ETL4LinkedProv: Managing Multigranular Linked Data Provenance

Rogers Reiche de Mendonça[1], Sérgio Manuel Serra da Cruz[2], Maria Luiza Machado Campos[3]

[1] Petróleo Brasileiro S.A, Brazil
rogers.mendonca@petrobras.com.br
[2] Universidade Federal Rural do Rio de Janeiro, Brazil
serra@ufrrj.br
[3] Universidade Federal do Rio de Janeiro, Brazil
mluiza@ppgi.ufrj.br

Abstract. This article presents the ETL4LinkedProv approach to manage the collection and publication of provenance with distinct levels of granularity as Linked Data. The proposed approach uses ETL-workflows and a component named Provenance Collector Agent to collect two kinds of provenance (prospective and retrospective) integrating them with domain data. The component also set the granularity of the provenance to be captured. Furthermore, ETL4LinkedProv is evaluated in a real world scenario where governmental Brazilian agencies produce and publish public data sources as Linked Data. In this article we also measure the amount of the provenance generated in the runtime of ETL-workflows and in the number of published RDF triples.

Categories and Subject Descriptors: H.2 [**Database Management**]: Miscellaneous; H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: ETL, linked data, LOD2, RDF, provenance, workflows

1. INTRODUCTION

In recent years an increasing number of data providers has adopted technologies and good practices of the Semantic Web [Berners-Lee et al. 2001] to publish and link structured data on the Web, forming the "Web of Data" [Bizer et al. 2009]. The Web of Data is a giant global graph consisting of billions of RDF (Resource Description Framework) triples from numerous sources (including prominent examples, such as DBpedia, YAGO, and DBLP) covering all sorts of subjects [Bizer et al. 2009]. According to Bizer et al. [2009], Linked Data is simply about using the Web to create typed links between data from different sources. Technically, Linked Data refers to data published on the Web in a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can be linked to and from external data sets. The principles of Linked Data were adopted by the Web of Data providing both simplicity and flexibility for data to be represented, interconnected and interoperated [Heath and Bizer 2011].

The accelerated growth of the Web of Data has encouraged the emergence of a new generation of added-value services and innovative uses to explore their full potential of published data [Heath and Bizer 2011]. For instance, national and international government initiatives [Sheridan and Tennison 2010; Breitman et al. 2012] fostered by transparency and collaboration agreements, began to publish open data following the principles of Linked Data. These initiatives, together with the growth and spread of mobile applications and public access services, brought a new perspective of empowerment and participation to citizens [Cordeiro et al. 2011] as well as improvements to governments in decision

making [Ding et al. 2012].

Considering the openness of the Web of Data and technical challenges to trace several non-invertible operators (*e.g.*, aggregation) involved in an data transformation scenarios, several issues related to the quality of Linked Data and its provenance started to play an increasingly important role [Selis et al. 2007a] [Hartig and Zhao 2010], [Zhao et al. 2011], [Zhao and Hartig 2012]. For instance, if a governmental linked dataset is replicated at different locations, different dataset copies from the same data sources can be created outside the boundaries of the governmental organization. These copies of RDF triples, that correspond to the same set of entities, might be in conflict or have variable quality, because the dataset copies might be interconnected by different RDF links or be freely changed by different publishers. Thus, finding a piece of data about a specific entity may result in multiple triples identifying this entity and linking to objects from different sources. In this case, a data consumer may have questions that are hard to be answered. *Which of these links should be followed? Which of the Linked Data objects provides more trustworthy or more up-to-date information about the entity?* Answering the above mentioned questions is not trivial, we need not only provenance about the origins of Linked Data but also the provenance of the interconnecting and the publishing processes. In this scenario, a challenge is dealing with the distinct kinds of provenance generated during the Linked Data lifecycle [Auer et al. 2012]. The provenance has a key role on enhancing trustworthy, quality and reliability of the published data and the publishing process.

This article presents the ETL4LinkedProv approach for collecting, linking and publishing Linked Data provenance data. The approach considers distinct types of provenance and that level of granularity, taking into account the impact on Linked Data publishing process carried out within the boundaries of an organization in the so called Data Preparation and Transformation Process, a stage of the LOD2 Project proposed by Auer et al. [2012]. ETL4LinkedProv uses the concept of Extraction, Transformation and Loading workflows (ETL-workflows) to publish Linked Data in the Web of Data presented by de Mendonça et al. [2013]. The approach includes a component called Provenance Collector Agent (PCA), which captures, stores and interconnects provenance data of distinct granularity, related both to the composition (prospective provenance) and to the execution (retrospective provenance) of the publishing process. Such provenance is also published as Linked Data and integrated with the linked datasets, enabling not only the joint exploitation of domain data and provenance via SPARQL queries, as well as investigations on the trustworthiness and quality of Linked Data.

This article is organized as follows: section 2 present the background; section 3 discusses related work; section 4 introduces the ETL4LinkedProv and its implementation in a ETL tool; section 5 illustrates a use case and discusses it according to qualitative and quantitative analyses with different levels of provenance granularity settings; finally, section 6 presents conclusion and future work.

2. BACKGROUND

Provenance is defined as a term which provides the history and origins of data and the processes by which data are derived and existed in the database [Buneman et al. 2001]. Its original usage was to interpret and reproduce the results of scientific experiments [Freire et al. 2008]. Nowadays, the usage of provenance data extrapolated for other areas such as Linked Data [Hartig and Zhao 2010], security and privacy [Martib et al. 2012], business processes [Cruz et al. 2013], ETL tools [Selis et al. 2007a], among others. In general, provenance data provides the documentation to register quality and authorship of a piece of data and the ability to reproduce and validate results of business or scientific processes [Freire et al. 2008].

There are different types of provenance data [Freire et al. 2008]. Prospective provenance data comprises an abstract process specification as a recipe for future data derivation. Unlike a specification, prospective provenance is, in general, independent from a model and it is intended to capture the recipe in an abstract and informative form to allow further querying of this information. Retrospective

provenance data consists in a structured and detailed history of the execution of a concrete process (computational tasks) and data derivation information, *i.e.*, which tasks were performed and how data artifacts were derived. Provenance data can be automatically captured during a process or workflow execution by a software engine or third-part provenance systems.

According to [Lebo et al. 2012], provenance has distinct granularities. The granularity refers the level of detail of provenance data. In this work we consider two levels of provenance granularity (coarse or fine) [Cruz et al. 2009]. Fine granularity refers to provenance data that is at or near the transaction level (*e.g.* parameters, time stamps, and file names). Data that is at this level is usually referred to as atomic level data. Coarse granularity refers to provenance data that is summarized or aggregated (*e.g.* process and datasets names). Different granularities are facets that can be exploited in the context of quality of Linked Data. The higher the granularity, the more refined will research possibilities between the published Linked Data and their provenance.

