et al., 2011).

O Processo de Interligação dos Dados remete a dois problemas principais: o problema de encontrar termos correspondentes nos diferentes vocabulários utilizados e o problema de encontrar entidades correspondentes presentes em *datasets* distintos. Esses problemas são resolvidos com atividades de mapeamento de esquemas, que consistem no mapeamento de termos correspondentes dos diferentes vocabulários, e atividades de resolução de entidades, que consistem em encontrar e interligar entidades similares presentes em *datasets* distintos. Para isso, técnicas de mapeamento e alinhamento de ontologias (EUZENAT, MOCAN, SCHARFFE, 2008) são fortemente utilizadas, com emprego de um grande conjunto de parâmetros e produção de uma grande quantidade de resultados.

A parametrização utilizada e os resultados obtidos tanto no Processo de Preparação e Transformação dos Dados, quanto no Processo de Interligação dos Dados, configuram-se como

dados de proveniência que podem apoiar a avaliação da qualidade dos dados publicados. A abordagem proposta por este trabalho, descrita no capítulo 4, tem como foco os dados de proveniência produzidos pelo *workflow* ETL do Processo de Transformação e Preparação dos Dados.

2.5 A abordagem ETL

A natureza e a variedade das fontes de dados são os fatores principais que devem ser considerados para a escolha da estratégia mais apropriada de publicação de Linked Data (HEATH, BIZER, 2011). No caso do processo de publicação realizado dentro dos limites de uma organização, a atividade de extração dos dados de múltiplas fontes heterogêneas, seguida das atividades de limpeza, consolidação, agregação e integração dos mesmos, antes de transformá-los em triplas RDF para serem disponibilizadas no contexto de Linked Data, podem naturalmente ser orquestradas por uma abordagem ETL (CORDEIRO et al., 2011). Alguns benefícios imediatos da abordagem ETL são (i) a sistematização do processo de publicação de Linked Data; (ii) o monitoramento e gerenciamento de suas atividades de extração, transformação e carga e (iii) a oportunidade de reutilização do *workflow* para carregar novos dados e atualizar os dados previamente publicados no contexto de Linked Data. Além disso, a abordagem ETL provê recursos que possibilitam a coleta sistemática dos dados de proveniência tanto sobre a composição, quanto sobre a execução do processo de publicação de Linked Data.

O processo de ETL é o fundamento dos sistemas de *Data Warehousing* (DW), em que dados provenientes de diferentes fontes são extraídos, limpos, consolidados e, finalmente, disponibilizados através de um banco de dados construído a partir de um modelo dimensional, que permite a exploração e a análise dos dados sob diferentes perspectivas (KIMBALL, CASERTA, 2004). Baseados nas melhores práticas aplicadas em milhares de projetos de DW bem sucedidos, KIMBALL et al. (2008) definiram 34 subsistemas ou conjuntos de funcionalidades para orientar o entendimento e categorizar a implementação e o gerenciamento de uma solução ETL. Estes 34 subsistemas podem ser agrupados em 4 componentes principais de uma solução ETL, que possuem as seguintes relações com o processo de publicação de Linked Data:

1. Extração: componente em que os dados brutos são coletados de diversas fontes e persistidos em disco, antes que qualquer reestruturação ou manipulação significativa seja realizada nos dados;

2. Limpeza e Consolidação: componente em que os dados coletados são corrigidos, rejeitados ou transformados de acordo com os requisitos da organização. Tratamento de questões de duplicação de dados ao integrar fontes distintas; anotação de dados com base em vocabulários comuns, também conhecido como fusão de Linked Data (MENDES, MÜHLEISEN, BIZER, 2012), e transformação dos dados em triplas RDF são exemplos de tarefas realizadas por este componente;
3. Carga: componente em que os dados são agrupados em grafos RDF e carregados em um repositório de triplas RDF;
4. Operação e Gerenciamento: componente em que se localizam tarefas de agendamento do processo de publicação, controle de versão, controle de segurança e rastreamento dos dados de proveniência.

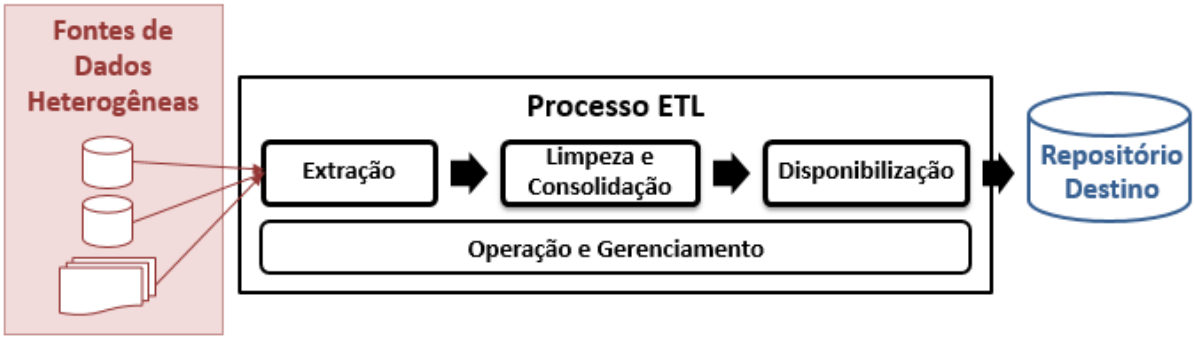

Figura 10. Visão geral dos componentes de uma solução ETL.

A abordagem que utiliza um *workflow* ETL para publicar Linked Data herda o potencial oferecido pelas técnicas e ferramentas de ETL, que foram desenvolvidas e refinadas durante anos, em desafiantes cenários de Inteligência de Negócio (BI - *Business Intelligence*). Por exemplo, a técnica *Slowly Change Dimensions*, utilizada em projetos de DW para gerenciar a atualização dos dados ao longo do tempo, trata uma questão investigada em recentes pesquisas no contexto de Linked Data (RULA et al., 2012; RULA, PALMONARI, MAURINO, 2012). Com relação às ferramentas de ETL, existe um conjunto significativo tanto de ferramentas proprietárias como SAP BusinessObjects Data Integrator[13], Microsoft SQL Server Integration

[13] http://www.sap.com/pc/tech/enterprise-information-management/software/data-integrator/index.html

Services[14], IBM InfoSphere DataStage[15] e Oracle Data Integrator (ODI)[16], quanto de ferramentas *open source* como Talend Studio for Data Integration[17] e Pentaho Data Integration[18] (também conhecida como Kettle). A ferramenta Pentaho Data Integration (Kettle) foi utilizada no projeto LinkedDataBR (CAMPOS, M. L. M., GUIZZARDI, 2010) e no exemplo de implementação da abordagem proposta por este trabalho.

2.5.1 Pentaho Data Integration (Kettle)

A ferramenta Pentaho Data Integration, também chamada de Kettle, é a ferramenta de *workflow* ETL integrante do conjunto de ferramentas da plataforma de BI Pentaho (CASTERS, BOUMAN, DONGEN, VAN, 2010). Além da característica comum das ferramentas de ETL, que é apoiar o processo de ETL em projetos de DW, o Kettle também pode ser utilizado para outros propósitos como, por exemplo, (i) migração de dados entre aplicações ou banco de dados; (ii) exportação e importação de dados entre diferentes tipos de repositórios; (iii) carga massiva de dados em um repositório de dados; (iv) limpeza de dados em um repositório e (v) integração entre aplicações. Entre os princípios sobre os quais o Kettle foi desenvolvido, encontram-se (i) a facilidade de instalação; (ii) a facilidade de desenvolvimento do *workflow* ETL; (iii) a disponibilização de interface gráfica intuitiva em detrimento da necessidade de programação por linhas de código; (iv) a extensibilidade das funcionalidades da ferramenta através de uma API e (v) a forte orientação e transparência na disponibilização de metadados sobre o *workflow* ETL.

No Kettle, o *workflow* ETL pode ser especificado através de dois tipos de componentes, denominados *transformations* e *jobs*. Um *transformation* consiste de um conjunto de passos conectados, onde cada passo, denominado *step*, é responsável por uma atividade de extração, transformação ou carga de dados. A conexão entre dois *steps* de um *transformation*, denominada *transformation hop*, permite que os dados fluam em um único sentido e de maneira assíncrona. Um *job* também consiste de um conjunto de passos conectados. No entanto, os passos de um *job*, denominados *job entries*, são responsáveis por executar um *transformation*, outro *job* ou atividades auxiliares como manipular e transferir arquivos, enviar e receber emails

[14] http://msdn.microsoft.com/en-us/library/ms141026.aspx

[15] http://www-03.ibm.com/software/products/us/en/ibminfodata

[16] http://www.oracle.com/technetwork/middleware/data-integrator/overview/index.html

[17] http://www.talend.com/products/talend-open-studio

[18] http://kettle.pentaho.com

e executar uma série de validações. A conexão entre dois *job entries*, denominada *job hop*, determina a ordem de execução dos passos do *job*, que, diferente dos passos do *transformation*, são executados de maneira síncrona. A Figura 11 ilustra o modelo conceitual do Kettle.

Tanto um *transformation*, quanto um *job* podem possuir notas para documentar informações sobre o *workflow* especificado por eles. Além disso, para fins de armazenamento da composição dos *workflows* ETL, o Kettle oferece dois tipos de repositórios. Sendo assim, as composições dos *jobs* e dos *transformations* podem ser armazenadas em tabelas de um repositório do tipo banco de dados ou em arquivos XML de um repositório do tipo sistema de arquivos.

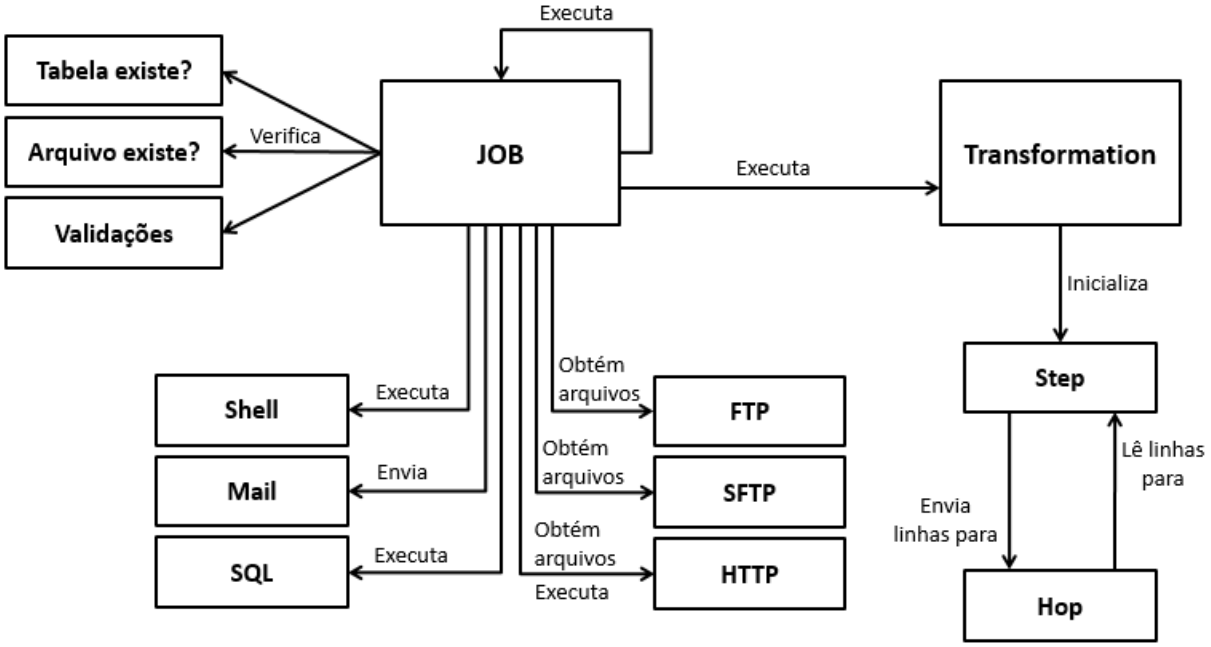

Figura 11. Modelo conceitual da ferramenta Pentaho Data Integration (Kettle).

O Kettle oferece ainda um conjunto de aplicativos para apoiar o desenvolvimento e a execução dos *workflows* ETL. Este conjunto inclui as seguintes ferramentas:

- *Spoon*: ferramenta de interface gráfica para desenvolvimento e execução do fluxo de dados, desde sua entrada até a saída, através da criação e execução de *jobs* e *tranformations*;
- *Pan*: ferramenta de linha de comando para execução dos *transformations* desenvolvidos no *Spoon*. Normalmente é utilizada para agendamento da execução de *transformations*;
- *Kitchen*: ferramenta de linha de comando para execução dos *jobs* desenvolvidos no *Spoon*. Normalmente é utilizada para agendamento da execução dos *jobs*;

- *Carte*: servidor web que possibilita o agrupamento em cluster do processo de ETL, através da execução remota de *transformations* e *jobs*.

2.5.2 ETL4LOD - Componentes do Kettle relacionados a Linked Data

Apesar de oferecer um conjunto significativo de *steps* e *job entries* que realizam uma série de atividades de extração, transformação e carga, o Kettle não oferece componentes específicos para o contexto de Linked Data. Para suprir esta carência do Kettle, o projeto LinkedDataBR desenvolveu e disponibilizou 4 novos *steps* utilizando a API Java do Kettle: Sparql Endpoint, Sparql Update Output, Data Property Mapping e Object Property Mapping. Este conjunto de *steps*, denominado ETL4LOD, executa atividades relevantes relacionadas com as tecnologias relacionadas com Linked Data, que possibilitam a publicação como triplas RDF tanto dos dados de domínio, quanto dos dados de proveniência. Assim, para concluir este capítulo, uma breve apresentação destes 4 *steps* será realizada.

O *step* Sparql Endpoint oferece a capacidade de extrair dados de um SPARQL *Endpoint*, a partir da especificação da URL relacionada ao SPARQL *Endpoint* e da definição de uma consulta SPARQL. Em conjunto com a API do Kettle, a API Jena[19] foi intensamente utilizada para implementação das operações relacionadas com RDF e SPARQL. A Figura 12 ilustra a interface de configuração do *step* Sparql Endpoint, com a especificação dos parâmetros necessários para extrair do *dataset* da DBPedia, as propriedades nome, ano de fundação e produto da empresa Petrobras.

[19] http://jena.apache.org

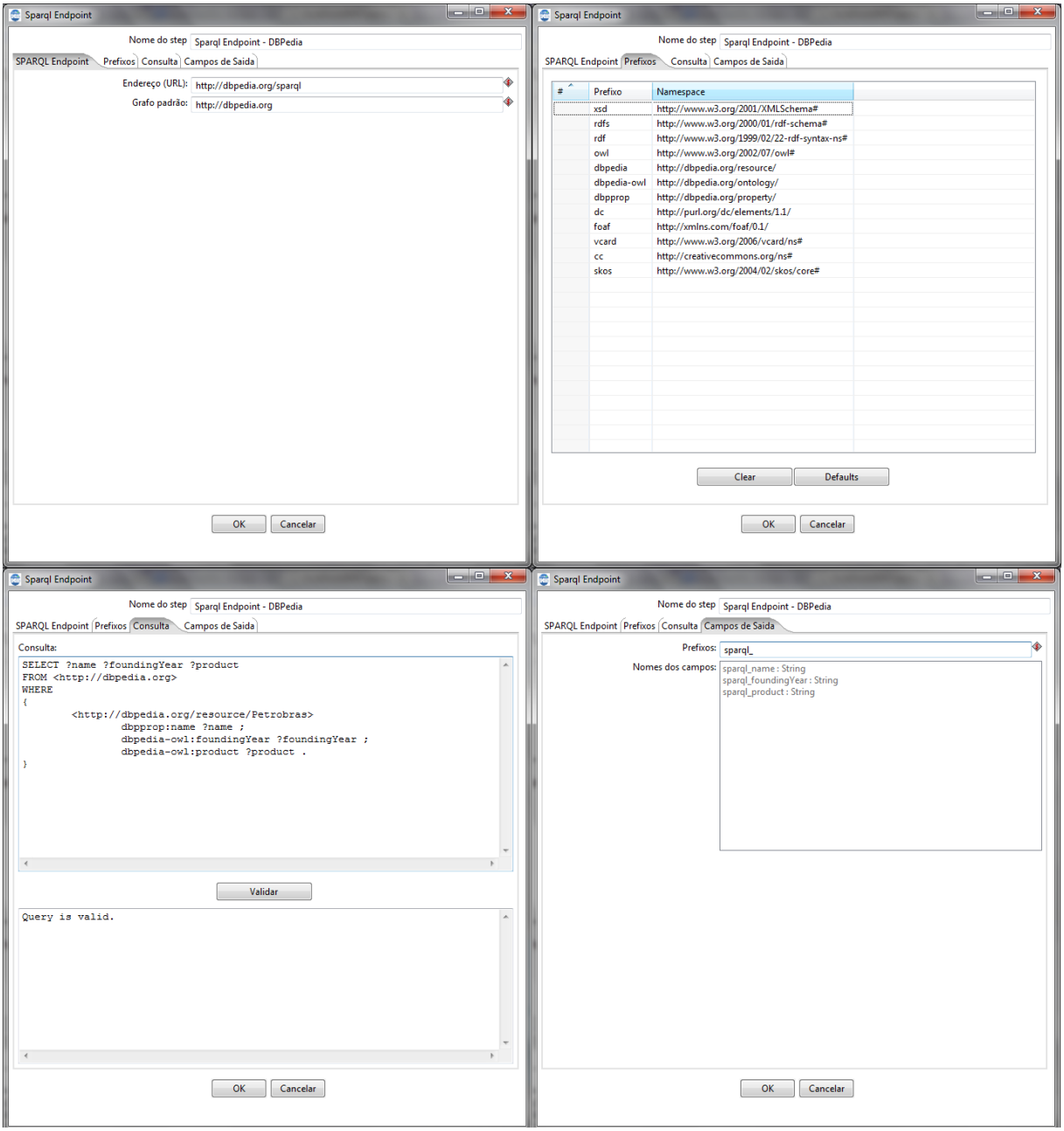

Figura 12. Interface de configuração do *step* Sparql Endpoint.

O *step* Sparql Update Output oferece a capacidade de carregar triplas RDF em um banco de triplas. Para isso, é necessário informar a URL do SPARQL *Endpoint* que provê suporte a operações de atualização, disponibilizadas na versão 1.1 do SPARQL; o usuário e a senha de autenticação no banco de triplas; a URI do grafo onde se deseja carregar as triplas RDF e o campo que contém a tripla RDF no formato NTriple. A Figura 13 ilustra a interface de configuração do *step* Sparql Update Output, com a especificação dos parâmetros necessários para inserir triplas RDF sobre a empresa Petrobras no grafo *http://petrobras.br*.

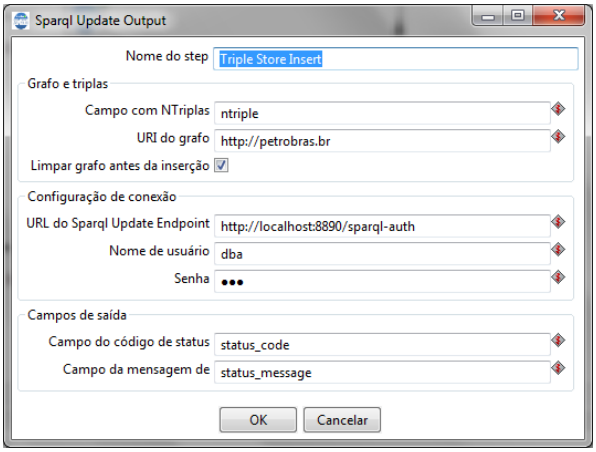

Figura 13. Interface de configuração do *step* Sparql Update Output.

O *step* Data Property Mapping oferece a capacidade de mapear, a partir das linhas do fluxo de entrada, os componentes de uma tripla RDF (sujeito, predicado e objeto) nas linhas do fluxo de saída, sendo o objeto um valor literal. Cada linha de entrada deve conter um campo com a URI que identifica o recurso do sujeito. Na configuração do *step*, é necessário definir uma ou mais URIs especificando o tipo do recurso e um mapeamento de propriedades do tipo literal, informando a URI da propriedade e o campo da linha de entrada que contém o valor da propriedade. Além disso, é possível definir, no mapeamento, o tipo do literal e a marcação de idioma.

O *step* Object Property Mapping é similar ao *step* Data Property Mapping, com a diferença de que o valor do objeto enviado no fluxo de saída é uma URI de um recurso. A configuração deste *step* é mais simples e consiste em indicar o campo da linha de entrada que contem a URI do sujeito, o campo da linha de entrada que contem a URI do objeto e definir a URI da propriedade. A Figura 14 ilustra a interface de configuração do *step* Data Property Mapping, com a especificação do tipo e do mapeamento das propriedades nome e ano de fundação do recurso Petrobras e a Figura 15 ilustra a interface de configuração do *step* Object Property Mapping, com a especificação do mapeamento da propriedade produto do recurso Petrobras.

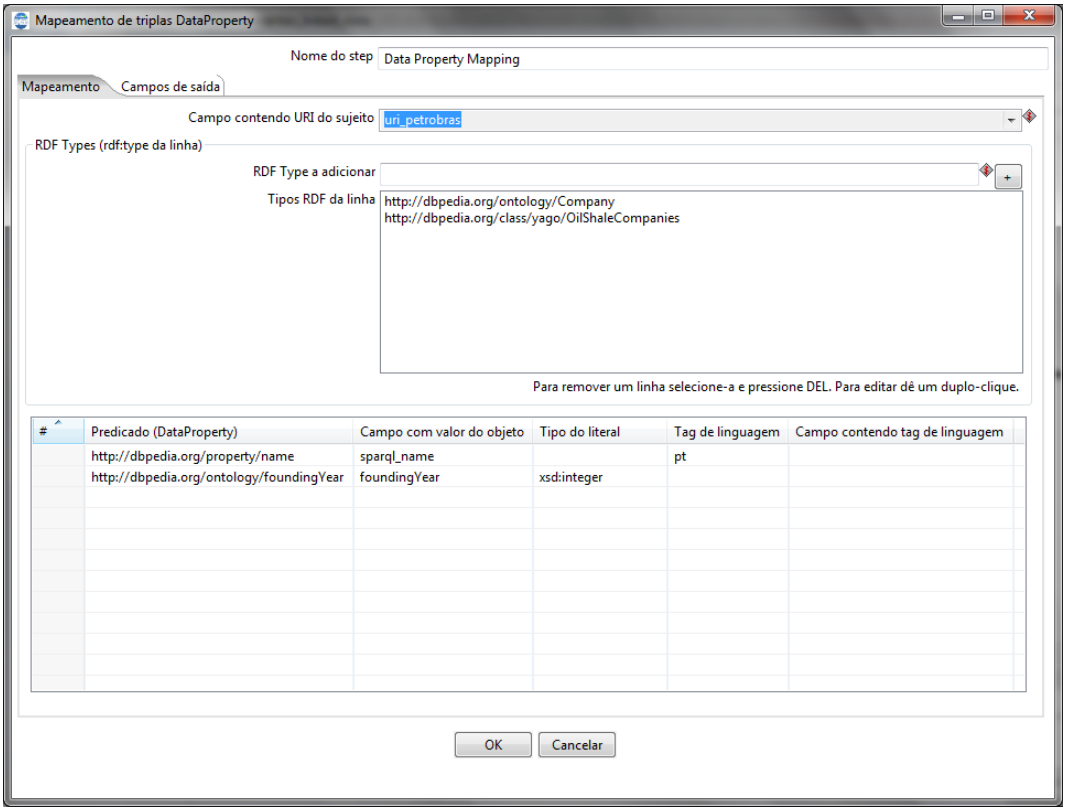

Figura 14. Interface de configuração do *step* Data Property Mapping.

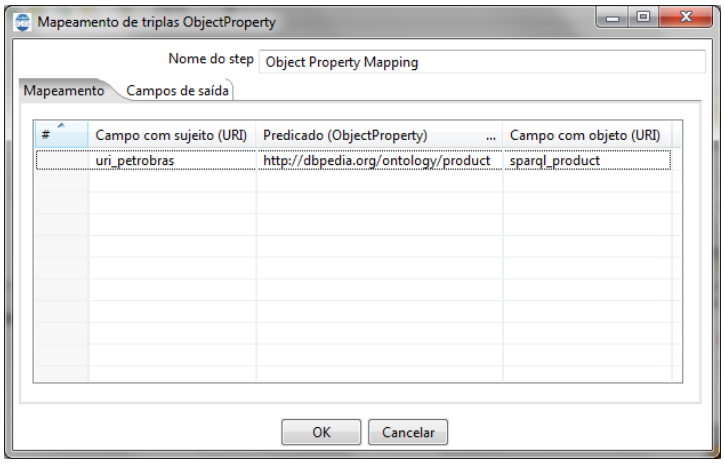

Figura 15. Interface de configuração do *step* Object Property Mapping.

3 Proveniência

3.1 Introdução

As definições do termo proveniência (*provenance*) encontradas nos dicionários Oxford English Dictionary[20] e Merriam-Webster Online Dictionary[21] indicam-no como sendo todo tipo de informação que descreve a origem de um objeto e o processo de derivação do mesmo até determinado estado (MOREAU, 2010). O conceito de proveniência é aplicado em diversos campos de estudo (BEARMAN, LYTLE, 1985; MILLAR, 2002; FABANI et al., 2009; MOREAU, 2010) e, mesmo dentro de determinado campo de estudo, possui diferentes significados dependendo do contexto abordado.

Por exemplo, na área da Ciência da Computação, a literatura especializada apresenta diferentes visões de proveniência: (i) proveniência como a documentação do processo que resultou um conjunto de dados (GROTH, MILES, MOREAU, 2009); (ii) proveniência representada como um Grafo Acíclico Dirigido (DAG - *Directed Acyclic Graph*) (MOREAU et al., 2008); (iii) proveniência como os locais dos quais foram extraídos cada resultado de uma consulta em um banco de dados (*Where-Provenance*) (BUNEMAN, KHANNA, TAN, W., 2001); (iv) proveniência como as origens envolvidas no cálculo de cada resultado de uma consulta em um banco de dados (*Why-Provenance*); (v) a maneira como estas origens foram trabalhadas (*How-Provenance*) (BUNEMAN, KHANNA, TAN, W., 2001; GREEN, KARVOUNARAKIS, TANNEN, 2007); (vi) proveniência como anotações apoiadas por estruturas semânticas como ontologias, taxonomias ou vocabulários controlados (HARTIG, 2009) e (vii) proveniência como uma visão orientada a eventos (HASAN, SION, WINSLETT, 2009).

Este trabalho utiliza a definição de proveniência estabelecida pelo grupo de trabalho do W3C Provenance Incubator Group[22], que considera como proveniência de um recurso no contexto de Linked Data, o conjunto de informações relacionadas às entidades, aos processos e aos agentes envolvidos na produção e disponibilização do recurso (GIL et al., 2010). Os

[20] http://www.oed.com

[21] http://www.merriam-webster.com/dictionary/provenance

[22] http://www.w3.org/2005/Incubator/prov/wiki/W3C_Provenance_Incubator_Group_Wiki

metadados de um recurso só se tornam parte de sua proveniência, quando indicam uma característica de sua origem ou do seu processo de produção. Por exemplo, o metadado que informa o tamanho de um arquivo não é considerado um metadado de proveniência, uma vez que não indica uma característica do processo de produção do arquivo, apesar de ser uma consequência deste processo. Já o metadado que informa a data de criação do arquivo é considerado um metadado de proveniência relevante.

A compilação bibliográfica sobre proveniência realizada por MOREAU (2010) na área da Ciência da Computação mostra que a publicação de trabalhos sobre o tema iniciou-se em 1986 e teve um crescimento acelerado nos últimos anos. Os trabalhos apresentam pesquisas relevantes que abordam aspectos da proveniência relacionados com disciplinas como Segurança (TAN, V. et al., 2006; BRAUN, SHINNAR, SELTZER, 2008), Recuperação da Informação (LYNCH, 2001), Banco de Dados (CHENEY, CHITICARIU, TAN, W., 2009) e e-Ciência (SIMMHAN, PLALE, GANNON, 2005), entre outras. Embora situadas em disciplinas diferentes, as pesquisas frequentemente investigam características em comum relacionadas com captura, representação e consumo da proveniência. Contudo, de maneira geral, os resultados das pesquisas tendem a se concentrar dentro de suas respectivas comunidades, sem ampla disseminação e extensão para o campo da Web e, especificamente, da Web Semântica.

No contexto da Web de Semântica, conforme citado no capítulo anterior, a proveniência dos dados é crucial para decidir se os dados são confiáveis, como eles devem ser integrados com outras fontes de informação e como dar crédito aos seus autores. Em situações em que as informações são contraditórias ou questionáveis, aplicativos da Web Semântica podem se beneficiar da representação explícita da proveniência para realizar o julgamento da qualidade e da confiabilidade das informações consumidas (GIL et al., 2010).

3.2 Taxonomia de proveniência

Uma taxonomia é, por definição, um esquema de classificação hierárquica de conceitos, ou seja, é um esquema que organiza um conjunto de conceitos de determinado domínio de conhecimento através de relacionamentos de generalização e especialização (CAMPOS, M. L. DE A., GOMES, 2008). O uso de taxonomia configura-se como uma importante estratégia de organização da informação, devido sua característica de atuar como um mapa conceitual e possibilitar a recuperação das informações através da navegação pelos níveis hierárquicos.

Na área da Ciência da Computação, a taxonomia tem sido amplamente adotada, por exemplo, para classificar os conteúdos de um sistema de Gerenciamento Eletrônico de Documentos (GED), para organizar o cardápio de funcionalidades de um Portal Web ou para complementar o acesso às informações recuperadas por um mecanismo de busca. Além disso, através de sua atuação como um mapa conceitual, a taxonomia pode ser empregada como uma ferramenta importante para orientar o entendimento de determinado domínio de conhecimento (CRUZ, 2011). Seguindo esta direção, este trabalho adotou a taxonomia de proveniência definida por CRUZ, CAMPOS, M. L. M. e MATTOSO (2009) como referência para alcançar uma melhor compreensão da proveniência de dados em geral e, especificamente, da proveniência no contexto de Linked Data.

3.2.1 Taxonomia de proveniência para o domínio da e-Ciência

A taxonomia definida por CRUZ, CAMPOS, M. L. M. e MATTOSO (2009) representa um esforço para unificar classificações de proveniência no domínio da e-Ciência propostas anteriormente (SIMMHAN, PLALE, GANNON, 2005; YU, BUYYA, 2005; DEELMAN et al., 2009). Deste modo, a taxomonia visa oferecer um panorama sistemático e conceitual de um grande conjunto de características de proveniência, correlacionando-as com os sistemas de proveniência que suportam experimentos científicos. Entre os principais objetivos desta taxonomia encontram-se (i) a classificação dos estilos de arquitetura dos sistemas de proveniência; (ii) a classificação da proveniência de acordo com o momento em que a mesma é capturada por determinado mecanismo e (iii) a classificação dos métodos existentes para armazenar e acessar os dados de proveniência capturados.

A Figura 16 ilustra as diversas facetas da taxonomia de proveniência adotada como referência por este trabalho. A ontologia *Open proVenance Ontology* (OvO), desenvolvida por CRUZ (2011) para representar a proveniência prospectiva e retrospectiva de experimentos científicos em larga escala, contempla as facetas ilustradas. A descrição de cada uma das facetas da taxonomia será realizada nas próximas subseções. Embora tenha sido desenvolvida para cobrir características de proveniência ligadas ao domínio da e-Ciência e possuir facetas de classificação específicas a este domínio, um grande subconjunto de facetas pode ser utilizado para classificar também a proveniência no contexto de Linked Data. Sendo assim, durante a descrição das facetas, é realizada uma análise da relação das facetas com a proveniência no contexto de Linked Data e, na subseção 3.2.2, uma adaptação da taxomonia é proposta para classificar a proveniência no contexto de Linked Data.

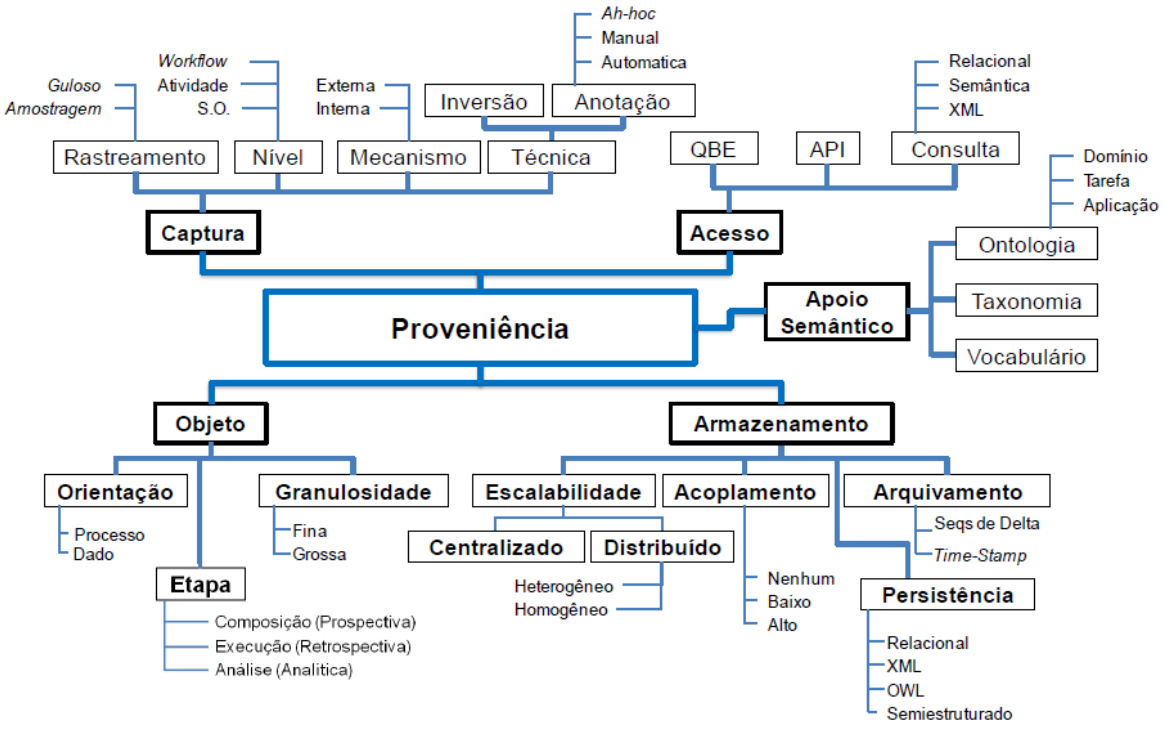

Figura 16. Taxonomia de proveniência para o domínio da e-Ciência. CRUZ (2011).

3.2.1.1 Faceta de captura

A faceta de captura (Figura 17) classifica a proveniência de acordo com as maneiras com que ela pode ser coletada pelos sistemas de proveniência no contexto de experimentos científicos. Possui quatro subcategorias utilizadas para classificar a proveniência segundo a abordagem de rastreamento, assim como os níveis, os mecanismos e as técnicas de captura.

