Journal of Computer Science & Technology, Volume 19, Number 1, April 2019

- ORIGINAL ARTICLE -

Improving Open Science Using Linked Open Data: *CONICET Digital* Use Case

Mejorando la Ciencia Abierta Usando Datos Abiertos Enlazados: Caso de Uso CONICET Digital

Marcos Zárate[1,2], Carlos Buckle[2,3], Renato Mazzanti[2,4], and Gustavo Samec[2,4]

[1] *Centre for the Study of Marine Systems, Patagonian National Research Center, (CENPAT-CONICET), Puerto Madryn, Argentina*
zarate@cenpat-conicet.gob.ar
[2] *Laboratorio de Investigación en Informática (LINVI), Universidad Nacional de la Patagonia San Juan Bosco, Puerto Madryn, Argentina*
{gsamec, renato}@cenpat-conicet.gob.ar
[3] *Departamento de Informática, Facultad de Ingeniería, Universidad Nacional de la Patagonia San Juan Bosco, Puerto Madryn, Argentina*
cbuckle@unpata.edu.ar
[4] *Unidad de Gestión de la Información, CCT CONICET-CENPAT, Puerto Madryn, Argentina*

Abstract

Scientific publication services are changing drastically, researchers demand intelligent search services to discover and relate scientific publications. Publishers need to incorporate semantic information to better organize their digital assets and make publications more discoverable. In this paper, we present the on-going work to publish a subset of scientific publications of *CONICET Digital* as Linked Open Data. The objective of this work is to improve the recovery and reuse of data through Semantic Web technologies and Linked Data in the domain of scientific publications. To achieve these goals, Semantic Web standards and reference RDF schema's have been taken into account (Dublin Core, FOAF, VoID, etc.). The conversion and publication process is guided by the methodological guidelines for publishing government linked data. We also outline how these data can be linked to other datasets DBLP, WIKIDATA and DBPEDIA on the web of data. Finally, we show some examples of queries that answer questions that initially *CONICET Digital* does not allow.

Keywords: CONICET Digital, Linked Open Data, Open Science, RDF, SPARQL.

Resumen

Los servicios de publicación científica están cambiando drásticamente, los investigadores demandan servicios de búsqueda inteligentes para descubrir y relacionar publicaciones científicas. Los editores deben incorporar información semántica para organizar mejor sus activos digitales y hacer que las publicaciones sean más visibles. En este documento, presentamos el trabajo en curso para publicar un subconjunto de publicaciones científicas de CONICET Digital como datos abiertos enlazados. El objetivo de este trabajo es mejorar la recuperación y la reutilización de datos a través de tecnologías de Web Semántica y Datos Enlazados en el dominio de las publicaciones científicas. Para lograr estos objetivos, se han tenido en cuenta los estándares de la Web Semántica y los esquemas RDF (Dublin Core, FOAF, VoID, etc.). El proceso de conversión y publicación se basa en las pautas metodológicas para publicar datos vinculados de gobierno. También describimos cómo estos datos se pueden vincular a otros conjuntos de datos como DBLP, Wikidata y DBPedia. Finalmente, mostramos algunos ejemplos de consultas que responden a preguntas que inicialmente no permite CONICET Digital.

Palabras claves: CONICET Digital, Datos Abiertos Enlazados, Ciencia Abierta, RDF, SPARQL.

1 Introduction and motivation

Open Science [1] is a movement whose objective is the accessibility of scientific research for all citizens. Open science increases and stimulates the production of scientific knowledge, because it includes different types of knowledge and knowledge, innovates with the use of technologies, promotes the value of sharing, reuses and allows data, reports and other parts of the research process to be available for everyone. In this context, *CONICET Digital* is the Open Access Institutional Repository belonging to National Council of Scientific and Technical Research (CONICET). It is a digital platform that makes the scientific and technological production of the country available to society. CONICET is the main organization dedicated to the

promotion of science and technology in Argentina for more than 50 years and is one of the most important assets of the national capital in science and technology, *CONICET Digital* is created with the objective of gathering, registering, disclosing, preserve and give public access to the scientific-technological production carried out by researchers, fellows and other CONICET personnel. The repository is a free access to information service that allows all those interested in the disciplines of knowledge, the recovery of scientific-technological production, both for the teaching field, as for research and study. Currently (July 2018) the repository has 39.942 available titles, 63.294 authors and 6 areas of knowledge. *CONICET Digital* adopted the well-known DSpace platform to implement the repository, like most software to create repositories, DSpace supports OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting) [2] as an interface to expose the stored metadata. Although OAI-PMH is well known in the field of repositories, it is rarely known in other areas what makes integration with information from other domains difficult.

Given that most Internet sites are oriented to human consumption (*CONICET Digital* is not the exception) in many cases the information can not be interpreted by machines, this situation has some drawbacks with respect to the integration and automatic retrieval of information. While providing tools for browsing and visualising data is another important means for making that data useful to a broad variety of people, another very important aspect is ensuring that the data has context to be exploited by machines. All data has some relevant context, for example, *Who has published the data? How was it collected? Are there any caveats that are important to its reuse and interpretation? To what does the data refer, and how do those things relate to one another?*.