There has been several efforts to promote the interoperability of provenance models. The OPM model has recently been supplanted by the W3C PROV recommendation, whose model [**?**] is based in three OPM key concepts *Entity* (immutable state of an object), *Activity* (action performed on or entities) and *Agent* (responsible for acting in an activity). PROV-O is a PROV extension developed to add explicit representation of prospective provenance with semantic technologies [Erickson et al. 2016].

The Linked Data lifecycle was defined by Auer et al. [2012] as a set of different stages which include: Storage, Authoring, Interlinking, Classification, Quality, Evolution and Exploration. The stages, however, should not be tackled in isolation, but by investigating methods which facilitate the development of novel applications to solve challenges such as: offering trustworthiness and quality for data consumers. The project contemplates an stack of integrated aligned tools which support the whole life cycle of Linked Data from extraction/preparation/transformation, authoring/creation via enrichment, interlinking, fusing/publishing to maintenance. The stack was designed to be versatile having clear interfaces, enabling the plugging in of alternative third-party implementations such as data warehouse ETL (extract-transform-load) tools, RDF databases and workflow engines.

Even though the term ETL is traditionally related to data warehousing, it may be used in a wider sense to refer to any process of exchanging and transforming data between data stores [Selis et al. 2007a]. An ETL-workflow is a design blueprint for the ETL process, it share several characteristics of a data-oriented workflow in which input data are fed into an acyclic graph of transformations to produce output data [Selis et al. 2007b]. However, ETL-workflows are not modeled and executed in a scientific workflow engine (such as Panda [Ikeda et al. 2012], Kepler, among others). ETL-workflows can be designed by ETL tools, they use ETL components and are executed in ETL tools (such as Pentaho Kettle, Oracle Warehouse Builder, Sagent Data Flow, Powercenter, among others) which usually do not offer built-in provenance harvesting support. Thus, to circumvent this limitation, [de Mendonça et al. 2013] investigated the problem of how to generate and capture provenance to support data quality and trustworthiness in Linked Data using ETL-workflows enacted in traditional ETL tools.

3. RELATED WORK

One of the first works that deals with semantic support of provenance in Linked Data was VOID [Alexander et al. 2009], which is a vocabulary that allows metadata description of RDF datasets (general metadata access, structural and linksets) and the OPM model [Moreau et al. 2011], designed to for allow interoperability of provenance among different systems.

Among the major works that discuss provenance to Linked Data, we highlight the works of [Hartig and Zhao 2010], [Zhao et al. 2011] and [Zhao and Hartig 2012]. The authors developed a vocabulary in order to describe provenance of Linked Data with RDF. They also provide a way of publishing

the provenance description as Linked Data on the web. The authors stress the importance of making provenance interoperable and accessible on the Web to achieve a trustworthy Web of Data. For this, they suggest the use of vocabularies and provenance models. Besides,they consider that PROV-O provides the constructs for expressing some more complicated provenance patterns, such as describing additional attributes of relationships between entities and activities. However, they do not consider the role of Linked Data lifecycle or the LOD2 Project stack in their works. Thus, as far as we are concerned, they underestimate the importance of integrating the distinct kinds and granularities of provenance to the Linked Data to the transformation and publishing process.

Up to now, few provenance-related works consider tracking data transformations that occur within the boundaries of an organization (a governmental agency, a public or private company, for example) before a dataset is exposed as Linked Data in the Web of Data. Often, in the context of organizations, they are heterogeneous datasets, not aligned with the formats desired by analysts and therefore need to go through processes of preparation and pre-integration before being triplified. Related to this shortcoming, Faria et al. [2011] and de Mendonça et al. [2014] proposed an approach to support the exhibition and sharing of datasets in the form of Linked Data. Besides, they advocate the use of workflows to orchestrate the Linked Data publishing process and, have described several strategies to capture the provenance of the data transformation tasks. Freitas et al. [2012] and Omitola et al. [2012] investigated complementary perspectives of this approach. Freitas et al. [2012] defined a three-tier model to represent the provenance of the Linked Data publishing process and Omitola et al. [2012] emphasized the potential of using interactive tools of data transformation to support data transformation efforts and defined a provenance management architecture based representation model defined by Freitas et al. [2012].

Finally, it is important to note that the above mentioned works, neither consider the use of PROV recommendation for publishing and linking Linked Data with its provenance data or the interlinked nor consider a complete lifecycle of Linked Data or the stages defined by Auer et al. [2012] for the LOD2 Project.

4. ETL4LINKEDPROV

ETL4LinkedProv collects, links and publishes Linked Data provenance of ETL steps implemented as ETL-workflows enacted by datawarehouse ETL tools. The approach considers the multiple kinds of provenance data and granularities of provenance in the Linked Data publishing process carried out within the boundaries of an organization. Furthermore, it was designed to be compatible with LOD2 Project. Among the most closely associated stages of LOD2, we highlight the classification and Data Enrichment - they are aimed at recording the data published in ontologies to facilitate the integration, fusion and subsequent searches and also the Quality Analysis stage - that exposes that the analytical support can benefit from provenance, context and structure.

Figure 1 shows the overview of the architecture of our proposed approach. The ETL4LinkedProv introduces the component named Provenance Collector Agent (PCA), which captures, stores and interconnects provenance, related both to the composition (prospective provenance) and to the execution (retrospective provenance) of the publishing process. Such provenance is also published as Linked Data, enabling not only the joint exploitation of the data and the provenance via SPARQL queries.

ETL4LinkedProv works as an extension of the extraction stage of the LOD2 stack. As mentioned earlier, in the context of organizations, additional cleanup activities, consolidation, aggregation and pre-integration of data quite often need to be executed among the data extraction and triplification of the Linked Data. Thus, the ETL4LinkedProv uses ETL as a dataflow engine to enact ETL-workflows that publish data from heterogeneous data sources, also supporting the capture and publication of provenance data through RDF triples semantically annotated on existing provenance ontologies.

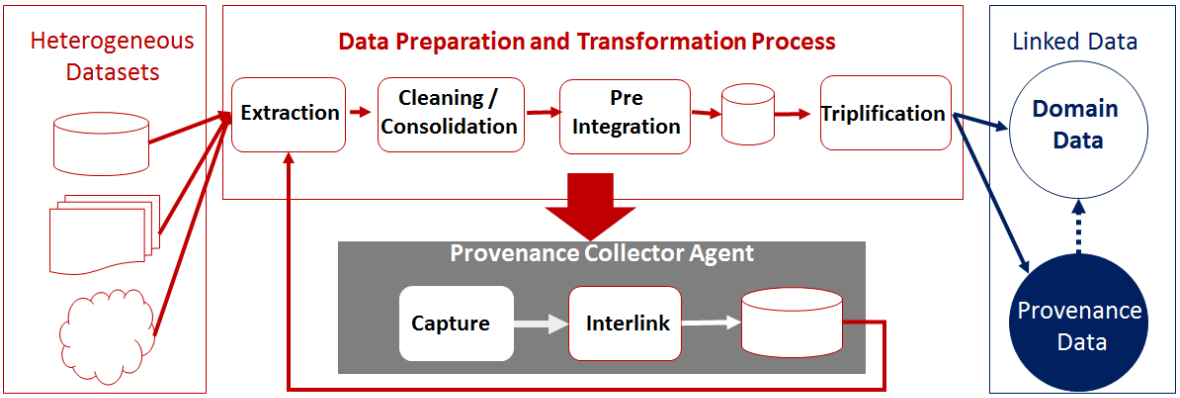

Fig. 1. ETL4LinkedData Architecture Overview.