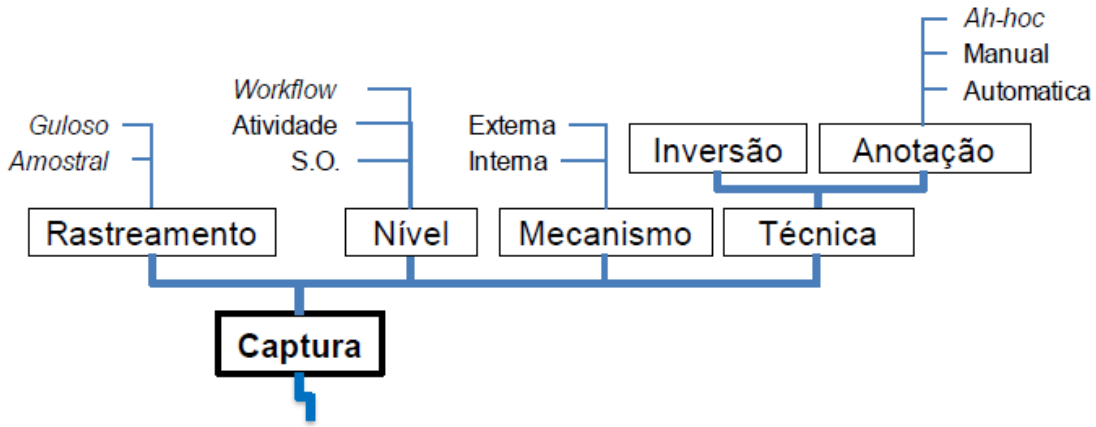

Figura 17. Faceta de captura de proveniência. CRUZ (2011).

Com relação ao tipo de rastreamento, existem duas abordagens: gulosa e amostragem (BUNEMAN, TAN, W., 2007 apud CRUZ, CAMPOS, M. L. M., MATTOSO, 2009). A

abordagem gulosa consiste na coleta da proveniência de maneira intensa à medida que os eventos monitorados ocorrem; enquanto que, na abordagem amostragem, a coleta é menos intensa e somente alguns dados de proveniência necessários para determinado objetivo são coletados. A principal vantagem da abordagem amostragem é o menor volume de dados coletados e, consequentemente, a menor sobrecarga do sistema e o menor consumo de espaço para armazenamento. Por outro lado, o maior volume de metadados de proveniência coletados pela abordagem gulosa e a disponibilização destes metadados imediatamente após a execução de um evento oferecem melhores condições para análises diversificadas da qualidade e da confiabilidade dos dados gerados. Este tipo de classificação pode ser aplicada para a proveniência no contexto de Linked Data.

Com relação aos níveis de captura, a proveniência dos experimentos científicos pode ser coletada em três níveis distintos: nível de *workflow*, nível de atividade ou nível de sistema operacional (DAVIDSON, FREIRE, 2008 apud CRUZ, CAMPOS, M. L. M., MATTOSO, 2009). O nível de *workflow* consiste na coleta da proveniência através de mecanismos integrados ou conectados a um sistema de *workflow* que gerencia o experimento científico. O nível de atividade consiste na coleta de proveniência desacoplado do Sistema de Gerenciamento de *Workflow* Científico (SGWfC), isto é, cada subprocesso é responsável por capturar suas próprias informações de proveniência. Por último, o nível de sistema operacional coleta metadados de proveniência de baixo nível através da API do sistema operacional. O nível de captura da proveniência no contexto de Linked Data também pode ser classificado nestas três categorias.

Com relação aos mecanismos de captura, existem dois mecanismos principais utilizados pelos sistemas de proveniência no contexto da e-Ciência. Alguns sistemas de proveniência (CALLAHAN et al., 2006; ZHAO et al., 2008) usam estruturas internas do SGWfC para coletar a proveniência, normalmente em ambientes centralizados, e outros (ALTINTAS, BARNEY, JAEGER-FRANK, 2006; BITON, BOULAKIA, DAVIDSON, 2007; MARINHO, MURTA, WERNER, 2011) utilizam serviços externos, que são mais genéricos e voltados para ambientes distribuídos e heterogêneos. Os mecanismos de captura da proveniência em Linked Data também podem ser classificados como internos e externos.

Com relação às técnicas de captura, os sistemas de proveniência relacionados a experimentos científicos utilizam dois tipos de técnicas: captura por anotação e captura por inversão. A captura de proveniência por anotação consiste na adição de marcações nos dados produzidos pelo SGWfC. Estas anotações podem ser realizadas manualmente pelos usuários ou

automaticamente por aplicações, algumas vezes de maneira *ad hoc*. A técnica de captura por anotação pode ser estendida para classificar também a proveniência em Linked Data e, assim como no contexto da e-Ciência, a técnica de anotação automática configura-se normalmente como a melhor opção, pois a técnica manual tende a ser demorada e propensa a erros. O outro tipo de técnica, captura de proveniência por inversão, tem uma relação mais específica com o contexto de *workflows* científicos e consiste no registro das operações executadas sobre os artefatos da experimentação, permitindo que os produtos de dados sejam recriados posteriormente.

3.2.1.2 Faceta de objeto

A faceta de objeto (Figura 18) classifica os dados de proveniência de acordo com os níveis de detalhamento. Possui três subcategorias utilizadas para classificar os dados de proveniência segundo sua orientação, sua granulosidade e a etapa em que é coletada.

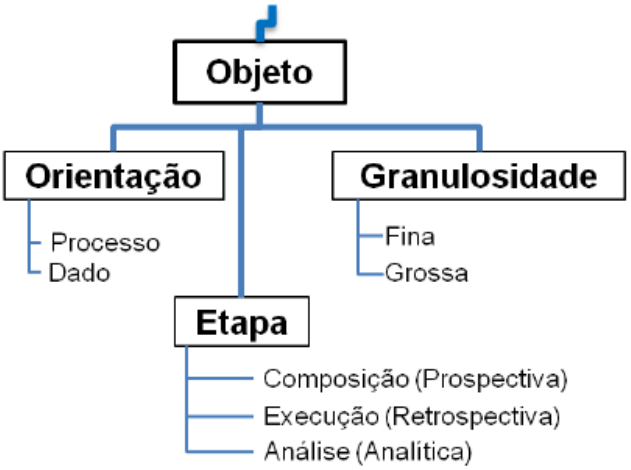

Figura 18. Faceta de objeto de proveniência. CRUZ (2011).

Existem três etapas distintas do ciclo de vida de um experimento científico, em que dados de proveniência podem ser coletados: composição, execução e análise (CLIFFORD et al., 2008; FREIRE et al., 2008 apud CRUZ, CAMPOS, M. L. M., MATTOSO, 2009). A proveniência relacionada com a composição dos passos e ações de um experimento científico é chamada de proveniência prospectiva e a proveniência relacionada com a execução do experimento científico é chamada de proveniência retrospectiva. Estes dois tipos também podem ser utilizados para classificar a proveniência dos dados publicados em Linked Data. Por exemplo, os metadados que informam quando e por quem o processo de publicação de determinado recurso no contexto de Linked Data foi composto configuram-se como metadados

de proveniência prospectiva; enquanto que os metadados que informam quando e por quem o processo foi executado configuram-se como metadados de proveniência retrospectiva. Há ainda um terceiro tipo, denominado proveniência analítica, que está relacionado com o exame dos resultados obtidos pelo experimento científico. Apesar de também poder ser utilizado no contexto de Linked Data para classificar a proveniência relacionada com a análise dos resultados do processo de publicação, a proveniência analítica não será considerada no escopo deste trabalho.

Assim como no contexto da e-Ciência, a estruturação da proveniência em Linked Data como prospectiva e retrospectiva permite a normalização dos dados de proveniência e, consequentemente, evita armazenamentos redundantes e diminui os custos de manutenção. Entretanto, de maneira análoga à dificuldade de interligar os dados de proveniência com os demais dados publicados no contexto de Linked Data, também configura-se como um problema, a dificuldade de interligar os dados de proveniência retrospectiva com seus respectivos dados de proveniência prospectiva. A abordagem proposta por este trabalho e descrita no próximo capítulo visa resolver não só a dificuldade de interligação entre os dados de proveniência com os demais dados publicados no contexto de Linked Data, como também visa resolver a dificuldade de interligação entre os dados de proveniência retrospectiva com seus respectivos dados de proveniência prospectiva.

Com relação às subcategorias orientação e granulosidade, elas também podem ser aplicadas no contexto da Linked Data. A subcategoria orientação indica se a proveniência está relacionada com o dado ou com o processo do experimento científico; no contexto de Linked Data, a subcategoria pode indicar se a proveniência está relacionada com o dado publicado ou com o processo de publicação dos dados. A subcategoria granulosidade classifica a proveniência de acordo com o seu nível de detalhamento: a granulosidade fina indica um alto nível de detalhamento e a granulosidade grossa, um nível de detalhamento mais genérico. Assim como no contexto da e-Ciência, o nível de granulosidade adotado para os dados de proveniência em Linked Data afeta o volume de armazenamento e o tipo de exploração conjunta que se pretende realizar entre os dados publicados e seus dados de proveniência. Apesar de permitir melhores condições de exploração conjunta, os dados de proveniência com granulosidade fina demandam estratégias mais aprimoradas para lidar com grandes volumes de dados, trazendo assim a necessidade de uma investigação mais aprofundada na área de Big Data (MENDONÇA et al., 2013).

3.2.1.3 Faceta de armazenamento

A faceta de armazenamento (Figura 19) classifica os sistemas de proveniência de acordo com os meios e técnicas utilizados para armazenar os dados de proveniência. Possui quatro subcategorias utilizadas para classificar o armazenamento da proveniência segundo sua escalabilidade, seu acoplamento, a maneira de persistência e a maneira de arquivamento.

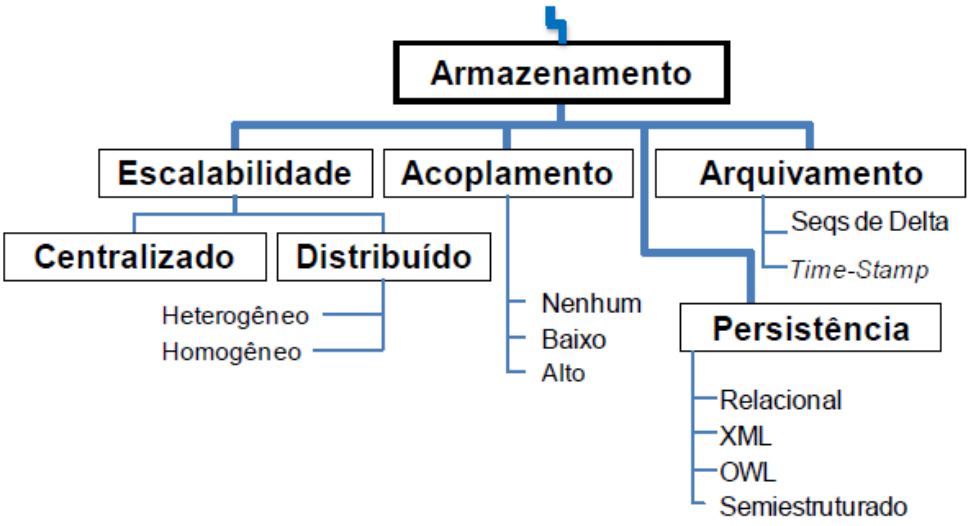

Figura 19. Faceta de armazenamento da proveniência. CRUZ (2011).

Com relação à escalabilidade do armazenamento, os metadados de proveniência podem ser armazenados de modo centralizado, isto é, em um repositório local e único, ou de modo distribuído, isto é, espalhados por um conjunto de repositórios logicamente inter-relacionados em uma rede de computadores. O armazenamento distribuído pode ainda ser classificado como homogêneo ou heterogêneo, conforme a utilização de um mesmo tipo ou de tipos diferentes de sistemas de armazenamento pelos repositórios distribuídos. O armazenamento centralizado é mais simples de gerenciar, manter e controlar. Em contrapartida, apresenta um ponto único de falha e menores condições de flexibilidade da arquitetura do que o armazenamento distribuído.

Com relação ao acoplamento, existem três estratégias: acoplamento nulo, acoplamento alto e acoplamento baixo. No acoplamento nulo, os dados de proveniência são armazenados de maneira separada e sem associação com os dados aos quais eles se referem; no acoplamento alto, os dados de proveniência são armazenados no mesmo repositório e associados com os dados aos quais eles se referem; por fim, no acoplamento baixo, os dados e seus metadados de proveniência são armazenados no mesmo repositório, mas separados em esquemas lógicos diferentes. Devido à característica de interligação dos dados no contexto de Linked Data, as

estratégias de acoplamento da proveniência podem ser altas ou baixas, mas não podem ser nulas.

Com relação à persistência, diversas maneiras são utilizadas pelos sistemas de proveniência no contexto da e-Ciência para armazenar os dados de proveniência: tabelas relacionais, documentos XML, repositórios relacionados a tecnologias da Web Semântica como RDF e OWL e documentos semi-estruturados. No contexto de Linked Data, os dados de proveniência são persistidos como triplas RDF em repositórios da Web Semântica, possibilitando assim a interligação dos dados de proveniência com os demais dados publicados em Linked Data.

Com relação à maneira de arquivamento, existem duas estratégias: sequência de delta e *timestamp* (TAN, W., 2004 apud CRUZ, CAMPOS, M. L. M., MATTOSO, 2009). As duas estratégias diferem apenas na forma de armazenamento das alterações entre as versões de um conjunto de dados. A estratégia sequência de delta armazena uma versão completa de referência, normalmente a versão inicial ou final, e uma sequência de diferenças entre as demais versões sucessivas; já a estratégia *timestamp* armazena todas as versões e utiliza marcas de tempo para indicar as inclusões e exclusões.

3.2.1.4 Faceta de acesso

A faceta de acesso (Figura 20) classifica a maneira como os usuários podem consumir os dados de proveniência no contexto de experimentos científicos. A taxonomia destaca três mecanismos de acesso: através de uma interface *Query-By-Example* (QBE), através de uma API ou através de consultas.

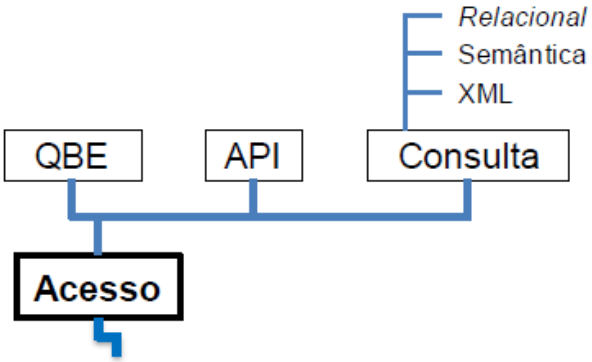

Figura 20. Faceta de acesso à proveniência. CRUZ (2011).

Os SGWfC possibilitam o acesso aos dados de proveniência de um experimento científico através de consultas submetidas por interfaces gráficas baseadas na especificação

QBE (RAMAKRISHNAN, GEHRKE, 2002), através de consultas submetidas por mecanismos desenvolvidos com as funcionalidades de uma API ou através de consultas submetidas diretamente por meio de uma linguagem de consulta. Com relação a este último modo de acesso, diversas linguagens de consulta, com diferentes níveis de expressividade e eficiência, vêm sendo utilizadas no contexto da e-Ciência. Estas linguagens de consultas podem ser subdivididas em três tipos, relacionados com a forma de armazenamento dos dados de proveniência: relacional, XML e Web Semântica. A linguagem SQL, as linguagens XPath/XQuery (CLARK, J., DEROSE, 1999; BOAG et al., 2010) e a linguagem SPARQL são exemplos de linguagens de consulta utilizadas para consultas relacionais, consultas baseadas em XML e para consultas semânticas, respectivamente.

No contexto de Linked Data, o acesso à proveniência pode ser obtido através de consultas SPARQL submetidas diretamente para um SPARQL *Endpoint* ou submetidas indiretamente por meio de uma API da Web Semântica como, por exemplo, a API Jena[23]. Adicionalmente, o uso de técnicas de visualização da informação pode oferecer uma contribuição importante à exploração conjunta entre os dados publicados como Linked Data e seus dados de proveniência.

3.2.1.5 Faceta de apoio semântico

A faceta de apoio semântico (Figura 21) classifica os sistemas de proveniência de acordo com os mecanismos semânticos utilizados para representar os dados de proveniência. As três subcategorias desta faceta, vocabulário, taxomomia e ontologia, estão relacionadas, respectivamente, com o uso de vocabulários controlados, taxonomias e ontologias de proveniência.

Embora muitas vezes considerados como conceitos equivalentes, vocabulário controlado, taxonomia e ontologia oferecem níveis crescentes de apoio semântico. Um vocabulário controlado consiste apenas de um conjunto de termos padrão para determinado domínio ou contexto; uma taxonomia consiste de termos organizados hierarquicamente através de relacionamentos de generalização e especialização e uma ontologia consiste de termos organizados através de uma gama maior de relacionamentos (generalização, especialização, parte-de e outros relacionamentos específicos de domínio) e da determinação de axiomas que definem os termos e trazem restrições de interpretação. As ontologias podem ainda ser

[23] http://jena.apache.org

classificadas como ontologias de domínio, que são direcionadas para um domínio genérico; como ontologias de tarefa, que são direcionadas para uma tarefa genérica realizada ou não em um mesmo domínio; ou ontologias de aplicação, que são direcionadas para uma aplicação específica.

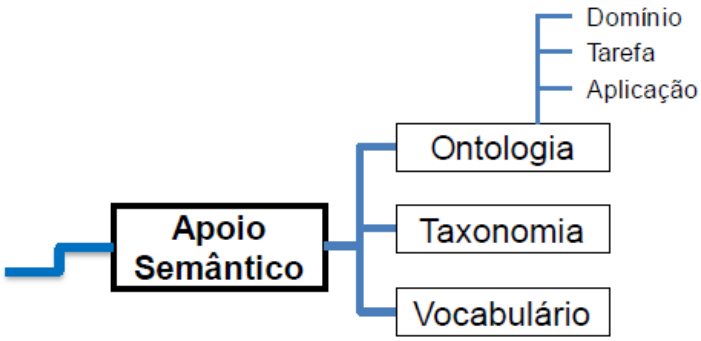

Figura 21. Faceta de apoio semântico para a proveniência. CRUZ (2011).

Esta é a faceta da taxonomia definida por CRUZ, CAMPOS, M. L. M. e MATTOSO (2009) com relacionamento mais estreito com a Web Semântica. Os principais modelos de dados, ontologias e vocabulários desenvolvidos para representar a proveniência no contexto de Linked Data serão abordados na subseção 3.3.

3.2.2 Taxonomia de proveniência para o contexto de Linked Data

A taxonomia proposta por este trabalho para classificar a proveniência no contexto de Linked Data, além de ser uma adaptação da taxonomia desenvolvida por CRUZ, CAMPOS, M. L. M. e MATTOSO (2009), contém uma nova faceta, denominada faceta de Linked Data. A faceta de Linked Data é baseada no trabalho de HARTIG e ZHAO (2010) e classifica a disponibilização dos dados de proveniência como Linked Data em três subcategorias: a partir de objetos Linked Data, a partir de *dumps* RDF ou a partir de SPARQL *Endpoints*. Apesar de não serem excludentes, estas subcategorias exigem o emprego de diferentes abordagens de publicação de proveniência.

A primeira subcategoria consiste em adicionar ao grafo RDF que descreve um objeto Linked Data, as triplas RDF dos dados de proveniência. Assim, ao recuperar o objeto Linked Data derreferenciando a URI HTTP que o identifica, o grafo RDF retornado conterá também os dados de proveniência.

A segunda subcategoria consiste em enriquecer os *dumps* RDF, que são documentos RDF que contêm o conteúdo completo de um *dataset* de Linked Data, com as triplas RDF dos

dados de proveniência. Normalmente, um *dump* RDF representa o *dataset* de Linked Data como um único grafo RDF, mas também é possível serializar um *dataset* de Linked Data como uma coleção de Grafos Nomeados (*Named Graphs*). Neste caso, cada um dos Grafos Nomeados pode conter os seus dados de proveniência ou então um Grafo Nomeado adicional pode ser criado com os dados de proveniência sobre os demais Grafos Nomeados da coleção.

Por fim, a terceira subcategoria consiste em utilizar um SPARQL *Endpoint* para obter os dados de proveniência armazenados no *dataset* de Linked Data acessível por ele. Além da possibilidade de utilização de uma consulta SPARQL definida explicitamente para obter os dados de proveniência, há também a possibilidade de estender o motor de consulta SPARQL para que os dados de proveniência possam ser adicionados automaticamente ao retorno de qualquer consulta submetida.

A Figura 22 ilustra as facetas da taxonomia proposta por este trabalho para classificar a proveniência no contexto de Linked Data. Esta taxonomia será utilizada para classificar a abordagem descrita no capítulo 4.

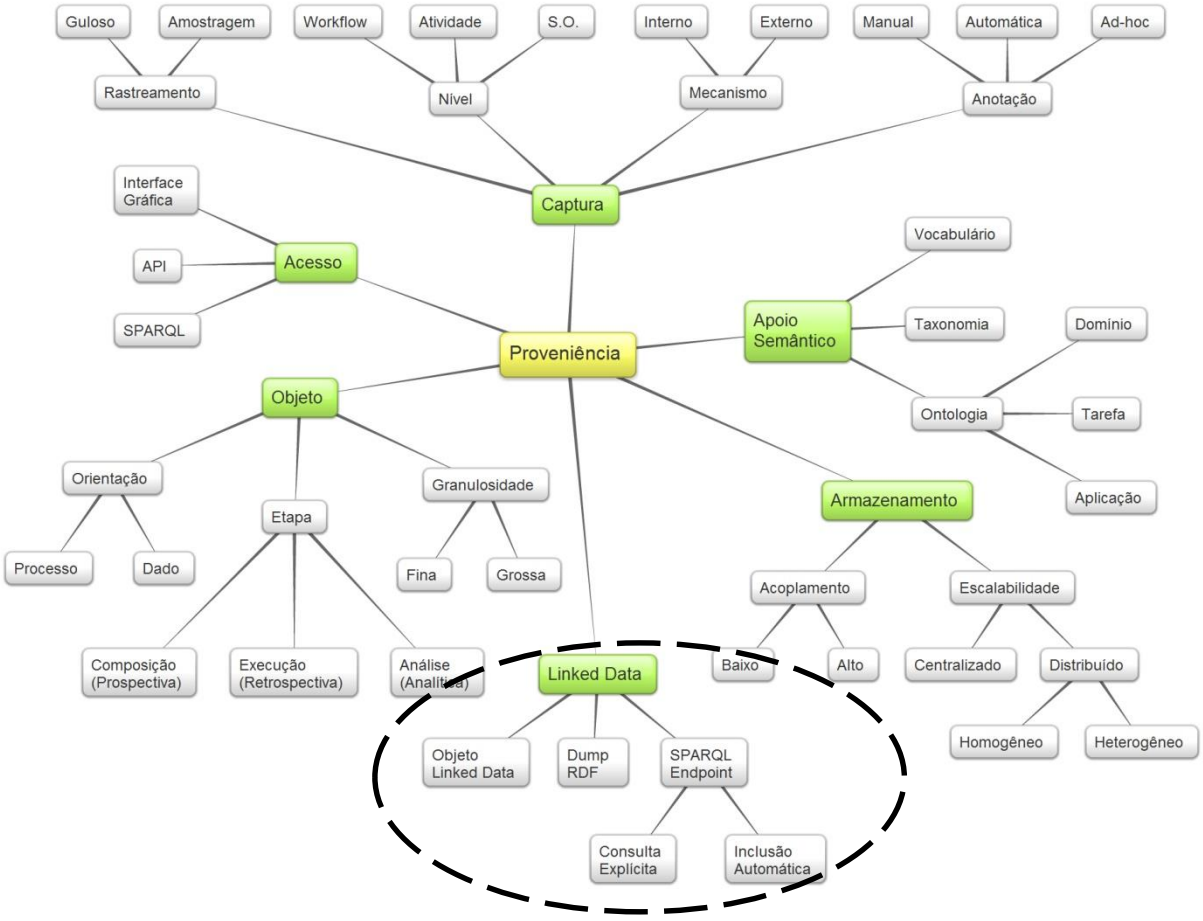

Figura 22. Taxonomia de proveniência para o contexto de Linked Data. Em destaque, a faceta de Linked Data, baseada no trabalho de HARTIG e ZHAO (2010).

3.3 Apoio semântico para proveniência

Para suprir a necessidade de representação das informações de proveniência na Web Semântica, muitos modelos de dados, ontologias e vocabulários de proveniência foram desenvolvidos nos últimos anos, para diversos domínios de aplicação (SAHOO, SHETH, 2009; HARTIG, ZHAO, 2010; MOREAU et al., 2011; MOREAU, MISSIER, 2013). Apesar de possuírem pontos em comum, estas estruturas semânticas de proveniência apresentam especificidades que refletem os requisitos da comunidade de usuários que as desenvolveram. Por exemplo, o modelo *Open Provenance Model* (OPM), base do vocabulário *Open Provenance Model Vocabulary* (OPMV), foi desenvolvido para representar de maneira genérica a proveniência de *workflows*, enquanto que o modelo Provenir foi desenvolvido como uma ontologia de topo para representar conceitos e relações gerais de proveniência em aplicações da Web Semântica no contexto da e-Ciência.

A reutilização de termos de vocabulários bem conhecidos e amplamente utilizados é considerada uma boa prática na Web Semântica (BIZER, HEATH, BERNERS-LEE, 2009). Novos termos só devem ser definidos se os termos necessários não puderem ser encontrados em vocabulários existentes. Desta maneira, para subsidiar a definição da melhor estratégia de representação da proveniência na Web Semântica, torna-se útil não só conhecer os modelos de dados, ontologias e vocabulários mais indicados na literatura especializada, como também saber as semelhanças e diferenças entre as estruturas semânticas disponibilizadas por eles. Assim, na próximas subseções, serão apresentados os principais modelos de dados, ontologias e vocabulários desenvolvidos para representar a proveniência no contexto de Linked Data.

Quadro 2. Vocabulários utilizados para descrever a proveniência no contexto de Linked Data.

Vocabulário	Prefixo	Namespace URI
Changeset	cs	http://purl.org/vocab/changeset/schema#
Cogs	cogs	http://vocab.deri.ie/cogs#
Dublin Core	dc	http://purl.org/dc/terms/
FOAF	foaf	http://xmlns.com/foaf/0.1/
OPMO	opmo	http://openprovenance.org/model/opmo#
OPMV	opmv	http://purl.org/net/opmv/ns#
OPMW	opmw	http://www.opmw.org/ontology/
PAV	pav	http://purl.org/pav/
PML-P	pmlp	http://inference-web.org/2.0/pml-provenance.owl#
PREMIS	premis	http://www.loc.gov/premis/rdf/v1#
Provenir	provenir	http://knoesis.wright.edu/provenir/provenir.owl#
Provenance Vocabulary	prv	http://purl.org/net/provenance/ns#
PROV-O	prov	http://www.w3.org/ns/prov#
VoID	void	http://rdfs.org/ns/void#
VoIDP	voidp	http://purl.org/void/provenance/ns/
WOT	wot	http://xmlns.com/wot/0.1/

3.3.1 OPM, OPMV, OPMO e OPMW

Open Provenance Model (OPM) é um modelo de dados de proveniência que foi desenvolvido para (i) permitir que informações de proveniência fossem compartilhadas entre sistemas, através de um modelo de dados comum; (ii) permitir que ferramentas fossem desenvolvidas, operando em tal modelo de proveniência; (iii) permitir que a proveniência fosse definida de maneira independente de uma tecnologia específica; (iv) permitir que a proveniência de um item fosse representada digitalmente, sendo este item produzido ou não por sistemas de computador; (v) permitir que múltiplos níveis de descrição de proveniência coexistissem e (vi) permitir que fossem identificadas as inferências válidas a partir das

informações de proveniência, por meio da definição de um conjunto básico de regras (MOREAU et al., 2011).

Para atender estes requisitos, OPM apresenta um modelo simples composto por um conjunto de dependências entre três entidades básicas: artefato, processo e agente. A entidade artefato representa um pedaço imutável do estado da materialização física de um objeto ou da representação digital do mesmo em um sistema de computador. A entidade processo representa a ação ou a série de ações executadas ou causadas pelos artefatos, resultando em novos artefatos. Por fim, a entidade agente representa o catalisador de um processo, habilitando, facilitando, controlando ou afetando a execução do processo. A Figura 23 ilustra os cinco tipos de dependências causais entre as entidades definidas pelo modelo OPM. Os nós representam um artefato, um processo ou um agente e as arestas, direcionadas do efeito para a causa, representam as dependências causais.

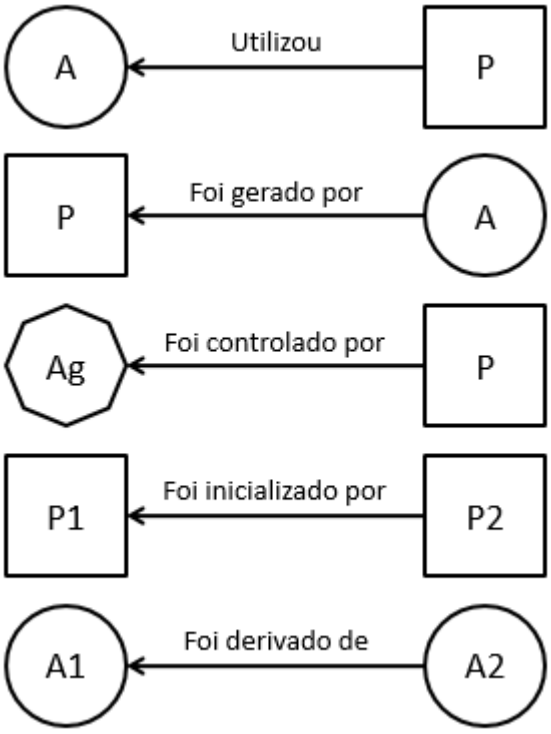

Figura 23. Os cinco tipos de dependências causais entre as entidades definidas pelo modelo OPM. O nó "A" representa um artefato, o nó "P" representa um processo e o nó "Ag" representa um agente. As arestas, direcionadas do efeito para a causa, representam as dependências causais. Adaptado de MOREAU et al. (2011).

Open Provenance Model Vocabulary (OPMV) (ZHAO, 2010) é um vocabulário leve de proveniência fundamentado no modelo OPM e definido como uma ontologia OWL-DL. Por estar fundamentado no modelo OPM, o OPMV visa auxiliar a interoperabilidade entre as informações de proveniência na Web Semântica. O vocabulário OPMV é particionado em uma

ontologia núcleo (Figura 24), que implementa somente as estruturas definidas pelo modelo OPM, e alguns módulos suplementares como OPMV Common, OPMV XSLT e OPMV SPARQL, que oferecem suclasses e subpropriedades especializadas para domínios específicos. Além disso, o OPMV pode ser utilizado em conjunto com outros vocabulários e ontologias de proveniência como Dublin Core, FOAF e Provenance Vocabulary.

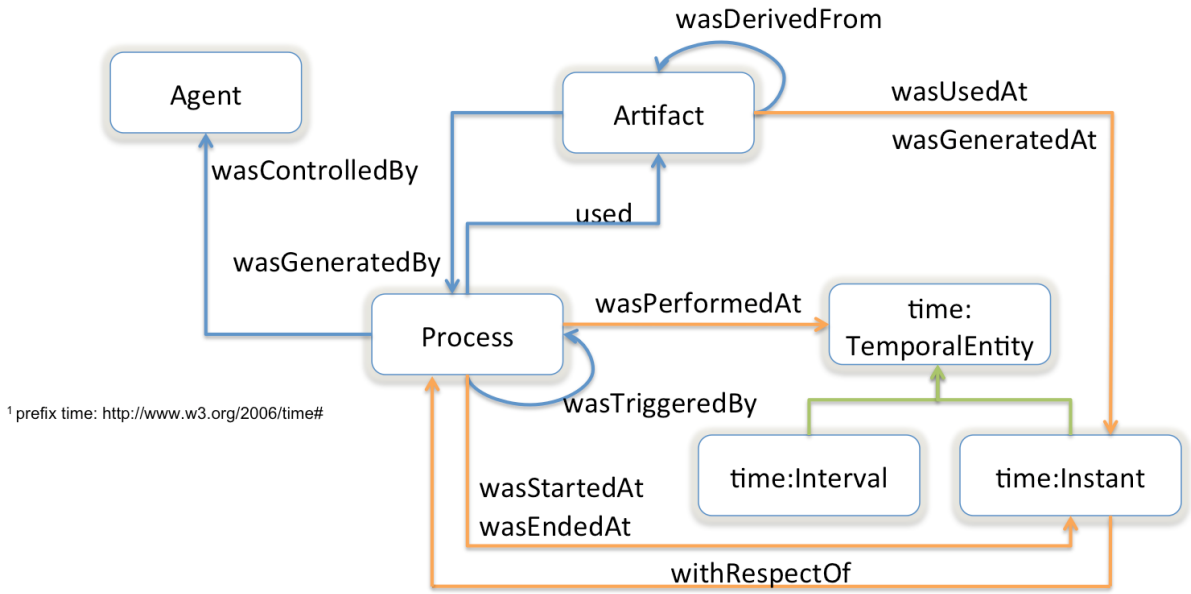

Figura 24. Ontologia núcleo implementada pelo OPMV. As arestas azuis representam propriedades que descrevem as dependências causais do modelo OPM, as arestas laranjas representam propriedades não definidas diretamente no modelo OPM e as arestas verdes representam relacionamentos de especialização. ZHAO (2010).

Open Provenance Model Ontology (OPMO) (MOREAU et al., 2010) estende o vocabulário OPMV com novas classes e propriedades que permitem a codificação dos conceitos do modelo OPM com maior expressividade, assim como a realização de inferências a partir das informações de proveniência descritas. OPMO e OPMV apresentam diferenças na forma como as informações de proveniência são codificadas. OPMV usa propriedades para codificar as arestas que representam as dependências causais do modelo OPM, enquanto que OPMO usa classes, pois não representa as dependências causais como propriedades binárias. Além disso, para definir uma visão particular à qual as informações de proveniência estão relacionadas, OPMV faz uso extensivo de Grafos Nomeados (*Named Graphs*) (CARROLL et al., 2005), uma técnica cuja semântica ainda não está totalmente padronizada. Para esta mesma finalidade, OPMO explicita uma classe (*opmo:Account*) e agrupa os membros de uma determinada visão de proveniência através da propriedade *opmo:account*. De maneira análoga, a ontologia OPMO também representa os papéis do modelo OPM através de classes explícitas, enquanto que o OPMV codifica os papéis implicitamente através dos nomes das propriedades. Em suma,

OPMV consiste de um vocabulário leve conveniente para codificar a maioria dos conceitos do modelo OPM e OPMO consiste de uma extensão útil para codificar o modelo OPM com maior expressividade e possibilitar a realização de inferências.

Para possibilitar não só a descrição da proveniência retrospectiva sobre a execução, mas também a descrição de dados de proveniência prospectiva sobre a composição abstrata de *workflows* científicos, GARIJO e GIL (2012) definiram uma extensão da ontologia OPMO, denominada *Open Provenance Model for Workflows* (OPMW). Além de manter a interoperabilidade dos dados de proveniência por meio do modelo OPM e possilitar a publicação da proveniência de um *workflow* de acordo com os princípios de Linked Data na Web, os termos da ontologia OPMW permitem descrever de maneira não ambígua os dados de proveniência retrospectiva e os dados de proveniência prospectiva de um *workflow*. Desde o estabelecimento da família de especificações do W3C para proveniência, denominada PROV, a ontologia OPMW também estende a ontologia PROV-O[24], que será abordada na próxima subseção. A Figura 25 ilustra um exemplo de como a OPMW representa a proveniência sobre a composição e sobre a execução de um *workflow*.

