LDWPO – A Lightweight Ontology for Linked Data Management

Sandro Rautenberg[1], Ivan Ermilov[2], Edgard Marx[2], Sören Auer[3]

[1]Computer Science Department – Midwestern State University (UNICENTRO)
PO Box 730 – Postal Code 85.015-430 – Guarapuava – PR – Brazil

[2]AKSW, Institute of Computer Science, University of Leipzig
Leipzig, Germany.

[3]University of Bonn and Fraunhofer IAIS
Bonn, Germany.

srautenberg@unicentro.br

Abstract. *Managing the lifecycle of RDF datasets is a cumbersome activity. Substantial efforts are spent on reproducing datasets over time. But, these efforts can be reduced by a data management workflow framework. We present the Linked Data Workflow Project ontology as the knowledge model for such a workflow framework. The ontology is centered on the Plan, Method, and Execution classes, facilitating the description of: i) the methodological process that guides the lifecycle of RDF datasets, ii) the complete plan of the RDF dataset production workflow, and iii) the executions of workflow. As a result, our approach enables the reproducibility and repeatability of Linked Data processing steps over time.*

1. Introduction

In the context of the Web of Data, the management of data collections encoded according to the Resource Description Framework (RDF dataset[1]) has been mainly focused on developing tools for supporting individual aspects of Linked Data Management (extraction, mapping/transformation, quality assessment/repairing, linking, and publishing/visualization). With this in mind, managing the complete lifecycle of RDF datasets over time can become a problem, due to the myriad of tools, environments, and data sources. Thus, that lifecycle requires substantial management effort for detailing provenance, ensuring reproducibility, and dealing with repeatability issues.

To facilitate the data management, workflow and provenance ontologies (or vocabularies) can be used to describe and automatize the linked data lifecycle. *Scufle2* [Hull et al. 2006] and *Kepler* [Ludäscher et al. 2006] are examples of such ontologies used as knowledge models in some Workflow Management Systems. With regard to ontology engineering best practices, those ontologies reveal important limitations. *Scufle2* is not available[2] and *Kepler* ontologies do not detail their elements with human-readable descriptions. These limitations hinder the adoption of those ontologies, mainly, for: i)

[1]Formally, it is a dataset "used to organize collections of RDF graphs, and comprise a default graph and zero or more named graphs [W3C 2014].

[2]The ontology is not published `http://taverna.incubator.apache.org/documentation/scufl2/ontology` 27-10-2015 17:00

reusing them as knowledge sources in other ontology developments; ii) extending them for sharing information among systems. Taking the provenance perspective into account, the *PROV ontology* (PROV-O) [Lebo et al. 2015] and the *Open Provenance Model Vocabulary* (OPMV) [Moreau et al. 2011] can be adopted. However, they lack crucial concepts to describe the plan and execution perspectives of a workflow. In a nutshell, PROV-O and OPMV are insufficient for describing, at the same time, the strategy (plan) and operation (execution) aspects of (re)producing RDF datasets.

Tackling the limitations of existing approaches, we model a lightweight ontology for orchestrating linked data processing workflows, dubbed the *Linked Data Workflow Project ontology* (LDWPO). To develop LDWPO, we applied artifacts and best practices from *On-to-Knowledge* [Sure and Studer 2002], *METHONTOLOGY* [Gomez-Perez et al. 2004], and the *Ontology Development 101 Guide* [Noy and McGuinness 2001]. Inspired on other knowledge sources, LDWPO standardizes the `Method`, `Plan`, and `Execution` concepts for guiding the production and maintenance of RDF datasets. It is noteworthy that the LDWPO is already used as the knowledge model in *LODFlow* [Rautenberg et al. 2015], an environment for planning, executing, and documenting workflows for linked data. LDWPO was verified in large-scale real-world use cases, expressing the: i) creation of RDF datasets according to a methodological process; ii) planning of RDF dataset maintenance on an high level of abstraction, thus, enabling provenance extraction and reproducibility over time; and iii) execution of the workflows for RDF dataset (re)production in a (semi-)automatized way, using *Linked Data Stack* technologies [Auer et al. 2012].

The article is structured as follows: The LDWPO scope and purposes are presented in Section 2. Section 3 discusses the main concepts of LDWPO. Section 4 describes the LDWPO evaluation with two large-scale real-world use cases for promoting the knowledge management in a Brazilian university. Section 5 presents related work that is complementary to LDWPO. Finally, Section 6 outlines conclusions and some directions for future work.

2. Preliminaries

LDWPO's scope is limited to the linked data domain, extending concepts for methodologically planning and executing the (re)production of RDF datasets. The main requirements addressed by the LDWPO are:

1. describing the methods which establish the process, activities, and tasks for producing plans of RDF datasets;
2. representing the plans as workflows for (re)producing RDF datasets over time. It is achieved by specifying a list of steps, where each step corresponds to a tool invocation using a specific tool configuration, as well as input and output datasets;
3. supporting the reuse of workflows for guiding the (re)production of RDF datasets over time;
4. mediating the automation of workflows, which involves a controlled environment for a plan execution. It should be achieved by running tools with tool configurations over input datasets as previously planned;
5. preserving workflow execution logs for checking the correctness or repeatability of the results; and

6. reporting projects of RDF datasets, its workflow plans and executions in human-readable formats.