4.1 Provenance Collector Agent

The Provenance Collector Agent (PCA) is a software component that encapsulates the ETL-workflow responsible for publishing Linked Data. It captures prospective and retrospective provenance and interconnects them to the domain data during the publishing process. The execution of an ETL-workflow consists of the execution of a set of interconnected steps, through which data flows in a one-way direction.

Each step corresponds to the *Activity* concept of PROV recommendation and is correspondent to an operation for implements the data extraction, transformation and loading process. A step can also contemplate the execution of ancillary activities such as transferring files, send e-mails or running a sub-workflow. The tasks carried out by the PCA component can be grouped into three distinct stages: capture, interconnection and temporary storage of provenance.

During the capture stage, the PCA monitors the events related to execution of the ETL-workflow and, whenever one of these events occurs, PCA performs the collection of multigranular provenance. The events monitored by PCA are: (i) the beginning and end of the main workflow or one of its sub-workflows (if exists); (ii) the beginning and end of each executed step; and (iii) reading of the data generated for a given step. A subset of step types can be selected, so that only the steps related to the selected type have the provenance with fine granularity level captured by the PCA. This selection of step types aims at minimizing the impact of the large amount of data generated by the provenance collection strategy with fine granularity. It occurs without neglecting the collection of provenance of the most important activities of the Linked Data publishing process.

In the interconnection stage, the PCA connects provenance collected during the capture stage. Interconnected provenance refer both to the publishing process and also to the data of the published domain. Furthermore, retrospective provenance are connected to the prospective provenance as RDF triples. Finally, in the temporary storage stage, the data are stored to be further extracted, processed and published as RDF triples by the Data Preparation and Transformation Process. The conceptual model of the temporary data repository used by the PCA (Figure 2) is composed of 13 entities, that comprise provenance data about the composition and on the publishing process.

To support the semantic publication of provenance temporarily stored by the PCA, our approach (ETL4LinkedProv) use a set of existing ontologies, which extend the details of provenance representation in the context of Linked Data. The PROV-O ontology [1] is used as the basis to represent the

[1] www.w3.org/ns/prov

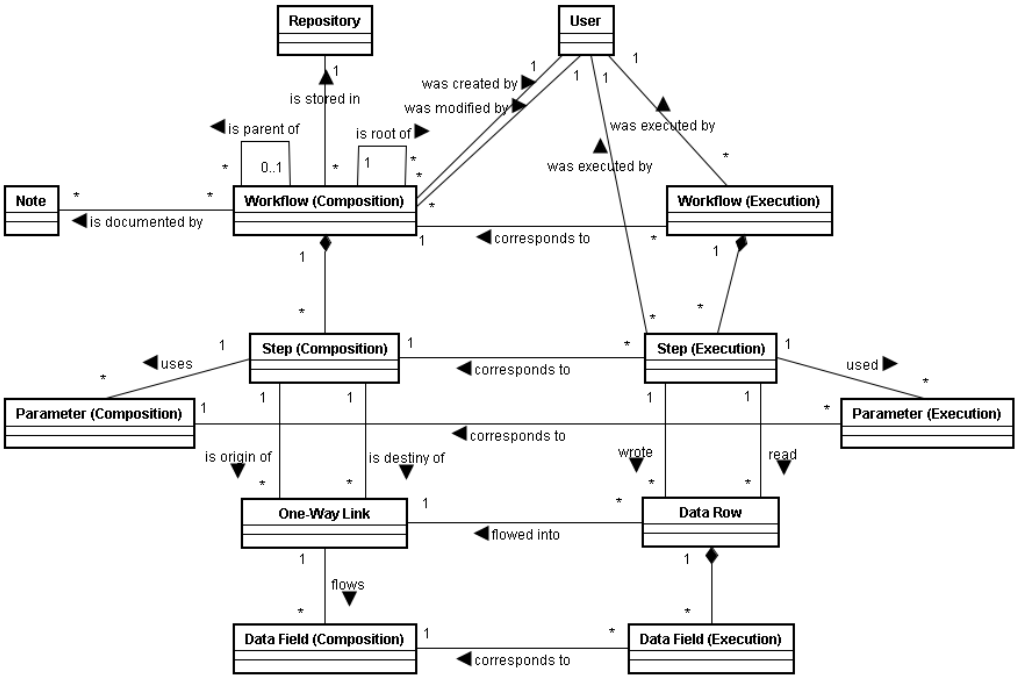

Fig. 2. Conceptual model of the temporary repository used by the PCA to store provenance.

semantics of the ETL-workflow and to enable the interoperability of provenance in the Web of Data.

The OPMW ontology [2] is used as an extension of PROV-O to distinguish the semantics of the workflow composition (prospective provenance) and the semantics of the execution of the workflow (retrospective provenance). The COGS ontology [3] is used as an extension of the PROV-O ontology to represent key concepts of the ETL process. Additionally, a specific vocabularies as Dublin Core (DC) and FOAF vocabularies [4] can also be used to complement the representation of the published provenance data.

4.2 Implementation of ETL4LinkedProv in Pentaho Kettle ETL tool

ETL4LinkedProv was conceived to be implemented independent of the particularities of an ETL tool. In this work, Pentaho Kettle [Matt Casters 2010] was used as the ETL tool that enacts to ETL-workflows. Pentaho Kettle has two types of components (transformations and jobs) that may be used to specify the ETL-workflows. A transformation component consists of a set of connected activities, where each one, called *step*, is responsible for an extraction activity or data load processing. The connection (named as *hop*) between two steps allows data to flow asynchronously in one direction. A job also consists of a set of connected steps. However, the steps of a job, called job entries, are responsible for performing a transformation, another job or auxiliary activities, such as transferring files or performing validation operations. The connection *hop* between two job entries determines their execution order, but running synchronously, unlike the transformation steps. Both transformations and jobs may have documentation and metadata, their specifications can be stored in repositories, for example as relational tables or XML files.

