Research Paper

JIS

Journal of Information Science
1–15

DOI: 10.1177/0165551520944351
journals.sagepub.com/home/jis
SAGE

REDI: Towards knowledge graph-powered scholarly information management and research networking

José Ortiz Vivar
Department of Computer Science, University of Cuenca, Ecuador

José Segarra
Department of Computer Science, University of Cuenca, Ecuador

Boris Villazón-Terrazas
Tinámica, Madrid, Spain

Víctor Saquicela
Department of Computer Science, University of Cuenca, Ecuador

Abstract
Academic data management has become an increasingly challenging task as research evolves over time. Essential tasks such as information retrieval and research networking have turned into extremely difficult operations due to an ever-growing number of researchers and scientific articles. Numerous initiatives have emerged in the IT environments to address this issue, especially focused on web technologies. Although those approaches have individually provided solutions for diverse problems, they still can not offer integrated knowledge bases nor flexibility to exploit adequately this information. In this article, we present REDI, a Linked Data-powered framework for academic knowledge management and research networking, which introduces a new perspective of integration. REDI combines information from multiple sources into a consolidated knowledge base through state-of-the-art procedures and leverages semantic web standards to represent the information. Moreover, REDI takes advantage of such knowledge for data visualisation and analysis, which ultimately improves and simplifies many activities including research networking.

Keywords
Academics; data integration; Linked Data; research networking; scholarly data; text mining

1. Introduction

Scientific production has been growing exponentially for several decades, thousands of papers are reviewed, published and cited on a daily bases worldwide. In order to deal with this immense amount of data, numerous information systems (e.g. digital libraries and search engines) have been deployed on the web, which handle, store and access valuable metadata and full text of academic contents. Although these systems meet their main objective of preserving digitally bibliographic resources, it is specially challenging to obtain a clean overall view of the information in these sources as a unified knowledge base due to access, format and ambiguity issues.

In this context, important academic activities such as research networking, bibliometric analysis and data visualisation are hampered because digital literature is scattered among multiple sources, noisy or disorganised. Overcoming these drawbacks frequently result in the need of manual labour from human experts to consolidate data and construct

Corresponding author:
José Ortiz Vivar, Department of Computer Science, University of Cuenca, April 12 Avenue, Cuenca, EC 010203, Ecuador.
Email: jose.ortizv@ucuenca.edu.ec

knowledge bases, which can be used to accomplish the multiple information requirements of researchers, universities and government agencies. Therefore, the manual integration approach has become the most reliable strategy to tackle heterogeneity in real use cases.

Besides, multiple attempts of addressing those issues have emerged in recent years ranging from universal identifiers for bibliographic resources, centralised bibliographic repositories to even researchers' social networks. The new technologies have effectively reduced the amount of manual work in academic processes; however, they constitute at most partial solutions for specific problems and do not provide a completely integrated knowledge base, which could be used for broader use cases and applications. Moreover, most of them do not take into account interoperability nor expandability, which are important aspects of modern systems.

In this work, we present REDI[1] a centralised scholarly framework, which provides both a Linked Data-powered consolidated knowledge base and research networking features. REDI harvests, standardises and integrates information from multiple academic information systems into one single repository, and also provides diverse tools for data analysis and visualisation such as similar authors' clustering and knowledge area classification. In addition, the integration approach implemented in the presented framework heavily relies on semantic web's technologies as Resource Description Framework (RDF), SPARQL Protocol and RDF Query Language (SPARQL) and well-known bibliographic ontologies (e.g. BIBO, FOAF and DCTERMS), which allows seamless integration with other scholarly focused Web 3.0 solutions.

The remaining of this article is organised as follows: in section 2, the related work is covered. REDI's architecture is presented in section 3, where its main components and workflows are outlined. In section 4, the successful use case of the Ecuadorian implementation of REDI is described. Finally, in section 5, we present conclusions and future work.

2. Related work

The idea of centralising and analysing scholarly data is not new at all; there are plenty of online tools and research work focused on this objective. Tools such as search engines, central repositories and business/scientific social networks have diversified and specialised according to the information they handle and their audience's needs over time. In addition, research on data integration and knowledge management in the academic context has contributed to new technologies and tendencies. This section introduces the most relevant related works, their advantages and limitations from the perspective of knowledge bases and Linked Data [1].

Search engines are probably the most commonly used alternatives when researchers look for recent scientific literature, colleagues and references. Websites such as Google Scholar, Microsoft Academic and Baidu Scholar ease access to millions of papers and researcher profiles through crawling the web and organising the information in a human-friendly manner. However, their reliability from a knowledge management point of view is limited due to the fact that their information is usually obtained through web scrapping and it is prone to data duplication, ambiguity and erroneous metadata field detection. Moreover, accessing the information from an external application can be challenging because these websites usually do not provide programmatic access through web services, and if they do, the representation formats and vocabularies are not standardised.