The main objective of Linked Data (LD) [3, 4] proposed by Tim-Berners Lee, is to publish and connect structured data on the Web through a set of good practices [5] for that purpose. Thus, new documents will be understandable by the machines, will have an explicitly defined meaning and will be linked with others, transforming the Web into a collection of RDF triples [6] referenced by URIs in the different namespaces. This ability to publish and connect data proposed by LD is fundamental for the implementation of the Semantic Web (SW) [7]. Linked Data is a rather big chance for repositories to present their content in a way that can easily be accessed, interlinked and (re)used. Among the advantages we find in using LD we can mention:

- ***Integration***: the use of OAI-PMH has not been widely accepted in other areas that they are not repositories, limiting the integration of them to other data sources.

- ***Semantics***: the data are no longer ambiguous and can be interpreted and understood both by humans and by other software applications.

- ***Visibility***: exposing the data as RDF graphs interconnected with other datasets in the LOD cloud facilitates its detection and visibility.

- ***Expressivity***: Data is recorded in a repository following a tree structure. On the contrary, RDF allows a description at the level of the graph, improving the expressiveness in describing the information.

- ***Queries***: Usually the options are limited to searches by keywords or by certain attributes on text strings. In RDF, the SPARQL query language [8] works on graphs and allows queries of greater scope and complexity. A user can perform searches in several repositories, from a SPARQL endpoint. You can also download part of the data and combine it with other data and processes according to your needs.

- ***Reusability***: other applications can make use of the data in their systems

Within the scope of this paper we propose to analyze different RDF vocabularies and ontologies of scientific publications that are required to offer a linked and open data source, with scientific information from *CONICET digital*, as well as approaches for its integration and tools to consume this data that will benefit of its standardized form of representation and the possibility of linking new data sources. In particular, We will develop a prototype application to retrieve information from the scientific domain integrating data from different sources. Along this process we follow the guidelines defined in [9]. This application will allow viewing the updated information associated with the available data and make queries in the SPARQL language [8] for RDF. The development of a tool with these characteristics will have a great impact on our research teams, since incorporating them into the data Web, will increase visibility, fostering scientific collaboration among interdisciplinary groups.

The rest of the paper is organized as follows: Section 2 describes the main linked datasets of scientific publications accessible via SPARQL. Section 3 explains the stages of the life cycle chosen for this work, while Section 4 presents case studies that allow retrieving information from different datasets. Section 5 describes the main layers of the proposed architecture. In Section 6 we discussed the problems we had and about the little development that our country has in terms of scientific publication as Linked Open Data. Finally in Section 7 we draw conclusions and suggest some future improvements.

2 Related work

While Linked Data is being embraced in various sectors, we are currently witnessing a substantial increase

in universities and platforms for the scholarly domain adopting the Linked Data initiative. For example Springer Nature SciGraph [10] a Linked Open Data platform for the scholarly domain which aggregates data sources from Springer Nature and key partners from the scholarly domain. The Linked Open Data platform collates information from across the research landscape, for example funders, research projects, conferences, affiliations and publications. The data in Springer Nature SciGraph is projected to contain 1.5 to 2 billion triples (January 2018).

DBLP [11] computer science bibliography contains the metadata of over 1.8 million publications, written by over 1 million authors in several thousands of journals or conference proceedings series. DBLP provide a SPARQL Query Interface, this interface allows queries to be made over the information held within the repository, using the SPARQL Query Language.

WIKIDATA [12] is a free, collaborative, multilingual, secondary database, collecting structured data to provide support for Wikipedia, Wikimedia Commons, the other wikis of the Wikimedia movement, and to anyone in the world. The WIKIDATA repository consists mainly of items, each one having a label, a description and any number of aliases. Items are uniquely identified by a `Q` followed by a number, such as *Pascal Hitzler* (`Q30103406`). Statements describe detailed characteristics of an Item and consist of a property and a value. Properties in WIKIDATA have a `P` followed by a number, such as sex (`P21`). Properties can also link to external databases, a property that links an item to an external database, such as an authority control database used by libraries and archives, is called an identifier. For example property `P2456` references an external identifier in DBLP. All this information can be displayed in any language, even if the data originated in a different language. When accessing these values, client wikis will show the most up-to-date data.

Another of the works that are relevant in this area is *Semantic Publishing and Referencing Ontologies* (SPAR Ontologies) [13], which forms a suite of orthogonal and complementary ontology modules for the creation of comprehensive machine-readable RDF metadata for every aspect of semantic publishing and referencing: document description, bibliographic resource identifiers, types of citations and related contexts, bibliographic references, document parts and status, agents' roles and contributions, bibliometric data and workflow processes. SPAR Ontologies have been already adopted by different communities and in several projects for describing data related to the publishing domain.

OpenCitations [14] the main work of OpenCitations is the creation and current expansion of the Open Citations Corpus (OCC), an open repository of scholarly citation data made available under a Creative Commons public domain dedication, which provides in RDF accurate citation information (bibliographic references) harvested from the scholarly literature. These are described using the SPAR Ontologies according to the OCC metadata model, and are made freely available so that others may freely build upon, enhance and reuse them for any purpose, without restriction under copyright or database law. The OCC is being continuously populated from the scholarly literature. As of January 2018, the OCC has ingested the references from 302.758 citing bibliographic resources and contains information about 12.830.347 citation links to 6.549.665 cited resources. The whole OCC is available for querying via SPARQL endpoint and for browsing by means of a very simple Web interface that shows only the data about bibliographic entities.