[2] www.opmw.org/ontology
[3] vocab.deri.ie/cogs
[4] www.foaf-project.org

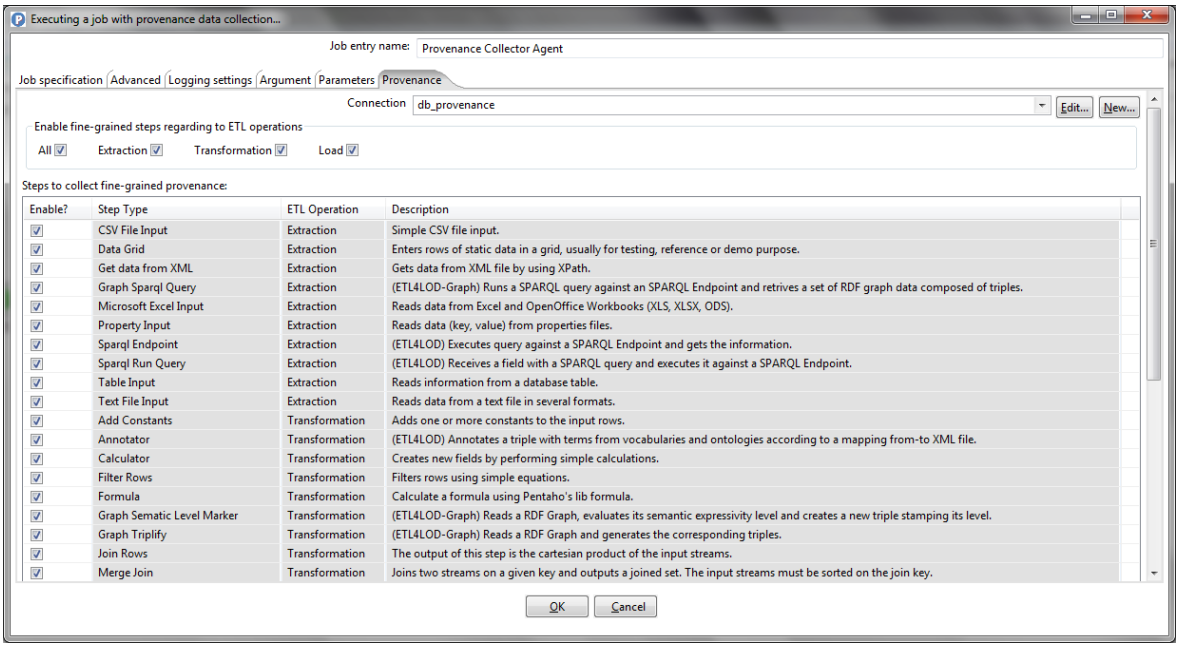

Fig. 3. Graphic Interface of PCA in Pentaho Kettle.

Despite offering a significant number of steps and job entries that perform a series of extraction, transformation and data loading operations, Pentaho Kettle does not contemplate in its default installation specific steps to work in the Linked Data context or collect provenance data. Thus, to overcome this and to enable implementation of ETL4LinkedProv, an additional set of steps, called ETL4LOD, has been developed, which allows the publication of domain data, as well as provenance in a database of RDF triples.

The PCA component has also been implemented with the Pentaho Kettle Java API, but as a job entry, whose type was called Provenance Collector Agent. Its configuration and graphic interface (Figure 3) has a tab called "Provenance". It enables the setup of the connection with the relational database used for temporary storage of provenance and the selection of the types of steps whose fine level granularity provenance will be collected. The ETL4LOD steps and PCA job entry are available at the ETL4LinkedProv[5] website.

5. USE CASE: INTEGRATING GOVERNMENTAL AGENCIES LINKED DATA

This section presents the use case discussed in this paper. In the scenario of today's national scientific research one can find difficult to reuse, consume and explore data from different research funding organizations. For example, actions such as auditing and correlating funding provided by different organizations with productivity in scientific publications usually are faced with many difficulties. Considering this real world issue, an example of application of ETL4LinkedProv was developed involving real data extracted from the National Scientific and Technological Development Council (CNPq) and the National Network of Education and Research (RNP).

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

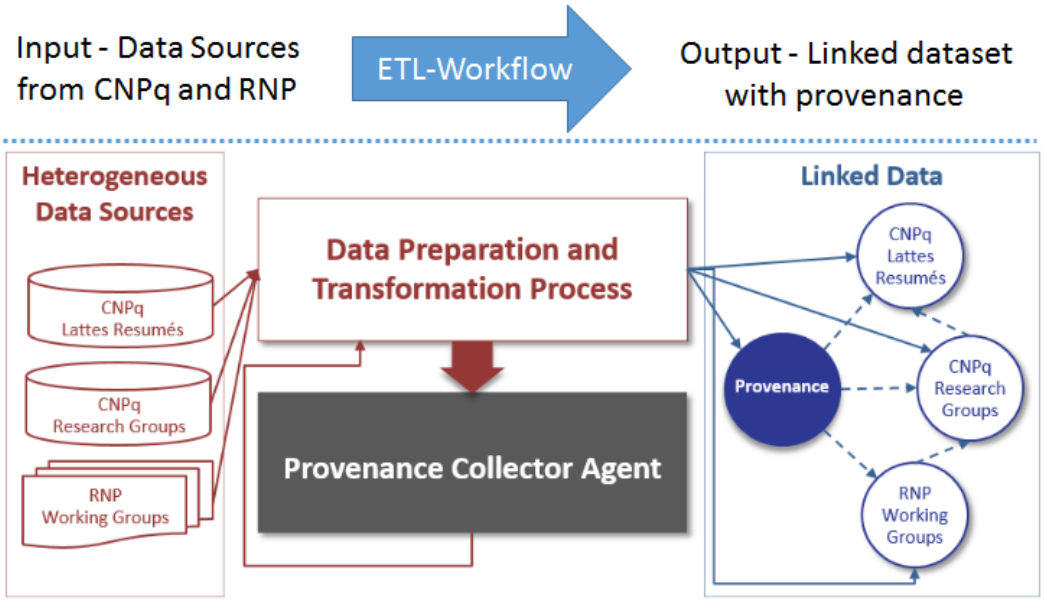

Fig. 4. General View of CNPq and RNP Integration Use Case.

5.1 Integrating CNPq and RNP datasets as Linked Data

CNPq is a Brazilian governmental organization responsible for fostering scientific and technological research groups and researchers in Brazil. RNP is another agency that finances research working groups through specific projects. The proposed scenario consists of integrating heterogeneous data from both CNPq and RNP data sources, gathering and publishing domain data and provenance about the steps and data involved in that integration through the use of ETL4LinkedProv.

Let us consider that the CNPq and RNP agencies have to publish their information as Linked Data. There are several benefits and advantages of having their data interlinked and to enable the joint consumption of these data by third-party Semantic Web applications. This verification would be a key advantage to support the validation, trustworthiness and check of consistency between data published by CNPq and RNP. Furthermore, it may promote a better understanding of this data domain and allow joint exploration of provenance data with increased quality and reliability.

In this use case, ETL4LinkedProv is able to encapsulate an ETL-workflow that integrates and publishes data from two heterogeneous data sources from both governmental agencies. The data sources of CNPq, both stored in a relational database, provide information about the Researchers Resumés and about the Research Groups respectively. The data source of RNP is stored in an XML repository. It provides information on research working groups that had projects financed by this agency. Figure 4 provides an overview of the use case integration scenario.