Considering the scope, purposes, and competence questions[3], we listed adherent ontologies and vocabularies, aiming the reusing of existing concepts and properties. We identified the *Publishing Workflow Ontology* (PWO) [Gangemi et al. 2014], the *Open Provenance Model Vocabulary* (OPMV) [Moreau et al. 2011], and the *PROV Ontology* (PROV-O) [Lebo et al. 2015]. These approaches are suitable for describing the *execution* of an RDF dataset and, therefore, can answer the questions about *what was done* during the RDF dataset maintenance. However, these works do not include important concepts such as *method* and *plan*. In particular, the *method* concept can answer questions about *how or why to proceed*. Instances of this concept support a knowledge engineering perspective of linked data, where an established process is related to the knowledge level of lifecycle, standards, and best practices [Bourque and Fairley 2004]. The *plan* concept answers questions about the actions related to a workflow or simply *what to do with something over time*. Instances of *plan* are related to the knowledge level for scheduling the tools, steps, and resources [WFMC 1995], supporting the lifecycle of RDF datasets in a systematic way.

In such way, we are proposing the LDWPO[4] as a new and complementary ontology to support the *method*, *plan*, and *execution* concepts for better representing and implementing the RDF dataset maintenance.

3. LDWPO in a Nutshell

In LDWPO (Figure 1), the main concepts are dubbed with the prefix "LDW", specializing some general concepts to the context of workflows for RDF dataset (re)production.

The starting point in LDWPO is the `LDWProject` concept, a description of a project for creating/maintaining RDF datasets. Among its properties, `LDWProject` is associated with a set of `LDWorkflows`. An `LDWorkflow` embodies the plan necessary to (re)produce `RDFDatasets`, encapsulating a linked list of `LDWSteps`. `LDWStep` is a concept that represents an atomic and reusable unit of an `LDWorkflow`. It describes a set of procedures over a set of input `Datasets`, using a `Tool` with a `Tool Configuration`, in order to produce a set of output `Datasets`. An `LDWStep` can be reused, which means that the same `LDWStep` can be associated with one or more `LDWorkflows` within existing `LDWProjects`. In addition, an `LDWStep` can be automatically executed in a computational environment, on a user request. We exemplify the automatization in more detail in Section 4, with real-world use cases.

An `LDWorkflow` can be reused as a `Plan` in `Executions` at any particular point of time. In LDWPO, the concept for describing an `LDWorkflow` execution instantiation is `LDWorkflowExecution`. Each `LDWorkflowExecution` needs to aggregate the sequence of `LDWStepExecutions` close related to the sequence of `LDWSteps` of a given `LDWorkflow`. In other words, it meets a representation for automating the execution of workflows, by running tools with tool configurations over datasets as it is previously

[3]A detailed technical report is available at: `https://github.com/AKSW/ldwpo/blob/master/misc/technicalReport/LDWPO_technicalReport.pdf`

[4]The ontology is available at: `https://github.com/AKSW/ldwpo/blob/master/1.0/ldwpo.owl`.