However, the aim of providing more structured scholarly information has triggered the development of digital repositories, which harvest directly journals and digital libraries. These repositories leverage on open access technologies (e.g. OAI-PMH[2]) in order to build a central repository of metadata that can be accessed and searched more conveniently. For instance, initiatives such as BASE [2] and CORE [3] share information of millions of open access documents in one single endpoint. Furthermore, the use of standard guidelines (e.g. OpenAire and CERIF) and protocols/formats (OAI-PMH, DCTERMS) for storing and accessing publications helped to the development of search Application Programming Interfaces (APIs) and harvest mechanisms. Nevertheless, the document-centric focus of those systems limits their usage on broader applications. In other words, the information regarding author profiles, affiliations or conferences is stored as fields of the document rather than independent resources, which induces to consistency and ambiguity issues, and limits the processable information that can be obtained from these important resources.

Recently, more sophisticated solutions have emerged, which take into consideration important features such as proper identifiers for resources, collaborative maintenance of metadata and advanced data analysis/visualisation tools. The main scholarly services worldwide (e.g. Scopus and Web of Science) have developed complete platforms that offer well organised and standardised data, strong resource identification mechanisms (e.g. ORCID, DOI and ScopusID) and useful visualisation interfaces. Similar features could be also found in Semantic Scholar [4], a platform that harvests information from other relevant sources (e.g. ArXiv). In addition, social networks (e.g. ResearchGate) open up new opportunities for direct interaction among researchers and research work spreading. However, there are still plenty of shortcomings on these approaches, especially regarding openness and interoperability. Restrictions that limit the ability to carry out

specific studies on certain contexts, especially if they come from third parties. Studies that can be very useful for organisations, governments and other stakeholders.

Numerous independent alternatives have also evolved in Universities, Libraries and Research Centres in order to fulfil the scholarly information requirements of its stakeholders. Updated reports of research results and metrics are turning increasingly common as institutions keep improving its internal processes and governmental control agencies continuously audit and categorise them. Therefore, commercial or ad hoc solutions for managing domestic-produced research assets inside universities are an increasingly growing trend in modern academics. For instance, the U-papers platform in the University of Chile allows researchers to share their results among them and with the community [5]. Rexplore is another relevant work that aims to analyse research trends on scholarly data leveraging on semantics and knowledge-based exploration [6]. Similarly, an Information Retrieval–based system is proposed in Atanassova and Bertin [7], which suggests the application of more innovative procedures (e.g. semantic annotations and ontologies). However, these proposals either have a very specific field of action or they are still at a conceptual stage that lacks applicable tools.

In recent years, several studies based on the application of semantic technologies have been carried out in the educational context. For instance, these technologies have been successfully implemented in real-world systems in the context of e-Education [8] and scholarly activity management [9]. Furthermore, researchers in the bibliographic field have also worked with these technologies, especially in ontology-related topics that have led to the creation of ontologies/vocabularies (e.g. Dublin Core and BIBO), and discussions about their applicability in the Semantic Web [10].

Concretely in the Ecuadorian context, some research works and prototypes have been developed to address the academic integration problem from a Linked Data approach. These works primarily focus on porting academic systems of Ecuadorian universities to the Linked Data world. For instance, frameworks for institutional repository integration are presented in Piedra et al. [11] and Segarra et al. [12], which lists a series of steps for publishing this kind of information as Linked Data. Similarly, a Linked Data–based integration use case for the library systems of the University of Cuenca is presented in Suma et al. [13]. However, the generation of collaboration networks of Ecuadorian researchers is analysed in Piedra et al. [14], where the authors classified a subset of Computer Science researchers (obtained from Scopus) into research topics through an inference process based on the DBpedia concepts. Next, the authors of this work expanded this idea to Computer Science researchers from the Andean region countries in a more recent paper [15]. Nonetheless, all the aforementioned approaches focus on individual data sources; therefore; they do not provide an entire view of the research status in Ecuador.

In this context, the present work outlines a new framework for scholarly information management and research networking, which relies on semantic technologies. The REDI framework offers an actual trade-off between information scope and quality for both researchers and institutions. On one hand, the knowledge base of REDI is bounded to researchers of a certain country, which diminish the amount irrelevant information for its stakeholders in a domestic ambit. On the other hand, scholarly information from diverse sources (institutional, regional and international) is recollected and processed for those researchers in order to warranty a concise and complete database. Then, advanced operations are performed on this knowledge base, such as automatic authors' categorisation by research area and collaborative networking discovery. This last feature reveals implicit associations among authors and organisations, and ultimately encourages researchers to establish new ones. Finally, all that knowledge is served through simple but pragmatic user interfaces that allow to explore and discover information of the regional-produced research that raises interest on scientific community (local and foreign), governmental institutions and public in general.