3 Methodology

The application of Linked Data principles to government datasets brings enormous potential [3]. However, this potential is currently untapped mostly because of the lack of resources required to transform raw data into high-quality Linked Data on a large scale [15]. While is true that Linked Data generation and publication does not follow a set of common and clear guidelines to scale out the generation and publication of Linked Data, the Methodological Guidelines for Publishing Government Linked Data proposed in [9] established that the process of publishing datasets as Linked Data must have a life cycle, in the same way of Software Engineering, in which every development project has a life cycle. This process has an iterative incremental life cycle model, which is based on the continuous improvement and extension of the Linked Data resulted from performing several iterations. Because of its similarity to with the software development process, we decided that this approach is the one we adopt for this paper.

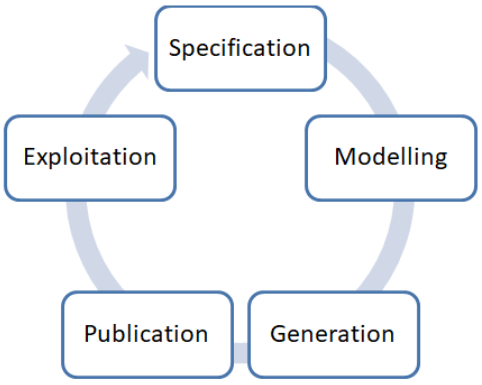

Fig. 1. Linked Data life cycle according to Villazon-Terrazas et al. [9]

(i) Specify: include URI design, define/describe the provenance information and analyze the data sources. (ii) Model: This stage includes the search for suitable ontologies/vocabularies that model the data sources,

create the model by reusing the ontologies/vocabularies selected. (iii) Generate: This is perhaps one of the most important stages, here we transform the data source to RDF, clean the Data and Link with other bibliographics datasets. (iv) Publish: Publish dataset and enable effective discovery. (v) Exploit: Make use of the data and applications that consume this data.

Fig. 1 shows the proposed life cycle and then each stage is explained in detail in the followings sections.

3.1 Specification: URI strategy

URIs (Uniform Resource Identifiers) are very important, providing both the core of the platform itself and the link between RDF and the Web. Currently, URIs for the resources pertaining to *CONICET Digital* follow the pattern:

`http://data.cd.gob.ar/{type}/{concept}/{ID}`

- The domain follows the two recommendations formulated by [16]: solely be used for the publication of *CONICET Digital* information and not include the name of any organization, as they may evolve over time.
- {`type`} can take any of the following values: `resource` for the HTTP URI of a resource, and `page` and `data` for that resources HTML and RDF documents respectively.
- {`concept`} it gives us a hint as to what this resource is about by referring to the class to which that resource belongs. For example `Person`,`Publication`, etc.
- {`ID`} for the unique identifiers we use the ones provided in the original datasets, normally identified with a URI like a DOI.

3.2 Modeling: vocabularies

After the specification activity, we need to determine the ontology to be used for modeling the domain of this data source. Several ontologies exist that can be used to represent references, including SPAR Ontologies [13], and others like Publishing Roles Ontology (PRO) [17] an ontology for the characterization of the roles of agents, people, corporate bodies and computational agents in the publication process. At the basis of a lot of these efforts is the Dublin Core metadata schema [18] which represents a common ground for the description of resources and documents. Other ontologies have been created that focus on more specific aspects of bibliographical references, such as the FRBR-aligned Bibliographic Ontology (FaBiO) [19], is an ontology for recording and publishing on the Semantic Web descriptions of entities that are published or potentially publishable, and that contain or are referred to by bibliographic references, or entities used to define such bibliographic references. Finally we took into account the most recent DBLP scheme are defined term to represent information on the types of publications, relationships between them, this scheme is interesting because it is mainly oriented to the field of computer science. Table 1 summarizes the main vocabularies and ontologies used to create the dataset.

Table 1: Ontologies and Vocabularies used to generate the RDF dataset.

Prefix	Description
`cd`	CONICET Digital Base URI
`fabio`	Bibliographic Ontology
`pro`	Publishing Roles Ontology
`dblp`	Computes science bibliography terms
`wd`	Wikidata entities
`wdt`	Properties in Wikidata
`dbo`	DBPedia Ontology
`foaf`	Friend of a Friend
`dc`	Dublic Core
`void`	Metadata about RDF datasets

3.3 Generation

RDF is the standard data model in which the government information has to be made available, according to the Linked Data principles. Therefore, in this activity we have to take the data sources selected in the specification activity (see Section 3.1), and transform them to RDF according to the vocabulary created in the modeling activity (see Section 3.2).

3.3.1 Data extraction

We based the pipeline on OpenRefine [20], a data workbench that has powerful capabilities for data massaging and tidying up. We extended OpenRefine with Linked Data capabilities using extensions like RDF Refine. Metadata of scientific publication are manually extracted from *CONICET Digital* repository and their content are processed. There, the records are cleaned and converted to standardised data types such as dates, numerical values, etc. and empty columns are removed. OpenRefine has powerful data cleaning and transformation capabilities. It also has an expressive expression language called GREL. The built-in clustering engine facilitates identifying duplicates.