Figure 5 depicts an excerpt of the implemented ETL-workflow in Pentaho Ketlle ETL tool that sequences the Provenance Collector Agent (PCA) component and other five ETL-subworkflows. The PCA encapsulates the CNPq and RNP Data Preparation and Transformation Process and the other five ETL-subworkflows publishes the collected provenance data as RDF triples based on the PROV-O, OPMW and Cogs ontologies.

After the execution of the ETL-workflow, the CNPq and RNP data are published and interconnected through three RDF graphs: the first contains the RDF triples of the researchers' curricula; the second contains the RDF triples of the CNPq Research Groups and the third contains the RDF triples of

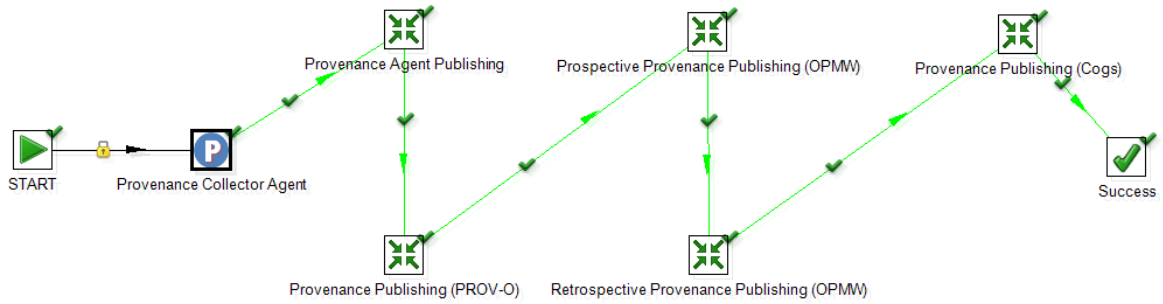

Fig. 5. Excerpt of ETL-workflow of integration of CNPq and RNP data sources.

the working groups that had projects with RNP. A fourth graph contains the RDF triples regarding provenance data captured by the PCA. The provenance of RDF triples is linked to the three RDF graphs of the domain. The approach allows the execution of SPARQL queries that use the data from the CNPq and RNP and also provenance associated to the transformations.

The experimental apparatus used in the execution of the use case experiments included the community edition of MySQL DBMS (version 5.0.51a), triple store Virtuoso (version 7.0.0), Kettle (4.3.0-stable version) and Java Standard Edition (version 1.7 .013). These platforms were installed locally on a notebook with Intel Core i3-2370M processor 2.4 GHz and 6 GB of main memory, running Windows 7 operating system in 64-bit version. The four published graphs were named respectively as *http://lattes.cnpq.br*, *http://www.cnpq.br*, *http://www.rnp.br* and *http://greco.ppgi.ufrj.br/provenance*.

5.2 Qualitative analysis of RDF triples and SPARQL queries

In this subsection we discuss a qualitative analysis using domain and provenance data results obtained from the execution of the ETL-workflow presented in Figure 5. The rationale behind this subsection is to show how ETL4LinkedProv users can take advantage of the integration of the different kinds of provenance and external domain data to execute SPARQL queries that otherwise would not be able to be performed. Due to space restriction we present three representative queries about our experiments.

The first SPARQL query in Listing 1 illustrates how provenance published by our approach can help to assess data consistency from different data sources. The second SPARQL query in Listing 2 shows how to retrieve both retrospective and prospective provenance data of a given process. Finally, the third SPARQL query in Listing 3 shows how to, starting from a given domain data, recover the prospective provenance of the related ETL-workflow.

```
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_wg ?rnp_wg_name ?iri_lattes_rschr ?lattes_rschr_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   GRAPH <http://greco.ppgi.ufrj.br/provenance> {
```

```
18      ?step_comp dc:title "Merge␣RNP␣Working␣Groups␣x␣CNPq␣Research␣Groups" .
19      ?field_iri_cnpq_rschr opmw:isGeneratedBy ?step_comp .
20      ?field_iri_cnpq_rschr dc:title "cnpq_rschr_uri" .
21      ?field_iri_cnpq_rg opmw:isGeneratedBy ?step_comp .
22      ?field_iri_cnpq_rg dc:title "cnpq_rg_uri" .
23      ?field_iri_rnp_wg opmw:isGeneratedBy ?step_comp .
24      ?field_iri_rnp_wg dc:title "uri_wg" .
25      ?step_exec opmw:correspondsToTemplateProcess ?step_comp .
26      ?data_iri_cnpq_rschr prov:wasGeneratedBy ?step_exec .
27      ?data_iri_cnpq_rschr opmw:correspondsToTemplateArtifact ?field_iri_cnpq_rschr .
28      ?data_iri_cnpq_rschr prov:value ?value_iri_cnpq_rschr .
29      ?data_iri_cnpq_rg prov:wasGeneratedBy ?step_exec .
30      ?data_iri_cnpq_rg opmw:correspondsToTemplateArtifact ?field_iri_cnpq_rg .
31      ?data_iri_cnpq_rg prov:value ?value_iri_cnpq_rg .
32      ?data_iri_rnp_wg prov:wasGeneratedBy ?step_exec .
33      ?data_iri_rnp_wg opmw:correspondsToTemplateArtifact ?field_iri_rnp_wg .
34      ?data_iri_rnp_wg prov:value ?value_iri_rnp_wg .
35      BIND(IRI(?value_iri_cnpq_rschr) AS ?iri_cnpq_rschr) .
36      BIND(IRI(?value_iri_cnpq_rg) AS ?iri_cnpq_rg) .
37      BIND(IRI(?value_iri_rnp_wg) AS ?iri_rnp_wg) .
38    }
39    GRAPH <http://www.cnpq.br> {
40      ?iri_cnpq_rg a cnpq:ResearchGroup .
41      ?iri_cnpq_rg cnpq-prop:name "GRECO-Grupo␣Engenharia␣do␣Conhecimento"@pt .
42      ?iri_cnpq_rschr owl:sameAs ?iri_lattes_rschr .
43    }
44    GRAPH <http://lattes.cnpq.br> {
45      ?iri_lattes_rschr a lattes:Researcher .
46      ?iri_lattes_rschr lattes-prop:name ?lattes_rschr_name .
47    }
48    GRAPH <http://www.rnp.br> {
49      ?iri_rnp_wg a rnp:WorkingGroup .
50      ?iri_rnp_wg rnp-prop:name ?rnp_wg_name .
51    }
52      }
```

Listing 1. SPARQL query that returns the RNP Working Groups associated with the CNPq Research Group "GRECO"

The first SPARQL query returns the RNP Working Groups associated with the CNPq Research Group "GRECO" and the associated researchers involved. The "Merge RNP Working Groups x CNPq Research Groups" step is the process that associates the RNP Working Groups and CNPq Research Groups, through researchers common to both.