In order to contrast this approach, in Table 1 we present the features comparison between other existing platforms and REDI. The first three characteristics focuses on openness and access of data; in these items, REDI excels due to its strict adherence to the open data principles. Moreover, REDI offers access to its knowledge base from both levels final user and programmatic. Next, REDI heavily relies on automatic processes for data population, and therefore, it dispenses with manual profiles creation, an essential feature on real-world applications. Subsequently, the scope of the Knowledge Base (KB) is reviewed on the fifth item, where the inclusion of institutional and regional sources is rather unusual on other platforms, which is an actual drawback on domestic research analysis. Finally, the research networking capabilities are analysed on the two last items where most platforms do not provide direct linking mechanism among authors nor a standardised taxonomy for an adequate thematic classification. All this makes REDI a more suitable tool for scholarly information handling, specially from a domestic perspective.

3. REDI architecture

In this section, we describe the overall operation of REDI including its focus, main components and workflows. REDI is built on top of the Apache Marmotta platform in order to facilitate its support for Linked Data and Semantic Web standards [16]. Moreover, all its components have been designed to work independently as collaborative modules that coexist

Table 1. Academic platform feature comparison.

Features/ platform	Research Gate	Google Scholar	Scopus	Semantic Scholar	Microsoft Academics	U-papers	REDI
Accessibility	Open access	Open access	Access under payment	Open access	Open access	Open access	Open access
Linked Data supported	N/D	N/D	Partial support	N/D	N/D	N/D	Full support
Access to developer	N/D	N/D	API under payment	REST API	API under payment	N/D	API SPARQL/ SOLR
Generation of author profiles	Semi-automatic	Manual	Semi-automatic	Automatic	Automatic	Manual	Automatic
Coverage	External sources	Internal/ external sources	External sources	External sources	Internal (partial)/ external sources	Internal sources	Internal/ external sources
Thematic classification of authors	Yes	Yes	Yes	No	Yes	No	Yes (two levels)
Collaboration networks	Yes	No	No	No	No	No	Yes
Multi-level data visualisation	Partial	No	Yes	No	Partial	No	Yes

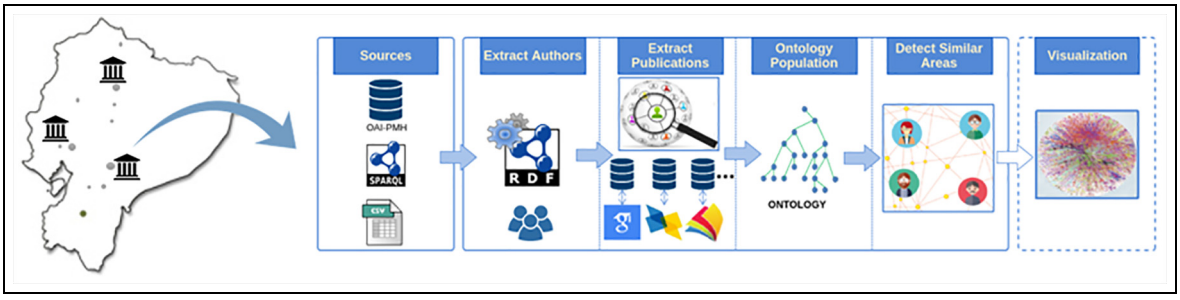

Figure 1. Main components of REDI.

together in the Marmotta environment. In Figure 1, we present the general architecture of our system and its core workflow. Broadly speaking, REDI obtains author profiles from universities, harvests publications from multiple online scholarly sources, consolidates the information and offers visualisation interfaces. A more descriptive explanation of each module is exhibited in the subsequent sections.

3.1. Authors' sources

The first stage of the workflow is gathering basic authors' information. REDI follows a university-centric approach of searching for authors in order to limit the scope and provide an affiliation-based focusing of the data that handles. For instance, the information of active researchers of universities is obtained from digital repositories, manually generated reports and journal management systems.

The process starts by building up a universities' registry (a.k.a. Organisations' list) on the repository, also including essential information such as acronyms, complete name, location (longitude, latitude), country, city and web page. Afterwards, the data sources of those institutions are added as endpoints, which will be processed for extracting author profiles. Currently, there is support for the most common systems and formats used in the Latin American context, which are listed in Table 2. Nevertheless, the addition of new extractor components could be easily implemented if needed in the future.

Table 2. Supported formats for extracting authors.

Type	Format	Description
Researcher list	CSV	Manually created lists of researchers.
Institutional repository	OAI-PMH	Repositories containing internal documents such as thesis and postgraduate works.
Journal systems	OAI-PMH	Local journals maintained by universities.
Semantic repository	RDF SPARQL	Scholarly data sets produced internally by universities or researchers following Linked Data principles.