3.3.2 Linking

Following the fourth Linked Data Principle [5], include links to other URIs, so that they can discover more things, the next task is to create links between the *CONICET Digital* dataset and external datasets. This task involves the discovery of relationships between data items. We can create these links manually, which

is a time consuming task, or we can rely on automatic or supervised tools.

OpenRefine allows adding reconciliation services based on SPARQL endpoints, which return candidate resources from external datasets to be matched to fields in the local datasets. In our process, we use WIKIDATA endpoint to reconcile names of authors with the `Q5` (humans) resource in WIKIDATA. Also the reconciliation service provided by ORCID (Open Researcher and Contributor ID) was used in conjunction with WIKIDATA and the names of the journals were reconciled using the endpoint provided by DBPedia. The link between the resources is made through the property `owl:sameAs`.

3.3.3 Converting raw data into RDF

After defining the URIs and generate links between external datasets, data are converted to RDF using RDF Refine which allows users to go through a graphical interface describing the RDF scheme alignment skeleton to be shared among different datasets. The RDF skeleton specifies the Subject, Predicate and the Object of the triples to be generated. The next step in the process is to set up prefixes defined in Section 3.2.

After skeleton definition, it remains to generate the RDF in some of the serializations that RDF Refine supports. A RDF turtle serialization of an article extracted from *CONICET Digital* whose identifier is `http://hdl.handle.net/11336/6964` is shown in Fig. 2 for reasons of simplicity the complete record is not shown, to consult the entire record see the corresponding link[1]. This article contains information about title, authors, date of publication, ISBN and affiliations among others. As you can see one of the authors (`Pascal Hitzler`) has a link to his corresponding URI on WIKIDATA, this was possible due to the process explained in Section 3.3.2

OpenRefine logs all the operations applied to the data. It explicitly represents these operations in JSON and enables extracting and (re)applying them. The RDF related operations added to OpenRefine are no exception. Both the RDF modeling and reconciling are recorded and saved in the project history. To consult all the operations that we carry out in the process of conversion and mapping of vocabulary, we recommend to seeing the following link[2].

3.4 Publishing

The transformed data have been published, and can to be accessed, through GraphDB [21] which is a highly efficient and robust graph database with RDF and SPARQL support. It allows users to explore the hierarchy of RDF classes (*Class hierarchy*), where each class can be browsed to explore its instances. Similarly, relationships among these classes also can be explored giving an overview about how many links exist between instances of the two classes (*Class relationship*). Each link is a RDF statement where its subject and object are class instances and its predicate is the link itself. Lastly, users also can explore resources providing URIs representing any of the subject, predicate or object of a triple (*View resource*).

Finally, the user can visually explore the dataset, accessing to the GraphDB interface with the user and password (user: `guest` password: `cd.lod`). Bulk download is possible at the following link[3]. Table 2 summarised the main links to access the data.

Table 2: Main features of *CONICET Digital* dataset. `BASE` is the abbreviation of the real URL `http://web.cenpat-conicet.gob.ar:7200/`

Repository	CONICET-DIGITAL
Login	user: **guest** pass: **cd.lod**
SPARQL endpoint	`BASE:sparql`
Class hierarchy	`BASE:hierarchy`
Class relationship	`BASE:relationships`
View resource	`BASE:resource/find`
Vocabularies	14
External links	36
No. Classes	6
No. Properties	48
No. Triples	1127

4 Exploitation

In order to validate the understandability, applicability and usability of *CONICET Digital* dataset, we conducted five experiments in real case scenarios. We propose the use of SPARQL queries against DBLP allows processing data in many ways. With a simple query one can find relations between authors: e.g. show all coauthors of a particular author or even further: show second degree co-authors (i.e. co-authors of co-authors) of a particular author; show all conferences two authors attended (or books they published papers in). The same way relations between papers (e.g. show papers with co-occurring keywords), books (e.g. group books by co-editors) or conferences (e.g. show conferences by time, place) can be analysed. Each SPARQL query in the following examples assumes the prefix defined in Table 1.

4.1 Retrieving publications from a specific topic.

The first query (See Algorithm 1) allows us to retrieve the publications of a certain topic, in our case we are

[1] `https://github.com/cenpat/conicet-digital/blob/master/scripts/hdl6964.ttl`

[2] `https://github.com/cenpat/conicet-digital/blob/master/scripts/mapping.json`

[3] `https://github.com/cenpat/conicet-digital/blob/master/dataset/cd-dataset.ttl`