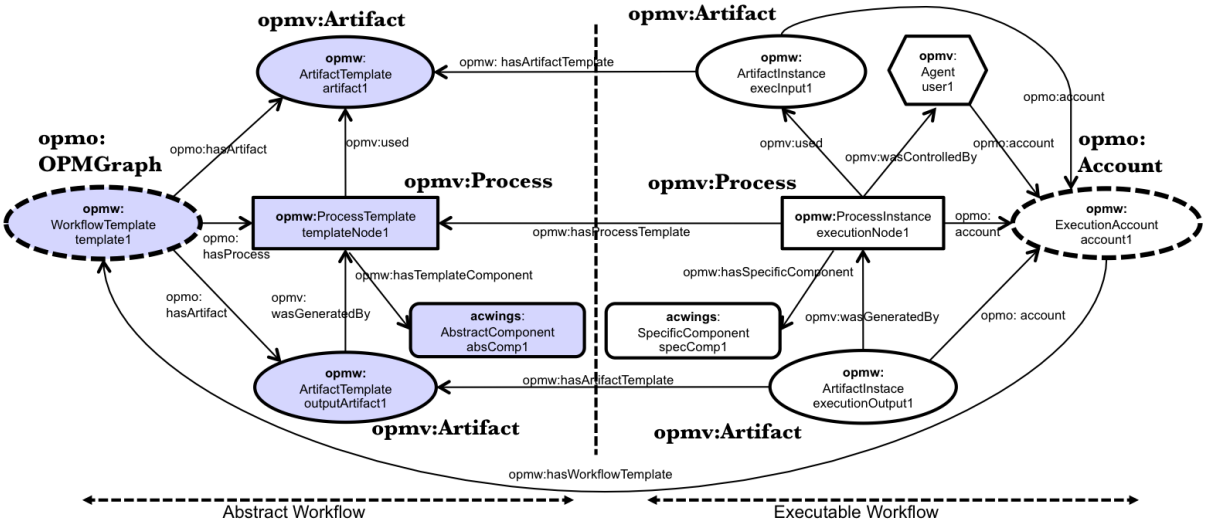

Figura 25. Exemplo de representação da proveniência sobre a composição e sobre a execução de um *workflow* através da ontologia OPMW. GARIJO e GIL (2012).

3.3.2 PROV-DM e PROV-O

PROV Data Model (PROV-DM) (MOREAU, MISSIER, 2013) é o modelo de dados que estabelece o fundamento da família de especificações do W3C, denominada PROV (Figura

[24] http://www.opmw.org/node/8

26), para representar e possibilitar a interoperabilidade da proveniência na Web e em outros sistemas de informação. O modelo PROV-DM, assim como a ontologia PROV-O que o codifica, foram estabelecidos como recomendações pelo W3C para representação da proveniência na Web.

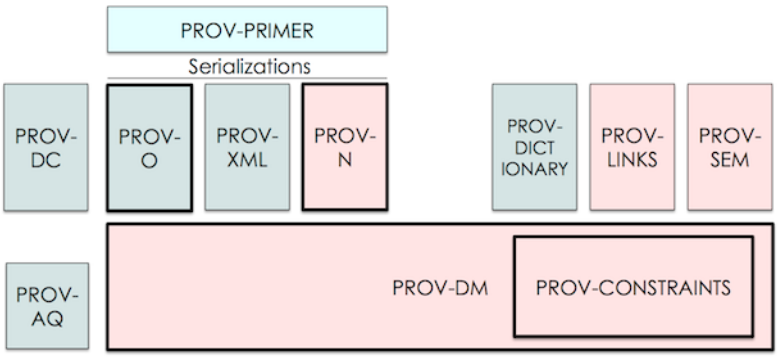

Figura 26. Família de especificações PROV do W3C. GROTH e MOREAU (2013).

No modelo PROV-DM, há a diferenciação entre estruturas centrais e estruturas estendidas. De forma análoga ao modelo OPM, as estruturas centrais do modelo PROV-DM modelam os conceitos essenciais para a representação da proveniência através de três tipos básicos: entidade, atividade e agente. Já as estruturas estendidas são direcionadas para apoiar a representação de conceitos mais avançados como proveniência de proveniência e coleções. Além disso, o modelo PROV-DM organiza seus tipos e relações em seis componentes, listados no Quadro 3.

Quadro 3. Os seis componentes do modelo PROV-DM e seus respectivos tipos e relações.

Componente	Conceito	Definição
Entidades / Atividades	Entidade	Física, digital, conceitual ou outro tipo de coisa com alguns aspectos fixos; entidades podem ser reais ou imaginárias.
	Atividade	Alguma coisa que ocorre durante um período de tempo e age em cima de ou com entidades.
	Geração	Conclusão da produção de uma entidade.
	Utilização	Início da utilização de uma entidade por uma atividade.
	Comunicação	Troca de uma entidade entre duas atividades.
	Início	Momento em que uma atividade é iniciada por uma entidade.

Componente	Conceito	Definição
	Fim	Momento em que uma atividade é finalizada por uma entidade.
	Invalidação	Início da destruição, término ou expiração de uma entidade por uma atividade.
Derivações	Derivação	Atualização de uma entidade, resultando em uma nova entidade ou a contrução de uma nova entidade baseada em uma entidade pré-existente.
	Revisão	Tipo de derivação em que a entidade resultante é uma revisão da entidade original.
	Repetição	Tipo de derivação em que a entidade resultante é uma repetição do todo ou de parte da entidade original.
	Fonte primária	Tipo de derivação que relaciona uma entidade resultante com suas fontes primárias.
Agentes / Responsabilidade / Influência	Agente	Algo que possui alguma forma de responsabilidade em uma atividade.
	Atribuição	Relação entre uma entidade e um agente, indicando que a entidade foi gerada por uma atividade não especificada que estava associada ao agente.
	Associação	Relação entre um agente e uma atividade, indicando que o agente possuía uma função na atividade.
	Delegação	Relação entre agentes, indicando que um agente age no lugar de outro agente responsável por uma atividade.
	Plano	Conjunto de ações ou passos destinados a um ou mais agentes para atingir metas.
	Pessoa	Especialização de agente para representar um agente humano.
	Organização	Especialização de agente para representar uma instituição social ou legal.
	Agente de software	Especialização de agente para representar um agente de *software*.

Componente	Conceito	Definição
	Influência	Capacidade de uma coisa (entidade, atividade ou agente) de ter um efeito sobre o caráter, desenvolvimento ou comportamento de outra coisa.
Pacote	Construtor do pacote	Especificação do nome e conteúdo de um conjunto de descrições de proveniência.
	Tipo do pacote	Conjunto nomeado de descrições de proveniência, que permite determinar a proveniência da proveniência.
Alternativa	Alternativa	Relação entre duas entidades que apresentam aspectos de uma mesma coisa.
	Especialização	Relação entre duas entidades, em que uma compartilha todos os aspectos de outra e adiciona outros aspectos específicos.
Coleções	Coleção	Entidade que fornece uma estrutura de agrupamento de componentes, em que cada componente, denominado membro, é uma entidade.
	Coleção vazia	Coleção sem membros.
	Membresia	Relação entre uma coleção e seus membros.

PROV Ontology (PROV-O) (LEBO, SAHOO, MCGUINNESS, 2013) é uma ontologia leve, ou seja, uma ontologia com pouca consideração sobre definições rigorosas de conceitos, que codifica as estruturas definidas pelo modelo de proveniência PROV-DM. Consequentemente, a ontologia PROV-O define classes e propriedades que podem ser usadas para representar as informações de proveniência diretamente ou, então, serem especializadas para representar informações de proveniência específicas a diversos de domínios de aplicação.

Os usuários da PROV-O podem empregar a ontologia completa ou somente determinadas partes, dependendo dos requisitos de proveniência e do nível de detalhes com que se deseja codificá-la. Assim, as classes e propriedades da PROV-O são agrupadas em três categorias, que possibilitam um nível de detalhamento incremental: categoria ponto de partida, categoria expandida e categoria qualificada. A categoria ponto de partida proporciona a base para os demais termos da PROV-O, através de classes e propriedades que codificam as

estruturas centrais do modelo PROV-DM (Figura 27). A categoria expandida proporciona termos adicionais para descrever, de maneira mais detalhada, a proveniência relacionada às entidades, atividades e agentes. Finalmente, a categoria qualificada é o resultado da aplicação do padrão de modelagem RDF denominado Relação Qualificada (DODDS, DAVIS, 2012) nas propriedades oferecidas pela categoria ponto de partida e pela categoria expandida.

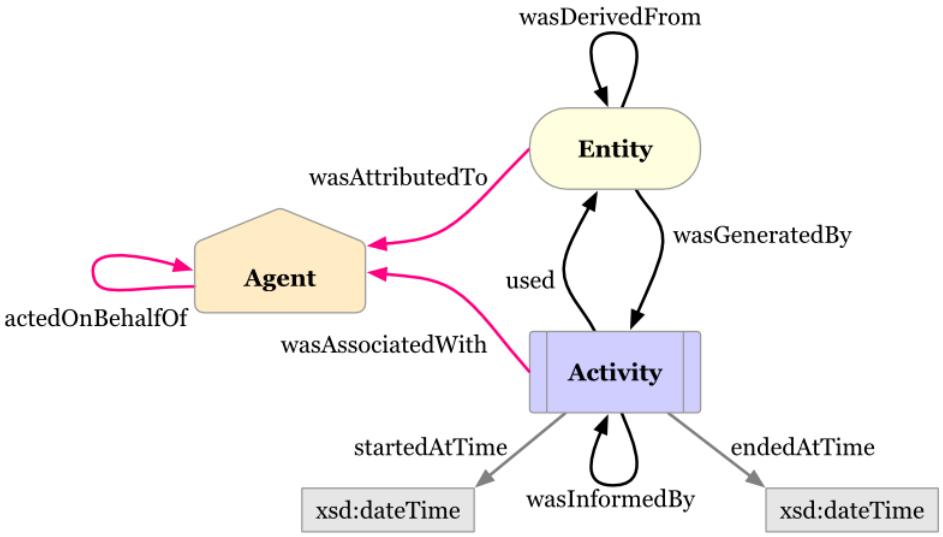

Figura 27. Classes e propriedades da categoria ponto de partida da ontologia PROV-O. LEBO, SAHOO, MCGUINNESS (2013).

3.3.3 Provenir

A ontologia Provenir (SAHOO et al., 2011) é uma ontologia de topo fundamentada em conceitos das ontologias *Suggested Upper Merged Ontology* (SUMO) (NILES, PEASE, 2001), *Basic Formal Ontology* (BFO) (GRENON, SMITH, B., GOLDBERG, 2004) e *Descriptive Ontology for Linguistic and Cognitive Engineering* (DOLCE) (GANGEMI et al., 2002), para representar a proveniência retrospectiva de *workflows* científicos, por meio de tecnologias da Web Semântica. A Provenir foi modelada com OWL-DL e pode ser estendida utilizando os termos do RDF-S como *rdfs:subClassOf* e *rdfs:subPropertyOf* para criação de novas ontologias de proveniência específicas a um domínio. Por estarem baseadas em um modelo consistente e no uso uniforme de termos, as novas ontologias estendidas da Provenir mantêm a capacidade dos sistemas de interoperarem os dados de proveniência.

De maneira análoga às ontologias OPMV e PROV-O, a ontologia Provenir é composta por três classes básicas (dado, processo e agente), no entanto, especializa a classe dado em cinco subclasses (data_collection, parameter, spatial_parameter, domain_parameter e temporal_parameter). As oito classes da ontologia Provenir são interligadas por dez

relacionamentos adaptados da ontologia *Relation Ontology* (RO) (SMITH, B. et al., 2005), que permitem ao Provenir capturar e representar explicitamente a semântica das conexões entre os termos. Esta semântica explícita das conexões pode ser usada por ferramentas de inferência para uma interpretação consistente da proveniência. O Quadro 4 exibe o mapeamento entre os termos do Provenir e as classes e as propriedades centrais dos vocabulários OPMV e PROV-O.

Quadro 4. Mapeamento entre as classes e propriedades centrais dos vocabulários OPMV, PROV-O e Provenir.

	OPMV	**PROV-O**	**Provenir**
Classes	opmv:Artifact	prov:Entity	provenir:data
	opmv:Process	prov:Activity	provenir:process
	opmv:Agent	prov:Agent	provenir:agent
Propriedades	opmv:used	prov:used	provenir:has_participant
	opmv:wasControlledBy	prov:wasAssociatedWith	provenir:has_agent
	opmv:wasDerivedFrom	prov:wasDerivedFrom	provenir:derives_from
	opmv:wasGeneratedBy	prov:wasGeneratedBy	provenir:has_participant
	opmv:wasTriggeredBy		provenir:preceded_by
		prov:wasInformedBy	provenir:preceded_by
		prov:actedOnBehalfOf	
		prov:wasAttributedTo	

3.3.4 Provenance Vocabulary

Provenance Vocabulary (HARTIG, ZHAO, 2012) é uma ontologia OWL-DL fundamentada em um modelo de dados proposto por HARTIG (2009) e direcionada para descrever dados de proveniência sobre a criação e também sobre o acesso aos dados publicados na Web segundo os princípios de Linked Data. Ela é dividida em uma ontologia núcleo simples e três módulos suplementares (Tipos, Arquivos e Verificação de Integridade), que apresentam conceitos utilizados com menor frequência e uma série de especializações para os conceitos da ontologia núcleo. Assim como a ontologia Provenir, a Provenance Vocabulary apresenta similaridades com a OPMV e a PROV-O (ZHAO, HARTIG, 2012).

Na última revisão da ontologia Provenance Vocabulary, ela foi posicionada como uma especialização da ontologia PROV-O (Figura 28). Deste modo, os dados de proveniência descritos com os termos da Provenance Vocabulary herdam a capacidade de serem interpretados e interoperados na Web de acordo com a família de especificações PROV do W3C.

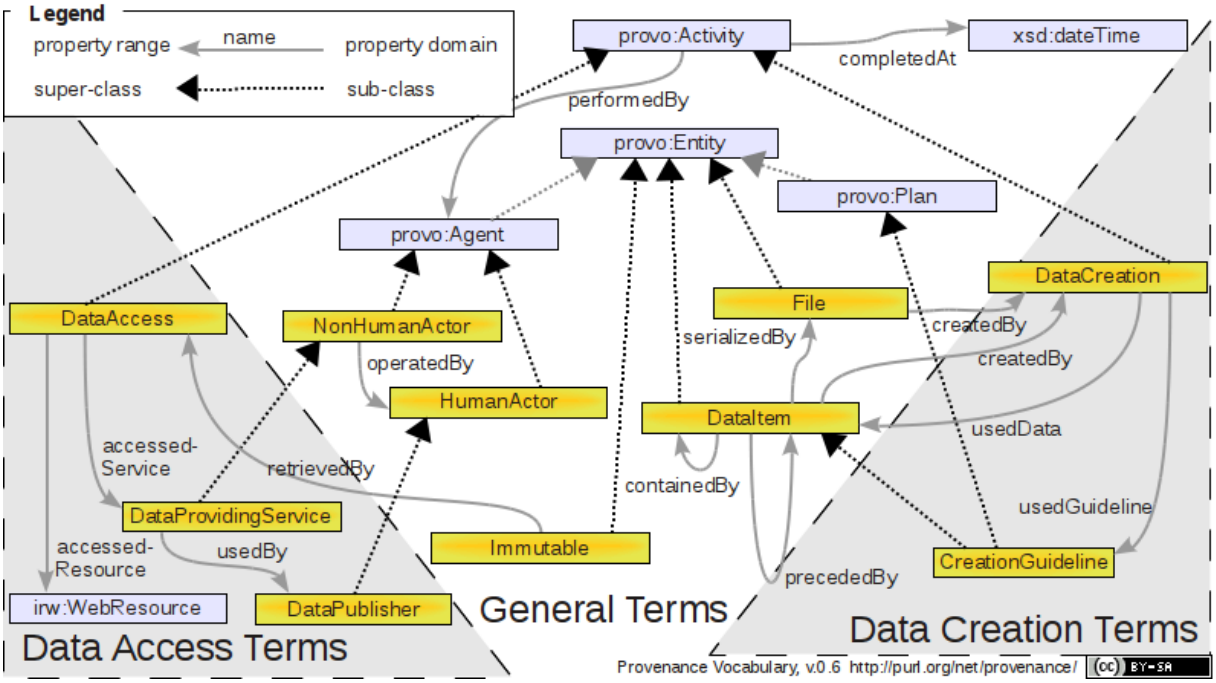

Figura 28. Classes e propriedades da ontologia núcleo da Provenance Vocabulary e os relacionamentos de especialização entre as classes da Provenance Vocabulary e as classes da PROV-O. HARTIG e ZHAO (2012).

3.3.5 Cogs

Cogs (FREITAS, KÄMPGEN, CURRY et al., 2012) é um vocabulário que estende a semântica de *workflows* proporcionada pela OPMV, por meio de uma taxonomia de 150 classes aproximadamente, para possibilitar a descrição detalhada da proveniência de um processo ETL. Por apresentar esta característica, Cogs foi considerado como mais um módulo suplementar da ontologia OPMV, assim como OPMV Common, OPMV XSLT e OPMV SPARQL (ZHAO, 2010).

O objetivo do vocabulário Cogs é possibilitar uma representação expressiva da proveniência da transformação dos dados em um *workflow* ETL e, ao mesmo tempo, manter a interoperabilidade semântica dos descritores de proveniência, fundamentando-se nos esforços de padronização do modelo OPM e, mais recentemente, do modelo PROV-DM. A taxonomia de classes do Cogs é estruturada a partir dos seguintes conceitos:

- Execução (*cogs:Execution*): Representa a execução da instância de um *workflow* ETL.
- Estado: Representa a observação de um indicator da execução de um *workflow* ETL. Pode variar de estados de execução, capturadas por subclasses da classe *cogs:ExecutionStatus*, a estatísticas de execução, capturadas por subclasses da classe *cogs:PerformanceIndicator*.
- Extração (*cogs:Extraction*): Representa operações da primeira fase do processo ETL, que envolve extração de dados de diferentes tipos de fontes. A classe *cogs:Extraction* é uma especialização da classe *opmv:Process*.
- Transformação (*cogs:Transformation*): Representa operações da fase de transformação do processo ETL. Tipicamente, esta é a fase que abrange a maior parte da semântica do *workflow* ETL, por isso, a classe *cogs:Transformation* possui um grande número de subclasses. Exemplos destas subclasses são *cogs:RegexFilter*, *cogs:MergeRow*, *cogs:DeleteColumn*, *cogs:SplitColumn*, *cogs:Trim* e *cogs:Round*. A classe *cogs:Transformation* é uma especialização da classe *opmv:Process*.
- Carga (*cogs:Loading*): Representa operações da última fase do processo ETL, quando os dados são carregados no destino final. Exemplos de subclasses de *cogs:Loading* são *cogs:ConstructiveMerge* e *cogs:IncrementalLoad*. A classe *cogs:Loading* é uma especialização da classe *opmv:Process*.
- Objeto (*cogs:Object*): Representa as fontes e os resultados das operações do *workflow* ETL. As subclasses de *cogs:Object*, como *cogs:ObjectReference*, *cogs:Cube* ou *cogs:File*, visam oferecer uma definição mais precisa do conceito artefato do modelo OPM e, junto com os tipos de operações que os geram e os consomem, capturam a semântica dos passos do *workflow*. A classe *cogs:Object* é uma especialização da classe *opmv:Artifact*.
- Camada (*cogs:Layer*): Representa as diferentes camadas nas quais os dados residem durante o processo ETL. *cogs:PresentationArea* e *cogs:StagingArea* são algumas das subclasses de *cogs:Layer*.

3.3.6 Dublin Core e FOAF

Dublin Core é um vocabulário criado em 1995 e mantido pelo *Dublin Core Metadata Initiative* (DCMI)[25], um fórum aberto que se dedica ao desenvolvimento de normas de interoperabilidade de metadados online para suportar uma ampla gama de propósitos e modelos de negócios. O conjunto de 15 propriedades originais do vocabulário Dublin Core, denominado *Dublin Core Metadata Element Set* (*DC Elements*)[26], foram posteriormente refinados e novos termos foram incluídos. Este vocabulário expandido é denominado como *DCMI Metadata Terms* (*DC Terms*)[27] e consiste atualmente de 55 propriedades.

Os termos do vocabulário Dublin Core possibilitam a descrição de uma série de informações de proveniência sobre os recursos do contexto de Linked Data: há propriedades que informam quando um recurso foi afetado, quem afetou e como ele foi afetado. As demais propriedades do vocabulário Dublin Core registram metadados de descrição sobre o que foi afetado. Sendo assim, o Quadro 5 propõe uma classificação das propriedades do vocabulário Dublin Core, com exceção de *dc:provenance*, em quatro categorias, que correspondem aos tipos de questões que a propriedade responde: o quê, quem, quando e como. A propriedade *dc:provenance*, definida como uma "sentença de quaisquer mudanças de propriedade ou custódia dos recursos desde a sua criação, que são significativos para sua autenticidade, integridade e interpretação"[28], não foi classificada nas categorias, pois representa um termo muito genérico e pode ser utilizado para ligar qualquer informação de proveniência a um recurso.

[25] http://dublincore.org

[26] http://dublincore.org/documents/dces

[27] http://dublincore.org/documents/dcmi-terms

[28] http://dublincore.org/documents/dcmi-terms/#terms-provenance

Quadro 5. Categorização das propriedades do vocabulário Dublin Core, de acordo com questões relacionadas à proveniência de recursos do contexto de Linked Data. Adaptado de GARIJO e ECKERT (2013).

Tipo de metadado	Questão	Propriedades do Dublin Core
Descrição	O quê?	abstract, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, bibliographicCitation, conformsTo, coverage, description, educationLevel, extent, hasPart, isPartOf, format, identifier, instructionalMethod, isRequiredBy, language, mediator, medium, relation, requires, spatial, subject, tableOfContents, temporal, title, type
Proveniência	Quem?	contributor, creator, publisher, rightsHolder
Proveniência	Quando?	available, created, date, dateAccepted, dateCopyrighted, dateSubmitted, issued, modified, valid
Proveniência	Como?	accessRights, isVersionOf, hasVersion, isFormatOf, hasFormat, license, references, isReferencedBy, replaces, isReplacedBy, rights, source

Friend of a Friend (FOAF) (BRICKLEY, DANIEL, MILLER, 2010) é um vocabulário que oferece classes e propriedades para descrever e interligar pessoas e informações na Web. Os termos do vocabulário FOAF podem ser utilizados em conjunto com outras ontologias de proveniência para descrever informações sobre as entidades ou agentes como, por exemplo, nomes, membros de grupo, endereço de email e identificação de contas online. Além disso, a propriedade *foaf:maker* e sua propriedade inversa *foaf:made* podem ser utilizadas para relacionar as entidades com os recursos produzidos por elas. Assim, as propriedades *foaf:maker* e *foaf:made* podem ser usadas para identificar o criador de um item de dados.

3.3.7 Changeset, PAV, PML-P, PREMIS, VoID, VoIDP e WOT

Além dos modelos de dados, ontologias e vocabulários de proveniência abordados nas subseções anteriores, o *dataset Linked Open Vocabularies* (LOV)[29] indica outros vocabulários e ontologias que são utilizados na núvem de LOD para representar informações de proveniência. Esta subseção irá descrever brevemente estes vocabulários e ontologias de

[29] http://lov.okfn.org/dataset/lov

proveniência indicados no LOV, concluindo assim a apresentação dos mecanismos de apoio semântico à proveniência utilizados no contexto de Linked Data.

Changeset Vocabulary (TUNNICLIFFE, DAVIS, 2005) descreve alterações em recursos descritos em triplas RDF. A alteração da descrição de um recurso é representado pela entidade *cs:ChangeSet* que encapsula as diferenças entre duas versões da descrição. Estas diferenças são representadas por inclusões e exclusões de triplas RDF.

Provenance, Authoring and Versioning (PAV) (CICCARESE, SOILAND-REYES, 2013) é uma ontologia leve, que especializa a ontologia PROV-O com um conjunto de propriedades para descrição da autoria, versionamento e proveniência de recursos publicados na Web. PAV não define classes explícitas nem domínio/escopo das propriedades, pois cada propriedade se destina a ser usada diretamente no recurso online descrito.

PML-P é um dos três componentes da *Proof Markup Language* (PML) versão 2 (MCGUINNESS et al., 2007), ontologia desenvolvida para representar e compartilhar conhecimento sobre como as informações publicadas na Web foram definidas ou derivadas por agentes de inteligência artificial, a partir de outras fontes de informação. PML-P oferece termos que possibilitam a descrição de metadados de proveniência de entidades do mundo real (*pmlp:IdentifiedThing*) como, por exemplo, informações (*pmlp:Information*), linguagens (*pmlp:Language*) e fontes (*pmlp:Source*).

Preservation Metadata Implementation Strategies (PREMIS) (GUENTHER et al., 2012) é um dicionário de dados para apoiar a preservação a longo prazo de documentos em bibliotecas. PREMIS define um conjunto básico de unidades semânticas que devem ser registrados sobre materiais físicos, como livros e partituras, ou objetos digitais, a fim de suportar as funções de preservação dos mesmos.

Vocabulary of Interlinked Datasets (VoID) (ALEXANDER et al., 2009) é um vocabulário desenvolvido para descrever metadados sobre datasets RDF, que são coleções de dados estruturados como RDF e publicados e mantidos por um mesmo provedor de dados. Os metadados descritos pelo VoID podem ser metadados gerais, descritos em conjunto com Dublin Core; metadados de acesso; metadados estruturais e metadados sobre interligações entre datasets. O objetivo do vocabulário VoID é ser uma ponte entre os publicadores e os usuários de um *dataset* RDF, apoiando desde a descoberta dos dados até a catalogação e arquivamento de *datasets*. VoIDP (OMITOLA et al., 2011) é uma extensão do vocabulário VoID, com reutilização de elementos das ontologias Provenance Vocabulary, Time Ontology (HOBBS,

PAN, F., 2006) e *Semantic Web Publishing* (SWP) (BIZER, 2006), para permitir a especificação de metadados de proveniência sobre os datasets RDF.

Web of Trust Schema (WOT) (BRICKLEY, DANIEL, 2004) é uma ontologia desenvolvida para representar o relacionamento entre documentos RDF, chaves de criptografia, como *Pretty Good Privacy* (PGP) ou *GNU Privacy Guard* (GPG), e usuários; oferecendo informações para a avaliação da confiabilidade dos itens de dados publicados na Web. Entre as informações disponibilizadas pela WOT, encontram-se informações de proveniência como quando e por qual chave de criptografia, um item de dados foi assinado.

3.4 Utilização de *workflow* para coletar e publicar proveniência no contexto de Linked Data

Como visto na seção 2.4, a utilização de um *workflow* cuidadosamente planejado para gerenciar o processo de publicação de Linked Data é uma importante estratégia de governança de dados, especialmente dentro dos limites de uma organização. Entre os benefícios desta estratégia, encontram-se a possibilidade de coletar sistematicamente dados de proveniência e publicá-los também como Linked Data. Neste sentido, encontram-se alguns trabalhos relacionados.

O projeto LinkedDataBR (CAMPOS, M. L. M., GUIZZARDI, 2010) propôs uma abordagem inovadora para apoiar a exposição, o compartilhamento e a associação de dados abertos governamentais na forma de Linked Data. Este projeto desenvolveu uma arquitetura interativa e de usabilidade amigável para estimular a publicação de dados heterogêneos como Linked Data, associando-os com outros dados governamentais existentes e realizando a coleta e publicação de dados de proveniência durante o processo de transformação dos dados. A ferramenta de *workflow* ETL Pentaho Data Integration, também conhecida como Kettle, foi utilizada para apoiar o processo de publicação de LOD, assim como a coleta e publicação dos dados de proveniência. A escolha do Kettle foi motivada pelo fato da ferramenta atender uma série de critérios como ser *open source*, possuir uma ampla comunidade de usuários, ter uma forte orientação a metadados e ser extensível através de API. A Figura 29 ilustra a arquitetura desenvolvida pelo projeto LinkedDataBR.

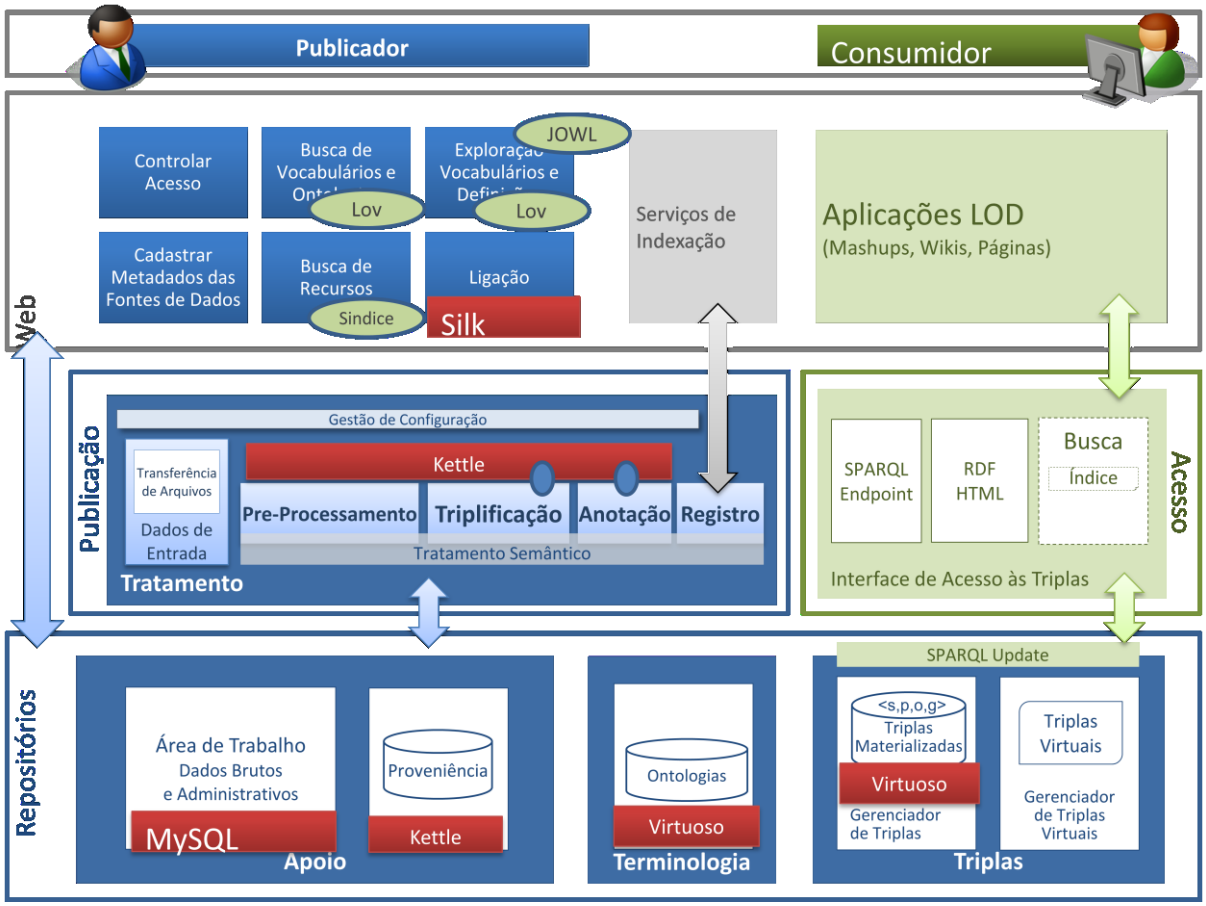

Figura 29. Arquitetura desenvolvida pelo projeto LinkedDataBR para apoiar a publicação e o consumo de dados abertos governamentais na forma de Linked Data. CAMPOS, M. L. M. e GUIZZARDI (2010)

Posteriormente, FREITAS, KÄMPGEN, OLIVEIRA et al. (2012) e OMITOLA et al. (2012) investigaram perspectivas complementares da abordagem proposta pelo projeto LinkedDataBR. FREITAS, KÄMPGEN, OLIVEIRA et al. (2012) definiram um modelo de três camadas para representação da proveniência do processo de publicação de Linked Data, com objetivo de proporcionar uma representação rica dos conceitos do *workflow* ETL e, ao mesmo tempo, manter a interoperabilidade dos dados de proveniência. Assim, a ontologia OPMV é utilizada na primeira camada para representar a semântica básica do *workflow* e possibilitar a interoperabilidade da proveniência; Cogs é utilizado na segunda camada para representar de maneira expressiva os conceitos do processo de ETL e uma ontologia complementar específica ao domínio da aplicação é utilizada na terceira camada para estender a representação da proveniência realizada nas camadas anteriores.

OMITOLA et al. (2012), depois de formalizar e enfatizar o potencial de ferramentas interativas de transformação de dados (IDT - *Interactive Data Transformation*) para apoiar as tarefas de transformação de dados no contexto de LOD, definiram uma arquitetura de gerenciamento dos dados de proveniência baseada no modelo de representação definido por

FREITAS, KÄMPGEN, OLIVEIRA et al. (2012). A arquitetura organizava o gerenciamento da proveniência prospectiva e retrospectiva em três camadas. Na primeira camada, denominada Camada de Captura de Eventos de Proveniência, a proveniência era capturada por eventos disparados durante o processo de transformação de dados executado pela ferramenta IDT. Então, os dados de proveniência capturados eram triplificados na segunda camada, denominada Camada de Representação da Proveniência, e armazenados em uma base de conhecimento na terceira camada, denominada Camada de Armazenamento de Proveniência. Google Refine[30], uma aplicação IDT fornecida pelo Google para limpeza e transformação de dados, foi utilizada para implementar e validar a arquitetura proposta.

A abordagem proposta neste trabalho e descrita no próximo capítulo também utiliza *workflows* ETL para orquestrar o processo de publicação de Linked Data, com publicação dos dados de proveniência. Para apoiar semanticamente a representação da proveniência, uma estratégia em camadas similiar à utilizada por FREITAS, KÄMPGEN, OLIVEIRA et al. (2012) foi adotada pela abordagem.

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

4 *ETL4LinkedProv* - Abordagem de coleta e publicação de proveniência como Linked Data dentro das organizações

4.1 Introdução

Conforme descrito no capítulo Web Semântica e Linked Data, a publicação de Linked Data realizada dentro dos limites de uma organização pode ser monitorada e gerenciada através de dois processos complementares, denominados neste trabalho de Processo de Preparação e Transformação dos Dados e Processo de Interligação dos Dados. Em ambos os processos, a descrição das etapas, os agentes e entidades envolvidos, os parâmetros e variáveis utilizados e os resultados obtidos configuram-se como dados de proveniência relevantes para a avaliação da qualidade e da confiabilidade dos dados publicados. A abordagem proposta por este trabalho, denominada *ETL4LinkedProv*, focaliza-se nos dados de proveniência produzidos pelo Processo de Preparação e Transformação dos Dados e faz parte de um trabalho maior realizado em conjunto com DE LA CERDA e CAVALCANTI (2012), cuja abordagem abrange os dados de proveniência do Processo de Interligação dos Dados.

A abordagem *ETL4LinkedProv* baseia-se na experiência de trabalhos anteriores (CAMPOS, M. L. M., GUIZZARDI, 2010; FREITAS, KÄMPGEN, OLIVEIRA et al., 2012; OMITOLA et al., 2012) que utilizaram *workflows* para publicar dados de fontes heterogêneas no contexto de Linked Data, com captura e publicação dos dados de proveniência através de triplas RDF apoiadas semanticamente por ontologias bem estabelecidas. Na abordagem *ETL4LinkedProv*, um novo componente, denominado Agente Coletor de Proveniência, é introduzido com a função de capturar, interligar e armazenar temporariamente os dados de proveniência prospectiva e retrospectiva, para posterior publicação dos mesmos como triplas RDF.