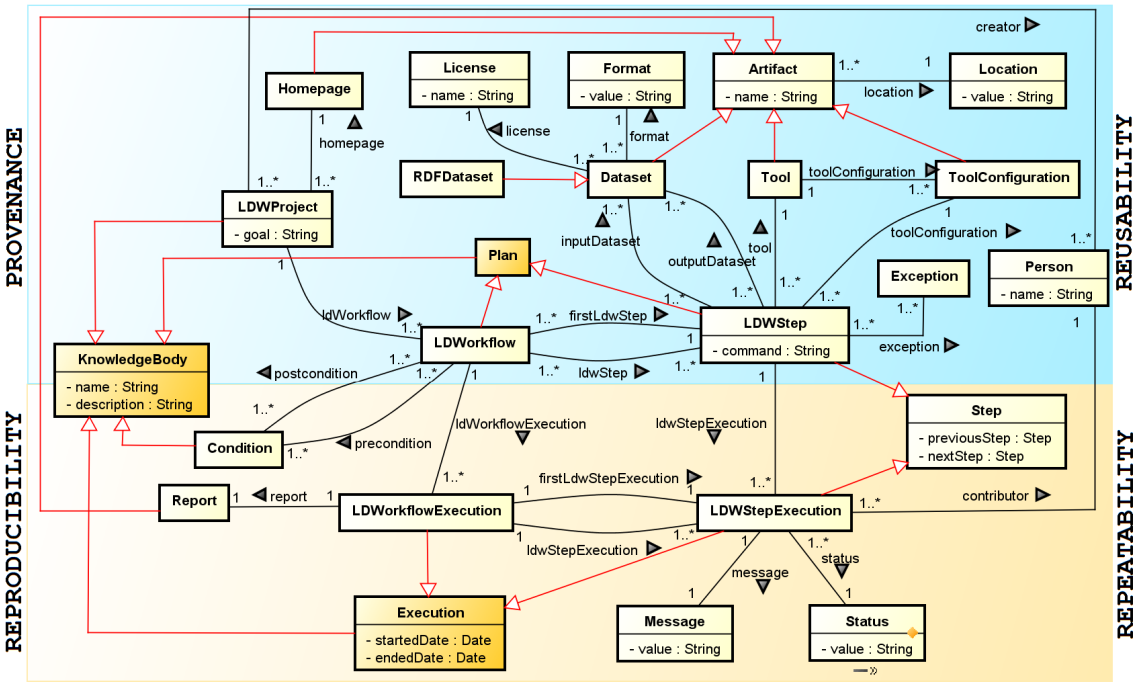

Figure 1. The LDWPO model.

planned. During the execution, the `LDWStepExecutions` can generate `Messages` such as logging report and `Statuses` such as successful finished, unsuccessful finished, aborted, etc. In this way, a whole `LDWorkflowExecution` can register the reproducibility information of an `LDWorkflow` for checking the correctness or repeatability of `RDFDatasets`. Forward, this kind of information can be used for reproducing the same result over time.

Another important concept in the LDWPO is `Task`. This class is an atomic unit of `Method` abstract concept and represents a best practice covered by the `LDWSteps`. When an `LDWStep` is planned, it can be related to a set of `Tasks`, which is necessary to accomplish during `LDWStepExecutions`. Examples of `Tasks` are: a) reusing established vocabularies, b) describing RDF dataset with the *Vocabulary of Interlinked Datasets* (VoID), c) establishing open license, d) keeping RDF dataset versioning information, or e) providing human-readable data description. Relating `Tasks` to `LDWSteps` can be useful to the data engineers for describing `LDWorkflows` in an upper level of abstraction like in software development process from Software Engineering context. This methodological perspective of the LDWPO is depicted in Figure 2. As it is illustrated, the *Linked Data Lifecycle* [Auer et al. 2012] is instantiated as the `Process`. Additionally, the *Extraction* `Activity` of that `Process` is related to the *Reusing vocabularies* `Task`. In a given `LDWProject`, an instance of `Task` is associated to an `LDWStep` instance, making explicit a relationship between an `LDWorkflow` unit and a best practice. As consequence, considering that the lifecycle of resources can be followed in a particular `LDWorkflow`, when describing `LDWSteps` with LDWPO, we can understand how an RDF dataset is (re)produced in the level of methodological approaches.