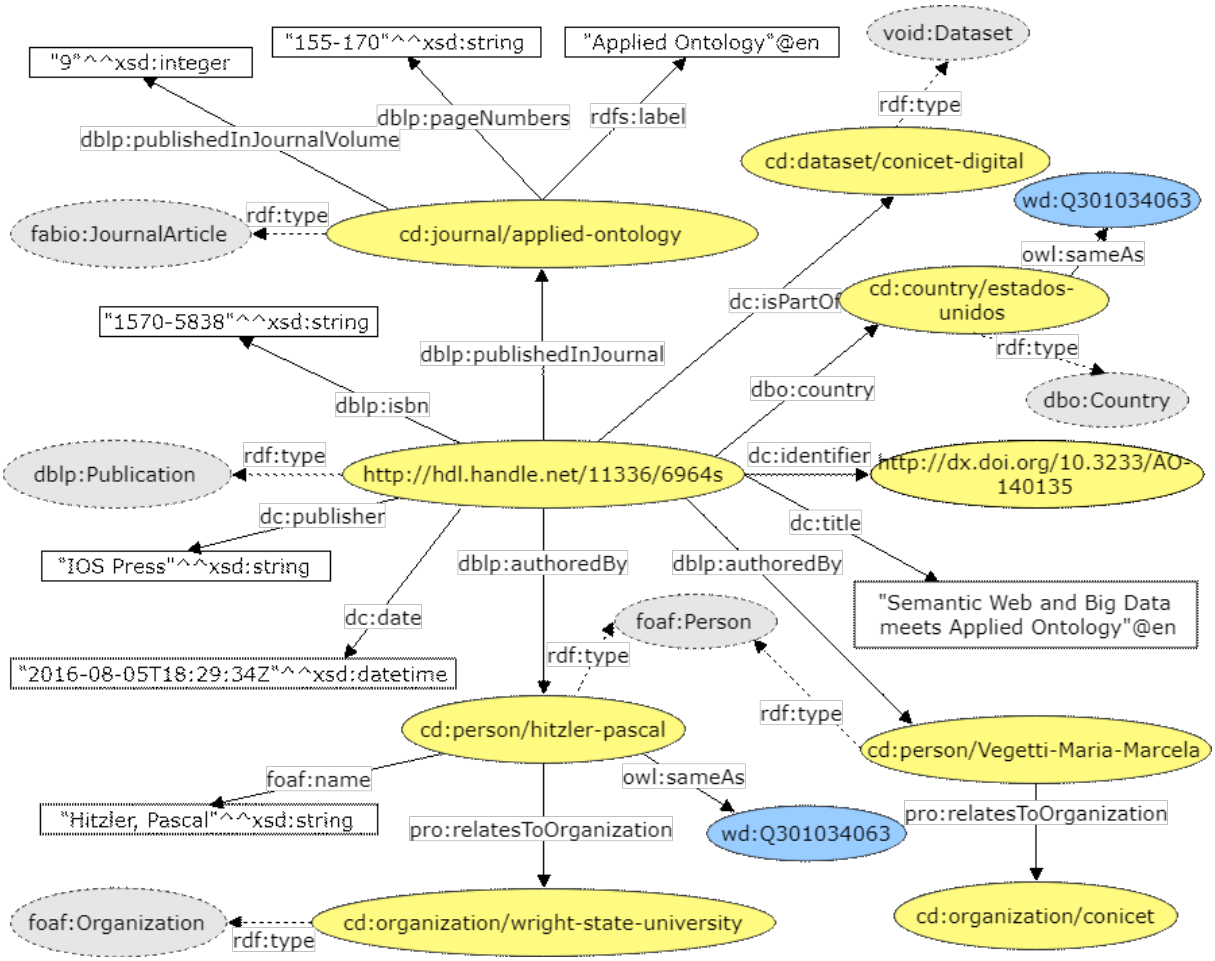

Fig. 2. Figure shows links between instances of classes in yellow colour, `rdf:type` assertions are shown in light gray. In blue color you can see the reconciled values from external datasets.

Algorithm 1: Publications by topic.

```
SELECT ?title
WHERE {
  ?s a dblp:Publication.
  ?s dc:subject ?sub.
  ?s dc:title ?title
  FILTER regex(STR(?sub),"ontology")
}
```

Algorithm 2: Publications by journal.

```
SELECT ?tit (COUNT(?jour) as ?count))
WHERE {
    ?s a dblp:Publication.
    ?s +dblp:publishedInJournal ?jour.
    ?jour rdfs:label ?tit
}
GROUP BY ?title
```

interested in retrieving all the publications related to ontologies.

4.2 Publications by journal.

The following SPARQL query (See Algorithm 2) allows counting the number of publications of each journal.

4.3 Impact factor by journal.

Checking the impact factor of a journal is essential, since this information is not visible in *CONICET Digital*, it is interesting to obtain it from another source. In this case DBPEDIA has information for some Journals. The properties we used were `dbo:impactFactor` and `dbo:impactFactorAsOf` as can be seen in Algorithm 3.