A Figura 30 ilustra a arquitetura da abordagem *ETL4LinkedProv*, que promove a publicação de Linked Data através do Processo de Preparação e Transformação dos Dados, com a atuação do Agente Coletor de Proveniência para subsidiar a captura e a publicação dos dados de proveniência. A aresta tracejada que liga o *dataset* de proveniência ao *dataset* de domínio,

ambos publicados como Linked Data, indica que os dados de proveniência encontram-se interligados aos seus respectivos dados de domínio.

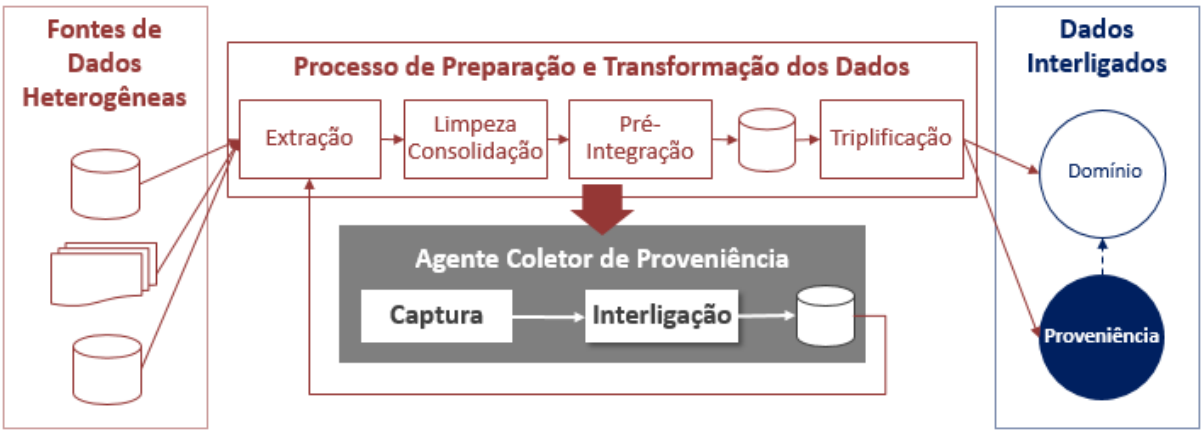

Figura 30. Arquitetura da abordagem *ETL4LinkedProv*.

4.2 Classificação da abordagem *ETL4LinkedProv* segundo a taxonomia de proveniência para o contexto de Linked Data

Para obter uma melhor compreensão da abordagem *ETL4LinkedProv*, proposta e implementada neste trabalho, torna-se útil classificá-la de acordo com as facetas da taxonomia de proveniência para o contexto de Linked Data, definida na seção 3.2.2. A Figura 31 destaca os itens da taxonomia relacionados com a abordagem.

Com relação à captura, a abordagem *ETL4LinkedProv* realiza um rastreamento guloso, ou seja, os dados de proveniência são coletados de maneira intensa, à medida que os eventos monitorados ocorrem durante o processo de publicação. Além disso, a captura é realizada no nível de *workflow*, com utilização das estruturas internas da ferramenta de gerenciamento do *workflow* para efetuar a anotação automática da proveniência.

Com relação à faceta de objeto, tanto a proveniência sobre a composição do *workflow* (proveniência prospectiva), quanto a proveniência sobre a execução do *workflow* (proveniência retrospectiva) são coletadas e a proveniência coletada é relacionada tanto ao dado publicado, quanto ao processo de publicação. A abordagem *ETL4LinkedProv* promove ainda uma coleta de granulosidade fina, a fim de permitir melhores condições de exploração conjunta entre os dados publicados e seus dados de proveniência.

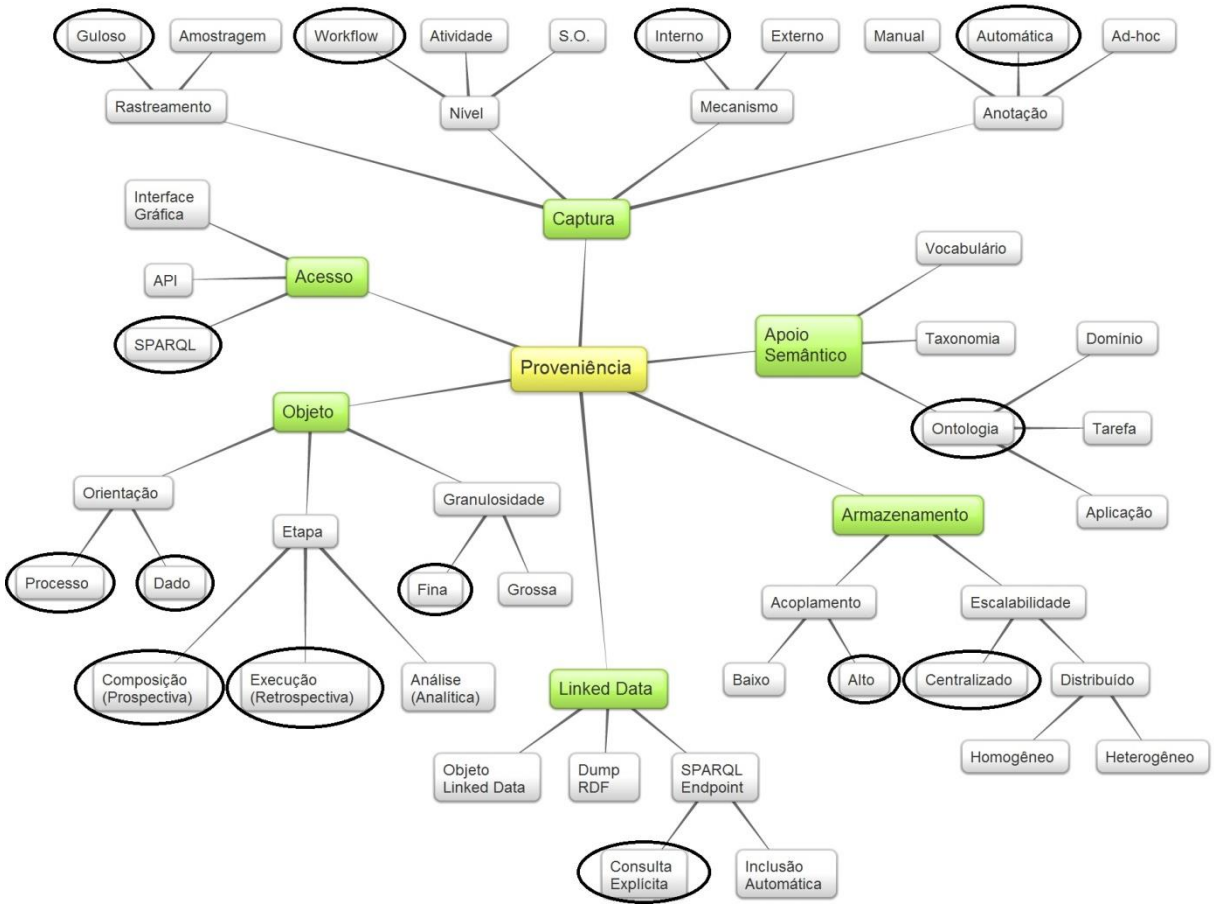

Figura 31. Classificação da abordagem *ETL4LinkedProv* segundo a taxonomia de proveniência para o contexto de Linked Data.

Com relação ao armazenamento dos dados de proveniência, apesar da abordagem *ETL4LinkedProv* possibilitar um armazenamento distribuído, a implementação descrita na seção 4.5 utilizou o modo de armazenamento centralizado, devido à simplicidade de gerenciamento, manutenibilidade e controle. Além disso, o acoplamento é alto, ou seja, os dados de proveniência são armazenados no mesmo repositório e associados com os dados aos quais eles se referem.

Com relação ao mecanismo de acesso, a implementação aplicada no cenário apresentado no capítulo 5 utilizou consultas SPARQL para acessar diretamente os dados de proveniência. No entanto, a abordagem *ETL4LinkedProv* não impede a utilização de uma interface gráfica ou a utilização de uma API da Web Semântica como meio para acessar os dados de proveniência publicados.

Com relação ao apoio semântico, a abordagem *ETL4LinkedProv* utiliza uma estratégia de quatro camadas, descrita na seção 4.4, com emprego de ontologias que apresentam níveis crescentes de detalhamento sobre a proveniência do processo de publicação.

Por fim, com relação à faceta de Linked Data, a abordagem *ETL4LinkedProv* disponibiliza os dados de proveniência por meio de um SPARQL *Endpoint* padrão. Assim, os dados de proveniência são obtidos a partir de consultas SPARQL definidas explicitamente e submetidas a um SPARQL *Endpoint*.

4.3 Agente Coletor de Proveniência

O Agente Coletor de Proveniência é um componente de software que encapsula o *workflow* ETL responsável pela publicação de Linked Data, a fim de capturar e interligar os dados de proveniência prospectiva e os dados de proveniência retrospectiva durante a execução do processo de publicação. A execução do *workflow* ETL consiste na execução de um conjunto de passos interligados, por onde os dados fluem de maneira unidirecional. Cada passo corresponde ao conceito Atividade da ontologia PROV-O e é responsável pela execução de um processo de extração, transformação ou carga de dados; pela execução de atividades auxiliares como transferir arquivos e enviar *emails* ou pela execução de um *sub-workflow*. Conforme ilustrado na Figura 30, as tarefas realizadas pelo Agente Coletor de Proveniência podem ser agrupadas em 3 etapas distintas: captura, interligação e armazenamento temporário dos dados de proveniência.

Na etapa de captura, o Agente Coletor de Proveniência monitora um conjunto de eventos relacionados ao *workflow* ETL encapsulado e, sempre que um dos eventos ocorre, realiza a coleta dos dados de proveniência. Os eventos monitorados pelo Agente Coletor de Proveniência são (i) o início e o término do *workflow* ETL principal ou de um dos seus *sub-workflows*; (ii) o início e o término de cada passo executado e (iii) a leitura de linhas de dados por determinado passo. Um subconjunto de tipos de passos pode ser definido, a fim de que somente os passos relacionados aos tipos definidos tenham a proveniência com nível de granulosidade fina capturada pelo Agente Coletor de Proveniência. Esta seleção de tipos de passos visa minimizar os impactos do grande volume de dados gerados pela estratégia de coleta da proveniência com granulosidade fina, sem, no entanto, negligenciar a coleta dos dados de proveniência das atividades mais relevantes do processo de publicação de Linked Data.

Na etapa de interligação, o Agente Coletor de Proveniência interliga os dados de proveniência coletados na etapa de captura. Os dados de proveniência interligados referem-se tanto ao processo de publicação, quanto aos dados de domínio publicados. Além disso, nesta

etapa, os dados de proveniência retrospectiva são interligados aos seus respectivos dados de proveniência prospectiva.

Por fim, na etapa de armazenamento temporário, os dados de proveniência são armazenados em um repositório, para, posteriormente, também serem extraídos, processados e publicados como triplas RDF pelo Processo de Preparação e Transformação dos Dados. A Figura 32 ilustra o modelo conceitual de dados do repositório temporário utilizado pelo Agente Coletor de Proveniência. O modelo conceitual é formado por 13 entidades, que separam e relacionam os dados sobre a composição e os dados sobre a execução do Processo de Preparação e Transformação dos Dados.

Com relação à composição do Processo de Preparação e Transformação dos Dados, tem-se:

(i) a composição de um *workflow* ETL está armazenada em um repositório, podendo ou não ser *sub-workflow* de outro *workflow* ETL e está relacionada a um *workflow* ETL raiz, que representa o *workflow* ETL principal;
(ii) a composição de um *workflow* ETL é criada ou modificada por um usuário, pode ser documentada através de anotações e é formada por um conjunto de passos interligados por ligações unidirecionais, por onde os campos de dados serão fluídos;
(iii) a composição de um passo utiliza um conjunto de parâmetros e pode ser origem ou destino de uma ou mais ligações unidirecionais.

Com relação à execução do Processo de Preparação e Transformação dos Dados, tem-se:

(i) a composição de um *workflow* ETL pode ser executada várias vezes;
(ii) a execução de um *workflow* ETL consiste na execução dos seus respectivos passos, que escrevem e lêem linhas de dados. Estas linhas de dados são compostas por um conjunto de campos, que fluem os dados através das ligações unidirecionais entre os passos;
(iii) na execução dos passos do *workflow* ETL, são utilizados valores específicos para os parâmetros definidos na composição do passo.

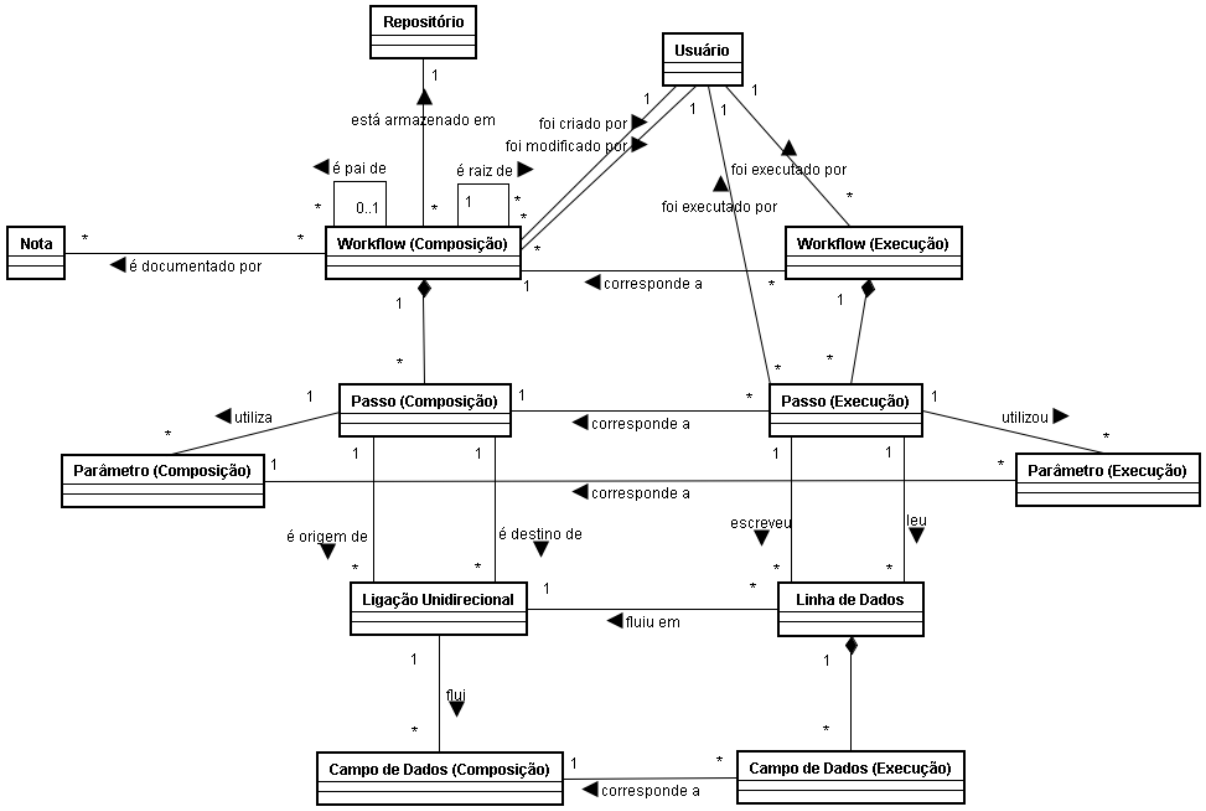

Figura 32. Modelo conceitual do repositório temporário utilizado pelo Agente Coletor de Proveniência para armazenar os dados de proveniência do Processo de Preparação e Transformação dos Dados.

A implementação do repositório temporário como um banco de dados relacional resultou na criação de 14 tabelas para separar e relacionar os dados de proveniência prospectiva e os dados de proveniência retrospectiva. A Figura 33 apresenta o modelo lógico de dados do repositório temporário no Sistema de Gerenciamento de Banco de Dados (SGBD) MySQL[31] e o Quadro 6 lista os respectivos tipos de dados armazenados em cada tabela.

[31] http://www.mysql.com

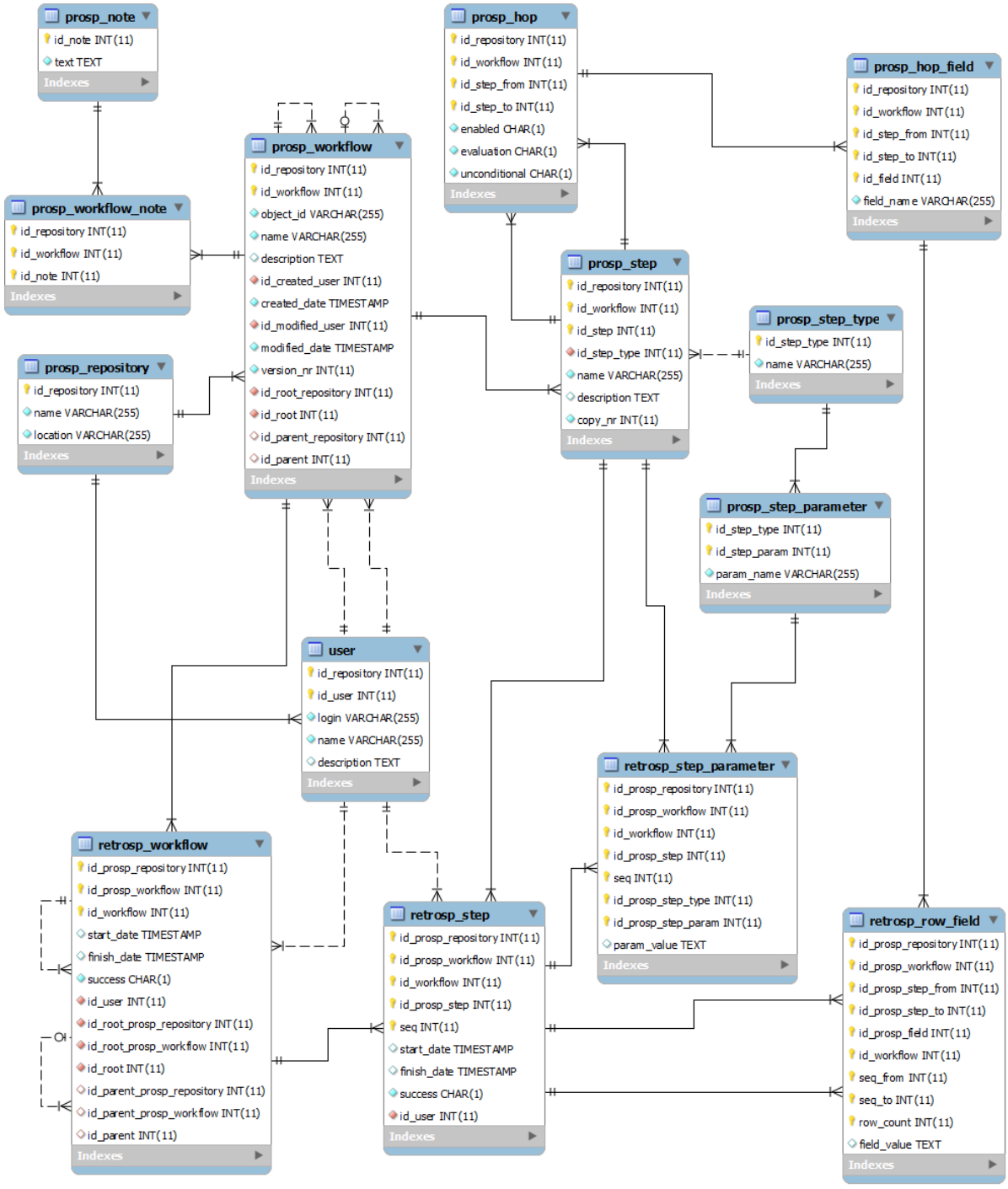

Figura 33. Modelo lógico de dados no MySQL do repositório temporário utilizado pelo Agente Coletor de Proveniência.

Quadro 6. Tipos de dados armazenados em cada tabela do banco de dados relacional utilizado pelo Agente Coletor de Proveniência para armazenar temporariamente os dados de proveniência.

Tabela	Tipo de dado armazenado
user	Dados sobre os usuários que atuam como agentes de um *workflow*.
prosp_repository	Dados sobre os repositórios em que a composição de um *workflow* está armazenada.
prosp_workflow	Dados da composição de um *workflow*.
prosp_note	Anotações de documentação de um *workflow*.
prosp_workflow_note	Relacionamento entre os dados da composição de um *workflow* e suas anotações de documentação.
prosp_step	Dados da composição dos passos de um *workflow*.
prosp_step_type	Dados da composição sobre os tipos dos passos de um *workflow*.
prosp_step_parameter	Dados da composição dos parâmetros utilizados nos passos de um *workflow*.
prosp_hop	Dados da composição das ligações entre os passos de um *workflow*.
prosp_hop_field	Dados da composição dos campos de dados que fluirão entre os passos de um *workflow*.
retrosp_workflow	Dados da execução de um *workflow*.
retrosp_step	Dados da execução dos passos de um *workflow*.
retrosp_step_parameter	Valor utilizado em um parâmetro na execução dos passos de um *workflow*.
retrosp_row_field	Valor de um campo em uma linha de dados que fluiu entre dois passos na execução de um *workflow*.

4.4 Estratégia de apoio semântico à proveniência

Para apoiar semanticamente a representação dos dados de proveniência capturados e armazenados temporariamente pelo Agente Coletor de Proveniência, este trabalho adotou uma estratégia de representação em camadas (Figura 34). A estratégia adotada é similiar à utilizada por FREITAS, KÄMPGEN, OLIVEIRA et al. (2012), mas sem empregar as mesmas ontologias de proveniência. Na presente estratégia, a ontologia PROV-O foi adotada por ter sido estabelecida como recomendação pelo W3C para possibilitar a interoperabilidade da

proveniência na Web e em outros sistemas de informação. Já a ontologia OPMW foi adotada devido à sua característica de estender a ontologia PROV-O para permitir a descrição, de maneira não ambígua, dos dados de proveniência prospectiva e dos dados de proveniência retrospectiva.

Sendo assim, a estratégia de representação utiliza, na primeira camada, a ontologia PROV-O para representar a semântica básica do *workflow* e possibilitar a interoperabilidade da proveniência. Na segunda camada, a ontologia OPMW é utilizada para distinguir a semântica sobre a composição do *workflow* (proveniência prospectiva) da semântica sobre a execução do *workflow* (proveniência retrospectiva). Na terceira camada, Cogs é utilizado para representar de maneira expressiva os conceitos do processo de ETL e, na quarta camada, uma ontologia complementar específica ao domínio da aplicação pode ser utilizada para estender a representação da proveniência realizada nas camadas anteriores. Além disso, os vocabulários Dublin Core e FOAF são utilizados para complementar as ontologias de proveniência anteriormente citadas.

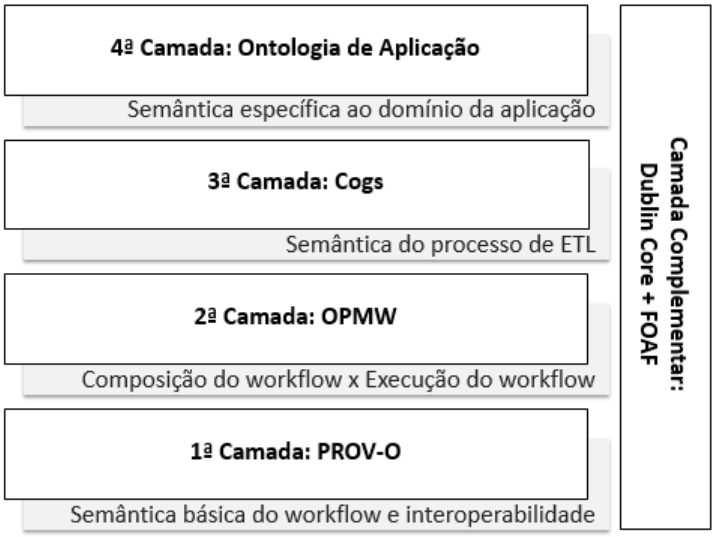

Figura 34. Camadas da estratégia de apoio semântico à proveniência adotada na presente abordagem.

O Quadro 7 apresenta o mapeamento entre 10 entidades do modelo conceitual do repositório temporário utilizado pelo Agente Coletor de Proveniência e as respectivas classes das ontologias PROV-O, OPMW e Cogs. As demais entidades do modelo conceitual do repositório temporário não são representadas por classes, mas seus atributos são utilizados como objetos de propriedades das mesmas ontologias de proveniência. Por exemplo, o atributo *localização* da entidade *Repositório* é utilizado na composição do objeto da propriedade *opmw:hasNativeSystemTemplate* da classe *opmw:WorkflowTemplate*; o atributo *texto* da entidade *Nota* é utilizado como objeto da propriedade *opmw:hasDocumentation* da classe

opmw:WorkflowTemplate e os atributos *origem* e *destino* da entidade *Ligação Unidirecional* são utilizados para definir a ordem dos passos do *workflow*, através da propriedade *cogs:precededBy*.

Quadro 7. Mapeamento entre entidades do repositório temporário utilizado pelo Agente Coletor de Proveniência e as classes das ontologias de proveniência da estratégia de apoio semântico adotada.

Repositório temporário	**Estratégia de apoio semântico à proveniência**		
	1ª Camada	**2ª Camada**	**3ª Camada**
Entidade	**PROV-O**	**OPMW**	**Cogs**
Usuário	prov:Agent	-	-
Workflow (Composição)	prov:Plan	opmw:WorkflowTemplate	-
Passo (Composição)	-	opmw:WorkflowTemplateProcess	cogs:Extraction (ou uma de suas subclasses) cogs:Transformation (ou uma de suas subclasses) cogs:Loading (ou uma de suas subclasses)
Parâmetro (Composição)	-	opmw:ParameterVariable	cogs:Object (ou uma de suas subclasses)
Campo de Dados (Composição)	-	opmw:DataVariable	cogs:Object
Workflow (Execução)	prov:Bundle	opmw:ExecutionAccount	-
Passo (Execução)	prov:Activity	opmw:ExecutionProcess	-
Parâmetro (Execução)	prov:Entity	opmw:WorkflowExecutionArtifact	-
Campo de Dados (Execução)	prov:Entity	opmw:WorkflowExecutionArtifact	-
Linhas de Dados (Execução)	prov:Collection	-	cogs:Row

4.5 Implementação

4.5.1 Pentaho Data Integration (Kettle)

A abordagem *ETL4LinkedProv* para coleta e publicação de proveniência como Linked Data foi concebida para ser implementada independentemente de uma ferramenta de *workflow* ETL específica. Conforme exposto na seção 2.5, existe um conjunto significativo de ferramentas que gerenciam um processo ETL, incluindo tanto ferramentas proprietárias, quanto ferramentas *open source*. Seguindo a experiência do projeto LinkedDataBR (CAMPOS, M. L. M., GUIZZARDI, 2010), a ferramenta Pentaho Data Integration, também conhecida como Kettle, foi utilizada para implementar um protótipo da abordagem *ETL4LinkedProv* e do Processo de Preparação e Transformação dos Dados.

O conjunto de componentes ETL4LOD, apresentado na seção 2.5.2, também foi utilizado na implementação. Este conjunto consiste dos *steps* Sparql Endpoint, Sparql Update Output, Data Property Mapping e Object Property Mapping, que foram desenvolvidos pelo projeto LinkedDataBR para suprir a carência do Kettle de componentes relacionados com as tecnologias de Linked Data. A versão inicial dos *steps* ETL4LOD possibilitava somente o armazenamento dos metadados de composição dos *steps* em arquivos XML de um repositório Kettle do tipo sistema de arquivos. Este trabalho estendeu os *steps* ETL4LOD para possibilitar o armazenamento dos metadados de composição dos mesmos também em um repositório Kettle do tipo banco de dados. Além disso, este trabalho desenvolveu o *step* NTriple Generator, que pode ser utilizado juntamente com os 4 *steps* do ETL4LOD.

O *step* NTriple Generator oferece a facilidade de geração de sentenças RDF no formato NTriple. A configuração consiste em informar cada campo da linha de entrada relacionado aos respectivos componentes da tripla RDF, ou seja, sujeito, predicado e objeto. Se o objeto for literal, ainda é possível informar os campos relacionados com o tipo do literal e a marca de idioma. Assim sendo, o *step* NTriple Generator torna-se útil, por exemplo, para receber as linhas de dados enviadas por um *step* Data Property Mapping ou Object Property Mapping e gerar, no fluxo de saída, as linhas de dados com as triplas RDF a serem inseridas em um banco de triplas, por meio do *step* Sparql Update Output. A Figura 35 ilustra a interface de configuração do *step* NTriple Generator.

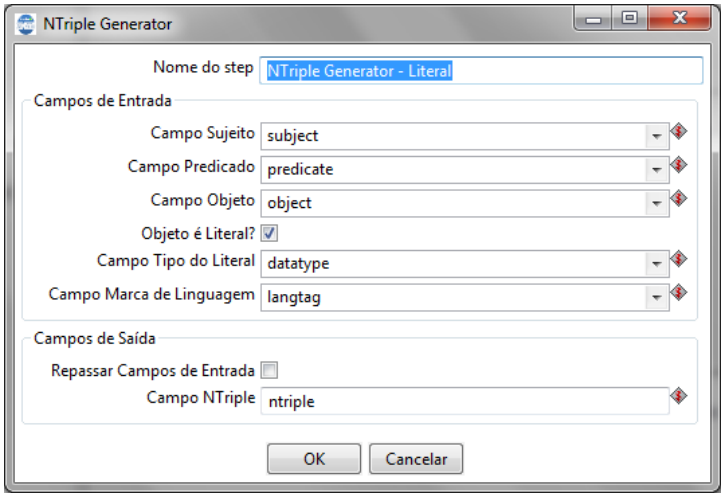

Figura 35. Interface de configuração do *step* NTriple Generator.

4.5.2 *Job entry* Provenance Collector Agent

Assim como os 4 *steps* do ETL4LOD e o *step* NTriple Generator, o Agente Coletor de Proveniência foi implementado através da API Java do Kettle. No entanto, o Agente Coletor de Proveniência foi implementado como um *job entry*, cujo tipo foi denominado Provenance Collector Agent. O *job entry* Provenance Collector Agent é uma extensão do *job entry* do tipo Job, que é responsável por executar um *job* dentro de outro *job*. Assim, um *job entry* do tipo Provenance Collector Agent, além de executar um *job* dentro de outro *job*, promove a coleta dos dados de proveniência prospectiva e dos dados de proveniência retrospectiva durante a execução do processo ETL encapsulado.

Além das abas do *job entry* Job, a interface de configuração do *job entry* Provenance Collector Agent apresenta uma aba adicional, denominada Proveniência. Esta aba possibilita a configuração da conexão com o banco de dados relacional utilizado para armazenamento temporário dos dados de proveniência e a seleção dos tipos de *steps*, cujos dados de proveniência com nível de granulosidade fina serão coletados. A Figura 36 ilustra a aba Proveniência da interface de configuração do *job entry* Provenance Collector Agent.

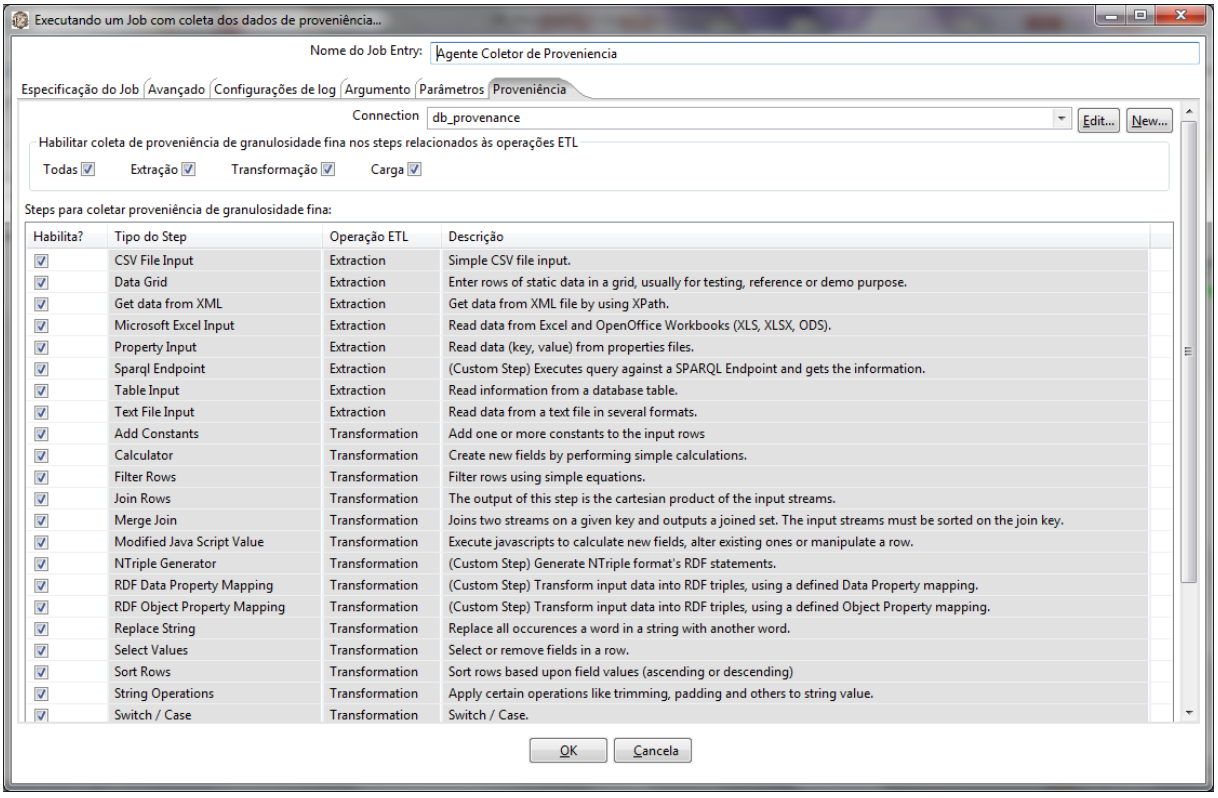

Figura 36. Aba Proveniência da interface de configuração do *job entry* Provenance Collector Agent.

Na implementação do *job entry* Provenance Collector Agent, os Padrões de Projeto de Software denominados Decorador (*Decorator*) e Observador (*Observer*) tiveram uma aplicação natural e fundamental. Um Padrão de Projeto de Software, também conhecido pelo termo original inglês *Design Pattern*, é uma solução genérica reutilizável para um problema recorrente no desenvolvimento de softwares orientados a objetos (FREEMAN, ERIC et al., 2004). Assim, em vez da reutilização de códigos de programação, os Padrões de Projeto de Software promovem a reutilização de boas práticas de desenvolvimento de software. GAMMA, HELM, JOHNSON e VILSSIDES (1995), conhecidos como a Guangue dos Quatro, tiveram um papel relevante no estabelecimento dos Padrões de Projeto de Software, pois foram responsáveis pela catalogação de 23 padrões. Estes padrões podem ser organizados em 3 categorias: (i) padrões de criação, que são relacionados à criação de objetos; (ii) padrões estruturais, que são relacionados à composição estrutural de classes e objetos; e (iii) padrões comportamentais, que são relacionados à comunicação entre objetos.