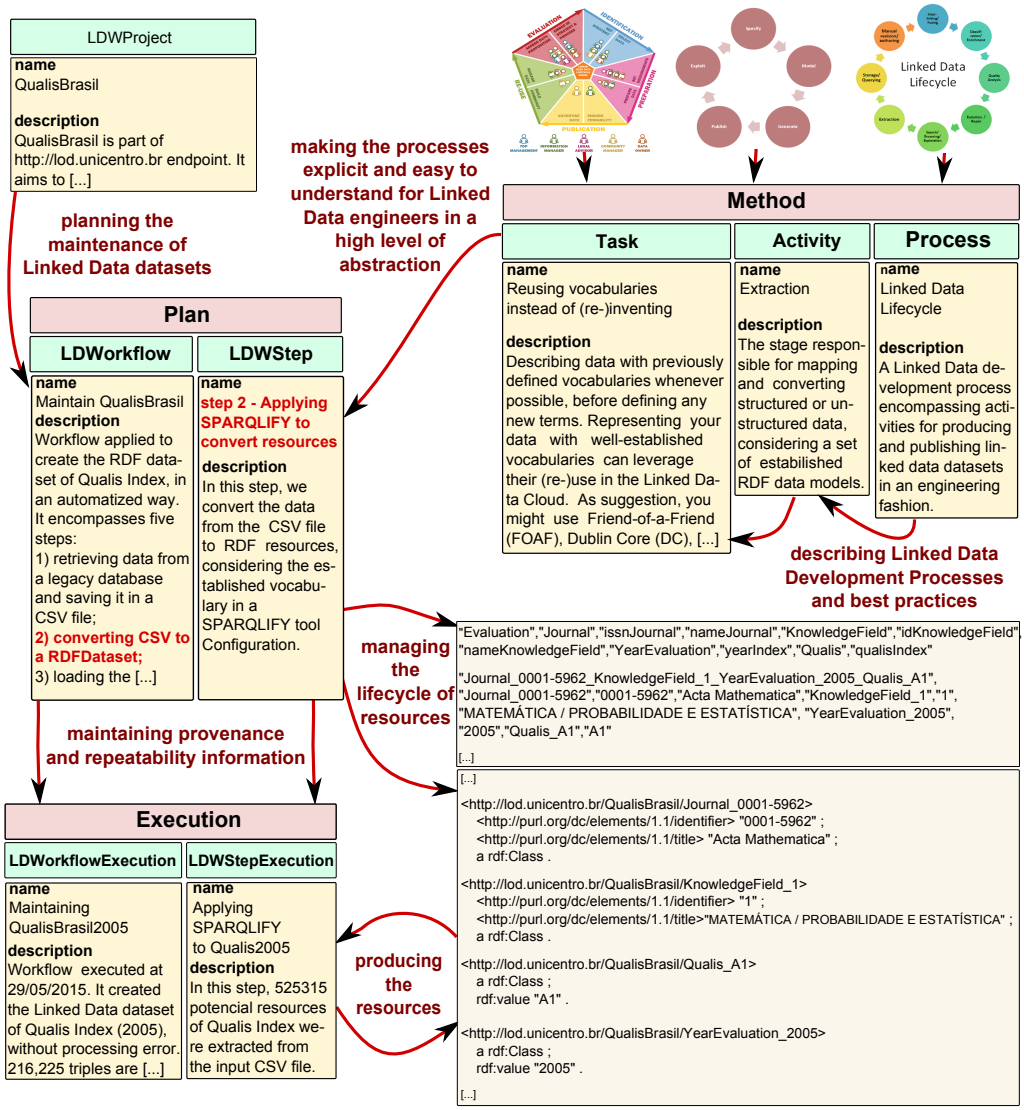

Figure 2. Exemplifying the LDWPO expressiveness.

4. LDWPO in Use

In this section, we describe how LDWPO supports the maintenance and publication of `5 star` RDF datasets[5]. In particular, these datasets support a Knowledge Management project in a Brazilian university.

4.1. Data Sources

4.1.1. Qualis dataset

One of the data sources originates from *Qualis*, a dataset created and used by the *Brazilian Research Community* and providing a complete view of research in and related to Brazil. *Qualis* dataset encompasses indirect scores[6] for research papers in journals, according to

[5]For more information, please see the data classification system proposed by Tim Berners-Lee at `http://5stardata.info/`

[6]A typical entry of *Qualis* consists of *ISSN*, journal name, related knowledge field, and qualified journal score.

the relevance of the journal to the knowledge fields (computer science, chemistry, medicine, among others). *Qualis* is used in bibliometric/scientometric assessments and for ranking post-graduate programs, research proposals, or individual research scholarships.

Although a web interface[7] is publicly available for querying *Qualis* data, it has several limitations: i) historical data is not available, making it difficult to perform time series studies; ii) in the early years, the data was available only as `1 Star Data` (i.e. Portable Document Format - PDF) in an outdated web interface; iii) only the last versions of the dataset are available for downloading as spreadsheets (MS Excel file extension - XLS); and iv) the data is not linked to other datasets, which makes its use challenging.

4.1.2. Lattes Curriculum dataset

Another data source is the Lattes Platform[8], an integrated information system maintained by the Ministry of Science, Technology and Innovation of Brazil. It is used to manage public information of individual researchers, groups, and institutions settled in Brazil. Lattes Curriculum[9] (*CVLattes*) is the core component of Lattes Platform. *CVLattes* contains information about personal activities and achievements such as teaching, research projects, patents, technological products, publications, and awards. The maintenance of such information requires manual input via web interface by individual researchers. *CVLattes* is used to evaluate competence of researchers/institutions for funding research proposals.

CVLattes is available publicly via graphical web interface, which implements security measures (e.g. CAPTCHA[10]) preventing crawlers to extract the data from the platform. Therefore, automatic data extraction from *CVLattes* requires sophisticated crawling mechanisms. In knowledge management perspective, we consider the scenario in which a university can access *CVLattes* via formal request. On such a request, Brazilian universities can extract a view of its researchers for loading data into internal databases.

4.2. The Use Cases

In our vision the scientific knowledge management for universities will benefit from a knowledge management instrument called *Yellow Pages*. *Yellow Pages* facilitates identification of responsible parties *"who knows what"* (location and description) and creates opportunities for sharing organizational knowledge. The value of such system directly depends on the fact that the data (descriptions of skills, experiences of the groups/individuals etc.) is up-to-date [Keyes 2006].

To enable *Yellow Pages* for the Brazilian universities, we consider: a) an integration of *Qualis* and *CVLattes* datasets; and b) maintenance of the *Yellow Pages* knowledge base in a systematic way. To achieve these goals, we use LDWPO to support the orchestration of knowledge bases. For the integration of *Qualis* and *CVLattes* datasets, we instantiated two `LDWProjects`: *QualisBrasil* and *PeriodicalPapers* (depicted in Figure 3).