4.4 Retrieving DBLP identifiers from WIKIDATA

This query allows us to retrieve the identifiers of the authors associated with the publications. As mentioned in Section 3.3.3, the reconciliation service allowed us to find the co-authors in WIKIDATA, so we used the URIs to extract the DBLP identifier (`P2456`). To consult the WIKIDATA endpoint we use the SPARQL `SERVICE` clause that allows federated query a special type of SPARQL query that runs on more then one

Algorithm 3: Impact factor by journal.

```
SELECT DISTINCT ?title ?dplink
                ?i_factor ?date
WHERE {
 ?s a dblp:Publication.
 ?s dblp:publishedInJournal ?journal.
 ?journal rdfs:label ?title.
 ?journal owl:sameAs ?dplink
 FILTER regex(STR(?dplink), "dbpedia")

 SERVICE <https://dbpedia.org/sparql>
 {
   ?dplink dbo:impactFactor ?i_factor.
   ?dplink dbo:impactFactorAsOf ?date
 }
}
```

Algorithm 4: Identifiers from Wikidata.

```
SELECT DISTINCT ?s ?wd_page ?dblpID
WHERE {
?s a foaf:Person.
?s owl:sameAs ?wd_page.
FILTER regex(STR(?wd_page),"wikidata")

 SERVICE <http://wikidata.org/sparql>
 {
   ?wd_page  wdt:P2456 ?dblpID.
 }
}
```

SPARQL endpoint. It allows access to multiple linked data resources in a single query as can be seen in Algorithm 4.

4.5 Authors of different papers with at least three identical co-authors

The following query (See Algorithm 5) allows us to find information that can be used to detect certain patterns in the publications. For example, determine the authors of different publications that have at least three identical co-authors.

After seeing several examples of queries that are interesting for the user, and as a summary of this section, we give the links to each one (see Table 3) to execute them in the SPARQL interface of GraphDB.

Table 3: Links to queries

Query	Link
Query Section 4.1	CD-Q001
Query Section 4.2	CD-Q002
Query Section 4.3	CD-Q003
Query Section 4.4	CD-Q004
Query Section 4.5	CD-Q005

5 Proposed platform

An important architectural pattern used in systems development is the multitier architecture [22]. A multitier architecture separates functionality into a number of layers from low-level data storage through to user interaction components. This architecture is commonly used for many kinds of web application. As many Linked Data applications are also web applications, they tend to conform to this architectural approach [22]. An important advantage of the tiered architecture is that it logically separates the functionality of the system into a series of layers and specifies the communication between those layers. This separation makes it far easier to replace a layer of the architecture or reuse a layer of an existing architecture in a new application. The most commonly used multitier architecture is the three-tier architecture due to its simplicity and proven reliability [23], which is why we decided to base our architecture on this model. Fig. 3 illustrates the design of the *CONICET Digital* architecture and the following sections describe each of the layers.

Algorithm 5: Authors of different papers with at least three identical co-authors.

```
SELECT DISTINCT ?paper ?nAuthor {
{
  SELECT ?author1 ?author2 ?nAuthor
  {
   ?paper1 dblp:authoredBy ?author1;
           dblp:authoredBy ?author2;
           dblp:authoredBy ?author3;
           dblp:authoredBy ?nAuthor.

   ?paper2 dblp:authoredBy ?author1;
           dblp:authoredBy ?author2;
           dblp:authoredBy ?author3;
           dblp:authoredBy ?nAuthor.

   FILTER (?author1 != ?author2 &&
           ?author1 != ?author3 &&
           ?author3 != ?author2)

   FILTER (?paper1 != ?paper2)
  }
}.
  ?paper dblp:authoredBy ?nAuthor;
         dblp:authoredBy ?author1;
         dblp:authoredBy ?author2;
         dblp:authoredBy ?author3
}
```

5.1 Tier 1: input data

Data tier stores the underlying data independently of the business logic. In this case the datasets are transformed to RDF and subsequently exported in Turtle format as described in Section 3.3. After that they are imported into GraphDB triple store which supports different RDF serializations. GraphDB allows users to explore the hierarchy of RDF classes (`Class hierarchy`), where each class can be browsed to explore its instances. Similarly, relationships among these classes also can be explored giving an overview about how many links exist between instances of the two classes (`Class relationship`). Each link is a