Listing 1. Author extracted from the University of Cuenca.

```
@prefix foaf: < http://xmlns.com/foaf/0.1/ > .
@prefix dc: < http://purl.org/dc/terms/ > .
@prefix redia: < https://redi.cedia.edu.ec/authors/ > .
@prefix redip: < https://redi.cedia.edu.ec/publications/ > .
@prefix redie: < https://redi.cedia.edu.ec/endpoints/ > .
redia:CULLCAY__JOSE
foaf:publications redip:oai_ojs.pkp.sfu.ca_article_1455;
   a foaf:Person;
   dc:provenance redie:ojsUCUENCA;
   foaf:lastName 'Cullcay'^^xsd:string;
   foaf:firstName 'Jose'^^xsd:string;
   foaf:name 'Cullcay, Jose'^^xsd:string.
redip:oai_ojs.pkp.sfu.ca_article_1455 dc:title
      'Identificacion automatica de articulos
      indexados en Latindex'^^xsd:string;
```

3.2. Authors' extraction

The extraction of author profiles from organisations relies on ETL[3] processes that acquire metadata from the registered endpoints and transform it into Linked Data resources. The specific fields harvested from the sources heavily depend on the formats supported by the endpoint and the granularity level that its administrators have implemented. In order to automatise the operation of extraction, we embedded a Linked Open Data Generation Framework (LOD-GF [17]) inside REDI. This framework uses internally an ETL process manager that jointly with R2RML[4] mappings allows to generate RDF data from heterogeneous sources.

The information of authors is represented using the FOAF ontology, and if the source provides publications information that data are annotated using the BIBO and DCTERMS ontologies. As a final result of this stage, we obtain basic information of the researchers and their publications (if available). This information is stored in our repository as a separated graph called 'Authors' Graph', which is used as a starting point for the following stages. In listing 1, we give an example of the information recollected and processed from an Open Journal System (OJS) (OAI-PMH) endpoint of the University of Cuenca.

3.3. Publication extraction

In order to provide a complete scholarly knowledge base, REDI searches and harvests publications of the registered authors in multiple digital sources. As mentioned in the related work section, there are plenty of scholarly online services (i.e. Scopus, Google Scholar, etc.), which store millions of scientific papers and other relevant information. So, taking advantage that those services usually have flexible policies for the use of their data in research projects or are open access, REDI has implemented an adequate infrastructure for accessing and harvesting them through modular components named Linked Data Providers (LD-Ps). However, it is important to point out that the components and processes of REDI are essentially data provider-agnostic, which implies that it is possible to add or remove determinate providers according to a specific setup of REDI in order to comply with external requirements such as licencing or access to the academic services.

Table 3. Data providers of REDI.

Source	Available resources	Accessibility
Scopus	Authors. Affiliations. Publications. Journal/Conference.	Weekly quota (Subscription).
Microsoft Academic	Authors. Affiliations. Publications. Journal/Conference.	Limited free API key.
Springer	Publications.	Free API key.
Google Scholar	Authors. Publications.	Web Scrapping.
DOAJ	Publications.	Free API.
Scielo	Publications.	Free API.
Crossref	Publications.	Free API.
ORCID	Authors. Publications.	
DBLP	Authors. Publications. Journal/Conference.	Free API.

LD-Ps are just an implementation of the Linked Data Client interfaces, which are essentially flexible and modular RDFisers[5] devised within the Marmotta environment. These components access online services through multiple formats (e.g. JSON, XML and HTML), transform their records into Linked Data resources and store the information into independent RDF graphs.

The specific harvesting logic implemented within an LD-P depends on the service's features and restrictions. For instance, most of the services implement some kind of quota control and limit the number of requests in a certain period of time, thereby small time delays were added on purpose between requests. Moreover, the successfully harvested authors are marked in order to avoid repeated requests if a harvesting process fails halfway through due to the aforementioned restrictions and needs to be reset. Other sources, such as Scopus allows date filters in its requests, in this case the last successful harvest timestamp is used to query incremental additions of data.

The structure of the returned data from the LD-Ps has to be consistent; however, the responses of the academic services vary greatly from one source to another. For instance, whereas Scopus and DBLP provide independent resources for each academic entity (i.e. researcher, publications and journals), other services usually store all those resources as fields of a single master entity: Publications. Therefore, in some cases, the missing resources and relationships have to be made up from single fields so as to preserve a basic consistency.

In Table 3, we present a list of the currently implemented providers and details about them. The addition of new providers is periodically analysed based on the candidate services' features, data access policies and overlap of content with existing providers. For instance, Semantic Scholar and Web of Science are being assessed as new potential providers.

The name of each author within the REDI registry is searched on the LD-Ps, also including the author's affiliation (if supported). Next, the resulting records are stored as candidate profiles in independent graphs. It is important to notice that the harvested data at this stage have issues such as duplicate profiles, incomplete metadata and false positives (namesake), which are addressed at the ontology population stage.

3.4. REDI ontology

An indispensable intermediate task in REDI is the standardisation of the data from its sources. Even though most of the homogenisation process is carried out within the LD-Ps, the data models have to be unified at an ontological level. Therefore, the vocabularies from the sources are treated through an R2R [18] post-processing that transforms all resources' representation to the common ontology model defined within the REDI framework. The REDI ontology network for the representation of academic contents was discussed in more detail in previous work [19], and it essentially reuses well-known ontologies within the scholarly domain with the addition of some ad hoc components specific for REDI. In Table 4, for the sake of brevity, we only present the core classes and ontologies used by REDI.

Table 4. Ontology network of REDI.