The interval 17-38 in Listing 1 defines the graph pattern to match against the graph *http://greco.ppgi.ufrj.br/provenance* for the subgraph with the defined data fields (prospective provenance) and the generated values (retrospective provenance) in the execution of "Merge RNP Working Groups x CNPq Research Groups" step. The variables *?value_iri_cnpq_rschr*, *?value_iri_cnpq_rg* and *?value_iri_rnp_wg* contain respectively the corresponding texts of CNPq Researcher, CNPq Research Group and RNP Working Group resources IRIs, so, using IRI and BIND functions, the resources are bound in *?iri_cnpq_rschr*, *?iri_cnpq_rg* and *?iri_rnp_wg*. The graph patterns defined in lines 39-43, 44-47 and 48-51 connect the resources bound in *?iri_cnpq_rschr*, *?iri_cnpq_rg* and *?iri_rnp_wg* with the domain data and apply the filter to get information related to "GRECO-Grupo Engenharia do Conhecimento" Research Group.

If the list of "GRECO" projects retrieved from CNPq graphs does not contain a RNP Working Group returned by the SPARQL query in Listing 1, it is possible to investigate the reason for the inconsistency with the support of the recovered provenance. Jointly analyzing the domain and provenance data,

```
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␣RNP␣Working␣Groups␣x␣CNPq␣Research␣Groups" .
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
```

Listing 2. SPARQL query that returns prospective and retrospective provenance represented by the parameterization used in the execution of the "Merge RNP Working Groups x CNPq Research Groups" process performed by Rogers at 04/11/2013 at 14:30:06

one can evaluate whether the inconsistency was due to a problem with the Linked Data publishing process of CNPq and RNP data or, still, if it was just an error from a researcher when declaring his/her associations at CNPq.

The second SPARQL query returns the prospective and retrospective provenance represented by the parametrization used in the execution of the "Merge RNP Working Groups x CNPq Research Groups" process, when it was performed by an agent (user) named Rogers at 04/11/2013 at 14:30:06.

Line 10 in Listing 2 filters the composition of the step named "Merge RNP Working Groups x CNPq Research Groups". Line 11 retrieves all the executions of the step. Lines 12, 13 and 14 filters the execution performed by the agent named "Rogers" at a given moment in time (04/11/2013 at 14:30:06). Lines 16 and 17 retrieves all the parameters specified at the composition of the "Merge RNP Working Groups x CNPq Research Groups" step. Lines 18 and 19 retrieves all the instances of the parameters used by the step filtered in the lines 12-14. Finally, lines 20 and 21 retrieves the name and the value of each parameter and line 23 specifies the ascending sort order of the result by parameter names and values retrieved.

In the last example of SPARQL query, we can investigate a typical question that is answered with descriptors that were generated before the execution of an ETL-workflow. The third SPARQL query returns prospective provenance data collected by PCA that notify when the source code of a ETL-workflow named "TransfCNPq" was last modified and who made the change.

In Listing 3, lines 8 and 9 recover all versions of the ETL-workflow named "TransfCNPq". Line 10 retrieves the compositions (prospective provenance) of all steps of this ETL-workflow and line 11 retrieves the executions of such steps. The domain data generated during the execution of the steps are recovered by line 12 and line 13 filters this domain data using the text of IRI of the researcher "Maria Luiza". Lines 14 and 15 recover, through Dublin Core properties, the modification date and the agent who made the change in the composition of the ETL-workflow. Finally, lines 17 and 18

```
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX opmw: <http://www.opmw.org/ontology/>
PREFIX prov: <http://www.w3.org/ns/prov#>
SELECT ?w_comp ?w_version ?w_modifier ?w_modified
FROM <http://greco.ppgi.ufrj.br/provenance>
WHERE
{
        ?w_comp dc:title "TransfCNPq" .
        ?w_comp opmw:versionNumber ?w_version .
        ?step_comp opmw:isStepOfTemplate ?w_comp .
        ?step_exec opmw:correspondsToTemplateProcess ?step_comp .
        ?data prov:wasGeneratedBy ?step_exec .
        ?data prov:value "http://www.cnpq.br/resource/Researcher/MARIA_LUIZA".
        ?w_comp dc:contributor ?w_modifier .
        ?w_comp dc:modified ?w_modified .
}
GROUP BY ?w_comp ?w_version ?w_modifier ?w_modified
HAVING ?w_version = MAX(?w_version)
```

Listing 3. SPARQL query that returns prospective provenance that identifies when and by whom the source code of an ETL-workflow that published the resource "http://www.cnpq.br/resource/Researcher/MARIA_LUIZA" in the Web of Data was last modified

group the results by the selected variables and filter the group related to the latest version of the composition of "TransfCNPq" ETL-workflow.

5.3 Quantitative analysis and performance evaluations

In this subsection we present quantitative evaluations of the ETL4LikedProv. Our results are based on the relationship between the extracted data and the number of RDF triples published by the ETL4LinkedProv, it is possible to compute the metrics that show the impact of provenance granularity

Extraction						Load	Level 1	Level 2	Level 3
Source	**Entity**	**I**	**A**	**P**	**$\sum R$**	**T**	**t_1 mm:ss**	**t_2 mm:ss**	**t_3 mm:ss**
CNPq – Lattes Resumé	Researcher	7	1	5	478	2242	00:02	00:03	06:10
	Production	217	1	5					
	Project	42	1	9					
CNPq – Research Groups	Research Group	3	1	2	158	425	00:01	00:02	02:10
	Researcher	69	1	2					
	Coordinator	4	1	0					
	Student	47	1	0					
RNP – Working Groups	Working Group	63	1	6	652	2267	00:02	00:04	09:50
	Institution	18	1	2					
	Researcher	362	1	2					
	Researcher	34	1	0					
Total		866			1288	4934	00:05	00:09	18:10

Table I. Execution time of ETL-workflows used to publish CNPq and RNP data sources as Linked Data.

in performance and the volume of published data during the Data Preparation and Transformation Process. Thus, three different settings of the captured provenance granularity level were applied in the use case described in the previous sub-sections.

In the first experiment, the ETL-workflow that published data from CNPq-Lattes, CNPq-Research Group and RNP-Working Groups was executed without being encapsulated by the PCA. In the other configurations, the same workflow has been performed by the encapsulated PCA. In the second experiment, no step was selected to have the provenance collected with fine granularity. In the third experiment, all kinds of steps were selected to have their provenance collected with fine granularity.

The number of RDF triples published by ETL4LinkedProv (for both domain and provenance data) follows the model represented by the formula:

$$T = \sum_{i=1}^{n} (I_{\mathrm{i}} * A_{\mathrm{i}} + I_{\mathrm{i}} * P_{\mathrm{i}}) + \sum_{i=1,j=1}^{n} R_{\mathrm{ij}}$$

where: T is the number of published RDF triples; n is the number of extracted entities of the data source by the ETL-workflow; I_{i} is the number of extracted instances of entity i; A_{i} is the number of RDF types related to entity i; P_{i} is the number of properties extracted from entity i, R_{ij} is the number of relationships between the instances of the entity i and the instances of the entity j.