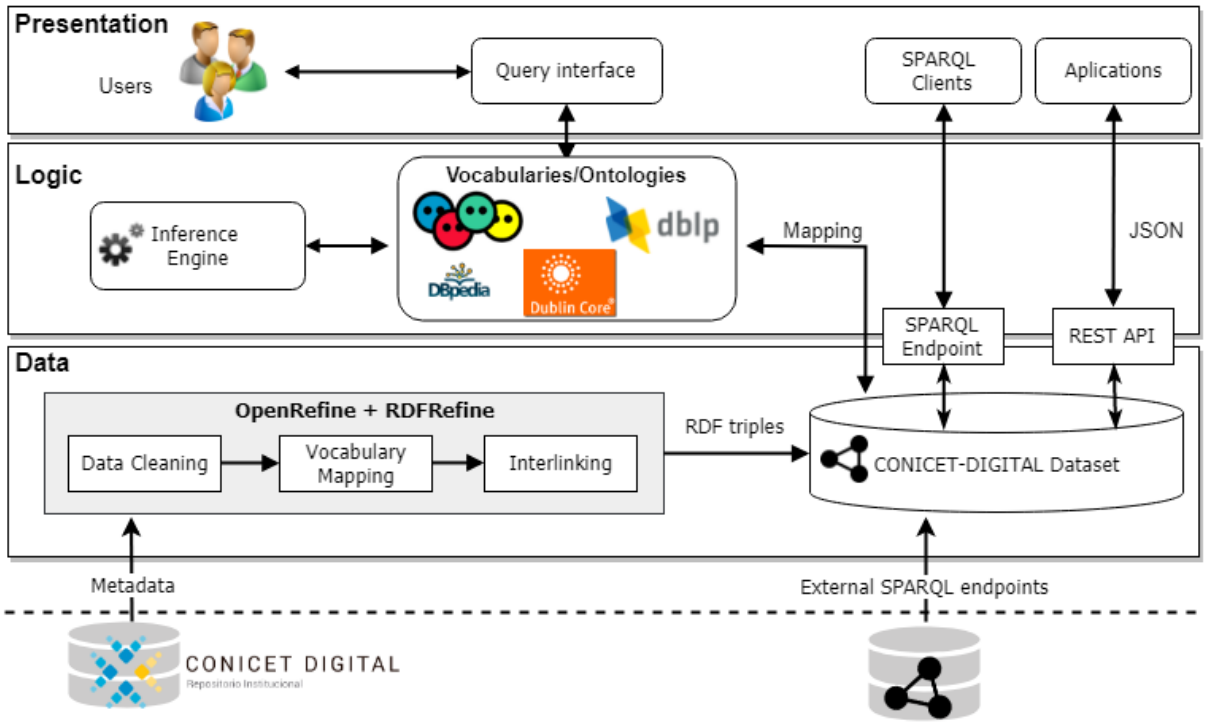

Fig. 3. Three-tier architecture proposed for *CONICET Digital*.

RDF statement where its subject and object are class instances and its predicate is the link itself. Lastly, users also can explore resources providing URIs representing any of the subject, predicate or object of a triple (`View resource`).

In case GraphDB becomes obsolete, *CONICET Digital* triple-based model is designed to live on, since it can be fully exported in RDF and imported into another RDF-compliant solution. Finally it is important to note that we can also import data to GraphDB from SPARQL endpoints allowing federated queries [24].

5.2 Tier 2: logic

Once the integrated data is available in GraphDB, it can be used and accessed by the logic and presentation layers. Some of the logic may be implemented in the data layer by reasoning over the triplestore although the reasoning is limited, that is why we need a higher level of expressiveness to reason. In this tier the ontologies allows the unequivocal identification of entities and the assertion of applicable named relationships that connect these entities. Specifically fulfills the following roles:

- *Explanation of content*: the ontologies allow the accurate interpretation of data from multiple sources through the explicit definition of terms and relationships.

- *Query model*: The query is formulated using the ontology as a global consultation scheme.

5.3 Tier 3: presentation

One of the features provided by GraphDB is an assistant-type interface that guides users in the creation of various RDF data visualizations with different starting points. You can set the default graphics display with the full expressiveness of the SPARQL language to control which graphics data you want to display.

GraphDB allows solving many of the complicated problems that arise when dealing with bibliographic data. This allows controlling the starting point of the visualization and creating more than one visualization on the same information. With this facility, the exploration of data, the analysis of data and the discovery of knowledge become easier and faster. So we can use the GraphDB facilities to infer relationships that are not explicitly established to get a complete picture of the data and gain additional knowledge about the links in our datasets. In addition GraphDB allows to visualize SPARQL queries using different types of charts, for example maps, bar charts, scatter charts, etc. Figure 4 shows a SPARQL query that groups by topic and visualized through a pie chart.

6 Discussion

The goal of exposing linked data is to make existing public data more accessible, reusable and exploitable. This can only be demonstrated through applications that make use of this data in innovative and/or cost-effective ways. With reference to the integration of scientific information, we have surveyed in similar

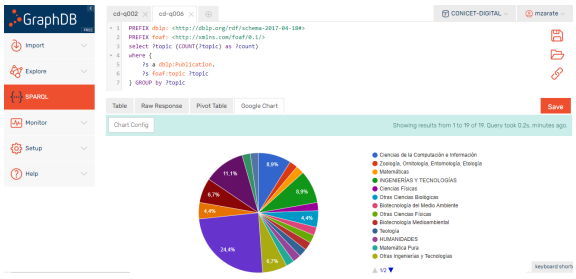

Fig. 4. SPARQL query visualization using a pie chart.

approaches, but we have found a single solution proposal in the field of the Argentine community [25]. In general, isues related to the scientific information management in Argentina, the infrastructure needed for Linked Data and how the ontology engineering could make use of this data are clearly open and the proposed solutions are scarce. To overcome partially this issue, it is possible to interlink scientific publications from some public datasets as DBLP or WIKIDATA. Nevertheless, an important number of publications are still being left out in this approach.

To sum up, issues related to the scientific information management in Argentina, the infrastructure needed for Linked Data and how the ontology engineering could make use of this data are clearly open and the proposed solutions are scarce. In order to start closing this gap, new applications should make use of linked open data.

7 Conclusions and future work

In this paper, we presents an overview of our initial efforts to create a linked open data repository using the information of a subset of scientific publications belonging to *CONICET Digital* to incorporate them into the web of data. We have detailed the transformation process and explained how to access and exploit them, promoting integration with other repositories. Moreover, we have depicted this process using queries extracted from the domain of application.

As a future work we plan to continue developing the followings aspects:

- Automate the process of extracting data from CONICET Digital using OpenRefine Python client libraries.
- In the future we intend to integrate a framework of automatic retrieval of connections, such as *Silk* [26].
- Develop a Web application for browsing scientific publications in the field of our researching groups. Through this application, we hope to establish connections with other educational institutions and information providers.