Entity	Ontology
Researchers	FOAF, vCard, SCoRO, BIO
Publications	BIBO, npg (nature), DC (Dublin Core)
Journals/Conferences	BIBO, DC (Dublin Core)
Organisations	FOAF, Schema
Clusters	SKOS, uc (ad hoc)
Subclusters	DBpedia, uc (ad hoc)

It is important to mention that, the construction and formalisation of this ontology network is an ongoing work. We are currently exploring its expansion to the so-called CRIS[6] entities, such as research projects, grants and research departments. The formalisation of this ontology network will be discussed and shared with the community in future work.

3.5. Ontology population

In the ontology population stage, a consolidated knowledge base is constructed from the information collected in the author and data provider graphs. There are mainly three internal tasks that are performed during this stage. First, a disambiguation algorithm is executed over all collected data, which remove namesake errors, merge equivalent resources and unveil hidden relations among authors. Next, the resulting graph is enriched using external sources such as journal ranking systems and regional indexing services, which enhance its overall content both in quality and quantity. Finally, the new knowledge base is loaded into the front end of REDI for data analysis and visualisation purposes (following stages).

The disambiguation process implemented by REDI is portrayed in more detail in a previous work [20]. Roughly, the authors' information obtained directly from universities (Author Graph) is used as a seed for a series of semantic comparisons upon the harvested information (Provider Graphs). During this procedure, metadata such as author's name, affiliation, publication's titles and keywords are compared and weighted in order to estimate a semantic distance between author profiles. That value is then used to determine whether or not two resources could be considered equivalent while generating the consolidated graph.

In the disambiguation process, we identified that the division of the data providers in two groups according to the quality of their data significantly improves the performance of the matching process. Academic services such as Scopus and ORCID offer trustworthy information about author profiles and co-authorship; therefore, we use them first in the matching process for enriching the locally extracted authors' information. Next, the additional information added to the local authors is used as context for the matching process with other providers, whose information is more prone to errors (i.e. Google Scholar and Microsoft Academic) or less structured (i.e. Scielo, DOAJ, etc.). According to our tests, this approach keeps a balance between both the quality and completeness of the disambiguation results.

Collecting academic data from multiples sources usually raises some uncertainty issues regarding correctness and completeness of the information. Each academic service usually provides a different set of fields for describing resources and still the values for the same fields may vary from a source to another, even though they refer to the same resource. In this context, REDI leverages on the number of its data providers and the content overlap among them to deal with these issues. Therefore, if a resource (researcher or publication) is identified in multiple sources, the metadata fields provided by each source are merged into a single REDI record. However, if there are conflicting values among data providers, only the most common value is preserved or, failing that, the entire field could be skipped if a minimum consensus cannot be reached.

Regarding external sources enrichment task, it is mostly focused on adding new information about journals and books. First, extra metadata is extracted from the Elsevier Metadata API that retrieves cover images, web sites and other fields. Second, the SCImago [21] database is used to obtain details such as SJR, best quartiles, country and scope. Finally, a regional-indexed identification process for publications is executed with the aim of establishing links between some articles and regional journal systems like Latindex (further details of this process are available in Ortiz et al. [22]).

3.6. Collaboration networks

One of the core objectives of REDI since its inception has been to facilitate the discovery of author's research clusters, as well as new possible collaboration groups. Therefore, we have been constantly working on new strategies for