O padrão Decorador é um padrão de projeto de software da categoria estrutural, que consiste em incluir ou sobrescrever dinamicamente responsabilidades em um objeto (GAMMA et al., 1995). No contexto da programação orientada a objetos, este padrão oferece uma alternativa à herança, para realizar a extensão das funcionalidades de um objeto. Desta forma,

o padrão Decorador foi naturalmente aplicado no desenvolvimento do *job entry* Provenance Collector Agent, para estender as funcionalidades do *job entry* Job com a inclusão das atividades de coleta e armazenamento temporário dos dados de proveniência.

O padrão Observador foi utilizado para a implementação do rastreamento guloso e da coleta dos dados de proveniência à medida que os eventos forem acontecendo durante a execução do processo ETL. O padrão Observador é um padrão de projeto de software da categoria comportamental, em que um objeto, denominado sujeito, mantém uma lista de objetos, denominados observadores, que são notificados automaticamente sempre que o estado do sujeito é alterado (GAMMA et al., 1995).

4.5.3 Protótipo do Processo de Preparação e Transformação dos Dados

A infraestrutura fundamental utilizada na implementação do protótipo do Processo de Preparação e Transformação dos Dados consistiu do sistema operacional Windows 7 versão 64 bits, da plataforma Java Standard Edition[32] (SE) versão 1.7.0_13 e da edição comunitária (Community Edition - CE) do Kettle[33] versão 4.3.0-stable. Além dos *steps* e *job entries* padrão do Kettle, os 4 *steps* do ETL4LOD, o *step* NTriple Generator e o *job entry* Provenance Collector Agent foram adicionados à ferramenta, para também serem utilizados no desenvolvimento dos *workflows* ETL.

A edição comunitária do SGBD MySQL[34] versão 5.0.51a foi utilizada para gerenciar o repositório temporário dos dados de proveniência coletados pelo *job entry* do tipo Provenance Collector Agent e a edição *open source* do banco de triplas Virtuoso[35] versão 7.0.0 foi utilizada para armazenar as triplas RDF geradas pelo processo de publicação. Tanto a edição comunitária do SGBD MySQL, quanto a edição *open source* do banco de triplas Virtuoso foram selecionadas por serem plataformas de licença aberta, além de serem estáveis, simples de instalar e administrar e possuírem uma ampla comunidade de usuários.

A Figura 37 ilustra o *workflow* principal do Processo de Preparação e Transformação dos Dados, implementado como um *job* do Kettle denominado Job Proveniência. Após o início da execução deste *job*, o *job entry* do tipo Provenance Collector Agent, denominado Agente

[32] http://www.oracle.com/technetwork/java/javase/overview/index.html

[33] http://sourceforge.net/projects/pentaho/files/Data%20Integration

[34] http://www.mysql.com/products/community

[35] http://www.openlinksw.com/dataspace/doc/dav/wiki/Main

Coletor de Proveniência, executa o *job* encapsulado, coletando seus dados de proveniência prospectiva e seus dados de proveniência retrospectiva. Os dados de proveniência coletados são armazenados no banco de dados MySQL configurado como repositório temporário. Ao final da execução do *job entry* do tipo Provenance Collector Agent, 3 dados são enviados para o *job* entry seguinte: o identificador do repositório onde a composição do *workflow* ETL encapsulado está armazenada (id_prosp_repository), o identificador da composição do *workflow* ETL encapsulado (id_prosp_workflow) e o identificador da execução do *workflow* ETL encapsulado (id_workflow).

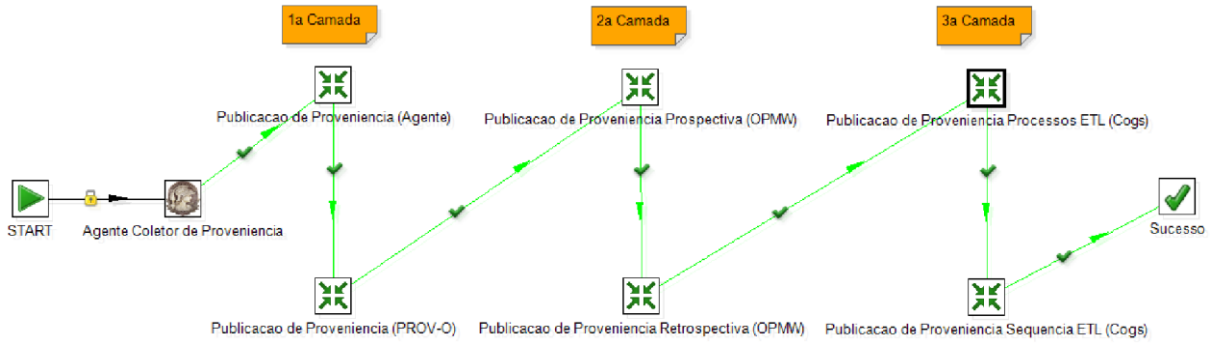

Figura 37. Job Proveniência: implementação do *workflow* principal do Processo de Preparação e Transformação dos Dados.

Os 6 *job entries* subsequentes ao *job entry* Agente Coletor de Proveniência são do tipo *Transformation*, tipo de *job entry* responsável por executar um *transformation*. Sendo assim, cada um destes 6 *job entries* executa um *transformation*, que implementa um *workflow* ETL responsável por extrair, do repositório temporário, os dados de proveniência coletados pelo *job entry* Agente Coletor de Proveniência para, em seguida, triplificar e carregar estes dados de proveniência no banco de triplas Virtuoso.

Cada um dos 6 *sub-workflows* ETL executados pelos *job entries* é responsável por uma representação específica dos dados de proveniência, de acordo com a estratégia de apoio semântico à proveniência adotada neste trabalho:

(i) o *transformation* executado pelo *job entry* Publicação de Proveniência (Agente) (Figura 38) representa o recurso do tipo Agente, por meio da ontologia PROV-O;

(ii) o *transformation* executado pelo *job entry* Publicação de Proveniência (PROV-O) (Figura 39) representa os recursos dos tipos Processo e Entidade, por meio da ontologia PROV-O;

(iii) o *transformation* executado pelo *job entry* Publicação de Proveniência Prospectiva (OPMW) (Figura 40) representa a proveniência prospectiva, por meio da ontologia OPMW;

(iv) o *transformation* executado pelo *job entry* Publicação de Proveniência Retrospectiva (OPMW) (Figura 41) representa a proveniência retrospectiva, por meio da ontologia OPMW;

(v) o *transformation* executado pelo *job entry* Publicação de Proveniência Processos ETL (Cogs) (Figura 42) representa os processos de ETL e seus parâmetros, por meio da ontologia Cogs;

(vi) o *transformation* executado pelo *job entry* Publicação de Proveniência Sequência ETL (Cogs) (Figura 43) representa a sequência dos processos de ETL no *workflow*, por meio da ontologia Cogs.

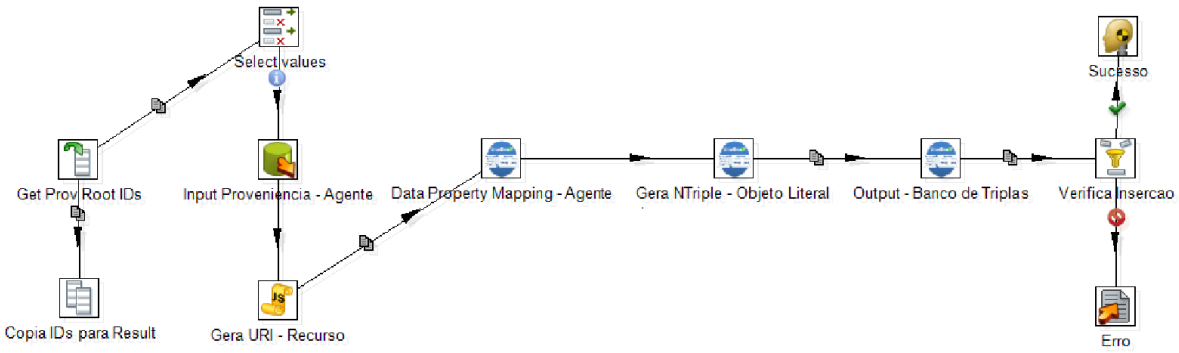

Figura 38. *Transformation* executado pelo *job entry* Publicação de Proveniência (Agente): implementação do *sub-workflow* ETL de publicação dos dados de proveniência do recurso do tipo Agente.

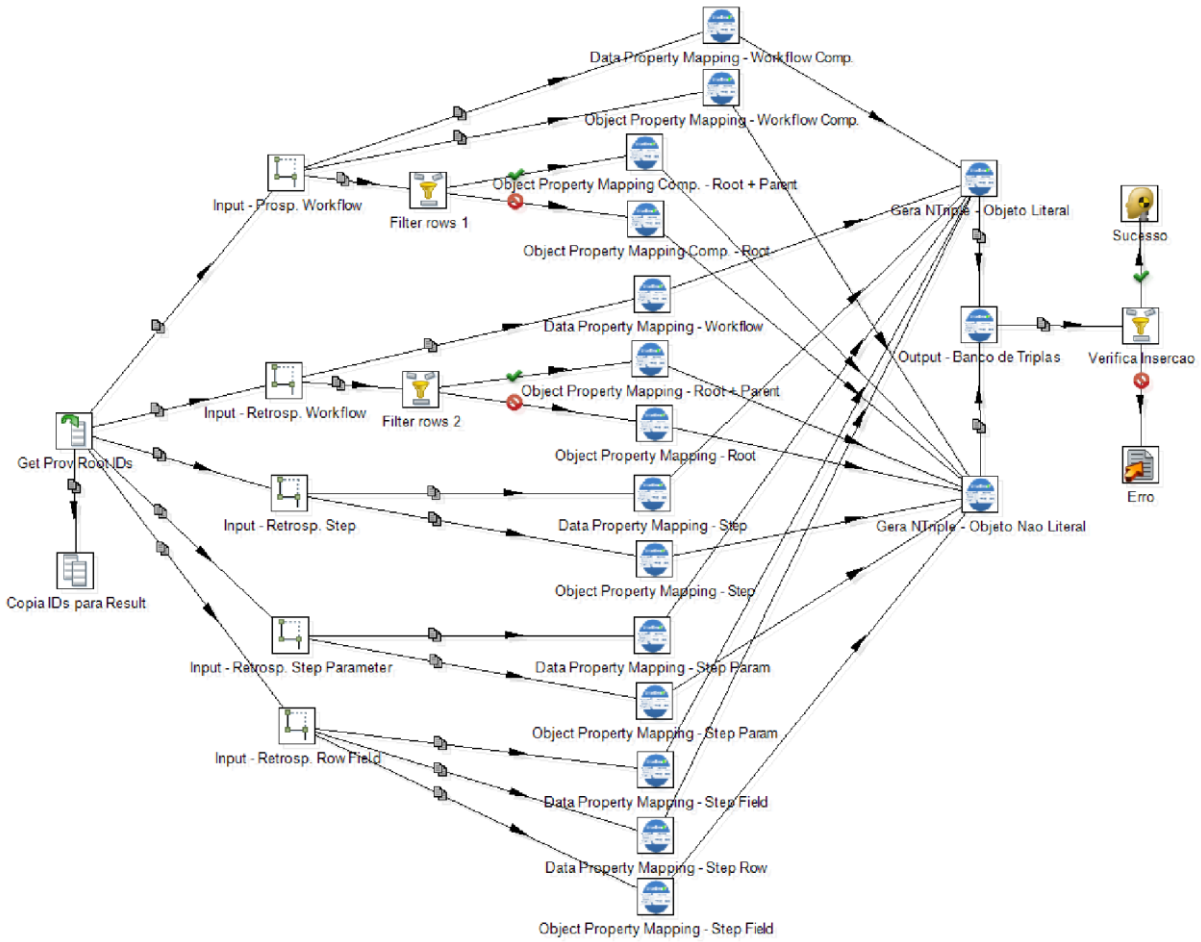

Figura 39. *Transformation* executado pelo *job entry* Publicação de Proveniência (PROV-O): implementação do *sub-workflow* ETL de publicação dos dados de proveniência com utilização da ontologia PROV-O.

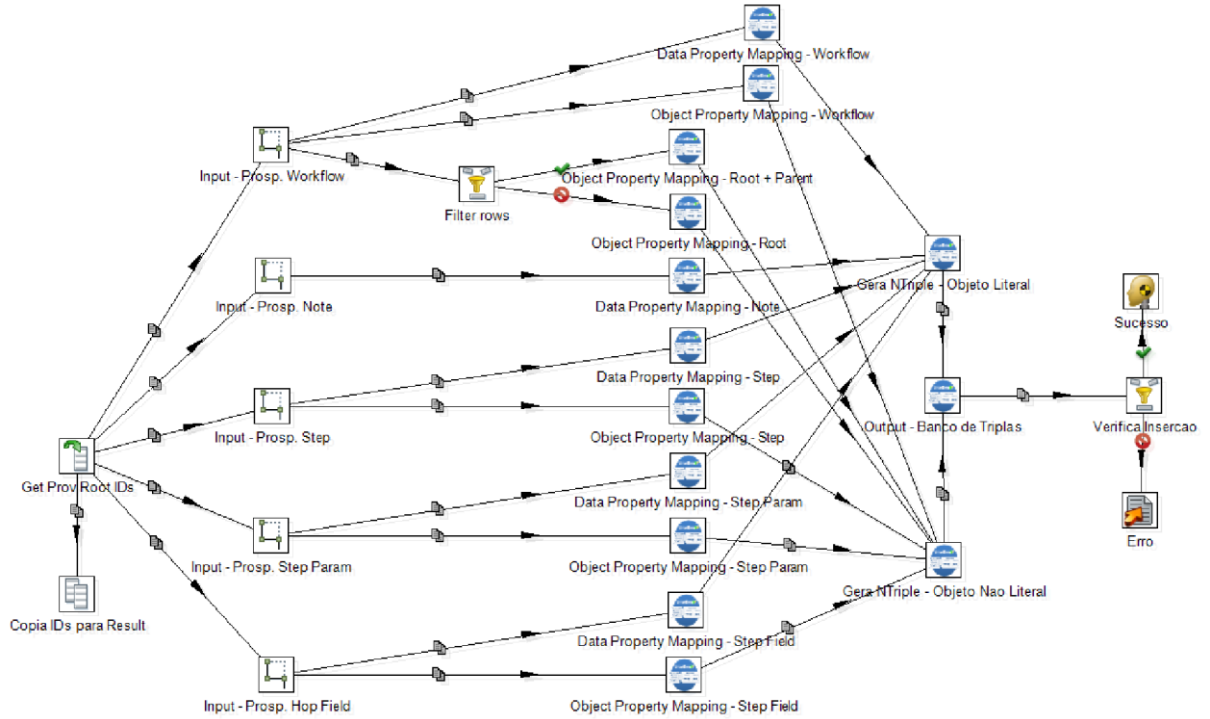

Figura 40. *Transformation* executado pelo *job entry* Publicação de Proveniência Prospectiva (OPMW): implementação do *sub-workflow* ETL de publicação dos dados de proveniência prospectiva com utilização da ontologia OPMW.

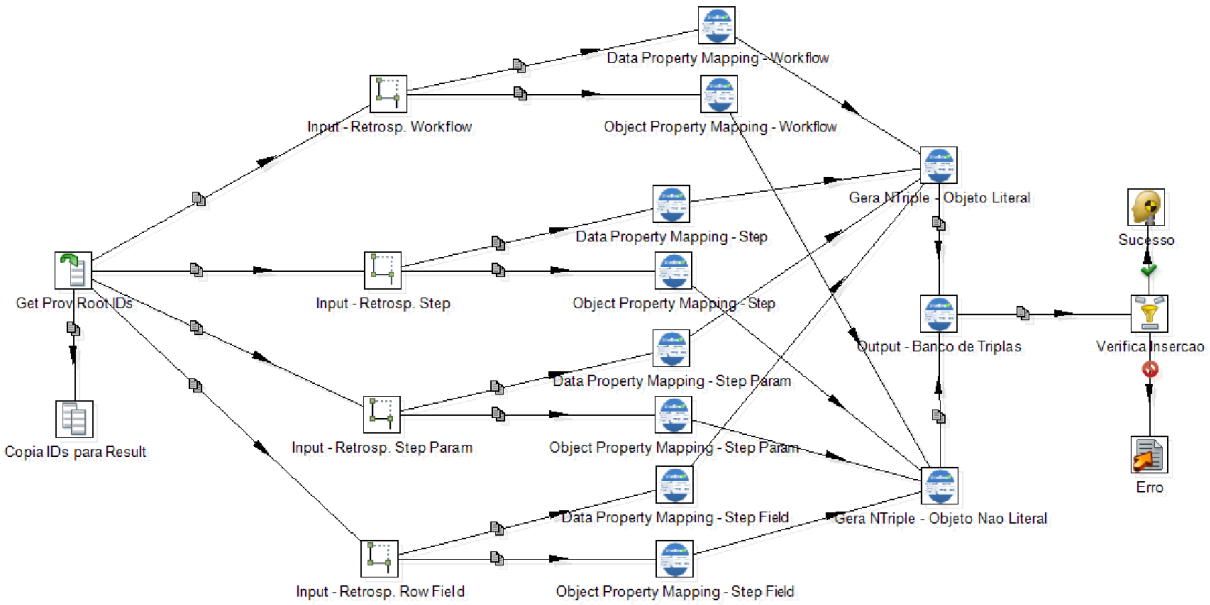

Figura 41. *Transformation* executado pelo *job entry* Publicação de Proveniência Retrospectiva (OPMW): implementação do *sub-workflow* ETL de publicação dos dados de proveniência retrospectiva com utilização da ontologia OPMW.

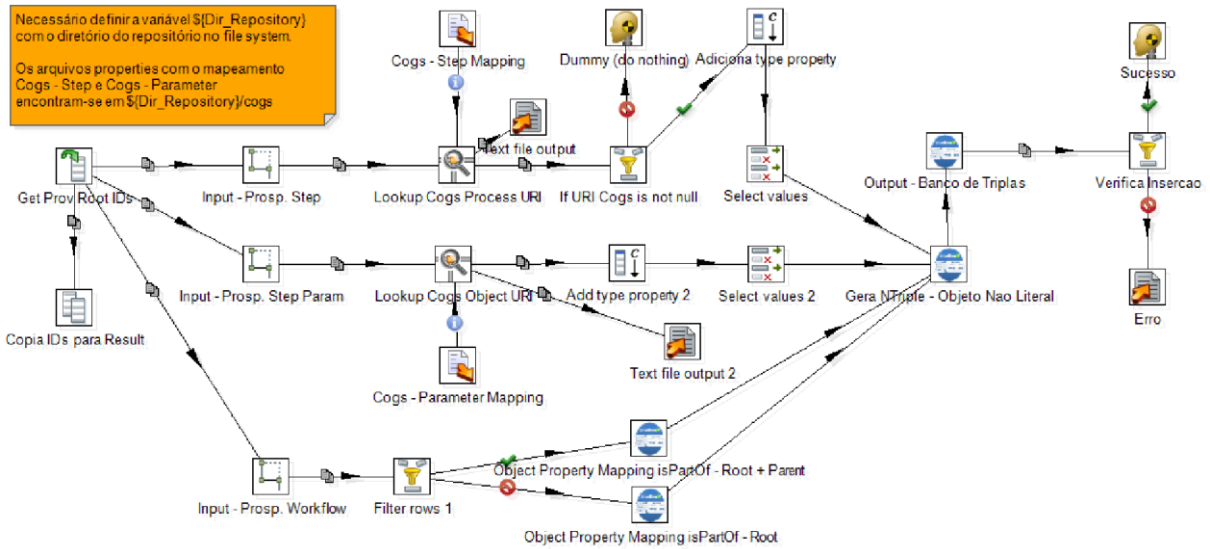

Figura 42. *Transformation* executado pelo *job entry* Publicação de Proveniência Processos ETL (Cogs): implementação do *sub-workflow* ETL de publicação dos dados de proveniência dos processos de ETL com utilização da ontologia Cogs.

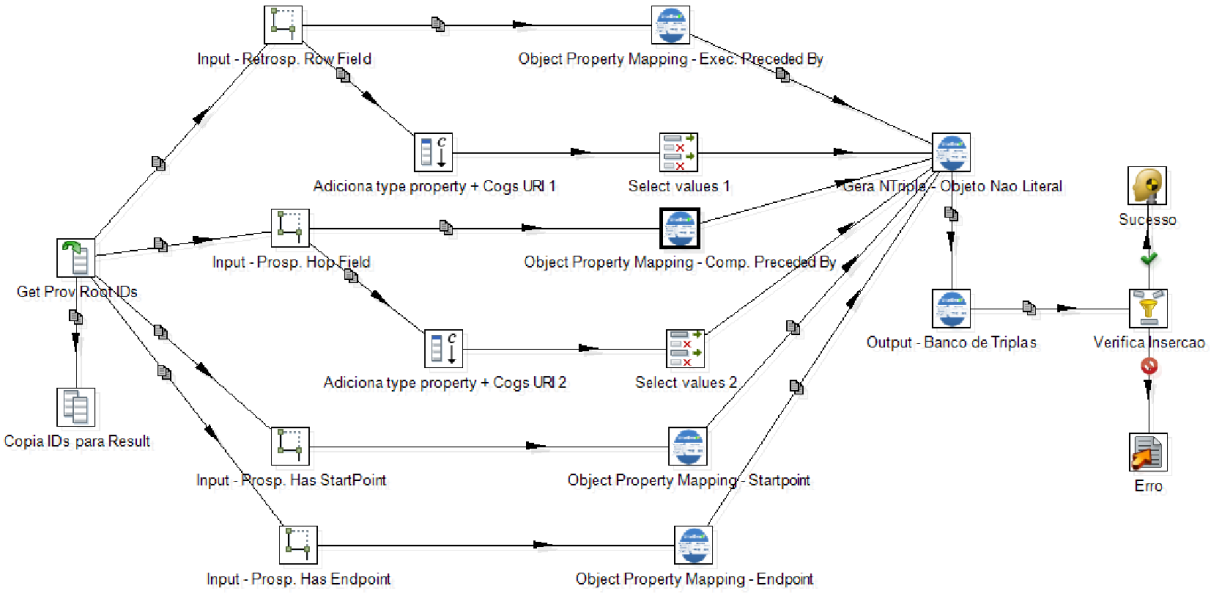

Figura 43. *Transformation* executado pelo *job entry* Publicação de Proveniência Processos ETL (Cogs): implementação do *sub-workflow* ETL de publicação dos dados de proveniência da sequência dos processos de ETL no *workflow* com utilização da ontologia Cogs.

Embora representem os dados de proveniência de maneira distinta, os 6 *sub-workflows* apresentam uma estrutura similar. Inicialmente, eles lêem os 3 identificadores enviados pelo *job entry* Agente Coletor de Proveniência (id_prosp_repository, id_prosp_workflow e id_workflow) e repassam os valores para os *steps* que executam os *subtransformations* responsáveis por extrair os dados de proveniência do repositório temporário, além de copiar os valores, por meio do *step* Copia IDs para Result, para serem lidos pelo próximo *job entry* do Job Proveniência.

Cada *subtransformation* extrai os dados de proveniência do repositório temporário por meio de um *step* do tipo Table Input e gera as URIs dos recursos através de um *step* do tipo JavaScript. Por exemplo, o *step* Input - Prosp. Step, utilizado tanto no *transformation* executado pelo *job entry* Publicação de Proveniência Prospectiva (OPMW), quanto no *transformation* executado pelo *job entry* Publicação de Proveniência Processos ETL (Cogs), executa o *subtransformation* transf_input_prosp_step (Figura 44), que contém o *step* Input Proveniencia - Prosp. Step do tipo Table Input e o *step* Gera URI - Recurso Step do tipo JavaScript.

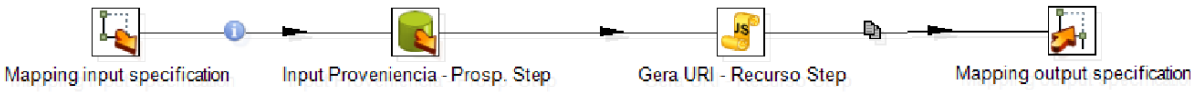

Figura 44. *Transformation* transf_input_prosp_step: *subtransformation* responsável por extrair os dados de proveniência prospectiva dos passos do workflow ETL do repositório temporário.

O *step* Input Proveniencia - Prosp. Step extrai, por meio da consulta SQL descrita a seguir, os dados de proveniência prospectiva dos passos do *workflow* ETL raiz e dos passos dos subsequentes *sub-workflows* executados pelo *job entry* Agente Coletor de Proveniência:

```
1  SELECT
2    t2.id_repository
3  , t2.id_workflow
4  , t2.id_step
5  , t3.name as type
6  , t2.name
7  , t2.description
8  , t2.copy_nr
9  , t4.name as w_name
10 FROM retrosp_workflow t1, prosp_step t2, prosp_step_type t3,
   prosp_workflow t4
11 WHERE t1.id_root_prosp_repository = ?
12 AND t1.id_root_prosp_workflow = ?
13 AND t1.id_root = ?
14 AND t4.id_repository = t1.id_prosp_repository
15 AND t4.id_workflow = t1.id_prosp_workflow
16 AND t2.id_repository = t4.id_repository
17 AND t2.id_workflow = t4.id_workflow
18 AND t3.id_step_type = t2.id_step_type
```

Em seguida, o *step* Gera URI - Recurso Step gera as URIs dos recursos que representam a composição de cada passo recuperado pela consulta SQL e também as URIs das composições dos *workflows* que contêm os passos. Os valores dos campos id_prosp_repository, id_prosp_workflow e id_workflow são utilizados como filtros na consulta para especificar o workflow ETL raiz executado pelo *job entry* Agente Coletor de Proveniência.

Após a execução do *subtransformation* de extração dos dados de proveniência do repositório temporário, *steps* do tipo Data Property Mapping e Object Property Mapping mapeiam os dados extraídos nos componentes sujeito, predicado e objeto da tripla RDF. Por fim, *steps* do tipo NTriple Generator geram e enviam as triplas RDF no formato NTriple para o *step* do tipo Sparql Update Output, denominado Output - Banco de Triplas, que insere as triplas RDF no banco de triplas Virtuoso.

No próximo capítulo, este protótipo da abordagem *ETL4LinkedProv* será aplicado em um cenário que envolve dados reais referentes a grupos de pesquisa e seus respectivos pesquisadores, publicações e projetos. O *job entry* do tipo Provenance Collector Agent irá executar e coletar os dados de proveniência de um *workflow* ETL que integra e publica duas fontes de dados do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e uma fonte de dados da Rede Nacional de Ensino e Pesquisa (RNP). Como forma de avaliação da abordagem *ETL4LinkedProv*, serão realizadas consultas SPARQL, explorando conjuntamente os dados do CNPq, os dados da RNP e os respectivos dados de proveniência prospectiva e retrospectiva.

5 Exemplo de aplicação

5.1 Publicação de fontes de dados do CNPq e da RNP, por meio da abordagem *ETL4LinkedProv*

No cenário da pesquisa científica e tecnológica nacional, encontramos hoje dificuldades de entender e explorar os financiamentos concedidos pelas diferentes organizações de incentivo à pesquisa do país. Estas organizações, a exemplo da Financiadora de Estudos e Projetos (FINEP) e das Fundações de Amparo à Pesquisa (FAP) estaduais, mantêm seus dados isolados, com pouca ou nenhuma interligação. Sendo assim, é difícil consumir conjuntamente e reutilizar os dados destas organizações para, por exemplo, correlacionar os financiamentos concedidos com a produtividade em publicações científicas e orientações de pesquisa ou correlacionar os projetos financiados com os Grupos de Pesquisa. Considerando este cenário, um exemplo de aplicação da abordagem *ETL4LinkedProv* é apresentado neste capítulo, envolvendo dados de duas organizações de incentivo à pesquisa: o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Rede Nacional de Ensino e Pesquisa (RNP).

O CNPq[36] é o órgão ligado ao Ministério da Ciência, Tecnologia e Inovação (MCTI) responsável por fomentar a pesquisa científica e tecnológica e incentivar a formação de pesquisadores no Brasil. O Currículo Lattes[37], padrão nacional de registro da vida pregressa e atual dos estudantes e pesquisadores do país, encontra-se entre os principais dados públicos disponibilizados pelo CNPq na Internet. A RNP[38] é uma outra organização de incentivo à pesquisa científica e tecnológica, criada em 1989 pelo MCTI com o objetivo de construir uma infraestrutura de rede Internet nacional para a comunidade acadêmica do Brasil. Entre as principais iniciativas da RNP, encontra-se o programa Grupos de Trabalho RNP[39], que financia e acompanha projetos de pesquisa, em temas específicos aos interesses dos usuários da RNP, propostos e desenvolvidos por Grupos de Pesquisa nacionais.

[36] http://www.cnpq.br

[37] http://lattes.cnpq.br

[38] http://www.rnp.br

[39] https://www.rnp.br/pd/gt.html

Se o CNPq e a RNP publicassem seus dados como Linked Data através da abordagem *ETL4LinkedProv*, seria possível utilizar os respectivos dados de proveniência para uma posterior verificação da linhagem dos dados publicados sobre as Instituições de Ensino e Pesquisa, sobre os Grupos de Pesquisa, sobre os projetos de pesquisa desenvolvidos e sobre os Currículos dos pesquisadores. Esta verificação seria um subsídio importante, por exemplo, para apoiar a validação da consistência entre os dados publicados pelo CNPq e pela RNP. Além de um melhor entendimento, de maneira geral, a consistência entre os dados mantidos por estas organizações de fomento permitiria a exploração conjunta dos dados com qualidade e confiabilidade.

Nesta direção, o exemplo de aplicação apresentado neste capítulo utiliza o protótipo da abordagem *ETL4LinkedProv*, descrito na seção 4.5.3, para encapsular um workflow ETL que integra e publica como Linked Data duas fontes de dados do CNPq e uma fonte de dados da RNP. As duas fontes de dados do CNPq, ambas armazenadas em banco de dados relacional, contêm informações sobre os Currículos Lattes e informações sobre os Grupos de Pesquisa respectivamente. Já a fonte de dados da RNP, armazenada em arquivos XML, contém informações sobre os Grupos de Trabalho do programa Grupos de Trabalho RNP. A Figura 45 oferece uma visão geral do cenário utilizado.

Figura 45. Visão geral do cenário utilizado no exemplo de aplicação.

O workflow ETL encapsulado pelo Agente Coletor de Proveniência inicialmente limpa, consolida e integra as duas fontes de dados do CNPq, para gerar a lista de projetos de cada Grupo de Pesquisa cadastrado no CNPq. Em seguida, o workflow ETL também limpa, consolida e integra a fonte de dados da RNP com as fontes de dados do CNPq, a fim de relacionar os Grupos de Trabalho da RNP e os Grupos de Pesquisa do CNPq. Os dados das três

fontes são publicados no banco de triplas Virtuoso, organizados em três grafos RDF: o primeiro grafo, nomeado *http://lattes.cnpq.br*, contém as triplas RDF dos Currículos Lattes; o segundo grafo, nomeado *http://www.cnpq.br*, contém as triplas RDF dos Grupos de Pesquisa do CNPq e o terceiro grafo, nomeado *http://www.rnp.br*, contém as triplas RDF dos Grupos de Trabalho da RNP. Um quarto grafo, nomeado *http://greco.ppgi.ufrj.br/provenance*, contém os dados de proveniência coletados e publicados pelo Agente Coletor de Proveniência. A Figura 46 ilustra o workflow ETL encapsulado, implementado com um *job* do Kettle, e a Figura 47, a Figura 48 e a Figura 49 ilustram os *subtransformations* responsáveis por publicar os dados das fontes de dados CNPq-Lattes, CNPq-Grupos de Pesquisa e RNP-Grupos de Trabalho.

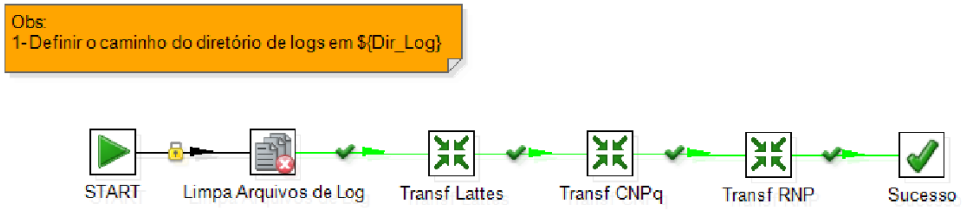

Figura 46. Job CNPq-RNP: Implementação do *workflow* ETL que integra e publica como Linked Data duas fontes de dados do CNPq e uma fonte de dados da RNP.

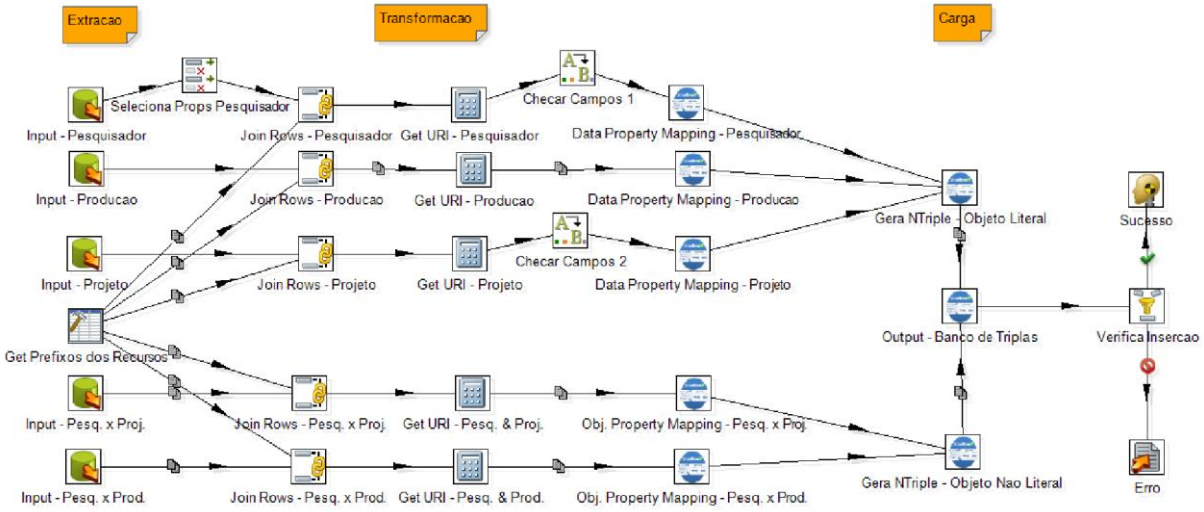

Figura 47. *Transformation* transf_lattes: Implementação do *sub-workflow* ELT responsável por publicar como Linked Data a fonte de dados dos Currículos Lattes.