Competing interests

The authors have declared that no competing interests exist.

Acknowledgments

The work presented in this paper is in partial fulfillment of the research objectives set by the project ***Infraestructura de acceso a Datos Primarios con aporte de semántica en Repositorios Digitales*** partially funded by Secretariat of Science and Technology of the National University of Patagonia San Juan Bosco (UNPSJB).

References

[1] J. C. Molloy, "The open knowledge foundation: open data means better science," *PLoS biology*, vol. 9, no. 12, p. e1001195, 2011.

[2] J.-M. Barrueco and I. Subirats-Coll, "Open archives initiative protocol for metadata harvesting (oai-pmh): descripción, funciones y aplicación de un protocolo," *El profesional de la información*, vol. 12, no. 2, pp. 99–106, 2003.

[3] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," *Synthesis lectures on the semantic web: theory and technology*, vol. 1, no. 1, pp. 1–136, 2011.

[4] L. Yu, *A developers guide to the semantic Web*. Springer Science & Business Media, 2011.

[5] K. Janowicz, P. Hitzler, B. Adams, D. Kolas, and C. Vardeman, "Five stars of Linked Data vocabulary use," *Semantic Web*, vol. 5, no. 3, pp. 173–176, 2014.

[6] M. Arenas, C. Gutierrez, and J. Pérez, "Foundations of rdf databases," in *Reasoning Web. Semantic Technologies for Information Systems*, pp. 158–204, Springer, 2009.

[7] T. Berners-Lee, J. Hendler, O. Lassila, *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[8] E. Prud'hommeaux and A. Seaborne, "SPARQL query language for RDF – W3C recommendation," tech. rep., W3C, 2008.

[9] B. Hyland, G. Atemezing, and B. Villazón-Terrazas, "Best practices for publishing linked data," *W3C Working Group Note*, 2014.

[10] "Springer Nature SciGraph." `http://www.springernature.com/gp/researchers/scigraph`. [Online; accessed 24-January-2018].

[11] M. Ley, "Dblp: some lessons learned," *Proceedings of the VLDB Endowment*, vol. 2, no. 2, pp. 1493–1500, 2009.

[12] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[13] D. Shotton, "Introduction the semantic publishing and referencing (spar) ontologies. october 14, 2010," *URL: http://opencitations. wordpress. com/2010/10/14/introducing-thesemantic-publishingand-referencing-spar-ontologies*, 2010.

[14] S. Peroni, A. Dutton, T. Gray, and D. Shotton, "Setting our bibliographic references free: towards open citation data," *Journal of Documentation*, vol. 71, no. 2, pp. 253–277, 2015.

[15] R. Cyganiak, F. Maali, and V. Peristeras, "Self-service linked government data with dcat and gridworks," in *Proceedings of the 6th International Conference on Semantic Systems*, p. 37, ACM, 2010.

[16] L. Sauermann, R. Cyganiak, and M. Völkel, "Cool uris for the semantic web," 2007.

[17] S. Peroni, D. Shotton, and F. Vitali, "Scholarly publishing and linked data: describing roles, statuses, temporal and contextual extents," in *Proceedings of the 8th International Conference on Semantic Systems*, pp. 9–16, ACM, 2012.

[18] D. C. M. Initiative *et al.*, "Dublin core metadata element set, version 1.1," 2012.

[19] S. Peroni and D. Shotton, "Fabio and cito: ontologies for describing bibliographic resources and citations," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 17, pp. 33–43, 2012.

[20] R. Verborgh and M. De Wilde, *Using OpenRefine*. Packt Publishing Ltd, 2013.

[21] B. Bishop, A. Kiryakov, D. Ognyanoff, I. Peikov, Z. Tashev, and R. Velkov, "OWLIM: A family of scalable semantic repositories," *Semantic Web*, 2011.

[22] H. Schuldt, "Multi-tier architecture," in *Encyclopedia of database systems*, pp. 1862–1865, Springer, 2009.

[23] W. W. Eckerson, "Three tier client/server architecture: Achieving scalability, performance, and efficiency in client server applications," *Open Information Systems*, 1995.

[24] B. Quilitz and U. Leser, "Querying distributed rdf data sources with sparql," in *European Semantic Web Conference*, pp. 524–538, Springer, 2008.

[25] G. Michelan, G. Braun, L. Cecchi, and P. R. Fillottrani, "Integration of scientific information through linked data," in *II Simposio Argentino de Ontologías y sus Aplicaciones (SAOA 2016)-JAIIO 45 (Tres de Febrero, 2016)*, 2016.

[26] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov, "Discovering and maintaining links on the web of data," in *International Semantic Web Conference*, pp. 650–665, Springer, 2009.

Citation: M. Zárate, C. Buckle, R. Mazzanti and G. Samec. "*Improving Open Science Using Linked Open Data:* CONICET Digital *Use Case"*, Journal of Computer Science & Technology, vol. 19, no. 1, pp 45-54, 2019.
DOI: 10.24215/116666038.19.e05
Received: July 16, 2018. **Accepted:** November 5, 2018.