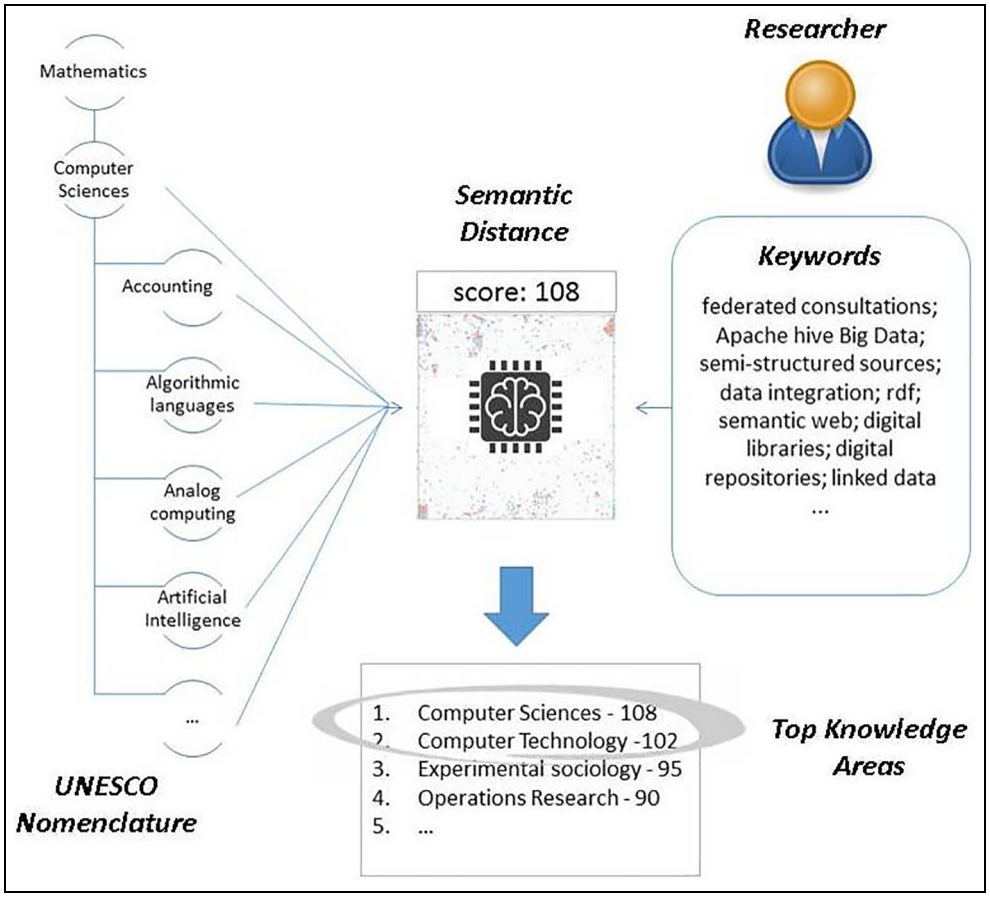

Figure 2. UNESCO semantic clustering.

grouping researchers according to the real-world requirements of academics. In former versions of the project, we discussed a machine-learning approach for automatically creating clusters and labelling them, which leveraged on some state-of-the-art techniques such as VSM[7] and LSI (Latent Semantic Indexing) and CVB (Collapsed Variational Bayes) [19]. Although that strategy showed to be correct from a theoretical perspective, the loss of control over the resulting groups due to this unsupervised strategy has caused the search of new alternatives. Therefore, a new clustering algorithm with controlled vocabularies has been developed in REDI, which relies on well-known and extensively used nomenclatures for science and technology maintained by UNESCO [23] and DBpedia [24].

First, the process classifies the authors into areas of knowledge based on the UNESCO classification. This new strategy compares semantically researchers' keywords against the nomenclature descriptions in order to discover the more suitable knowledge area. This process both implicitly clusters similar researchers into the same knowledge areas and forms a complete and standardised catalogue of researchers. In Figure 2, a basic graphical representation of the first level clustering process is depicted.

Subsequently, a second fine-grained grouping process is executed based on the previously formed clusters and the keywords of the most relevant works of the authors that uses the DBpedia knowledge base. In this process, concepts of DBpedia are extracted from the keywords using the DBpedia Spotlight entity detection service [25] and only research-related concepts are preserved. The result of this enrichment procedure is a list of research topics which offer a more detailed hierarchy for authors. In Figure 3, a graphical representation of the area of research detection process is depicted.

Finally, the results of both classification processes are stored in REDI as a separate graph called the Cluster Graph. The information of both clusters and subclusters is converted to RDF using the Simple Knowledge Organization System (SKOS) ontology [26]. The complete workflow and algorithms used in this classification process are explained in more detail in Segarra et al. [27]

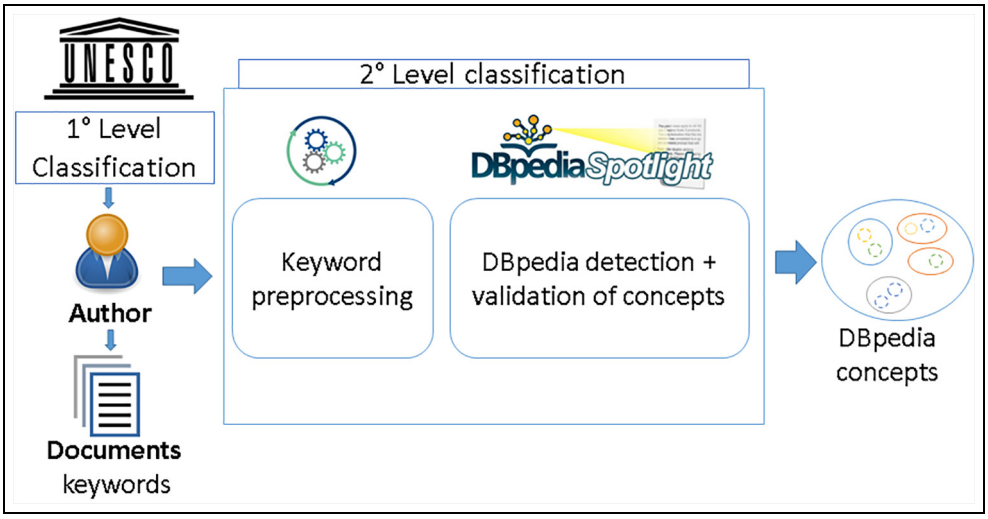

Figure 3. DBpedia concept detection.

Figure 4. REDI homepage.