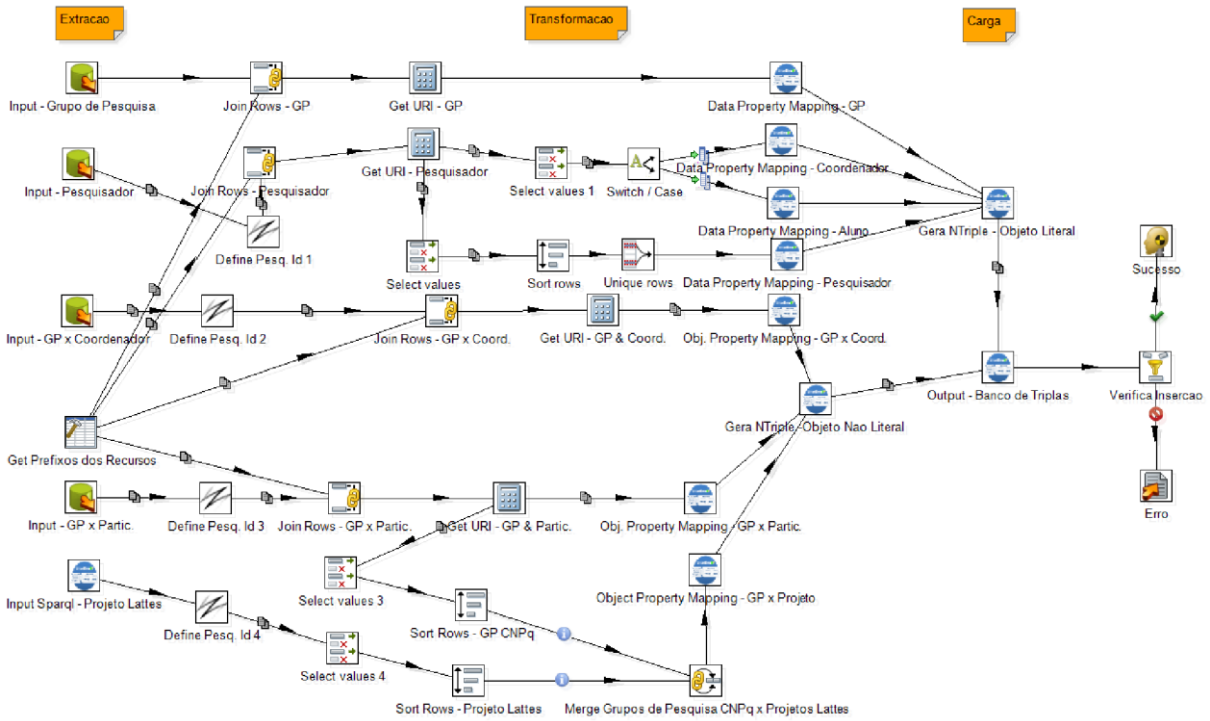

Figura 48. *Transformation* transf_cnpq: Implementação do *sub-workflow* ELT responsável por publicar como Linked Data a fonte de dados dos Grupos de Pesquisa do CNPq.

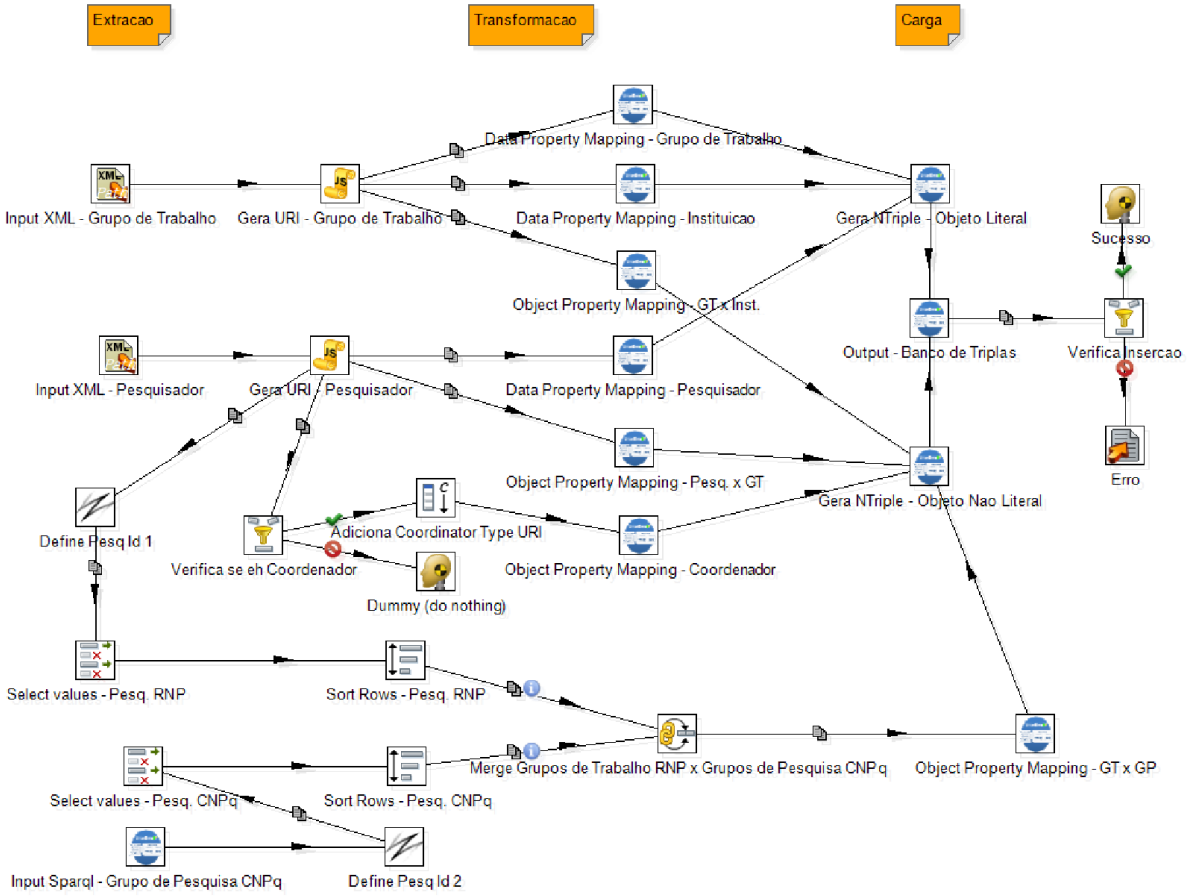

Figura 49. *Transformation* transf_rnp: Implementação do *sub-workflow* ELT responsável por publicar como Linked Data a fonte de dados dos Grupos de Trabalho da RNP.

Após a execução do processo de publicação dos dados do CNPq e da RNP por meio da abordagem *ETL4LinkedProv*; consultas SPARQL de exploração conjunta entre os dados do CNPq, os dados da RNP e seus respectivos dados de proveniência foram realizadas. Essas consultas SPARQL serão apresentadas na próxima subseção, a fim de evidenciar a contribuição da abordagem proposta neste trabalho.

5.2 Consultas SPARQL de exploração conjunta entre os dados do CNPq, os dados da RNP e seus respectivos dados de proveniência

Sete consultas SPARQL de exploração conjunta, realizadas após a publicação dos dados do CNPq e da RNP por meio do protótipo do Processo de Preparação e Transformação dos Dados, serão apresentadas nas próximas subseções, com a respectiva a análise dos resultados obtidos.

5.2.1 Consulta 1 - Recuperação de dados de domínio, a partir dos dados de proveniência retrospectiva

A consulta 1 é um exemplo de consulta SPARQL que tem por objetivo recuperar dados de domínio, a partir dos respectivos dados de proveniência retrospectiva. Neste exemplo, os dados de domínio são os Grupos de Pesquisa do CNPq publicados no grafo RDF *http://www.cnpq.br* e os dados de proveniência retrospectiva utilizados como filtros são o término da execução do processo de publicação e o agente que executou o processo. Como a consulta 1 envolve somente o tipo retrospectivo dos dados de proveniência, o emprego da PROV-O, ontologia da primeira camada da estratégia de apoio semântico à proveniência adotada pela aborgagem *ETL4LinkedProv*, foi suficiente para explorar os dados de proveniência publicados no grafo *http://greco.ppgi.ufrj.br/provenance*.

Consulta 1: Quais Grupos de Pesquisa do CNPq foram publicados por Rogers no dia 04/11/2013 ?

```
1  PREFIX cnpq: <http://www.cnpq.br/ontology/>
2  PREFIX cnpq-prop: <http://www.cnpq.br/property/>
3  PREFIX foaf: <http://xmlns.com/foaf/0.1/>
4  PREFIX prov: <http://www.w3.org/ns/prov#>
5
6  SELECT DISTINCT ?iri ?group_name
7  FROM NAMED <http://greco.ppgi.ufrj.br/provenance>
8  FROM NAMED <http://www.cnpq.br>
9  WHERE
10 {
11       GRAPH <http://greco.ppgi.ufrj.br/provenance>
12       {
13             ?step_exec prov:wasAssociatedWith ?agent .
14             ?agent foaf:name "ROGERS" .
15             ?step_exec prov:endedAtTime ?end .
16             FILTER (?end >= "2013-11-04"^^xsd:dateTime && ?end  <
   "2013-11-05"^^xsd:dateTime) .
17             ?data prov:wasGeneratedBy ?step_exec .
18             ?data prov:value ?value .
19             BIND(IRI(?value) AS ?iri) .
20       }
21       GRAPH <http://www.cnpq.br>
22       {
23             ?iri a cnpq:ResearchGroup .
24             ?iri cnpq-prop:name ?group_name .
25       }
26 }
```

Na consulta 1, as linhas 13 e 14 filtram os processos executados pelo agente determinado; as linhas 15 e 16 filtram os processos concluídos no dia 04/11/2013; as linhas 17 e 18 recuperam os dados gerados pelo processo; a linha 19 transforma os dados literais em IRIs por meio da função IRI do SPARQL 1.1 e coloca o valor transformado na variável *?iri* por meio da função BIND do SPARQL 1.1 e, por fim, as linhas 23 e 24 realizam a interligação entre as IRIs publicadas no grafo de proveniência e as IRIs que representam os Grupos de Pesquisa do CNPq, publicadas no grafo *http://www.cnpq.br*. O Quadro 8 apresenta uma amostra dos resultados obtidos pela Consulta 1.

Quadro 8. Amostra do resultado obtido pela Consulta 1.

iri	group_name
http://www.cnpq.br/resource/ResearchGroup/11	"GRECO - Grupo Engenharia do Conhecimento"@pt
http://www.cnpq.br/resource/ResearchGroup/37	"Ontologia e Taxonomia: aspectos teóricos e metodológicos"@pt
http://www.cnpq.br/resource/ResearchGroup/2	"Bioinformática e genética molecular evolutiva"@pt

5.2.2 Consulta 2 - Recuperação de dados de proveniência prospectiva, a partir dos dados de domínio

A consulta 2 é um exemplo de consulta SPARQL que tem por objetivo utilizar dados de domínio para recuperar os respectivos dados de proveniência prospectiva. Neste exemplo, o dado de domínio é a pesquisadora do CNPq representada pelo recurso *http://www.cnpq.br/resource/Researcher/MARIA_LUIZA_MACHADO_CAMPOS* e os dados de proveniência prospectiva são a data da modificação e o agente que fez a última modificação na composição do *workflow* que publicou o recurso. Como a consulta 2 envolve o tipo prospectivo dos dados de proveniência, além da PROV-O, foi necessário utilizar a OPMW, ontologia da segunda camada da estratégia de apoio semântico à proveniência adotada pelo Processo de Preparação e Transformação dos Dados, para explorar os dados de proveniência publicados no grafo *http://greco.ppgi.ufrj.br/provenance*.

Consulta 2: Por quem e quando foi feita a última modificação na composição do *workflow* ETL "transf_cnpq", que publicou a pesquisadora do CNPq representada pelo recurso *http://www.cnpq.br/resource/Researcher/MARIA_LUIZA_MACHADO_CAMPOS* ?

```
1  PREFIX dc: <http://purl.org/dc/terms/>
2  PREFIX opmw: <http://www.opmw.org/ontology/>
3  PREFIX prov: <http://www.w3.org/ns/prov#>
4
5  SELECT ?w_comp ?w_version ?w_modifier ?w_modified
6  FROM <http://greco.ppgi.ufrj.br/provenance>
7  WHERE
8  {
9        ?w_comp dc:title "transf_cnpq" .
10       ?w_comp opmw:versionNumber ?w_version .
11       ?step_comp opmw:isStepOfTemplate ?w_comp .
12       ?step_exec opmw:correspondsToTemplateProcess ?step_comp .
13       ?data prov:wasGeneratedBy ?step_exec .
14       ?data prov:value
   "http://www.cnpq.br/resource/Researcher/MARIA_LUIZA_MACHADO_CAMPOS".
15       ?w_comp dc:contributor ?w_modifier .
16       ?w_comp dc:modified ?w_modified .
17 }
18 GROUP BY ?w_comp ?w_version ?w_modifier ?w_modified
19 HAVING ?w_version = MAX(?w_version)
```

Na consulta 2, as linhas 9 e 10 recuperam todas as versões da composição do *workflow* denominado "transf_cnpq". A linha 11 recupera as instâncias que representam as composições dos passos do *workflow* (proveniência prospectiva) e a linha 12 recupera as execuções destas instâncias (proveniência retrospectiva). Os dados gerados na execução dos passos são

recuperados pela linha 13 e a linha 14 filtra estes dados, utilizando o valor da IRI que representa o dado de domínio. As linhas 15 e 16 recuperam, por meio de propriedades do Dublin Core, a data de modificação e o agente que fez a modificação na composição do *workflow*. Por fim, as linhas 18 e 19 agrupam os resultados pelas variáveis a serem selecionadas e filtra o grupo relacionado à última versão da composição do *workflow* "transf_cnpq". O Quadro 9 apresenta o resultado obtido pela Consulta 2.

Quadro 9. Resultado obtido pela Consulta 2.

w_comp	w_version	w_modifier	w_modified
http://greco.ppgi.ufrj.br/provenance/resource/WorkflowTemplate/TRANSF_CNPQ13	3	http://greco.ppgi.ufrj.br/provenance/resource/Agent/ROGERS1	2013-11-04 T10:42:26Z

5.2.3 Consulta 3 - Recuperação de dados de proveniência retrospectiva do processo de extração, a partir da especificação da fonte de extração dos dados

A consulta 3 é um exemplo de consulta SPARQL que tem por objetivo recuperar os dados de proveniência retrospectiva do processo de extração do *workflow* ETL, a partir da especificação do repositório de onde os dados foram extraídos. Neste exemplo, a fonte de dados é o banco de dados "db_lattes" e os dados de proveniência são o passo do *workflow* que executa o processo de extração de dados da fonte especificada e a data da última execução bem sucedida. Para explorar os dados de proveniência relacionados à semântica do processo de ETL, a ontologia Cogs, da terceira camada da estratégia de apoio semântico à proveniência adotada pelo Processo de Preparação e Transformação dos Dados, foi aplicada.

Consulta 3: Qual o passo da última execução bem sucedida do *workflow* "transf_lattes", que extraiu dados de pesquisadores do banco de dados "db_lattes" ?

```
1  PREFIX cogs: <http://vocab.deri.ie/cogs#>
2  PREFIX dc: <http://purl.org/dc/terms/>
3  PREFIX opmo: <http://openprovenance.org/model/opmo#>
4  PREFIX opmw: <http://www.opmw.org/ontology/>
5  PREFIX prov: <http://www.w3.org/ns/prov#>
6
7  SELECT ?step_exec ?step_title ?step_start ?step_end ?step_executor
8  FROM <http://greco.ppgi.ufrj.br/provenance>
9  WHERE
10 {
11       ?w_exec a opmw:WorkflowExecutionAccount .
12       ?w_exec opmw:correspondsToTemplate ?w_comp .
13       ?w_comp dc:title "transf_lattes" .
14
15       ?w_exec opmw:overallEndTime ?max_end .
16       {
17             SELECT MAX(?end_tmp) as ?max_end
18             FROM <http://greco.ppgi.ufrj.br/provenance>
19             WHERE
20             {
21                   ?w_exec_tmp a opmw:WorkflowExecutionAccount .
22                   ?w_exec_tmp opmw:correspondsToTemplate ?w_comp_tmp
   .
23                   ?w_comp_tmp dc:title "transf_lattes" .
24                   ?w_exec_tmp opmw:overallEndTime ?end_tmp .
25                   ?w_exec_tmp opmw:hasStatus "Y" .
26             }
27       }
28
29       ?step_exec opmo:account ?w_exec .
30       ?step_exec opmw:correspondsToTemplateProcess ?step_comp .
31       ?step_comp a cogs:Extraction .
32
33       ?step_exec prov:used ?param_exec1 .
34       ?param_exec1 opmw:correspondsToTemplateArtifact ?param_comp1 .
35       ?param_comp1 a cogs:Database .
36       ?param_exec1 opmw:hasValue "db_lattes" .
37
38       ?step_exec prov:used ?param_exec2 .
39       ?param_exec2 opmw:correspondsToTemplateArtifact ?param_comp2 .
40       ?param_comp2 a cogs:SelectQueryObject .
41       ?param_exec2 opmw:hasValue ?sql .
42       FILTER REGEX(?sql, "from pesquisador", "i")
43
44       ?step_exec dc:title ?step_title .
45       ?step_exec prov:startedAtTime ?step_start .
46       ?step_exec prov:endedAtTime ?step_end .
47       ?step_exec prov:wasAssociatedWith ?step_executor .
48 }
```

Na consulta 3, as linhas 11, 12 e 13 recuperam a instâncias da composição do *workflow* denominado "transf_lattes" e as instâncias de suas execuções. A linha 15 filtra as instâncias das execuções do *workflow* pela data da última execução bem sucedida, data que é recuperada pela subconsulta localizada entre as linhas 17 e 26. As linhas 29, 30 e 31 recuperam as instâncias das execuções dos passos da última execução bem sucedida do *workflow* "transf_lattes", cuja composição é um processo de extração. As linhas 33, 34, 35 e 36 filtram as instâncias recuperadas, cujo parâmetro do tipo *cogs:Database* está configurado com o banco de dados "db_lattes" e as linhas 36, 37, 38, 39 e 40 filtram as instâncias recuperadas, que estão extraindo informações da tabela "pesquisador". Por fim, as linhas 44, 45, 46 e 47 recuperam, por meio de propriedades da ontologias PROV-O e Dublin Core, o título, as datas de início e fim da execução do processo de extração e o agente que executou o processo. O Quadro 10 apresenta o resultado obtido pela Consulta 3.

Quadro 10. Resultado obtido pela Consulta 3.

step_exec	step_title	step_start	step_end	step_executor
http://greco.ppgi.ufrj.br/provenance/resource/WorkflowExecutionProcess/INPUT___PESQUISADOR121728	"INPUT___PESQUISADOR17_TRANSF_LATTES12 : Execution 28"	2013-11-05T13:32:48Z	2013-11-05T13:32:49Z	http://greco.ppgi.ufrj.br/provenance/resource/Agent/EXPEDITO1

5.2.4 Consulta 4 - Recuperação da fonte de dados de onde um recurso publicado no contexto de Linked Data foi derivado

A consulta 4 é um exemplo de consulta SPARQL que tem por objetivo recuperar, através dos dados de proveniência, a fonte de dados de onde um recurso publicado no contexto de Linked Data foi derivado. Neste exemplo, o recurso em questão é uma pesquisadora participante do programa Grupo de Trabalho RNP, representada pela IRI *http://www.rnp.br/resource/Researcher/MARIA_LUIZA_MACHADO_CAMPOS* e publicada no grafo *http://www.rnp.br*.

Consulta 4: De qual fonte de dados derivaram as informações da pesquisadora *http://www.rnp.br/resource/Researcher/MARIA_LUIZA_MACHADO_CAMPOS* do grafo *http://www.rnp.br* ?

```
1  PREFIX cogs: <http://vocab.deri.ie/cogs#>
2  PREFIX dc: <http://purl.org/dc/terms/>
3  PREFIX opmw: <http://www.opmw.org/ontology/>
4  PREFIX prov: <http://www.w3.org/ns/prov#>
5
6  SELECT ?w_title ?step_extract_title ?source
7  FROM <http://greco.ppgi.ufrj.br/provenance>
8  WHERE
9  {
10       ?step_load opmw:isStepOfTemplate ?w_comp .
11       ?step_load a cogs:InsertTriple .
12       ?step_load opmw:uses ?param_graph .
13       ?param_graph a cogs:RDFGraph .
14       ?step_load_exec opmw:correspondsToTemplateProcess ?step_load .
15       ?step_load_exec prov:used ?param_graph_exec .
16       ?param_graph_exec opmw:correspondsToTemplateArtifact
   ?param_graph .
17       ?param_graph_exec prov:value "http://www.rnp.br" .
18
19       ?step_transf opmw:isStepOfTemplate ?w_comp .
20       ?step_transf a cogs:RDFDataPropertyMapping .
21       ?step_transf_exec opmw:correspondsToTemplateProcess
   ?step_transf .
22       ?data prov:wasGeneratedBy ?step_transf_exec .
23       ?data prov:value
   "http://www.rnp.br/resource/Researcher/MARIA_LUIZA_MACHADO_CAMPOS" .
24
25       ?step_transf_exec cogs:precededBy+ ?step_extract_exec .
26       ?step_extract_exec opmw:correspondsToTemplateProcess
   ?step_extract .
27       ?step_extract a cogs:Extraction .
28       ?step_extract opmw:uses ?param_source .
29       ?param_source a cogs:Source .
30       ?param_source_exec opmw:correspondsToTemplateArtifact
   ?param_source .
31       ?param_source_exec prov:value ?source .
32
33       ?w_comp opmw:versionNumber ?w_version .
34       ?w_comp dc:title ?w_title .
35       ?step_extract dc:title ?step_extract_title .
36 }
37 GROUP BY ?w_version ?w_title ?step_extract_title ?source
38 HAVING ?w_version = MAX(?w_version)
```

Na consulta 4, as linhas 10 e 11 recuperam as composições dos passos de *workflow* do tipo *cogs:InsertTriple*, que são responsáveis pelo processo de carga das triplas RDF no banco de triplas. As linhas 12 e 13 recuperam o parâmetro de definição do grafo RDF, onde a tripla

deverá ser inserida. As linhas 14, 15, 16 e 17 recuperam as instâncias das execuções dos passos recuperados pelas linhas 10 e 11, cujo parâmetro grafo RDF foi definido com o valor do grafo "http://www.rnp.br". As linhas 19, 20 e 21 recuperam as composições dos passos de *workflow* e suas respectivas execuções, responsáveis pelo processo de mapeamento dos dados de entrada nos componentes de uma tripla RDF (sujeito, predicado e objeto), sendo o objeto um valor literal. Destes passos recuperados, as linhas 22 e 23 filtram aquele que gerou o valor da IRI da pesquisadora participante do programa Grupo de Trabalho RNP. A linha 25 utiliza o operador + do SPARQL 1.1 na propriedade *cogs:precededBy*, para recuperar todos os passos do caminho desde o início do *workflow* até o passo filtrado pelas linhas 22 e 23. O operador + é um dos operadores de caminho de propriedades (*property path*) suportados pelo SPARQL 1.1[40]. Deste conjunto de passos recuperados, as instâncias da execução dos passos responsáveis pelo processo de extração são filtrados pelas linhas 26 e 27. As linhas 28, 29, 30 e 31 recuperam o parâmetro de definição da fonte de dados e o valor utilizado nas execuções dos processos de extração filtrados. Por fim, a linha 33, 34 e 35 recuperam respectivamente a versão do *workflow*, o nome do *workflow* e o nome do passo responsável pelo processo de extração e as linhas 37 e 38, de maneira semelhante à Consulta 2, filtram a última versão do *workflow*. O Quadro 11 apresenta o resultado obtido pela Consulta 4.

Quadro 11. Resultado obtido pela Consulta 4.

w_title	step_extract_title	source
"transf_rnp"	"Input XML - Pesquisador"	"D:/projetos/provenance_collector/rnp/grupo_trabalho.xml"

5.2.5 Consulta 5 - Recuperação dos dados de proveniência prospectiva e de dados de proveniência retrospectiva da parametrização de um processo

A consulta 5 é um exemplo de consulta SPARQL que tem por objetivo recuperar tanto a proveniência prospectiva, quanto a proveniência retrospectiva da parametrização utilizada em um processo do *workflow* de publicação. Neste exemplo, o processo em questão é responsável pela integração entre os Grupos de Pesquisa do CNPq e os projetos cadastrados nos Currículos Lattes. A aplicação da ontologia OPMW foi fundamental para explorar de maneira não ambígua os dados de proveniência prospectiva e os dados de proveniência retrospectiva publicados e interligados no grafo *http://greco.ppgi.ufrj.br/provenance*.

[40] http://www.w3.org/TR/sparql11-property-paths/#complex_paths

Consulta 5: Quais os parâmetros utilizados no passo "Merge Grupos de Pesquisa CNPq x Projetos Lattes", que representa o processo que integra os projetos cadastrados nos Currículos Lattes aos Grupos de Pesquisa do CNPq, e quais os valores utilizados na execução realizada por Rogers em 04/11/2013 às 14:30:06 ?

```
1  PREFIX dc: <http://purl.org/dc/terms/>
2  PREFIX foaf: <http://xmlns.com/foaf/0.1/>
3  PREFIX opmw: <http://www.opmw.org/ontology/>
4  PREFIX prov: <http://www.w3.org/ns/prov#>
5
6  SELECT ?param_name ?param_value
7  FROM <http://greco.ppgi.ufrj.br/provenance>
8  WHERE
9  {
10       ?step_comp dc:title "Merge Grupos de Pesquisa CNPq x Projetos
   Lattes" .
11       ?step_exec opmw:correspondsToTemplateProcess ?step_comp .
12       ?step_exec prov:wasAssociatedWith ?agent .
13       ?agent foaf:name "ROGERS" .
14       ?step_exec prov:startedAtTime "2013-11-
   04T14:30:06Z"^^xsd:dateTime .
15
16       ?step_comp opmw:uses ?param_comp .
17       ?param_comp a opmw:ParameterVariable .
18       ?step_exec prov:used ?param_exec .
19       ?param_exec opmw:correspondsToTemplateArtifact ?param_comp .
20       ?param_comp dc:title ?param_name .
21       ?param_exec opmw:hasValue ?param_value .
22 }
23 ORDER BY ?param_name ?param_value
```

Na consulta 5, a linha 10 filtra a composição do passo do *workflow* denominado "Merge Grupos de Pesquisa CNPq x Projetos Lattes" e a linha 11 recupera as instâncias de suas execuções. As linhas 12, 13 e 14 filtram a execução associada ao agente determinado e iniciada em 04/11/2013 às 14:30:06. As linhas 16 e 17 recuperam os parâmetros definidos na composição do passo "Merge Grupos de Pesquisa CNPq x Projetos Lattes" e as linhas 18 e 19 recuperam as instâncias dos parâmetros utilizadas pela execução do passo filtrado pelas linhas 12, 13 e 14. Por fim, as linhas 20 e 21 recuperam, respectivamente, o nome e o valor de cada parâmetro. A linha 23 define a ordenação crescente da lista de resultados pelo nome e pelo valor dos parâmetros recuperados. O Quadro 12 apresenta o resultado obtido pela Consulta 5.

Quadro 12. Resultado obtido pela Consulta 5.

param_name	param_value
"join_type"	"INNER"
"key1#0"	"id"
"key1_count"	"1"
"key2#0"	"lattes_rschr_id"
"key2_count"	"1"
"step1"	"Sort Rows - GP CNPq"
"step2"	"Sort Rows - Projeto Lattes"

5.2.6 Consultas 6 e 7 - Utilização dos dados de proveniência para avaliar a consistência de dados de domínio provenientes de diferentes fontes

A Consulta 6 e a Consulta 7 são exemplos de consultas SPARQL que podem ser utilizadas em conjunto para avaliar a consistência de dados de domínio provenientes de diferentes fontes de dados. Neste caso, as consultas oferecem um apoio para avaliar a consistência entre os projetos de um Grupo de Pesquisa do CNPq e os Grupos de Pesquisa de um projeto desenvolvido por um Grupo de Trabalho RNP.

Consulta 6: Do currículo de qual pesquisador é proveniente cada projeto da lista de projetos do Grupo de Pesquisa "GRECO - Grupo Engenharia do Conhecimento" do CNPq ?

Consulta 7: Quais pesquisadores foram utilizados para relacionar os Grupos de Trabalho do RNP ao Grupo de Pesquisa GRECO ?

```
1  PREFIX cnpq: <http://www.cnpq.br/ontology/>
2  PREFIX cnpq-prop: <http://www.cnpq.br/property/>
3  PREFIX dc: <http://purl.org/dc/terms/>
4  PREFIX lattes: <http://lattes.cnpq.br/ontology/>
5  PREFIX lattes-prop: <http://lattes.cnpq.br/property/>
6  PREFIX opmw: <http://www.opmw.org/ontology/>
7  PREFIX prov: <http://www.w3.org/ns/prov#>
8
9  SELECT DISTINCT ?iri_lattes_prj ?lattes_prj_name ?iri_lattes_pesq
   ?lattes_pesq_name
10 FROM NAMED <http://greco.ppgi.ufrj.br/provenance>
11 FROM NAMED <http://www.cnpq.br>
12 FROM NAMED <http://lattes.cnpq.br>
13 WHERE
14 {
15       GRAPH <http://greco.ppgi.ufrj.br/provenance>
16       {
17             ?step_comp dc:title "Merge Grupos de Pesquisa CNPq x
   Projetos Lattes" .
18             ?field_iri_lattes_pesq opmw:isGeneratedBy ?step_comp .
19             ?field_iri_lattes_pesq dc:title "lattes_rschr_uri" .
20             ?field_iri_lattes_prj opmw:isGeneratedBy ?step_comp .
21             ?field_iri_lattes_prj dc:title "lattes_prj_uri" .
22             ?field_iri_cnpq_gp opmw:isGeneratedBy ?step_comp .
23             ?field_iri_cnpq_gp dc:title "uri_rg" .
24             ?step_exec opmw:correspondsToTemplateProcess ?step_comp
   .
25             ?data_iri_lattes_pesq prov:wasGeneratedBy ?step_exec .
26             ?data_iri_lattes_pesq opmw:correspondsToTemplateArtifact
   ?field_iri_lattes_pesq .
27             ?data_iri_lattes_pesq prov:value ?value_iri_lattes_pesq
   .
28             ?data_iri_lattes_prj prov:wasGeneratedBy ?step_exec .
29             ?data_iri_lattes_prj opmw:correspondsToTemplateArtifact
   ?field_iri_lattes_prj .
30             ?data_iri_lattes_prj prov:value ?value_iri_lattes_prj .
31             ?data_iri_cnpq_gp prov:wasGeneratedBy ?step_exec .
32             ?data_iri_cnpq_gp opmw:correspondsToTemplateArtifact
   ?field_iri_cnpq_gp .
33             ?data_iri_cnpq_gp prov:value ?value_iri_cnpq_gp .
34             BIND(IRI(?value_iri_lattes_pesq) AS ?iri_lattes_pesq) .
35             BIND(IRI(?value_iri_lattes_prj) AS ?iri_lattes_prj) .
36             BIND(IRI(?value_iri_cnpq_gp) AS ?iri_cnpq_gp) .
37       }
38       GRAPH <http://www.cnpq.br>
39       {
40             ?iri_cnpq_gp a cnpq:ResearchGroup .
41             ?iri_cnpq_gp cnpq-prop:name "GRECO - Grupo Engenharia do
   Conhecimento"@pt .
```

```
42        }
43        GRAPH <http://lattes.cnpq.br>
44        {
45              ?iri_lattes_pesq lattes-prop:name ?lattes_pesq_name .
46              ?iri_lattes_prj lattes-prop:name ?lattes_prj_name .
47        }
48  }
```

Na consulta 6, as linhas 17 a 26 recuperam a definição dos campos gerados pelo passo "Merge Grupos de Pesquisa CNPq x Projetos Lattes" que contém a URI do pesquisador do Currículo Lattes ("lattes_rschr_uri"), a URI do projeto cadastrado no currículo deste pesquisador ("lattes_prj_uri") e a URI do Grupo de Pesquisa CNPq ("uri_rg"). Conforme mencionao na Consulta 5, o passo "Merge Grupos de Pesquisa CNPq x Projetos Lattes" representa o processo que integra os projetos cadastrados nos Currículos Lattes aos Grupos de Pesquisa do CNPq.

As linhas 25 a 32 recuperam os valores dos campos lattes_rschr_uri, lattes_prj_uri e uri_rg nas instâncias de execução do passo "Merge Grupos de Pesquisa CNPq x Projetos Lattes" e as linhas 34, 35 e 36 transformam os valores literais recuperados em IRIs, colocando os valores transformados nas variáveis *?iri_lattes_pesq*, *?iri_lattes_prj* e *?iri_cnpq_gp*. As linhas 40 e 41 filtram os projetos do grafo *http://www.cnpq.br*, cujo nome é GRECO - Grupo Engenharia do Conhecimento e as linhas 45 e 46 recuperam o nome do pesquisador e o nome do projeto publicados no grafo *http://lattes.cnpq.br*. O Quadro 13 apresenta uma amostra do resultado obtido pela Consulta 6.

Quadro 13. Amostra do resultado obtido pela Consulta 6.

iri_lattes_prj	lattes_prj_name	iri_lattes_pesq	lattes_pesq_ name
http://lattes.cnpq.br/ resource/Project/407	"GT-LinkedDataBR: Exposição, compartilhamento e conexão de recursos de dados abertos na Web (Linked Open Data)"@pt	http://lattes.cnpq.br/resource/ Researcher/K4781460T3	"Maria Luiza Machado Campos"
http://lattes.cnpq.br/ resource/Project/482	"Identificação e Análise de Redes Sociais Complexas"@pt	http://lattes.cnpq.br/resource/ Researcher/K4761314U5	"Jonice de Oliveira Sampaio"
http://lattes.cnpq.br/ resource/Project/665	"Núcleo de Pesquisa de Sistemas Computacionais Complexos para a Gestão de Emergências"@pt	http://lattes.cnpq.br/resource/ Researcher/K4717449A7	"Kelli de Faria Cordeiro"

A estrutura da Consulta 7, é similar à estrutura da Consulta 6, com a diferença de que, no lugar do passo "Merge Grupos de Pesquisa CNPq x Projetos Lattes", é utilizado o passo

"Merge Grupos de Trabalho RNP x Grupos de Pesquisa CNPq". Este passo representa o processo o processo que relaciona os Grupos de Pesquisa do CNPq aos Grupos de Trabalho da RNP:

```
1  PREFIX cnpq: <http://www.cnpq.br/ontology/>
2  PREFIX cnpq-prop: <http://www.cnpq.br/property/>
3  PREFIX dc: <http://purl.org/dc/terms/>
4  PREFIX lattes: <http://lattes.cnpq.br/ontology/>
5  PREFIX lattes-prop: <http://lattes.cnpq.br/property/>
6  PREFIX opmw: <http://www.opmw.org/ontology/>
7  PREFIX prov: <http://www.w3.org/ns/prov#>
8  PREFIX rnp: <http://www.rnp.br/ontology/>
9  PREFIX rnp-prop: <http://www.rnp.br/property/>
10
11 SELECT DISTINCT ?iri_rnp_gt ?rnp_gt_name ?iri_lattes_pesq
   ?lattes_pesq_name
12 FROM NAMED <http://greco.ppgi.ufrj.br/provenance>
13 FROM NAMED <http://www.cnpq.br>
14 FROM NAMED <http://lattes.cnpq.br>
15 FROM NAMED <http://www.rnp.br>
16 WHERE
17 {
18       GRAPH <http://greco.ppgi.ufrj.br/provenance>
19       {
20             ?step_comp dc:title "Merge Grupos de Trabalho RNP x
   Grupos de Pesquisa CNPq" .
21             ?field_iri_cnpq_pesq opmw:isGeneratedBy ?step_comp .
22             ?field_iri_cnpq_pesq dc:title "cnpq_p_uri" .
23             ?field_iri_cnpq_gp opmw:isGeneratedBy ?step_comp .
24             ?field_iri_cnpq_gp dc:title "cnpq_gp_uri" .
25             ?field_iri_rnp_gt opmw:isGeneratedBy ?step_comp .
26             ?field_iri_rnp_gt dc:title "uri_gt" .
27             ?step_exec opmw:correspondsToTemplateProcess ?step_comp
   .
28             ?data_iri_cnpq_pesq prov:wasGeneratedBy ?step_exec .
29             ?data_iri_cnpq_pesq opmw:correspondsToTemplateArtifact
   ?field_iri_cnpq_pesq .
30             ?data_iri_cnpq_pesq prov:value ?value_iri_cnpq_pesq .
31             ?data_iri_cnpq_gp prov:wasGeneratedBy ?step_exec .
32             ?data_iri_cnpq_gp opmw:correspondsToTemplateArtifact
   ?field_iri_cnpq_gp .
33             ?data_iri_cnpq_gp prov:value ?value_iri_cnpq_gp .
34             ?data_iri_rnp_gt prov:wasGeneratedBy ?step_exec .
35             ?data_iri_rnp_gt opmw:correspondsToTemplateArtifact
   ?field_iri_rnp_gt .
36             ?data_iri_rnp_gt prov:value ?value_iri_rnp_gt .
37             BIND(IRI(?value_iri_cnpq_pesq) AS ?iri_cnpq_pesq) .
38             BIND(IRI(?value_iri_cnpq_gp) AS ?iri_cnpq_gp) .
```

```
39              BIND(IRI(?value_iri_rnp_gt) AS ?iri_rnp_gt) .
40        }
41        GRAPH <http://www.cnpq.br>
42        {
43              ?iri_cnpq_gp a cnpq:ResearchGroup .
44              ?iri_cnpq_gp cnpq-prop:name "GRECO - Grupo Engenharia do
   Conhecimento"@pt .
45              ?iri_cnpq_pesq owl:sameAs ?iri_lattes_pesq .
46        }
47        GRAPH <http://lattes.cnpq.br>
48        {
49            ?iri_lattes_pesq a lattes:Researcher .
50              ?iri_lattes_pesq lattes-prop:name ?lattes_pesq_name .
51        }
52        GRAPH <http://www.rnp.br>
53        {
54              ?iri_rnp_gt a rnp:WorkingGroup .
55              ?iri_rnp_gt rnp-prop:name ?rnp_gt_name .
56        }
57  }
```

O Quadro 14 apresenta uma amostra do resultado obtido pela Consulta 7.