Table I shows the result of the first quantitative analysis, which consisted of comparing the times of execution of the ETL-workflows during the publication of the CNPq and RNP data sources and RNP as Linked Data, considering the three different settings of granularity level of the source. t_1 is the runtime without PCA performance and t_2 and t_3, the runtime with PCA performance, respectively, with any and all types of selected steps with provenance capture with fine granularity.

Usually, the ETL-workflows would be executed at a high performance computational infrastructure and would extract a larger volume of data. However, the impact of PCA's role in encapsulated ETL-workflow runtime would also be related to the configuration of the granularity level of the provenance to be collected. Without the fine granularity level enabled, *i.e.* only with more generic collection coming from metadata about the publishing process, the duration (t_2) of the ETL-workflow is very close to the runtime performance without PCA (t_1). However, the execution time grows exponentially according to the number of steps configured to be collected at the provenance level of each data read

		PROV-O		OPMW		Cogs		DC		Level 2				Level 3			
	Entity	**A**	**P**	**A**	**P**	**A**	**P**	**A**	**P**	**I**	**∑ *R***	**T**	**t_4 mm:ss**	**I**	**∑ *R***	**T**	**t_5 mm:ss**
Composition	Workflow	1	2	1	4	0	0	0	4	4	480	1727	00:09	4	551396	1141388	20:13
	Step	0	0	1	0	1	0	0	2	105				105			
	Parameter	0	0	2	0	1	0	0	1	0				1428			
	Data Field	0	0	2	0	1	0	0	1	0				517			
Execution	Workflow	2	2	1	4	0	0	0	1	4				4			
	Step	1	2	1	2	0	0	0	1	105				105			
	Parameter	1	1	1	1	0	0	0	0	0				831			
	Data Field	2	0	0	0	1	0	0	0	0				31055			
	Data Row	1	1	1	1	0	0	0	0	0				121119			
	User	1	0	0	0	0	0	1	2	1				1			

Table II. Amount of provenance RDF triples and execution times of the ETL-workflows. .

and manipulated by the steps. As usually data repetitions occur and there is less relevant data to apply to ETL4LinkedProv, it is recommended to perform a preliminary analysis of the provenance grain that is best suited to be configured in PCA. This configuration must meet the requirements of joint exploration of domain data and its provenance. At the same time, it must not unnecessarily increase the runtime of the Linked Data publishing process.

Regarding the number of provenance data triples and the duration of the execution of the ETL-workflow that has published them, Table II presents the results considering the last two settings of provenance granularity level applied in the PCA. The number of RDF types (A) related to the temporary repository entities used by the PCA (Figure 1) and the number of properties (P) were recorded for each ontology adopted to represent the provenance in the context of Linked Data. The entities *Repository*, *Note* and *One-Way Link* in the conceptual model are not represented by classes of the provenance ontologies, but their attributes are used as objects of properties from other entities. For each level of granularity of provenance, there have been extracted the number of instances (I) of the temporary repository, the sum of relationships between instances ($\sum$R) and the run time of the ETL-workflow for provenance publication (t).

Similar to the results of the first analysis, the impact of PCA of the runtime of the ETL-workflow for provenance publication as well as to the number of published RDF triples is directly related to the level of configured provenance granularity. This impact is intensified in the analysis of retrospective provenance data on the implementation of the ETL-workflow.

Finally, in order to investigate the amount of RDF triples generated by ETL4LikedProv, we executed a third quantitative investigation where we compared the number of coarse and fine provenance RDF triples published by ETL4LinkedProv and the execution times of ETL-subworkflows. In this evaluation we considered two different experiments. In the first one, we configured the PCA to collected only coarse provenance data of the ETL-subworkflows (no steps were selected). In the second experiment, we did the opposite, we configured the PCA by selecting all types of steps of the ETL-subworkflows to collect fine granularity provenance. Table III shows the result of such investigation.

Table III shows that PCA component is capable to be configured to collect and publish provenance data with multiple granularity. In these experiments, we measured the execution times of each data transformation jobs (*i.e.* sub-workflows that execute a set of data transformations steps based on the PROV-O, OPMW and Cogs ontologies) and the amount of collected RDF provenance triples.

We observe that the total amount of coarse provenance triples collected is small (2,099 triples). It

	Experiment 1 (coarse provenance)		**Experiment 2 (fine provenance)**	
ETL-Subworkflow (Job Names)	**Numers of RDF triples**	**Execution time t_1 mm:ss**	**Numers of RDF triples**	**Execution time t_2 mm:ss**
Provenance Agent Publishing	10	00:00	10	00:00
Provenance Publishing (PROV-O)	714	00:02	1.086.114	12:13
Prospective Provenance Publishing (OPMW)	373	00:01	8.685	00:06
Retrospective Provenance Publishing (OPMW)	872	00:02	724.557	06:33
Provenance Publishing (COGS)	130	00:02	244.836	01:21
Totals	2.099	00:09	2.064.222	20:13

Table III. Comparing the amount of fine and coarse provenance data and sub-workflows executions time. .

is about a thousand times smaller than the total amount of fine grained provenance triples collected during the second experiment. We stress that the second and the fourth columns of Table III details the number of coarse and fine prospective and retrospective triples collected by PCA in each job. Table III also shows that the execution times of the jobs vary according to the amount of provenance collected. In the first experiment the total execution time (t_1) is about nine seconds, it is very small when compared with the second experiment, where (t_2) is about twenty minutes and thirteen seconds. However, in the second experiment more data transformations were executed. Thus, about 2,064.222 provenance triples were collected and published during this experiment. Finally, we can conclude that, as expected, the execution times of the sub-workflows increase as more fine grained provenance triples are collected by PCA component.

6. CONCLUSION

Provenance has been recognized as a key mechanism to support quality assessment and data consistency and on integration and interoperability efforts. In particular, in the context of Linked Data, where domain data from different sources are being manipulated, transformed and interconnected they certainly can benefit from the capture and subsequent publication of provenance data so that they can be published together with the corresponding data.

Provenance is essential in ETL processes and can be viewed at different levels of detail. In order to bring quality and trustworthiness to the Web of Data, in this article, we presented ETL4LinkedProv and PCA, a component supported by a ETL tool to capture and publish provenance data in the form of Linked Data designed to be compatible with LOD2 Project. We have explored the different levels of detail of these descriptors and their implications in terms of performance and number of complementary triples generated. We presented an use case and several experiments showing that it is possible to take advantage of the flexibility to configure the process of capturing provenance data in order to adjust the volume and performance needs of each situation.

As future work, we point out new strategies for linking data and provenance, exploring the possibilities of reification of triples and subgraphs. Index mechanisms may provide the basis for the recovery process of the information provided by data producers. In addition, as future works it is intended to test the distribution of SPARQL queries processing in the case of using fine granularity provenance data, taking advantage of the parallelism and higher processing capability.