3.7. Visualisation

Data visualisation is a core component of REDI, due to making accessible its knowledge base to final users through friendly graphical user interface (GUI) is indispensable for achieving its main objectives. So, the front end of REDI incorporates a website with several stats, reports, dynamic charts and search utilities that help users to browse through all resources. Currently, the REDI website is publicly available on https://redi.cedia.edu.ec/ and it is being used by students and researchers of all Ecuador. In Figure 4, there is a screenshot of the homepage of REDI, where users can search and access information of researchers, publications and collaboration networks.

As an academic management framework, REDI offers web interfaces for exploring its knowledge base in a traditional manner. Some of the functionalities that are present are researcher profiles, publication lists, statistical charts and downloadable reports. In addition, an ORCID authentication mechanism has been added to REDI, which allows a researcher to validate his or her profile within REDI and fix any potential issues in the knowledge base. In Figure 5, a researcher profile is shown with some relevant information such as name, e-mail, affiliation, in-line biography, co-authors, citation indicators and so on.

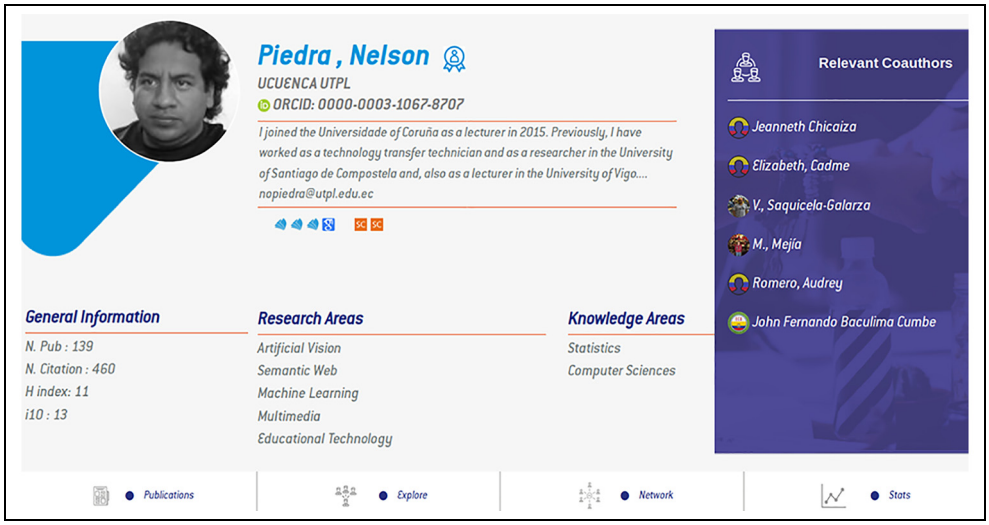

Figure 5. Researcher profile.

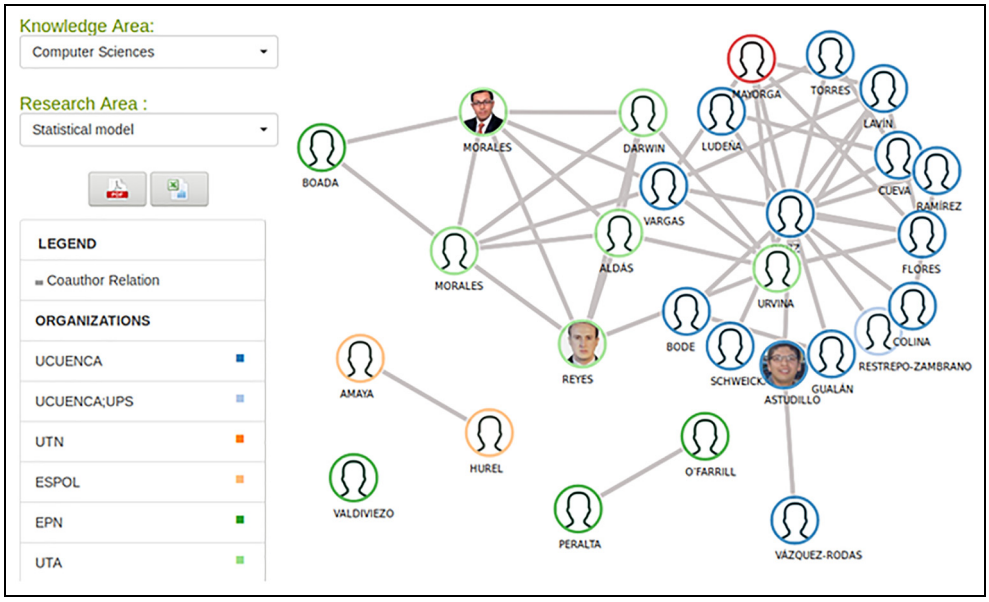

Figure 6. Collaboration networks.

Finally, the researcher classifications in knowledge areas and research areas are also present in the web interfaces of REDI, where the final users can perform multiple tasks such as filter authors by their research interests and obtain heat maps about research trends in the universities of Ecuador. In addition, there some dynamic charts that combine co-authorship information with the authors' classification in order to expose collaboration networks among researchers.