Quadro 14. Amostra do resultado obtido pela Consulta 7.

iri_rnp_gt	rnp_gt_name	iri_lattes_pesq	lattes_pesq_name
http://www.rnp.br/resource/WorkingGroup/EXPOSICAO_COMPARTILHAMENTO_E_CONEXAO_DE_RECURSOS_DE_DADOS_ABERTOS_NA_WEB411	"Exposição, Compartilhamento e Conexão de Recursos de Dados Abertos na Web"@pt	http://lattes.cnpq.br/resource/Researcher/K4781460T3	"Maria Luiza Machado Campos"
http://www.rnp.br/resource/WorkingGroup/EXPOSICAO_COMPARTILHAMENTO_E_CONEXAO_DE_RECURSOS_DE_DADOS_ABERTOS_NA_WEB411	"Exposição, Compartilhamento e Conexão de Recursos de Dados Abertos na Web"@pt	http://lattes.cnpq.br/resource/Researcher/K4761314U5	"Jonice de Oliveira Sampaio"
http://www.rnp.br/resource/WorkingGroup/EXPOSICAO_COMPARTILHAMENTO_E_CONEXAO_DE_RECURSOS_DE_DADOS_ABERTOS_NA_WEB411	"Exposição, Compartilhamento e Conexão de Recursos de Dados Abertos na Web"@pt	http://lattes.cnpq.br/resource/Researcher/K4717449A7	"Kelli de Faria Cordeiro"

Caso o resultado da Consulta 7 retorne um Grupo de Pesquisa relacionado a um projeto da RNP e este projeto não se encontre na lista de projetos do Grupo de Pesquisa no CNPq, é possível investigar o motivo desta inconsistência por meio dos dados de proveniência recuperados. Assim, com os dados de proveniência em mãos, é possível avaliar se a inconsistência deve-se a algum problema com o processo de publicação dos dados de domínio

como Linked Data ou, então, se não é um problema com os processos de publicação, mas simplesmente um erro cometido por um pesquisador do CNPq ao declarar seus projetos no Currículo Lattes.

5.3 Análise quantitativa dos tempos de execução e dos números de triplas RDF gerados pelos *workflows* ETL do exemplo de aplicação

Esta subseção conclui a apresentação do exemplo de aplicação da abordagem *ETL4LinkedProv* com análises quantitativas dos tempos de execução e dos números de triplas RDF geradas por cada *workflow* ETL, considerando diferentes níveis de granulosidade da proveniência.

A infraestrutura utilizada para a execução do exemplo de aplicação contemplou a edição comunitária do SGBD MySQL[41] versão 5.0.51a, a edição *open source* do banco de triplas Virtuoso[42] versão 7.0.0, a edição comunitária (Community Edition - CE) do Kettle[43] versão 4.3.0-stable e a plataforma Java Standard Edition[44] (SE) versão 1.7.0_13. Estas plataformas estavam instaladas localmente em um *notebook* com processador Intel Core i3-2370M de 2.4 GHz e 6 GB de memória principal, executando o sistema operacional Windows 7 na versão 64 bits.

A primeira análise quantitativa consistiu em comparar os tempos de execução dos *workflows* ETL de publicação dos dados de domínio, considerando 3 cenários distintos. No primeiro cenário, os 3 *workflows* ETL, que publicaram os dados das fontes CNPq-Lattes, CNPq-Grupos de Pesquisa e RNP-Grupos de Trabalho respectivamente, foram executados sem estarem encapsulados pelo Agente Coletor de Proveniência. No segundo e no terceiro cenário, os mesmos *workflows* foram executados encapsulados pelo Agente Coletor de Proveniência, com a diferença de que, no segundo cenário, o Agente Coletor de Proveniência foi configurado com nenhum tipo de *step* selecionado para ter a proveniência coletada com granulosidade fina e, no terceiro cenário, o Agente Coletor de Proveniência foi configurado com todos os tipos de *step* selecionados para terem a proveniência coletada com granulosidade fina. O Quadro 15 apresenta o resultado desta primeira análise quantitativa.

[41] http://www.mysql.com/products/community

[42] http://www.openlinksw.com/dataspace/doc/dav/wiki/Main

[43] http://sourceforge.net/projects/pentaho/files/Data%20Integration

[44] http://www.oracle.com/technetwork/java/javase/overview/index.html

Quadro 15. Tabela de comparação dos tempos de execução dos *workflows* ETL do exemplo de aplicação.

***Workflow* ETL**	**Fonte de dados**	**Tipo da fonte**	**Entrada - Registros**	**Saída - Triplas**	**t_1**[45]	**t_2**[46]	**t_3**[47]
transf_lattes	CNPq-Lattes	MySQL	527	2242	00:00:02	00:00:03	00:06:10
transf_cnpq	CNPq-Grupos de Pesquisa	MySQL	106	425	00:00:01	00:00:02	00:02:10
transf_rnp	RNP-Grupos de Trabalho	XML	577	2267	00:00:02	00:00:04	00:09:50
Total			1.210	4.934	00:00:05	00:00:09	00:18:10

A partir deste resultado, podemos concluir que o impacto da atuação do Agente Coletor de Proveniência no tempo de execução do *workflow* ETL encapsulado está diretamente ligado ao nível de granulosidade configurado para a realização da coleta dos dados de proveniência. Sem o nível de granulosidade fina habilitado, os tempos de execução dos *workflows* ETL com atuação do Agente Coletor de Proveniência (t_2) foram muito próximos dos tempos de execução sem a atuação do Agente Coletor de Proveniência (t_1). Em contrapartida, ao serem executados com o nível de granulosidade fina habilitado, os tempos de execução dos *workflows* ETL (t_3) sofrem um significativo aumento.

A segunda análise quantitativa consistiu em comparar o número de triplas RDF de proveniência publicadas e os tempos de execução dos *workflows* ETL de proveniência, considerando 2 cenários distintos. No primeiro cenário, o Agente Coletor de Proveniência foi configurado com nenhum tipo de *step* selecionado para ter a proveniência coletada com

[45] t_1 é o tempo (hh:mm:ss) de execução do *workflow* ETL sem a atuação do Agente Coletor de Proveniência.

[46] t_2 é o tempo (hh:mm:ss) de execução do *workflow* ETL com a atuação do Agente Coletor de Proveniência configurado com nenhum tipo de step selecionado para ter a proveniência coletada com granulosidade fina.

[47] t_3 é o tempo (hh:mm:ss) de execução do *workflow* ETL com a atuação do Agente Coletor de Proveniência configurado com todos os tipos de steps selecionados para terem a proveniência coletada com granulosidade fina.

granulosidade fina e, no segundo cenário, o Agente Coletor de Proveniência foi configurado com todos os tipos de *steps* selecionados para terem a proveniência coletada com granulosidade fina. O Quadro 16 apresenta o resultado da segunda análise quantitativa.

Quadro 16. Tabela de comparação da execução dos *workflows* ETL de publicação dos dados de proveniência com diferentes níveis de granulosidade.

	Execução 1[48]		**Execução 2**[49]	
***Workflow* ETL de Proveniência**	**Saída 1 - Triplas**	**t_1**[50]	**Saída 2 - Triplas**	**t_2**[51]
transf_provenance_agent	10	00:00:00	10	00:00:00
transf_provenance_provo	714	00:00:02	1.086.114	00:12:13
transf_prospective_provenance_opmw	373	00:00:01	8.685	00:00:06
transf_retrospective_provenance_opmw	872	00:00:02	724.577	00:06:33
transf_provenance_cogs_step_parameter	106	00:00:01	1.534	00:00:02
transf_provenance_cogs_etl_sequence	24	00:00:01	243.302	00:01:19
Total	2.099	00:00:09	2.064.222	00:20:13

De maneira semelhante ao resultado da primeira análise, podemos concluir que o impacto do Agente Coletor de Proveniência no tempo de execução dos *workflows* ETL de proveniência está diretamente ligado ao nível de granulosidade configurado para a realização da coleta dos dados de proveniência. Este impacto se intensifica nos *workflows* ETL responsáveis por publicar os dados de proveniência retrospectiva.

O número de triplas RDF publicadas, tanto de dados de domínio, quanto de dados de proveniência, segue o modelo representado pela fórmula:

[48] A Execução 1 publicou a proveniência coletada pelo Agente Coletor de Proveniência configurado com nenhum tipo de step selecionado para ter a proveniência coletada com granulosidade fina.

[49] A Execução 2 publicou a proveniência coletada pelo Agente Coletor de Proveniência configurado com todos os tipos de steps selecionados para ter a proveniência coletada com granulosidade fina.

[50] t_1 é o tempo (hh:mm:ss) da Execução 1.

[51] t_2 é o tempo (hh:mm:ss) da Execução 2.

$$T = \sum_{i=1}^{n} (I_i * A_i + I_i * P_i) + \sum_{\substack{i=1 \\ j=1}}^{n} I_i * R_{ij}$$

Nesta fórmula, *T* é o número de triplas RDF publicadas, *n* é o número de entidades extraídas da fonte de dados pelo *workflow* ETL, I_i é o número de instâncias extraídas da entidade *i*, A_i é o número de tipos RDF relacionados à entidade *i*, P_i é o número de propriedades extraídas da entidade *i* e R_{ij} é o número de relacionamentos entre a entidade *i* e a entidade *j* extraídos. O Quadro 17 apresenta os valores das variáveis A, P, R e I das entidades *Parâmetro (Execução)*, *Campo de Dados (Execução)* e *Linha de Dados (Execução)* nas duas execuções dos *workflows* ETL que publicaram os dados de proveniência retrospectiva utilizando a ontologia PROV-O (tranf_provenance_provo) e a ontologia OPMW (transf_retrospective_provenance_opmw). Estes *workflows* ETL foram os que apresentaram as maiores diferenças entre o número de triplas publicadas nas execuções com e sem o nível de granulosidade fina habilitado, sendo as entidades *Parâmetro (Execução)*, *Campo de Dados (Execução)* e *Linha de Dados (Execução)* as principais responsáveis pela diferença registrada.

Quadro 17. Valores das variáveis A, P, R e I das entidades Parâmetro (Execução), Campo de Dados (Execução) e Linha de Dados (Execução) na Execução 1 e na Execução 2 dos *transformations* tranf_provenance_provo e transf_retrospective_provenance_opmw.

***Workflow* ETL**	**Entidade**	**A**	**P**	**R**	**I (Execução 1)**	**I (Execução 2)**
tranf_provenance_ provo	Parâmetro (Execução)	1	1	1	0	831
	Campo de Dados (Execução)	1	1	2	0	101.116
	Linha de Dados (Execução)	2	0	3	0	31.055
transf_ retrospective_ provenance_opmw	Parâmetro (Execução)	2	1	2	0	831
	Campo de Dados (Execução)	2	1	3	0	101.116

6 Conclusão

Um conjunto de tecnologias e boas práticas da Web Semântica tem sido adotado por um número crescente de provedores de dados para publicar e interligar dados de diversas naturezas, utilizando os princípios de Linked Data. O projeto Linking Open Data[52] teve um papel relevante nesta atividade, ao identificar conjuntos de dados disponibilizados com licença aberta, para convertê-los e publicá-los na Web. Encontramos também organizações que adotaram os princípios de Linked Data para publicar e interligar seus dados, a exemplo das iniciativas governamentais no Reino Unido (SHERIDAN, TENNISON, 2010), nos Estados Unidos (HENDLER, J. et al., 2012) e no Brasil (BREITMAN et al., 2012).

O processo particular de publicação de Linked Data realizado dentro dos limites de uma organização consiste normalmente da extração de dados de múltiplas fontes heterogêneas, seguida da execução, através do uso de diversas ferramentas, de uma gama de operações de preparação, modificação e integração dos dados antes de carregá-los em um banco de triplas RDF (CORDEIRO et al., 2011). Desta maneira, coletar e disponibilizar dados de proveniência relacionados a este processo de publicação de Linked Data e a suas respectivas operações de extração, preparação, modificação, integração e carga apresentam um subsídio adicional importante para a posterior análise da qualidade e da confiabilidade dos dados publicados. Contudo, na análise deste processo, verifica-se uma carência de ferramental para coletar e publicar os dados de proveniência, além da dificuldade para interligar a proveniência da composição (prospectiva) com a proveniência da execução (retrospectiva) do processo de publicação e os dados publicados com seus respectivos dados de proveniência.

Visando tratar estes problemas, esta dissertação propôs a abordagem *ETL4LinkedProv* para coleta e publicação dos dados de proveniência do processo de publicação de Linked Data realizado dentro dos limites de uma organização. A abordagem *ETL4LinkedProv* baseia-se na experiência de trabalhos anteriores (CAMPOS, M. L. M., GUIZZARDI, 2010; FREITAS, KÄMPGEN, OLIVEIRA et al., 2012; OMITOLA et al., 2012), que utilizaram *workflows* ETL para publicar Linked Data, e introduz um novo componente, denominado Agente Coletor de Proveniência, que possui a função de capturar, interligar e armazenar temporariamente os dados

[52] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

de proveniência prospectiva e retrospectiva, para posterior publicação dos mesmos como triplas RDF.

A ferramenta Pentaho Data Integration (Kettle) foi utilizada para implementar um protótipo da abordagem *ETL4LinkedProv* e um exemplo de aplicação envolvendo dados reais relacionados às organizações CNPq e RNP foi desenvolvido. Com ele, buscou-se evidenciar a contribuição da abordagem para suprir a carência de ferramental para coletar e publicar os dados de proveniência durante o processo de publicação de Linked Data, assim como tratar a dificuldade de interligação entre os dados de domínio e os dados de proveniência publicados.

6.1 Contribuições

As contribuições deste trabalho que se destacam são:

- a definição da abordagem *ETL4LinkedProv*, de maneira independente de uma ferramenta de ETL específica, para coleta e publicação dos dados de proveniência do processo de publicação de Linked Data realizado dentro das organizações, por meio de um *workflow* ETL;
- a definição de uma estratégia de apoio semântico à proveniência, que:
  - (i) provê a interoperabilidade da proveniência, habilitada pela utilização do padrão PROV do W3C;
  - (ii) representa, de maneira não ambígua, a proveniência prospectiva e a proveniência retrospectiva, por meio da ontologia OPMW;
  - (iii) representa, de maneira expressiva, os conceitos de ETL, por meio da ontologia Cogs;
  - (iv) possibilita a representação da semântica específica ao domínio de uma aplicação, por meio de uma ontologia de aplicação complementar;
- a implementação de um protótipo da abordagem *ETL4LinkedProv* na ferramenta Pentaho Data Integration (Kettle), incluindo a produção dos seguintes artefatos disponíveis no GitHub[53] e no *site* do Grupo de Engenharia do Conhecimento (GRECO) do PPGI/UFRJ[54]:
  - (i) o componente *job entry* Provenance Collector Agent, que é a implementação do Agente Coletor de Proveniência por meio da API Java do Kettle;

[53] https://github.com/rogersmendonca

[54] http://greco.ppgi.ufrj.br/lodbr/index.php/principal/etl4linkedprov

(ii) o componente *step* NTriple Generator, que oferece a facilidade de geração de sentenças RDF no formato NTriple, podendo ser utilizado em conjunto com os 4 *steps* do ETL4LOD produzidos pelo projeto LinkedDataBR;

(iii) o conjunto de *transformations* (arquivos .ktr) do Kettle, que representam os *workflows* ETL de publicação dos dados de proveniência como Linked Data;

- a aplicação da abordagem *ETL4LinkedProv* em um cenário envolvendo a publicação de dados reais do CNPq, dados reais da RNP e dos respectivos dados de proveniência como Linked Data, com posterior exploração conjunta dos dados publicados por meio de consultas SPARQL;
- a adaptação da taxonomia definida por CRUZ, CAMPOS, M. L. M. e MATTOSO (2009), para classificar a proveniência no contexto de Linked Data;
- a integração da abordagem definida neste trabalho em uma arquitetura de suporte à publicação e interligação dos dados de proveniência na Web de Dados, que envolve também a abordagem de proveniência desenvolvida por DE LA CERDA e CAVALCANTI (2012).

Com relação à última contribuição, a arquitetura proposta envolve, além da proveniência sobre o Processo de Preparação e Transformação dos Dados abordado nesta dissertação, a proveniência do Processo de Interligação dos Dados, mencionado nos capítulos 2 e 4. Esta arquitetura foi descrita no artigo "LOP - *Capturing and Linking Open Provenance on LOD Cycle*", apresentado e publicado nos anais do 5º seminário *Semantic Web Information Management* (SWIM)[55] 2013, realizado em conjunto com a conferência *Conference on Management of Data* (SIGMOD) 2013.

6.2 Limitações

No decorrer do desenvolvimento deste trabalho, identificamos limitações que restringem a utilização da abordagem proposta para a coleta, publicação e exploração dos dados de proveniência no contexto de Linked Data. Entre as limitações identificadas, destacam-se:

- a falta de uma interface gráfica para apoiar a exploração dos dados de proveniência publicados. Embora a abordagem ofereça a possibilidade de exploração por meio de

[55] http://pamir.dia.uniroma3.it:8080/SWIM2013

consultas SPARQL definidas explicitamente, de maneira geral, as tecnologias relacionadas ao contexto de Linked Data ainda são pouco difundidas, mesmo entre o público mais técnico (RIBEIRO, 2012);

- a restrição do tratamento dos dados de proveniência limitados à etapa de extração do ciclo de vida de Linked Data. Esta limitação deveu-se ao foco do trabalho no processo de publicação de Linked Data realizado dentro das organizações, denominado Processo de Preparação e Tranformação dos Dados. A etapa de extração, no entanto, foi estendida, levando em consideração, as atividades de limpeza, consolidação, agregação e integração dos dados, além da atividade de extração;
- a necessidade de uma estratégia para gerenciar o grande volume de dados gerado pela publicação da proveniência de granulosidade fina.

6.3 Trabalhos futuros

Por fim, dentre as sugestões de trabalhos futuros a serem desenvolvidos dentro do tema desta dissertação, destacam-se:

- o desenvolvimento de uma ontologia, fundamentada em uma ontologia de topo, para substituir ou integrar o conjunto de ontologias de proveniência definido na estratégia de apoio semântico para a proveniência no contexto de Linked Data adotada neste trabalho;
- a análise de diferentes critérios e métricas de qualidade, com identificação de quais dados de proveniência apoiam cada critério/métrica de qualidade no contexto de Linked Data;
- a definição de uma estratégia na área de Big Data para gerenciar o grande volume de dados gerado pela publicação da proveniência de granulosidade fina no contexto de Linked Data.

Com relação ao último trabalho sugerido, a plataforma ProvBase[56] apresenta-se como uma ferramenta potencial para apoiar a estratégia na área de Big Data para gerenciar a proveniência com nível de granulosidade fina.

[56] http://portal.utpa.edu/utpa_main/daa_home/coecs_home/provbase_home

Referências

ALEXANDER, K.; CYGANIAK, R.; HAUSENBLAS, M.; ZHAO, JUN. **Describing Linked Datasets - On the Design and Usage of voiD, the "Vocabulary of Interlinked Datasets".** Proceedings of the 2nd Workshop on Linked Data on the Web LDOW2009. **Anais...** Citeseer, 2009. Disponível em: <http://events.linkeddata.org/ldow2009/papers/ldow2009_paper17.pdf>.

ALTINTAS, I; BARNEY, O.; JAEGER-FRANK, E. Provenance Collection Support in the Kepler Scientific Workflow System. (L. Moreau & I. T. Foster, Eds.)**IPAW**, v. 4145, p. 118–132, 2006. Springer. Disponível em: <http://dx.doi.org/10.1007/11890850_14>.

ARANDA, C. B.; CORBY, O.; DAS, S.; FEIGENBAUM, L.; GEARON, P. **SPARQL 1.1 Overview**. Disponível em: <http://www.w3.org/TR/sparql11-overview/>. Acesso em: 18/7/2013.

ARTZ, D.; GIL, YOLANDA. A survey of trust in computer science and the Semantic Web. **Web Semantics Science Services and Agents on the World Wide Web**, v. 5, n. 2, p. 58–71, 2007. Elsevier. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S1570826807000133>.

AUER, S.; BIZER, C.; KOBILAROV, G. et al. **DBpedia : A Nucleus for a Web of Open Data**. The 6th International Semantic Web Conference (ISWC 2007). **Anais...** Busan, Korea: Springer, 2007. p. 722–735.

AUER, S.; TRAMP, S.; NUFFELEN, B. VAN; et al. **Managing the Life-Cycle of Linked Data with the LOD2 Stack**. The Semantic Web–ISWC 2012. **Anais...** Boston, USA: Springer Berlin Heidelberg, 2012. p. 1–16.

BAUER, F.; KALTENBÖCK, M. **Linked Open Data : The Essentials**. mono/monoc ed. Vienna, Austria, 2011. 59 p.

BEARMAN, D. A.; LYTLE, R. H. The Power of the Principle of Provenance. **Archivaria**, v. 1, n. 21, p. 14–27, 1985.

BECKETT, D. **RDF/XML Syntax Specification (Revised)**. Disponível em: <http://www.w3.org/TR/rdf-syntax-grammar/>. Acesso em: 18/7/2013.

BECKETT, D.; BERNERS-LEE, T. **Turtle - Terse RDF Triple Language**. Disponível em: <http://www.w3.org/TeamSubmission/turtle/>. Acesso em: 18/7/2013.

BERNERS-LEE, T. Information Management: A Proposal. **Word Journal Of The International Linguistic Association**, v. February 2, n. May 1990, p. 1–10, 1989. CERN-March. Disponível em: <http://www.w3.org/History/1989/proposal.html>.

BERNERS-LEE, T. **Semantic Web Status and Direction**. Disponível em: <http://www.w3.org/2003/Talks/1023-iswc-tbl/slide26-0.html>. Acesso em: 11/7/2013.

BERNERS-LEE, T. **Linked Data - Design Issues**. Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 16/7/2013.

BERNERS-LEE, T. **Giant Global Graph**. Disponível em: <http://dig.csail.mit.edu/breadcrumbs/node/215>. Acesso em: 18/7/2013.

BERNERS-LEE, T.; CONNOLLY, D. **Notation3 (N3) A readable RDF syntax**. Disponível em: <http://www.w3.org/TeamSubmission/2011/SUBM-n3-20110328/>. Acesso em: 18/7/2013.

BERNERS-LEE, T.; FIELDING, R.; MASINTER, L. Uniform Resource Identifier (URI): Generic Syntax. **Network Working Group**, 2005. IETF. Disponível em: <http://www.ietf.org/rfc/rfc3986.txt>.

BERNERS-LEE, T.; HALL, W.; HENDLER, J. A. et al. A Framework for Web Science. **Foundations and Trends® in Web Science**, v. 1, n. 1, p. 1–130, 2006. Now Publishers. Disponível em: <http://eprints.soton.ac.uk/263347/>.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific American**, v. 284, n. 5, p. 34–43, 2001.

BITON, O.; BOULAKIA, S. C.; DAVIDSON, S B. **Zoom*UserViews: Querying Relevant Provenance in Workflow Systems**. VLDB. **Anais...** VLDB Endowment, 2007. p. 1366–1369. Disponível em: <http://www.vldb.org/conf/2007/papers/demo/p1366-biton.pdf>.

BIZER, C. **Semantic Web Publishing Vocabulary (SWP) User Manual**. Disponível em: <http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa/swp/SWP-UserManual.pdf>. Acesso em: 7/9/2013.

BIZER, C.; CYGANIAK, R. **D2R Server – Publishing Relational Databases on the Semantic Web**. World. **Anais...** Citeseer, 2006. p. 26. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.1329&rep=rep1&type=pdf>.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. **Linked Data: Principles and State of the Art**. Disponível em: <http://www.w3.org/2008/Talks/WWW2008-W3CTrack-LOD.pdf>. Acesso em: 17/7/2013.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data - The Story So Far. **International Journal on Semantic Web and Information Systems**, v. 5, n. 3, p. 1–22, 2009.

BIZER, C.; OLDAKOWSKI, R. Using context- and content-based trust policies on the semantic web. **Proceedings of the 13th international World Wide Web conference on Alternate track papers posters WWW Alt 04**, v. 5, n. 2, p. 228, 2004. ACM Press. Disponível em: <http://portal.acm.org/citation.cfm?doid=1013367.1013409>.

BOAG, S.; CHAMBERLIN, D.; FERNÁNDEZ, M. F. et al. **XQuery 1.0: An XML Query Language (Second Edition)**. Disponível em: <http://www.w3.org/TR/xquery/>. Acesso em: 16/8/2013.

BOLEY, HAROLD; HALLMARK, G.; KIFER, M. et al. **RIF Core Dialect (Second Edition)**. Disponível em: <http://www.w3.org/TR/rif-core/>. Acesso em: 18/7/2013.

BRAUN, U.; SHINNAR, A.; SELTZER, MARGO. Securing Provenance. **Proceedings of the 3rd conference on Hot topics in security**, p. 1–5, 2008. USENIX Association. Disponível em: <http://portal.acm.org/citation.cfm?id=1496675>.

BRAY, T.; HOLLANDER, D.; LAYMAN, A.; TOBIN, R.; THOMPSON, H. S. **Namespaces in XML 1.0 (Third Edition)**. Disponível em: <http://www.w3.org/TR/2009/REC-xml-names-20091208/>. Acesso em: 18/7/2013.

BRAY, T.; PAOLI, J.; SPERBERG-MCQUEEN, C. M.; MALER, E.; YERGEAU, F. **Extensible Markup Language (XML) 1.0 (Fifth Edition)**. Disponível em: <http://www.w3.org/TR/2008/REC-xml-20081126/>. Acesso em: 18/7/2013.

BREITMAN, K.; SALAS, P.; CASANOVA, M. A. et al. Open Government Data in Brazil. **IEEE Intelligent Systems**, v. 27, n. June, p. 45–49, 2012.

BRICKLEY, DAN; GUHA, R. V. **RDF Vocabulary Description Language 1.0: RDF Schema**. Disponível em: <http://www.w3.org/TR/rdf-schema/>. Acesso em: 18/7/2013.

BRICKLEY, DANIEL. **Web Of Trust RDF Ontology**. Disponível em: <http://xmlns.com/wot/0.1/>. Acesso em: 7/9/2013.

BRICKLEY, DANIEL; MILLER, L. **FOAF Vocabulary Specification 0.98**. Disponível em: <http://xmlns.com/foaf/spec/>. Acesso em: 8/9/2013.

BRUIJN, J. DE; WELTY, CHRIS. **RIF RDF and OWL Compatibility (Second Edition)**. Disponível em: <http://www.w3.org/TR/rif-rdf-owl/>. Acesso em: 18/7/2013.

BUNEMAN, P.; KHANNA, S.; TAN, W. Why and Where: A Characterization of Data Provenance. (J. Van Den Bussche & V. Vianu, Eds.)**ICDT**, v. 1973, n. D, p. 316–330, 2001. Springer. Disponível em: <http://link.springer.de/link/service/series/0558/bibs/1973/19730316.htm>.

BUNEMAN, P.; TAN, W. **Provenance in databases**. SIGMOD Conference. **Anais...** ACM Press, 2007. p. 1171–1173. Disponível em: <http://doi.acm.org/10.1145/1247480.1247646>.

CALLAHAN, S. P.; FREIRE, J.; SANTOS, EMANUELE; et al. **VisTrails: visualization meets data management**. SIGMOD Conference. **Anais...** ACM, 2006. p. 745–747. Disponível em: <http://doi.acm.org/10.1145/1142473.1142574>.

CAMPOS, M. L. DE A.; GOMES, H. E. Taxonomia e Classificação: o princípio de categorização. **DataGramaZero-Revista de Ciência da Informação**, 2008.

CAMPOS, M. L. M.; GUIZZARDI, G. **GT-LinkedDataBR – Exposição, compartilhamento e conexão de recursos de dados abertos na Web (Linked Open Data)**. Disponível em: <http://www.rnp.br/pd/gts2010-2011/gt_linkeddatabr.html>. Acesso em: 8/9/2013.

CARROLL, J. J.; BIZER, C.; HAYES, P.; STICKLER, P. Named graphs, provenance and trust. **Proceedings of the 14th international conference on World Wide Web WWW 05**, v. 14, p. 613, 2005. ACM Press. Disponível em: <http://portal.acm.org/citation.cfm?doid=1060745.1060835>.

CASTERS, M.; BOUMAN, R.; DONGEN, J. VAN. **Pentaho Kettle Solutions: Building Open Source ETL Solutions with Pentaho Data Integration**. Indianapolis, USA: Wiley Publishing, Inc., 2010. 674 p.

CHENEY, J.; CHITICARIU, L.; TAN, W. **Provenance in databases: Why, how, and where**. Hanover, USA: Now Publishers Inc, 2009. 99 p.

CICCARESE, P.; SOILAND-REYES, S. **PAV - Provenance, Authoring and Versioning**. Disponível em: <http://pav-ontology.googlecode.com/svn/trunk/pav.html>. Acesso em: 8/9/2013.

CLARK, J.; DEROSE, S. **XML Path Language (XPath) Version 1.0**. Disponível em: <http://www.w3.org/TR/xpath/>. Acesso em: 16/8/2013.

CLARK, K. G.; GRANT, K.; TORRES, E. **SPARQL Protocol for RDF**. Disponível em: <http://www.w3.org/TR/rdf-sparql-protocol/>. Acesso em: 18/7/2013.

CLIFFORD, BEN; FOSTER, IAN; VOECKLER, J.-S.; WILDE, MICHAEL; ZHAO, YONG. Tracking provenance in a virtual data grid. **Concurrency and Computation Practice and Experience**, v. 20, n. 5, p. 565–575, 2008. Wiley Online Library. Disponível em: <http://doi.wiley.com/10.1002/cpe.1256>.

CORDEIRO, K. DE F.; CAMPOS, M. L. M.; BORGES, M. R. DA S. **Empowering Citizens and Government with Collaboration on Linked Open Data**. ESWC/Workshop Semantics in Governance and Policy Modelling. **Anais...** Creta, Greece, 2011. p. 33–37.

CORDEIRO, K. DE F.; FARIA, F. F.; PEREIRA, B. DE O. et al. **An approach for managing and semantically enriching the publication of Linked Open Governmental Data**. Workshop de Computação Aplicada em Governo Eletrônico do Simpósio Brasileiro de Banco de Dados. **Anais...** , 2011. p. 82–95.

CROCKFORD, D. The application/json Media Type for JavaScript Object Notation (JSON). **RFC4627**, 2006. IETF. Disponível em: <http://www.ietf.org/rfc/rfc4627.txt>.

CRUZ, S. M. S. DA.**Uma estratégia de apoio à gerencia de dados de proveniência em experimentos científicos**. 2011. 214 f. Tese (doutorado) – Programa de Engenharia de Sistemas e Computação, COPPE, UFRJ, Rio de Janeiro, 2011.

CRUZ, S. M. S. DA; CAMPOS, M. L. M.; MATTOSO, M. **Towards a Taxonomy of Provenance in Scientific Workflow Management Systems**. 2009 Congress on Services I. **Anais...** DC, USA: Ieee, 2009. p. 259–266. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5190667>. Acesso em: 6/4/2013.

CYGANIAK, R.; JENTZSCH, A. **Linking Open Data cloud diagram**. Disponível em: <http://lod-cloud.net>. Acesso em: 19/7/2013.

DACONTA, M. C.; OBRST, L. J.; SMITH, K. T. **The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management**. Indianapolis, USA: Wiley Publishing, Inc., 2003. 281 p.

DAVIDSON, SUSAN B; FREIRE, J.. **Provenance and scientific workflows: challenges and opportunities**. IEEE Transactions on Signal Processing. **Anais...** ACM Press, 2008. p. 1345–1350. Disponível em: <http://doi.acm.org/10.1145/1376616.1376772>.

DAVIS, I.; STEINER, T. **RDF 1.1 JSON Serialisation (RDF/JSON)**. Disponível em: <https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-json/index.html>. Acesso em: 18/7/2013.

DEELMAN, EWA; GANNON, D.; SHIELDS, M.; TAYLOR, I. Workflows and e-Science: An overview of workflow system features and capabilities. **Future Generation Computer Systems**, v. 25, n. 5, p. 528–540, 2009. Disponível em: <http://www.sciencedirect.com/science/article/B6V06-4SYCPKX-2/2/bb631979e3dd7071ddede90bbff65a91>.

DE LA CERDA, J. F. S. M.; CAVALCANTI, M. C. **Registro de Procedência de Ligações RDF em Dados Ligados**. Proceedings of Joint V Seminar on Ontology Research in Brazil and VII International Workshop on Metamodels, Ontologies and Semantic Technologies. **Anais...** Recife, Brazil, 2012. p. 218–223.

DODDS, L.; DAVIS, I. **Qualified Relation**. Disponível em: <http://patterns.dataincubator.org/book/modelling-patterns.html>. Acesso em: 4/9/2013.

DUERST, M.; SUIGNARD, M. Internationalized Resource Identifiers (IRIs). **Internet Engineering Task Force IETF Request for Comments**, 2005. IETF. Disponível em: <http://www.ietf.org/rfc/rfc3987.txt>.

ERLING, O.; MIKHAILOV, I. RDF Support in the Virtuoso DBMS. (Sören Auer, C. Bizer, C. Müller, & A. Zhdanova, Eds.)**Proceedings of the 1st Conference on Social Semantic Web CSSW**, v. 221, n. ISBN 978-3-88579-207-9, p. 59–68, 2007. GI. Disponível em: <http://www.springerlink.com/index/w043243q2l729308.pdf>.