7. ACKNOWLEDGMENTS

We want to thank all agencies that supported this scientific research: FAPERJ (to M.L.M.C., process number E-26/110.492/2012) and CNPq (to M.L.M.C., process number 308934/2012-1). We especially thank all the team members from the group GRECO/UFRJ who helped in this work.

REFERENCES

Alexander, K., Cyganiak, R., Hausenblas, M., and Zhao, J. Describing Linked Datasets - On the Design and Usage of voiD, the 'Vocabulary of Interlinked Datasets'. In *Proceedings of the Linked Data on the Web*. Madrid, Spain, 2009.

Auer, S., Buhmann, L., Dirschl, C., Erling, O., Hausenblas, M., Isele, R., Lehmann, J., Martin, M., Mendes, P. N., van Nuffelen, B., Stadler, C., Tramp, S., and Williams, H. pp. 1–16. In , *Managing the Life-Cycle of Linked Data with the LOD2 Stack*. Lecture Notes in Computer Science, vol. 7650. Springer, pp. 1–16, 2012.

Berners-Lee, T., Hendler, J., and Lassila, O. The semantic web. *Scientific American* 284 (5): 28–37, 2001.

Bizer, C., T., H., and Berners-Lee, T. Linked data - the story so far. *International Journal on Semantic Web and Information Systems* 5 (3): 1–22, 2009.

Breitman, K., Salas, P., Casanova, M., Saraiva, D., Viterbo, J., Magalha es, R., Franzosi, E., and Chaves, M. Open government data in brazil. *Intelligent Systems* 27 (3): 45–49, 2012.

Buneman, P., Khanna, S., and Wang-Chiew, T. Why and where: A characterization of data provenance. In *Proceedings of International Conference Database Theory*. Berlin, Heidelberg, pp. 316–330, 2001.

Cordeiro, K. F., Campo, M. L. M., and Borges, M. R. S. Empowering Citizens and Government with Collaboration on Linked Open Data. In *Proceedings of Workshop Semantics in Governance and Policy Modelling* . Crete, Grece, pp. 33–37, 2011.

Cruz, S., Costa, R. J. M., Manhães, M., and Zavaleta, J. Monitoring SOA-based applications with business provenance. In *Proceedings of the Annual ACM Symposium on Applied Computing* . New York, USA, pp. 1927–1932, 2013.

Cruz, S. M. S., Campos, M. L. M., and Mattoso, M. Towards a taxonomy of provenance in scientific workflow management systems. In *Services - I, 2009 World Conference on*. Los Angeles, USA, pp. 259–266, 2009.

de Mendonça, R. R., Cruz, S. M. S., and Campos, M. L. M. Gerência de proveniência multigranular em linked data com a abordagem etl4linkedprov. In *Anais do Simpósio Brasileiro de Bancos de Dados*. Paraná, Brazil, 2014.

de Mendonça, R. R., Cruz, S. M. S., De La Cerda, J. F. S., M. Cavalcanti, M. L., Cordeiro, K. F., and Campos, M. L. M. Lop: Capturing and linking open provenance on lod cycle. In *Proceedings of the Workshop on Semantic Web Information Management*. New York, NY, USA, pp. 3:1–3:8, 2013.

Ding, L., Peristeras, V., and Hausenblas, M. Linked open government data - guest editors introduction. *Intelligent Systems* 27 (3): 11–15, 2012.

Erickson, J. S., Sheehan, J., Bennett, K. P., and McGuinness, D. L. Addressing scientific rigor in data analytics using semantic workflows. In *Proceedings of the Provenance and Annotation of Data and Processes*. Berlin, Heidelberg, pp. 187–190, 2016.

Faria, F. F., Pereira, B. O., Freitas, A., Ribeiro, C. E., Freitas, J. V. B., Bringuente, C., Arantes, L. O., Calhau, R., Zamborlini, V., Campos, M. L. M., and Guizzardi, G. An approach for managing and semantically enriching the publication of Linked Open Governmental Data. In *Proceedings of Workshop Semantics in Governance and Policy Modelling* . Santa Catarina, Brazil, pp. 82–95, 2011.

Freire, J., Koop, D., Santos, E., and Silva, C. T. Provenance for computational tasks: A survey. *Computing in Science and Engineering* 10 (2): 11–21, 2008.

Freitas, A., Kampgen, B., Oliveira, G., J., and Curry, E. Representing interoperable provenance descriptions for etl workflows. In *Proceedings of the International Workshop on Role of Semantic Web in Provenance Management*. Berlin, Heidelberg, pp. 43–57, 2012.

Hartig, O. and Zhao, J. pp. 78–90. In , *Publishing and consuming provenance metadata on the Web of Linkied Data*. Lecture Notes in Computer Science, vol. 6873. Troy, USA, pp. 78–90, 2010.

Heath, T. and Bizer, C. pp. 1–136. In , *Linked Data: Evolving the Web into a Global Data Space*. Morgan and Claypool, pp. 1–136, 2011.

Ikeda, R., Cho, J., Fang, C., and Widom, J. pp. 1249–1252. pp. 1249–1252, 2012.

Lebo, T., Wang, P., Graves, A., and McGuinness, D. L. Towards unified provenance granularities. In *Proceedings of the Provenance and Annotation of Data and Processes*. Berlin, Heidelberg, pp. 39–51, 2012.

Martib, A., Lyle, J., and Namilkuo, C. Provenance as a Security Control. In *Proceedings of Workshop on the Theory and Practice of Provenance* . Boston, USA, pp. 1–4, 2012.

Matt Casters, Roland Bouman, J. v. D. In , *Pentaho Kettle Solutions Building Open Source ETL Solutions with Pentaho Data Integration*. Wiley Publishing Inc, Indianapolis, 2010.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., and den Bussche, J. V. The open provenance model core specification (v1.1). *Future Generation Computer Systems* 27 (6): 743 – 756, 2011.

Omitola, T., Freitas, A., Curry, E., and Shadbolt, N. Capturing interactive data transformation operations using provenance workflows. In *Proceedings of the International Workshop on Role of Semantic Web in Provenance Management*. Crete, Greece, pp. 29–42, 2012.

Selis, T., Skoutas, D., and Simitsis, A. ETL workflows: from formal specification to optimization. In *Proceedings of the East European conference on Advances in databases and information systems*. pp. 1–11, 2007b.

Selis, T., Skoutas, D., Simitsis, A., and Vassiliadis, P. Data Provenance in ETL Scenarios. In *Proceedings of the workshop of principles of provenance*. New York, USA, pp. 1–3, 2007a.

Sheridan, J. and Tennison, J. Linking UK Government Data. In *Proceedings of the Linked Data on the Web*. North Carolina, USA, pp. 1–4, 2010.

Zhao, J. and Hartig, O. Towards interoperable provenance publication on the linked data web. In *Proceedings of the Linked Data on the Web*. Lyon, France, 2012.

Zhao, J., Sahoo, S. S., Missier, P., and Sheth, A.and Goble, C. Extending semantic provenance into the web of data. *Internet Computing* 15 (1): 40–48, 2011.