[7] `https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf`

[8] a web interface is available at: `lattes.cnpq.br`

[9] an example of *CVLattes* can be accessed at `http://buscatextual.cnpq.br/buscatextual/visualizacv.do?id=K4787027P5&idiomaExibicao=2`

[10] acronym for Completely Automated Public Turing test to tell Computers and Humans Apart. In

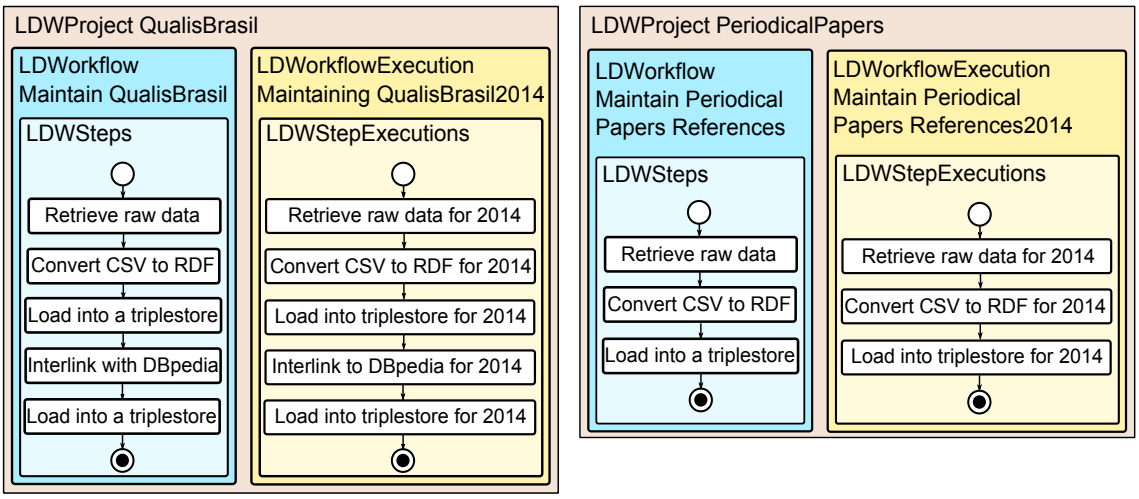

Figure 3. LDWProjects provide a pipeline for upgrading Qualis and CVLattes data sources up to 5 Stars Linked Data.

QualisBrasil `LDWProject` is based on *Maintain QualisBrasil* `LDWorkflow`, which is composed by five `LDWSteps` as follows:

1. ***LDWStep a*** retrieves data from a legacy database and saving it in a Comma Separated Values (CSV) format;
2. ***LDWStep b*** converts the CSV data to the *Qualis* `RDFDataset`, using the transformation tool *Sparqlify*[11];
3. ***LDWStep c*** updates a graph[12] in a triple store with the generated resources;
4. ***LDWStep d*** interlinks the resulting *Qualis* `RDFDataset` with *DBpedia*[13] data, using the link discovery tool *LIMES*[14]. For linking, it is considered the *International Standard Serial Number* (ISSN) and rdfs:seeAlso property; and
5. ***LDWStep e*** loads the acquired links into the triple store.

PeriodicalPapers is an `LDWProject`, which converts the paper references from scientific journals to linked data. *Maintain Periodical Papers References* `LDWorkflow` is constituted by three `LDWSteps`:

1. ***LDWStep a*** retrieves the data from a legacy database and saves it in a CSV format;
2. ***LDWStep b*** performs conversion of the CSV data to the *PeriodicalReferences* `RDFDataset` using the *Sparqlify*; and
3. ***LDWStep c*** updates a graph[15] in a triple store with the `RDFDataset`.

---

computing, it is used to check whether or not the user is human.

[11] `http://aksw.org/Projects/Sparqlify.html`

[12] published on datahub `http://datahub.io/dataset/qualisbrasil` and publicly available at `http://lodkem.led.ufsc.br:8890/sparql`, graph name: "http://lod.unicentro.br/QualisBrasil/".

[13] is a community effort to extract structured information from Wikipedia and to make this information accessible on the Web [Lehmann et al. 2009].

[14] `http://aksw.org/Projects/LIMES.html`

[15] published on datahub `https://datahub.io/dataset/lattes-production` and publicly available at `http://lodkem.led.ufsc.br:8890/sparql`, graph name: "http://lod.unicentro.br/LattesProduction/".