EUZENAT, J.; MOCAN, A.; SCHARFFE, F. Ontology Alignments: An Ontology Management Perspective. **Ontology management: semantic web, semantic web services, and business applications**. p.177–206, 2008. New York, USA: Springer.

FABANI, M. P.; TORO, M. E.; VÁZQUEZ, F.; DÍAZ, M. P.; WUNDERLIN, D. A. Differential absorption of metals from soil to diverse vine varieties from the Valley of Tulum (Argentina): consequences to evaluate wine provenance. **Journal of Agricultural and Food Chemistry**, v. 57, n. 16, p. 7409–7416, 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/19645479>.

FREEMAN, ERIC; FREEMAN, ELISABETH; SIERRA, K.; BATES, B. **Head First Design Patterns**. USA: O'Reilly Media, Inc., 2004. 638 p.

FREIRE, J.; KOOP, DAVID; SANTOS, EMANUELE; SILVA, C. T. Provenance for Computational Tasks : A Survey. **Computing in Science Engineering**, v. 10, n. 3, p. 11–21, 2008. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4488060>.

FREITAS, A.; KÄMPGEN, B.; CURRY, E.; O'RIAIN, S.; OLIVEIRA, J. G. **Cogs ETL Provenance Vocabulary**. Disponível em: <https://sites.google.com/site/cogsvocab/>. Acesso em: 7/9/2013.

FREITAS, A.; KÄMPGEN, B.; OLIVEIRA, J. G.; O'RIAIN, S.; CURRY, E. **Representing Interoperable Provenance Descriptions for ETL Workflows**. Proceedings of the 3rd International Workshop on Role of Semantic Web in Provenance Management (SWPM 2012), Extended Semantic Web Conference (ESWC). **Anais...** Heraklion, Crete, 2012.

GAMMA, E.; HELM, R.; JOHNSON, R.; VLISSIDES, J. **Design patterns: elements of reusable object-oriented software**. Addison-Wesley, 1995. 395 p.

GANGEMI, A.; GUARINO, N.; MASOLO, C.; OLTRAMARI, A.; SCHNEIDER, L. Sweetening Ontologies with DOLCE. (A. Gómez-Pérez & V. R. Benjamins, Eds.)**Lecture Notes in Computer Science**, v. 2473, p. 166–181, 2002. Springer. Disponível em: <http://www.springerlink.com/index/5p86jk323x0tjktc.pdf>.

GARIJO, D.; ECKERT, K. **Dublin Core to PROV Mapping**. Disponível em: <http://www.w3.org/TR/2013/WD-prov-dc-20130312/>. Acesso em: 8/9/2013.

GARIJO, D.; GIL, YOLANDA. **The OPMW Ontology**. Disponível em: <http://www.opmw.org/model/OPMW/>. Acesso em: 4/9/2013.

GIL, YOLANDA; ARTZ, D. Towards Content Trust of Web Resources. **Journal of Web Semantics**, v. 5, n. 4, p. 227–239, 2007.

GIL, YOLANDA; CHENEY, J.; GROTH, PAUL; et al. **Provenance XG Final Report**. Disponível em: <http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/>. Acesso em: 18/8/2013.

GRANT, J.; BECKETT, D. **RDF Test Cases**. Disponível em: <http://www.w3.org/TR/rdf-testcases>. Acesso em: 18/7/2013.

GREEN, T. J.; KARVOUNARAKIS, G.; TANNEN, V. Provenance semirings. (L. Libkin, Ed.)**Proceedings of the twentysixth ACM SIGMODSIGACTSIGART symposium on Principles of database systems PODS 07**, v. pages, p. 31, 2007. ACM Press. Disponível em: <http://portal.acm.org/citation.cfm?doid=1265530.1265535>.

GRENON, P.; SMITH, B.; GOLDBERG, L. Biodynamic ontology: applying BFO in the biomedical domain. (D M Pisanelli, Ed.)**Studies In Health Technology And Informatics**, v. 102, n. ii, p. 20–38, 2004. IOS Press. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15853262>.

GROTH, PAUL; MILES, SIMON; MOREAU, L. A Model of Process Documentation to Determine Provenance in Mash-ups. **Transactions on Internet Technology TOIT**, v. 9, n. 1, p. 1–31, 2009. ACM. Disponível em: <http://eprints.ecs.soton.ac.uk/20861/>.

GROTH, PAUL; MOREAU, L. **PROV-Overview: An Overview of the PROV Family of Documents**. Disponível em: <http://www.w3.org/TR/prov-overview/>. Acesso em: 4/9/2013.

GUARINO, N. **Formal Ontology and Information Systems**. Proceedings of FOIS98. **Anais...** IOS Press, 1998. p. 3–15. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.29.1776&rep=rep1&type=pdf>.

GUENTHER, R.; BRAMA, Y.; BREDENBERG, K. et al. **PREMIS Data Dictionary for Preservation Metadata, version 2.2**. Disponível em: <http://www.loc.gov/standards/premis/v2/premis-2-2.pdf>. Acesso em: 7/9/2013.

HARTIG, O. Provenance Information in the Web of Data. **Proceedings of the Linked Data on the Web LDOW Workshop at WWW**, v. 39, n. 27, p. 1–9, 2009. Ceur-Ws. Disponível em: <http://www.dbis.informatik.hu-berlin.de/fileadmin/research/papers/conferences/2009-ldow-hartig.pdf>.

HARTIG, O.; ZHAO, JUN. Publishing and Consuming Provenance Metadata on the Web of Linked Data. **Provenance and Annotation of Data and Processes**, v. 6378, n. 24, p. 78–90, 2010.

HARTIG, O.; ZHAO, JUN. **Provenance Vocabulary Core Ontology Specification**. Disponível em: <http://purl.org/net/provenance/ns>. Acesso em: 6/9/2013.

HASAN, R.; SION, R.; WINSLETT, M. The Case of the Fake Picasso: Preventing History Forgery with Secure Provenance. (M. I. Seltzer & R. Wheeler, Eds.)**FAST**, p. 1–14, 2009. USENIX Association. Disponível em: <http://www.usenix.org/events/fast09/tech/full_papers/hasan/hasan.pdf>.

HEATH, T.; BIZER, C. **Linked Data: Evolving the Web into a Global Data Space**. Morgan & Claypool, 2011. 1–136 p.

HEINO, N.; DIETZOLD, S.; MARTIN, M.; AUER, SOREN. Developing semantic web applications with the ontowiki framework. (T. Pellegrini, Sören Auer, K. Tochtermann, & S. Schaffert, Eds.)**Networked KnowledgeNetworked Media**, v. 221, p. 61–77, 2009. Springer. Disponível em: <http://www.springerlink.com/content/0j7673v3012p7156/>.

HENDLER, J.; HOLM, J.; MUSIALEK, C.; THOMAS, G. US Government Linked Open Data: Semantic.data.gov. **IEEE Intelligent Systems**, v. 27, n. 3, p. 25–31, 2012.

HOBBS, J. R.; PAN, F. **Time Ontology in OWL**. Disponível em: <http://www.w3.org/TR/owl-time/>. Acesso em: 8/9/2013.

HU, Y.-J. H. Y.-J.; BOLEY, H. **SemPIF: A Semantic Meta-policy Interchange Format for Multiple Web Policies**. Ieee, 2010.

JANIK, M.; SCHERP, A.; STAAB, S. EN. The Semantic Web: Collective Intelligence on the Web. **InformatikSpektrum**, 2011. Disponível em: <http://link.springer.com/article/10.1007/s00287-011-0535-x>.

KIFER, M. Rule Interchange Format: The Framework. (D. Calvanese & G. Lausen, Eds.)**Knowledge Engineering Review**, v. 13, n. 1, p. 1–11, 2008. Springer-Verlag New York Inc. Disponível em: <http://www.scopus.com/inward/record.url?eid=2-s2.0-0032024432&partnerID=40&md5=c93b09efebfc8256b8d208f58610e8bf>.

KIMBALL, R.; CASERTA, J. **The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning Conforming, and Delivering Data**. Indianapolis, USA: Wiley Publishing, Inc., 2004. 491 p.

KIMBALL, R.; ROSS, M.; THORNTHWAITE, W.; MUNDY, J.; BECKER, B. **The Data Warehouse Lifecycle Toolkit**. Indianapolis, USA: Wiley Publishing, Inc., 2008. 672 p.

KLYNE, G.; CARROLL, J. J. Resource Description Framework (RDF): Concepts and Abstract Syntax. **W3C recommendation**, 2004. World Wide Web Consortium. Disponível em: <http://www.w3.org/TR/rdf-concepts/>.

KUMAR, S.; PRAJAPATI, R. K.; SINGH, M.; DE, A. **Security enforcement using PKI in Semantic Web**. Computer Information Systems and Industrial Management Applications CISIM 2010 International Conference on. **Anais...** Ieee, 2010. p. 392–397. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5643507>.

LEBO, T.; SAHOO, S. S.; MCGUINNESS, D. **PROV-O: The PROV Ontology**. Disponível em: <http://www.w3.org/TR/prov-o/>. Acesso em: 4/9/2013.

LIM, C.; LU, S.; CHEBOTKO, A.; FOTOUHI, F. Prospective and Retrospective Provenance Collection in Scientific Workflow Environments. **2010 IEEE International Conference on Services Computing**, p. 449–456, 2010. Ieee. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5557202&tag=1>.

LYNCH, C. A. When Documents Deceive: Trust and Provenance as New Factors for Information Retrieval in a Tangled Web. **JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY**, v. 52, n. 1, p. 12–17, 2001.

MARINHO, A.; MURTA, L.; WERNER, C. ProvManager: a provenance management system for scientific workflows. **Concurrency and Computation Practice and Experience**, p. n/a–n/a, 2011. John Wiley & Sons, Ltd. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/cpe.1870/full>.

MARJIT, U.; SHARMA, K.; BISWAS, U. Provenance Representation and Storage Techniques in Linked Data : A State-of-the-Art Survey. **International Journal of Computer Applications**, v. 38, n. 9, p. 23–28, 2012.

MCGUINNESS, D.; DING, L.; SILVA, P. P. DA; CHANG, C. **PML 2: A Modular Explanation Interlingua**. Workshop on Explanation-aware Computing (ExaCt-2007). **Anais...** , 2007. p. 49–55.

MCGUINNESS, D. L.; HARMELEN, F. VAN. **OWL Web Ontology Language Overview**. Disponível em: <http://www.w3.org/TR/owl-features/>. Acesso em: 18/7/2013.

MEALLING, M.; DENENBERG, R. Report from the Joint W3C/IETF URI Planning Interest Group: Uniform Resource Identifiers (URIs), URLs, and Uniform Resource Names (URNs): Clarifications and Recommendations. (M. Mealling & R. Denenberg, Eds.)**Internet Engineering Task Force IETF Request for Comments**, 2002. IETF. Disponível em: <http://www.ietf.org/rfc/rfc3305.txt>.

MENDES, P. N.; MÜHLEISEN, H.; BIZER, C. **Sieve: Linked Data quality assessment and fusion**. ACM International Conference Proceeding Series. **Anais...** , 2012. p. 116–123.

MENDONÇA, R. R. DE; DE LA CERDA, J. F. S. M.; CORDEIRO, K. DE F. et al. **LOP - Capturing and Linking Open Provenance on LOD Cycle**. Proceedings of the Fifth Workshop on Semantic Web Information Management. **Anais...** New York, USA: ACM, 2013.

MILLAR, L. The Death of the Fonds and the Resurrection of Provenance : Archival Context in Space and Time. **Archivaria**, v. 1, n. 53, p. 1–15, 2002.

MOREAU, L. The Foundations for Provenance on the Web. **Foundations and Trends in Web Science**, v. 2, n. 2-3, p. 99–241, 2010. Now publisher. Disponível em: <http://eprints.soton.ac.uk/271691/>. Acesso em: 31/3/2013.

MOREAU, L.; CLIFFORD, BEN; FREIRE, J.; et al. The Open Provenance Model: Core Specification (v1.1). **Future Generation Computer Systems**, v. 27, n. 6, p. 743–756, 2011.

MOREAU, L.; DING, L.; FUTRELLE, JOE; et al. **Open Provenance Model (OPM) OWL Specification**. Disponível em: <http://openprovenance.org/model/opmo>. Acesso em: 4/9/2013.

MOREAU, L.; LUDÄSCHER, B.; ALTINTAS, ILKAY; et al. Special Issue: The First Provenance Challenge. **Concurrency and Computation: Practice and Experience**, v. 20, n. 5, p. 409–418, 2008.

MOREAU, L.; MISSIER, P. **PROV-DM: The PROV Data Model**. Disponível em: <http://www.w3.org/TR/prov-dm/>. Acesso em: 4/9/2013.

NILES, I.; PEASE, A. Towards a standard upper ontology. (Chris Welty & B. Smith, Eds.)**Proceedings of the international conference on Formal Ontology in Information Systems FOIS 01**, v. 2001, n. 3, p. 2–9, 2001. ACM Press. Disponível em: <http://portal.acm.org/citation.cfm?doid=505168.505170>.

OMITOLA, T.; FREITAS, A.; EDWARD, C. et al. **Capturing interactive data transformation operations using provenance workflows**. 3rd International Workshop on Role of Semantic Web in Provenance Management SWPM 2012. **Anais...** Heraklion, Crete, 2012.

OMITOLA, T.; GUTTERIDGE, C.; MILLARD, I. C. et al. **Tracing the Provenance of Linked Data using voiD**. The International Conference on Web Intelligence Mining and Semantics WIMS11. **Anais...** New York, USA: ACM, 2011. p. 17:1–17:7. Disponível em: <http://eprints.ecs.soton.ac.uk/21956/>.

RAMAKRISHNAN, R.; GEHRKE, J. **Database Management Systems**. McGraw-Hill, 2002. 177–192 p.

RIBEIRO, C. E.**Módulos de inferência na nuvem de dados ligados para apoiar o desenvolvimento de aplicações**. 2012. 203 f. Dissertação (mestrado) – Universidade Federal do Rio de Janeiro, Instituto de Matemática, Instituto Tércio Pacitti, Programa de Pós-Graduação em Informática, 2012.

RULA, A.; PALMONARI, M.; HARTH, A.; STADTMÜLLER, S.; MAURINO, A. **On the Diversity and Availability of TemporalInformation in Linked Open Data**. The Semantic Web–ISWC 2012. **Anais...** Springer Berlin Heidelberg, 2012. p. 492–507.

RULA, A.; PALMONARI, M.; MAURINO, A. **Capturing the Age of Linked Open Data: Towards a Dataset-Independent Framework**. 2012 IEEE Sixth International Conference on Semantic Computing. **Anais...** Ieee, 2012. p. 218–225. Disponível em: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6337107>.

SAHOO, S. S.; NGUYEN, V.; BODENREIDER, O. et al. A unified framework for managing provenance information in translational research. **BMC Bioinformatics**, v. 12, n. 1, p. 461, 2011. BioMed Central Ltd. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/22126369>.

SAHOO, S. S.; SHETH, A. P. Provenir ontology: Towards a Framework for eScience Provenance Management. **Microsoft eScience Workshop**, p. 15–17, 2009. Pittsburgh, USA.

SHERIDAN, J.; TENNISON, J. **Linking UK Government Data**. Proceedings of the Linked Data on the Web LDOW Workshop at WWW. **Anais...** Raleigh, USA, 2010.

SIMMHAN, YOGESH L.; PLALE, BETH; GANNON, D. A survey of data provenance in e-science. **ACM SIGMOD Record**, v. 34, n. 3, p. 31–36, 2005. Acm. Disponível em: <http://portal.acm.org/citation.cfm?doid=1084805.1084812>.

SIZOV, S. **What Makes You Think That? The Semantic Web's Proof Layer**. IEEE Computer Society, 2007.

SMITH, B.; CEUSTERS, W.; KLAGGES, B. et al. Relations in biomedical ontologies. **Genome Biology**, v. 6, n. 5, p. R46, 2005. BioMed Central. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/15892874>.

SMITH, B.; WELTY, CHRISTOPHER. Ontology: Towards a New Synthesis. (C Welty & B Smith, Eds.)**Expert Systems**, v. 2001, n. 3, p. iii – x, 2001. ACM Press. Disponível em: <http://portal.acm.org/citation.cfm?doid=505168.505201>.

TAN, V.; GROTH, PAUL; MILES, SIMON; JIANG, SHENG; MUNROE, STEVE. Security Issues in a SOA-based Provenance System. (L. Moreau & Ian Foster, Eds.)**Program**, v. 36, p. 203–211, 2006. Springer. Disponível em: <http://eprints.ecs.soton.ac.uk/12569/>.

TAN, W. Research Problems in Data Provenance. **Techniques**, v. 27, n. 4, p. 1–8, 2004. Citeseer. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.6963&rep=rep1&type=pdf>.

TUNNICLIFFE, S.; DAVIS, I. **Changeset**. Disponível em: <http://vocab.org/changeset/schema.html>. Acesso em: 7/9/2013.

VOLZ, J.; BIZER, C.; GAEDKE, M.; KOBILAROV, G. Silk – A Link Discovery Framework for the Web of Data. (C. Bizer, T. Heath, T. Berners-Lee, & K. Idehen, Eds.)**Language**, v. 135, n. 2, p. 1–6, 2009. Citeseer. Disponível em: <http://events.linkeddata.org/ldow2009/papers/ldow2009_paper13.pdf>.

YU, J.; BUYYA, R. A Taxonomy of Workflow Management Systems for Grid Computing. **Journal of Grid Computing**, v. 3, n. 3-4, p. 29, 2005. Springer. Disponível em: <http://arxiv.org/abs/cs/0503025>.

ZHAO, JUN. **Open Provenance Model Vocabulary Specification**. Disponível em: <http://open-biomed.sourceforge.net/opmv/ns.html>. Acesso em: 4/9/2013.

ZHAO, JUN; GOBLE, CAROLE; STEVENS, ROBERT; TURI, DANIELE. Mining Taverna's semantic web of provenance. **Concurrency and Computation Practice and Experience**, v. 20, n. August 2007, p. 463–472, 2008. John Wiley \& Sons. Disponível em: <http://www3.interscience.wiley.com/journal/115806853/abstract>.

ZHAO, JUN; HARTIG, O. **Towards Interoperable Provenance Publication on the Linked Data Web**. Proceedings of the 5th Linked Data on the Web (LDOW) Workshop at the World Wide Web Conference (WWW). **Anais...** Lyon, France, 2012.

Apêndices

Apêndice A – Dicionário de dados do banco de dados utilizado pelo Agente Coletor de Proveniência para armazenar temporariamente os dados de proveniência.

Quadro 18. Metadados das tabelas do banco de dados utilizado pelo Agente Coletor de Proveniência.

Tabela	Colunas	Comentários
prosp_hop	7	Ligação entre 2 passos (proveniência prospectiva).
prosp_hop_field	6	Campo de dados (proveniência prospectiva).
prosp_note	2	Anotação de documentação (proveniência prospectiva).
prosp_repository	3	Repositório (proveniência prospectiva).
prosp_step	7	Passo do workflow (proveniência prospectiva).
prosp_step_parameter	3	Parâmetro do passo (proveniência prospectiva).
prosp_step_type	2	Tipo de passo (proveniência prospectiva).
prosp_workflow	14	Workflow (proveniência prospectiva).
prosp_workflow_note	3	Relação entre workflow e anotação (proveniência prospectiva).
retrosp_row_field	10	Campo de uma linha de dados (proveniência retrospectiva).
retrosp_step	9	Passo do workflow (proveniência retrospectiva).
retrosp_step_parameter	8	Parâmetro do passo (proveniência retrospectiva).
retrosp_workflow	13	Workflow (proveniência retrospectiva).
user	5	Usuário.
14 Tabelas	**92**	

Quadro 19. Metadados das colunas de cada tabela do banco de dados utilizado pelo Agente Coletor de Proveniência. As colunas em negrito integram a chave primária das respectivas tabelas.

Tabela	Coluna	Tipo	Tam.	Nulo	Padrão	Comentários
prosp_hop	enabled	char	1		N	Indica se a ligação entre os 2 passos está habilitada (Y) ou não (N).
prosp_hop	evaluation	char	1		N	Se a utilização da ligação estiver condicionada ao resultado do passo de origem, indica se a ligação será utilizada se o passo de origem for concluído com sucesso (Y) ou com erros (N).

Tabela	Coluna	Tipo	Tam.	Nulo	Pa-drão	Comentários
prosp_hop	**id_repository**	int	10			Identificador do repositório onde a composição do workflow está armazenada.
prosp_hop	**id_step_from**	int	10			Identificador da composição do passo de origem da ligação.
prosp_hop	**id_step_to**	int	10			Identificador da composição do passo de destino da ligação.
prosp_hop	**id_workflow**	int	10			Identificador da composição do workflow.
prosp_hop	unconditional	char	1		N	Indica se a utilização da ligação não estará condicionada ao resultado do passo de origem (Y) ou se estará condicionada (N).
prosp_hop_field	field_name	varchar	255			Nome do campo de dados.
prosp_hop_field	**id_field**	int	10			Identificador do campo de dados.
prosp_hop_field	**id_repository**	int	10			Identificador do repositório onde a composição do workflow está armazenada.
prosp_hop_field	**id_step_from**	int	10			Identificador da composição do passo de origem da ligação.
prosp_hop_field	**id_step_to**	int	10			Identificador da composição do passo de destino da ligação.
prosp_hop_field	**id_workflow**	int	10			Identificador da composição do workflow.
prosp_note	**id_note**	int	10			Identificador da anotação.
prosp_note	text	text	65535			Texto da anotação.
prosp_repository	**id_repository**	int	10			Identificador do repositório.
prosp_repository	location	varchar	255			Localização do repositório.

Tabela	Coluna	Tipo	Tam.	Nulo	Padrão	Comentários
prosp_repository	name	varchar	255			Nome do repositório.
prosp_step	copy_nr	int	10			Número da cópia do passo.
prosp_step	description	text	65535	√	null	Descrição do passo.
prosp_step	**id_repository**	int	10			Identificador do repositório onde a composição do workflow está armazenada.
prosp_step	**id_step**	int	10			Identificador da composição do passo.
prosp_step	id_step_type	int	10			Identificador do tipo do passo.
prosp_step	**id_workflow**	int	10			Identificador da composição do workflow.
prosp_step	name	varchar	255			Nome do passo.
prosp_step_parameter	**id_step_param**	int	10			Identificador do parâmetro.
prosp_step_parameter	**id_step_type**	int	10			Identificador do tipo do passo.
prosp_step_parameter	param_name	varchar	255			Nome do parâmetro.
prosp_step_type	**id_step_type**	int	10			Identificador do tipo de passo.
prosp_step_type	name	varchar	255			Nome do tipo de passo.
prosp_workflow	created_date	time stamp	19			Data e hora de criação da composição do workflow.
prosp_workflow	description	text	65535	√	null	Descrição do workflow.
prosp_workflow	id_created_user	int	10			Identificador do usuário que criou o workflow (prosp.).
prosp_workflow	id_modified_user	int	10			Identificador do usuário que fez a última modificação na composição do workflow.
prosp_workflow	id_parent	int	10	√	null	Identificador da composição do workflow pai.
prosp_workflow	id_parent_repository	int	10	√	null	Identificador do repositório onde a composição do workflow pai está armazenada.

Tabela	Coluna	Tipo	Tam.	Nulo	Pa-drão	Comentários
prosp_workflow	**id_repository**	int	10			Identificador do repositório onde a composição do workflow está armazenada.
prosp_workflow	id_root	int	10			Identificador da composição do workflow raiz.
prosp_workflow	id_root_repository	int	10			Identificador do repositório onde a composição do workflow raiz está armazenada.
prosp_workflow	**id_workflow**	int	10			Identificador da composição do workflow.
prosp_workflow	modified_date	time stamp	19			Data e hora da última modificação na composição do workflow.
prosp_workflow	name	varchar	255			Nome do workflow.
prosp_workflow	object_id	varchar	255			Identificador da composição do workflow na ferramenta de ETL.
prosp_workflow	version_nr	int	10			Número da versão da composição do workflow.
prosp_workflow_note	**id_note**	int	10			Identificador da anotação.
prosp_workflow_note	**id_repository**	int	10			Identificador do repositório onde a composição do workflow está armazenada.
prosp_workflow_note	**id_workflow**	int	10			Identificador da composição do workflow.
retrosp_row_field	field_value	text	65535	√	null	Valor do campo de dados.
retrosp_row_field	**id_prosp_field**	int	10			Identificador do campo de dados.
retrosp_row_field	**id_prosp_repository**	int	10			Identificador do repositório onde a composição do workflow está armazenada.
retrosp_row_field	**id_prosp_step_from**	int	10			Identificador da composição do passo de origem da ligação.

Tabela	Coluna	Tipo	Tam.	Nulo	Padrão	Comentários
retrosp_row_field	**id_prosp_step_to**	int	10			Identificador da composição do passo de destino da ligação.
retrosp_row_field	**id_prosp_workflow**	int	10			Identificador da composição do workflow.
retrosp_row_field	**id_workflow**	int	10			Identificador da execução do workflow.
retrosp_row_field	**row_count**	int	10			Número da linha de dados.
retrosp_row_field	**seq_from**	int	10			Sequencial de execução do passo de origem no workflow.
retrosp_row_field	**seq_to**	int	10			Sequencial de execução do passo de destino no workflow.
retrosp_step	finish_date	time stamp	19	√	null	Data e hora do término da execução do passo.
retrosp_step	**id_prosp_repository**	int	10			Identificador do repositório onde a composição do workflow está armazenada.
retrosp_step	**id_prosp_step**	int	10			Identificador da composição do passo.
retrosp_step	**id_prosp_workflow**	int	10			Identificador da composição do workflow.
retrosp_step	id_user	int	10			Identificador do usuário que executou o passo.
retrosp_step	id_workflow	int	10			Identificador da execução do workflow.
retrosp_step	seq	int	10			Sequencial de execução do passo no workflow.
retrosp_step	start_date	time stamp	19	√	null	Data e hora do início da execução do passo.
retrosp_step	success	char	1		N	Indica se o passo foi executado com sucesso (Y) ou não (N).

Tabela	Coluna	Tipo	Tam.	Nulo	Padrão	Comentários
retrosp_step_parameter	**id_prosp_repository**	int	10			Identificador do repositório onde a composição do workflow está armazenada.
retrosp_step_parameter	**id_prosp_step**	int	10			Identificador da composição do passo.
retrosp_step_parameter	**id_prosp_step_param**	int	10			Identificador do parâmetro.
retrosp_step_parameter	**id_prosp_step_type**	int	10			Identificador do tipo do passo.
retrosp_step_parameter	**id_prosp_workflow**	int	10			Identificador da composição do workflow.
retrosp_step_parameter	**id_workflow**	int	10			Identificador da execução do workflow.
retrosp_step_parameter	param_value	text	65535	√	null	Valor do parâmetro.
retrosp_step_parameter	**seq**	int	10			Sequencial de execução do passo no workflow.
retrosp_workflow	finish_date	time stamp	19	√	null	Data e hora do término da execução do workflow.
retrosp_workflow	id_parent	int	10	√	null	Identificador da execução do workflow pai.
retrosp_workflow	id_parent_prosp_ repository	int	10	√	null	Identificador do repositório onde a composição do workflow pai está armazenada.
retrosp_workflow	id_parent_prosp_workflow	int	10	√	null	Identificador da composição do workflow pai.
retrosp_workflow	**id_prosp_repository**	int	10			Identificador do repositório onde a composição do workflow está armazenada.
retrosp_workflow	**id_prosp_workflow**	int	10			Identificador da composição do workflow.
retrosp_workflow	id_root	int	10			Identificador da execução do workflow raiz.
retrosp_workflow	id_root_prosp_repository	int	10			Identificador do repositório onde a composição do

Tabela	Coluna	Tipo	Tam.	Nulo	Padrão	Comentários
						workflow raiz está armazenada.
retrosp_workflow	id_root_prosp_workflow	int	10			Identificador da composição do workflow raiz.
retrosp_workflow	id_user	int	10			Identificador do usuário que executou o workflow.
retrosp_workflow	**id_workflow**	int	10			Identificador da execução do workflow.
retrosp_workflow	start_date	time stamp	19	√	null	Data e hora do início da execução do workflow.
retrosp_workflow	success	char	1		N	Indica se o workflow foi executado com sucesso (Y) ou não (N).
user	description	text	65535	√	null	Descrição do usuário.
user	**id_repository**	int	10			Identificador do repositório.
user	**id_user**	int	10			Identificador do usuário.
user	login	varchar	255			Login do usuário.
user	name	varchar	255			Nome do usuário.

Quadro 20. Chaves estrangeiras das tabelas do banco de dados utilizado pelo Agente Coletor de Proveniência.

Nome da Restrição	Chave Estrangeira	Coluna Referenciada
cstr_prosp_hop_field_ fk_hop	prosp_hop_field.id_repository	prosp_hop.id_repository
	prosp_hop_field.id_workflow	prosp_hop.id_workflow
	prosp_hop_field.id_step_from	prosp_hop.id_step_from
	prosp_hop_field.id_step_to	prosp_hop.id_step_to
cstr_prosp_hop_ fk_from	prosp_hop.id_repository	prosp_step.id_repository
	prosp_hop.id_workflow	prosp_step.id_workflow
	prosp_hop.id_step_from	prosp_step.id_step
cstr_prosp_hop_fk_to	prosp_hop.id_repository	prosp_step.id_repository
	prosp_hop.id_workflow	prosp_step.id_workflow
	prosp_hop.id_step_to	prosp_step.id_step
cstr_prosp_step_ fk_step_type	prosp_step.id_step_type	prosp_step_type.id_step_type
cstr_prosp_step_	prosp_step.id_repository	prosp_workflow.id_repository

Nome da Restrição	Chave Estrangeira	Coluna Referenciada
fk_workflow	prosp_step.id_workflow	prosp_workflow.id_workflow
cstr_prosp_step_param_ fk_step_type	prosp_step_parameter.id_step_type	prosp_step_type.id_step_type
cstr_prosp_workflow_ fk_created	prosp_workflow.id_repository	user.id_repository
	prosp_workflow.id_created_user	user.id_user
cstr_prosp_workflow_ fk_modified	prosp_workflow.id_repository	user.id_repository
	prosp_workflow.id_modified_user	user.id_user
cstr_prosp_workflow_ fk_parent	prosp_workflow.id_parent_repository	prosp_workflow.id_repository
	prosp_workflow.id_parent	prosp_workflow.id_workflow
cstr_prosp_workflow_ fk_repository	prosp_workflow.id_repository	prosp_repository.id_repository
cstr_prosp_workflow_ fk_root	prosp_workflow.id_root_repository	prosp_workflow.id_repository
	prosp_workflow.id_root	prosp_workflow.id_workflow
cstr_prosp_workflow_ note_fk_note	prosp_workflow_note.id_note	prosp_note.id_note
cstr_prosp_workflow_ note_fk_workflow	prosp_workflow_note.id_repository	prosp_workflow.id_repository
	prosp_workflow_note.id_workflow	prosp_workflow.id_workflow
cstr_retrosp_row_field_ fk_prosp_hop_field	retrosp_row_field.id_prosp_repository	prosp_hop_field.id_repository
	retrosp_row_field.id_prosp_workflow	prosp_hop_field.id_workflow
	retrosp_row_field.id_prosp_step_from	prosp_hop_field.id_step_from
	retrosp_row_field.id_prosp_step_to	prosp_hop_field.id_step_to
	retrosp_row_field.id_prosp_field	prosp_hop_field.id_field
cstr_retrosp_row_field_ fk_step_from	retrosp_row_field.id_prosp_repository	retrosp_step.id_prosp_repository
	retrosp_row_field.id_prosp_workflow	retrosp_step.id_prosp_workflow
	retrosp_row_field.id_workflow	retrosp_step.id_workflow
	retrosp_row_field.id_prosp_step_from	retrosp_step.id_prosp_step
	retrosp_row_field.seq_from	retrosp_step.seq
cstr_retrosp_row_field_ fk_step_to	retrosp_row_field.id_prosp_repository	retrosp_step.id_prosp_repository
	retrosp_row_field.id_prosp_workflow	retrosp_step.id_prosp_workflow
	retrosp_row_field.id_workflow	retrosp_step.id_workflow
	retrosp_row_field.id_prosp_step_to	retrosp_step.id_prosp_step
	retrosp_row_field.seq_to	retrosp_step.seq
cstr_retrosp_step_fk_ prosp_step	retrosp_step.id_prosp_repository	prosp_step.id_repository
	retrosp_step.id_prosp_workflow	prosp_step.id_workflow
	retrosp_step.id_prosp_step	prosp_step.id_step
cstr_retrosp_step_ fk_user	retrosp_step.id_prosp_repository	user.id_repository
	retrosp_step.id_user	user.id_user
cstr_retrosp_step_ fk_workflow	retrosp_step.id_prosp_repository	retrosp_workflow.id_prosp_repository
	retrosp_step.id_prosp_workflow	retrosp_workflow.id_prosp_workflow
	retrosp_step.id_workflow	retrosp_workflow.id_workflow
cstr_retrosp_step_ param_fk_prosp_step	retrosp_step_parameter.id_prosp_repository	prosp_step.id_repository
	retrosp_step_parameter.id_prosp_workflow	prosp_step.id_workflow

Nome da Restrição	Chave Estrangeira	Coluna Referenciada
	retrosp_step_parameter.id_prosp_step	prosp_step.id_step
	retrosp_step_parameter.id_prosp_step_type	prosp_step.id_step_type
cstr_ retrosp_step_param_ fk_prosp_step_param	retrosp_step_parameter.id_prosp_step_type	prosp_step_parameter.id_step_type
	retrosp_step_parameter.id_prosp_step_ param	prosp_step_parameter.id_step_param
cstr_retrosp_step_ param_fk_step	retrosp_step_parameter.id_prosp_repository	retrosp_step.id_prosp_repository
	retrosp_step_parameter.id_prosp_workflow	retrosp_step.id_prosp_workflow
	retrosp_step_parameter.id_workflow	retrosp_step.id_workflow
	retrosp_step_parameter.id_prosp_step	retrosp_step.id_prosp_step
	retrosp_step_parameter.seq	retrosp_step.seq
cstr_retrosp_workflow_ fk_prosp_workflow	retrosp_workflow.id_prosp_repository	prosp_workflow.id_repository
	retrosp_workflow.id_prosp_workflow	prosp_workflow.id_workflow
cstr_retrosp_workflow_ fk_retrosp_parent	retrosp_workflow.id_parent_prosp_repository	retrosp_workflow.id_prosp_repository
	retrosp_workflow.id_parent_prosp_workflow	retrosp_workflow.id_prosp_workflow
	retrosp_workflow.id_parent	retrosp_workflow.id_workflow
cstr_retrosp_workflow_ fk_retrosp_root	retrosp_workflow.id_root_prosp_repository	retrosp_workflow.id_prosp_repository
	retrosp_workflow.id_root_prosp_workflow	retrosp_workflow.id_prosp_workflow
	retrosp_workflow.id_root	retrosp_workflow.id_workflow
cstr_retrosp_workflow_ fk_user	retrosp_workflow.id_prosp_repository	user.id_repository
	retrosp_workflow.id_user	user.id_user
cstr_user_ fk_repository	user.id_repository	prosp_repository.id_repository

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

CCMN - Bloco C - Cidade Universitária - Ilha do Fundão
Rio de Janeiro - RJ CEP: 21941-916
www.ppgi.ufrj.br