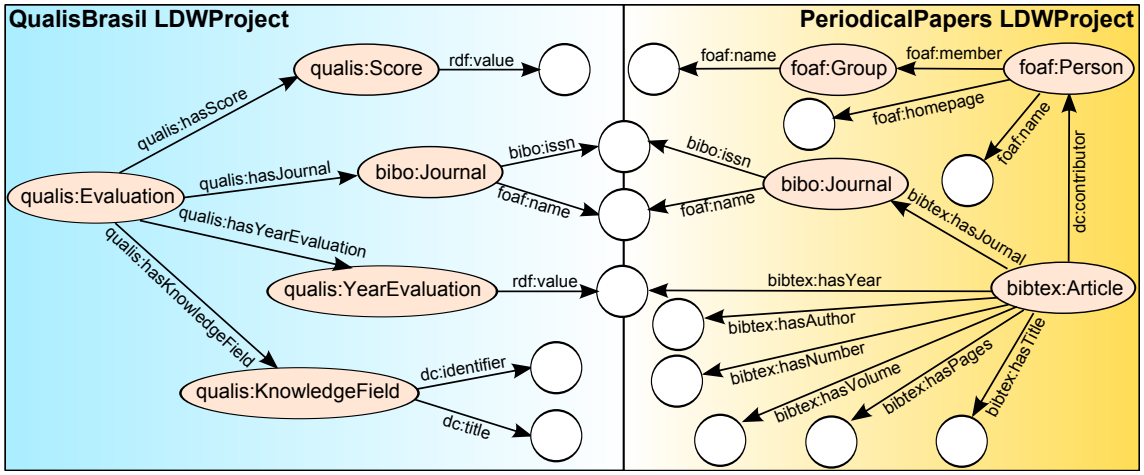

Figure 4. Representing the knowledge base for the *Yellow Pages* System.

For the execution of `LDWorkflows`, we developed the *Linked Data Workflow Execution Engine for LDWPO* (LODFlow Engine[16]). This tool retrieves the `LDWProjects` and `LDWorkflows` from the LDWPO knowledge base and manages the pipeline for producing the RDF datasets in an automated fashion. Using LODFlow Engine and the `LDWorkflow` definitions, we generated 698 668 interlinked entities for *Qualis* in an automated fashion. For *PeriodicalPapers* `LDWProject`, LODFlow Engine generated 5 557 entities, representing the periodical papers references of 630 researchers related to a Brazilian university.

The resulting RDF datasets of *Qualis* and *PeriodicalPapers* provide a foundation for the *Yellow Pages* system. In other words, the resulting knowledge base (depicted in Figure 4) integrates the data from heterogeneous sources, enabling new knowledge management perspectives. For example, there is a limitation on classifying the periodical papers according to the journal scores. Commonly, it requires manual effort and, generally, include one knowledge field. Using the resulting knowledge base and appropriated SPARQL[17] queries, the periodical papers can be classified more efficiently, considering the research group units and/or knowledge fields. In this case, the SPARQL query in the listing below can be customized for exploring new scientometric scenarios. These scenarios could include questions, such as:

- What are the main competences of my university in the specific knowledge fields?
- Which researchers in my university work together in a particular knowledge field?
- Which researchers in my university could possibly work together in a research project of a particular knowledge field? (finding possibilities of a collaboration)
- Which researchers should collaborate to improve the university key performance indicators?

Such questions are easily formulated by research supervisors inside universities, but are hardly answered by external researchers, who have university and institution web sites as main information sources. We argue that the use of *Yellow Pages*, supported by a knowledge base that evolves semantically, can be a cornerstone for sharing the knowledge inside and out of a university.

[16] `https://github.com/AKSW/LODFlow/tree/master/tools/LODFlowEngine`
[17] a recursive acronym for SPARQL Protocol and RDF Query Language.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX bibtex: <http://purl.org/net/nknouf/ns/bibtex/>
PREFIX prod: <http://lod.unicentro.br/LattesProduction/>
PREFIX qualis: <http://lod.unicentro.br/QualisBrasil/>

SELECT ?qualisYearEvaluationValue ?qualisKnowledgeFieldTitle ?
    qualisScoreValue COUNT(*) as ?qtde where {
  ?evaluation  rdf:type  qualis:Evaluation  .
  ?evaluation  qualis:hasJournal  ?qualisJournal  .
  ?evaluation  qualis:hasYearEvaluation  ?qualisYearEvaluation  .
  ?evaluation  qualis:hasKnowledgeField  ?qualisKnowledgeField  .
  ?evaluation  qualis:hasScore  ?qualisScore  .
  ?qualisJournal  bibo:issn  ?qualisJournalId  .
  ?qualisYearEvaluation  rdf:value  ?qualisYearEvaluationValue  .
  ?qualisScore  rdf:value  ?qualisScoreValue  .
  ?qualisKnowledgeField  dc:title  ?qualisKnowledgeFieldTitle  .
  ?paper  rdf:type  prod:PeriodicalPaper  .
  ?paper  bibtex:hasJournal  ?paperJournal  .
  ?paper  bibtex:hasTitle  ?paperTitle  .
  ?paper  bibtex:hasYear  ?qualisYearEvaluationValue  .
  ?paperJournal  bibo:issn  ?qualisJournalId  .
  ?paperJournal  foaf:name  ?journalName  .
}
GROUP BY ?qualisYearEvaluationValue ?qualisKnowledgeFieldTitle ?
    qualisScoreValue
```

5. Related Work

To the best of our knowledge, this work presents the first ontology focused on concepts of Method (process), Plan (provenance), and Execution (reproducibility) for publishing linked data. Although, there are works targeting the provenance and reproducibility.

For example, PWO [Gangemi et al. 2014] is an ontology for describing the workflows associated with the publication of a document. Using the core concepts of PWO, it is possible to: i) define the initial `Step` for a given `Workflow`, ii) relate next/previous `Steps` (therewith creating the `Workflow`) and iii) define the inputs and outputs for each `Step`. OPMV [Moreau et al. 2011] is recommended as a model for data provenance, which enables data publishing as well as data exchange between various systems. In OPMV: i) a `Process` is controlled by an `Agent`; ii) a `Process` uses `Artifacts` at certain time; iii) an `Artifact` is generated by a `Process`; iv) an `Artifact` can be derived from another `Artifact`; and v) to execute the workflow, a `Process` triggers a subsequent `Process`. However, OPMV does not define the concepts of `Plan` explicitly. PROV-O [Lebo et al. 2015] is the W3C recommendation for representing and interchanging provenance and reproducibility information generated by different systems and contexts. With the core concepts, in PROV-O: i) an `Activity` is associated with an `Agent`; ii) also, an `Entity` is attributed to an `Agent`; iii) an `Activity` uses `Entities` in an interval of time; iv) an `Entity` can be derived from another `Entity`; and v) to keep the workflow, an `Activity` is associated (wasInformedBy) to another `Activity`. As OPMV, the concept of `Plan` cannot be entirely formulated. To overcome this limitation, the *Ontology for Provenance and Plans* (P-Plan ontology) extends PROV-O enabling the publishing of workflow plan and its execution(s) as linked data [Garijo and Gil 2012].

Considering a different domain of Linked Data, the scientific community coined the term Scientific Workflow as "the automated process that combines data and pro-

cesses in a structured set of steps to implement computational solutions to a scientific problem" [Altintas et al. 2006]. To facilitate workflows for data and control sequences, *Scientific Workflow Management Systems* such as *Apache Taverna* [Hull et al. 2006] and *Kepler* [Ludäscher et al. 2006] were developed. These management systems employ ontologies for modeling the workflows, such as *Scufl2* and *Kepler* ontologies, respectively. At the time of writing, the *Scufl2* ontology is not available at the Taverna's homepage. *Kepler* ontologies are part of the *Kepler* framework and can be found in the source code. *Kepler* ontologies do not include human-readable descriptions for concepts, as we show in the following listing. Concept descriptions are required to facilitate the reuse of ontology resources. In our vision, the absence of such descriptions limits the adoption of *Kepler* ontologies. To leverage the limitations of *Scufl2* and *Kepler* ontologies, we designed LDWPO to support the LODFlow, a customized Workflow Management System for Linked Data Processing.

```
1  [...]
2    <owl:Class rdf:ID="Workflow">
3      <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
           Workflow</rdfs:label>
4    </owl:Class>
5
6    [...]
7
8    <owl:Class rdf:ID="WorkflowOutput">
9      <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
           Workflow Output</rdfs:label>
10     <rdfs:subClassOf>
11       <owl:Class rdf:about="#DataOutput"/>
12     </rdfs:subClassOf>
13   </owl:Class>
14 [...]
```

6. Conclusion, Limitations, and Future Work

In this paper, we presented *Linked Data Workflow Project Ontology* (LDWPO), an ontology for supporting the RDF dataset maintenance. In our vision, an established process should rule a workflow, which controls all computational procedures for maintaining an RDF dataset over time. Focusing on provenance, reusability, and reproducibility issues, LDWPO is aligning with existing vocabularies and ontologies, such as OPMV, PROV-O, and PWO.

The benefits of explicitness, reusability, and repeatability are observed when LDWPO is applied. In particular, with the ontology, it is possible to create comprehensive workflow descriptions, preserving provenance information for reproducing the `LDWorkflows` of an `LDWProject`. Moreover, technologically, it is possible to mediate the use of tools, enabling the automatized execution of `LDWorkflows` in the context of the *Linked Data Stack* and *Linked Data Lifecycle*.

With LDWPO we aimed to tackle one of the most pressing and challenging problems of Linked Data management – managing the lifecycle of RDF datasets over time, considering the myriad of tools, environments, and resources. Considering that the support to lifecycle of RDF datasets is currently a cumbersome activity, when applied more widely, LDWPO can provide a boost to the advancement and maturation of Linked Data technologies.

Thus, we see this work as an important step in a large research agenda, which

aims at providing comprehensive workflow support for RDF dataset (re)production and maintenance processes. As first contribution, LDWPO is already used in a real-world application for publishing scientometric resources in an automated fashion. More precisely, a scientific journal index and journal papers entries are maintained as linked open data, using LDWPO for promoting knowledge management in a Brazilian university. The datasets are publicly available at `http://lodkem.led.ufsc.br:8890/sparql`. Specially, the *Qualis* RDF dataset can be reused by the community in other studies in the Information Science field.

As future work, we aim to maintain the developed ontology, as well as, adopt it in further use cases. In the context of *Yellow Pages* system, LDWPO can assist the knowledge base expansion, considering the following scenarios:

1. Integration of new data sources, improving the knowledge base expressiveness (e.g. research project descriptions, technological products, patents, courses, coming from *CVLattes* or another bibliometric scores such as *Journal Citation Reports* (JCR), *SCImago Journal Rank* (SJR), and *Source Normalized Impact per Paper* (SNIP).
2. Maintenance of the existing RDF datasets (e.g. *CVLattes* and *Qualis*) via continuous execution of the `LDWorkflows` over time.
3. Data validation and debugging via repeating `LDWorkflowExecutions`, when necessary.
4. Generation of the documentation for `LDWProjects` to support data engineers in assessing quality issues.

In addition, we are working on incorporating the LDWPO into a *Linked Data Stack* tool, providing a full-integrated Workflow Management System for linked dataset (re)production.

Acknowledgment

This work was supported by the Brazilian Federal Agency for the Support and Evaluation of Graduate Education (CAPES/Brazil), under the program Sciences without Borders (Process number - 18228/12-7) and Araucaria Foundation (Project number 601/14).

References

[Altintas et al. 2006] Altintas, I., Barney, O., and Jaeger-Frank, E. (2006). Provenance collection support in the kepler scientific workflow system. In Moreau, L. and Foster, I. T., editors, *IPAW*, volume 4145 of *Lecture Notes in Computer Science*, pages 118–132. Springer.

[Auer et al. 2012] Auer, S., Bühmann, 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. (2012). Managing the life-cycle of linked data with the LOD2 stack. In *Proceedings of International Semantic Web Conference (ISWC 2012)*.

[Bourque and Fairley 2004] Bourque, P. and Fairley, R. E. (2004). Guide to software engineering body of knowledge. Retrieved October, 2014, from http://www.computer.org/portal/web/swebok.

[Gangemi et al. 2014] Gangemi, A., Peroni, S., Shotton, D., and Vitali, F. (2014). A pattern-based ontology for describing publishing workflows. In *Proceedings of the 5th Workshop on Ontology and Semantic Web Patterns (WOP2014) co-located with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 19, 2014.*, pages 2–13.

[Garijo and Gil 2012] Garijo, D. and Gil, Y. (2012). Augmenting prov with plans in p-plan: Scientific processes as linked data. In *Linked Science*.

[Gomez-Perez et al. 2004] Gomez-Perez, A., Fernandez-Lopez, M., and Corcho, O. (2004). *Ontological Engineering: With Examples from the Areas of Knowledge Management, E-Commerce and the Semantic Web, 1st Edition*. Springer-Verlag, Heidelberg.

[Hull et al. 2006] Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., and Oinn, T. (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*, 34(Web Server issue):729–732.

[Keyes 2006] Keyes, J. (2006). *Knowledge Management, Business Intelligence, and Content Management: The IT Practitioner's Guide*. Auerbach Publications, 1 edition.

[Lebo et al. 2015] Lebo, T., Sahoo, S., McGuinness, D., Belhajjame, K., Cheney, J., Corsar, D., Garijo, D., Soiland-Reyes, S., Zednik, S., and Zhao, J. (2015). PROV-O: The prov ontology. Retrieved from http://www.w3.org/TR/prov-o/ on 13.01.2015.

[Lehmann et al. 2009] Lehmann, J., Bizer, C., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., and Hellmann, S. (2009). DBpedia - a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165.

[Ludäscher et al. 2006] Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E. A., Tao, J., and Zhao, Y. (2006). Scientific workflow management and the kepler system. *Concurrency and Computation: Practice and Experience*, 18(10):1039–1065.

[Moreau et al. 2011] 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. (2011). The open provenance model core specification (v1.1). *Future Generation Computer Systems (FGCS)*, 27(6):743–756. [IF 1.978, CORE A].

[Noy and McGuinness 2001] Noy, N. F. and McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology. *Development*, 32(1):1–25.

[Rautenberg et al. 2015] Rautenberg, S., Ermilov, I., Marx, E., Auer, S., and Ngomo Ngonga, A.-C. (2015). Lodflow – a workflow management system for linked data processing. In *SEMANTiCS 2015*.

[Sure and Studer 2002] Sure, Y. and Studer, R. (2002). On-To-Knowledge methodology. In Davies, J., Fensel, D., and van Harmelen, F., editors, *On-To-Knowledge: Semantic Web enabled Knowledge Management*, chapter 3, pages 33–46. J. Wiley and Sons.

[W3C 2014] W3C (2014). RDF 1.1 Concepts and Abstract Syntax. http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[WFMC 1995] WFMC (1995). The workflow reference model. Technical report, The Workflow Management Coalition.