In Figure 6, we present an example of a collaboration network for the research topic 'Statistical model' within the Computer Science area that was obtained from REDI. These charts are useful for researchers and organisations due to that we can identify researchers' groups within certain knowledge areas, their organisations and how they interact (co-authorship). For instance, in the figure, we can infer the existence of two research groups working on Statistical Models: the University of Cuenca (UCUENCA) and the Technical University of Ambato (UTA). Also, we can establish that the

Table 5. Authors' sources of REDI – Ecuador.

Extraction method	Number of universities	Number of researchers
Researchers list	20	7,128
Journal system	20	7,459
Institutional repository	16	17,738

Table 6. Authors – REDI versus Scopus.

Repository	Number of universities	Number of researchers
REDI	48	30,141
Scopus	52	17,341

researchers Morales and Urvina act as the linkages between those two organisations in this specific topic. Moreover, further details about the researchers and publications can be accessed from this chart if the users navigate to their profiles.

4. Use cases

4.1. Use case A

The first use case of the REDI framework was implemented over scholarly data of academic institutions in Ecuador. The information of tens of high education institutions was registered in the framework and the aforementioned processes were applied. In each of these phases, a sample of the resulting data will be presented and compared with Scopus, one of the best known and most used scientific databases in the world. Due to the fact that Scopus, like REDI, contains data for entities such as authors, publications and organisations, and this information could be easily compared between them. This comparison intends to highlight the advantages of REDI in a real-world scenario.

4.1.1. Authors' extraction. In this phase, 48 organisations were registered in the repository of Ecuador, which comprise almost the totality of the universities of Ecuador. In order to extract authors from those organisations, lists of researchers with basic information were requested to their staff. The information requested included full names, identifiers such as ORCID, e-mails and optionally areas of expertise. However, it was not possible to get access to this information for all universities, so alternative sources such as journal systems and institutional repositories were used instead. In some cases, several types of author's sources were combined for the same university. In Table 5, we present a summary of the number of resources and the extraction methods used for this implementation of REDI.

The authors collected from different sources went through a disambiguation process in which many duplicated author records were unified under the same researcher profile. The total number of final authors found by REDI is presented in Table 6. In the same table, the number of researchers with Ecuadorian affiliations registered within Scopus was added for comparison. Obtaining this information required manual work and multiple requests to the Scopus website, because in Scopus there are no views of authors per country and the existing interfaces only return partial information. The data show that Scopus covers a larger number of organisations for Ecuador, but REDI has a larger number of authors for the common organisations, probably a consequence of having multiple sources to harvest authors.

4.1.2. Publications' extraction. Through the utilisation of the aforementioned mentioned LD-P modules of REDI, information about publications, journals and other details was harvested from multiple online sources. However, as stated before, the collected data have ambiguity issues, and therefore, they need to be cured. In Table 7, a summary of the harvesting process results is presented grouped by LD-P.

The comparison of the publications between REDI and Scopus follows the same trend shown in the case of authors. In the comparative Table 8, REDI has fewer institutions, but the number of documents is much larger than Scopus. This result is predictable because REDI harvests several sources, including Scopus itself. However, it helps to understand how REDI is capable of covering a large amount of information, which makes it more complete than their counterparts.

Table 7. Raw information harvested by REDI.

LD-P	Author profiles	Publications
Scopus	22,538	78,542
Microsoft Academic	103,970	300,381
Springer	8,062	1,939
Google Scholar	1,563	29,947
DOAJ	148,847	130,838
Scielo	363,257	260,298
DBLP	68,914	271,964
Crossref	985,986	906,623
ORCID	6,943	86,630

Table 8. Publications – REDI versus Scopus.

Repository	Number of universities	Number of publications
REDI	48	88,628
Scopus	52	24,545

Table 9. Resources of the knowledge base of REDI Ecuador.

Resource type	Instances
Affiliations	48
Researcher profiles	30,141
Publications	88,628
SCImago Journals	6,357
Latindex Journals	798
Other journals and conferences	1,051

4.1.3. Ontology population and enrichment. The ontology population process comprises two stages. First, in the disambiguation stage, similarity links among all the records (authors, publications, affiliations, etc.) are discovered, both between universities and within external sources (LD-Ps). Second, the information collected by from the LD-Ps and the universities is consolidated in a single graph, merging equivalent resources during the process. Finally, additional information about journals is collected from SCImago and Latindex and added. In Table 9, we present a summary of the resulting knowledge base of REDI Ecuador.

The final results of the aforementioned processes can be compared with respect to Scopus in Tables 6 and 8. Besides, REDI's approach of using Linked Data has multiple advantages in respect to the interoperability and reusability of the information. For instance, it has the capability of performing complex queries on its SPARQL endpoint (http://201.159.222.25:8180/sparql). However, Scopus has restrictive APIs and uses closed vocabularies which hampers integration and information retrieval tasks. These issues become noticeable even when executing simple tasks such as obtaining the number of researchers that belong to academic institutions of Ecuador.

4.1.4. Collaboration networks. In order to improve the knowledge base of REDI, and according to the aforementioned process, all the researchers of REDI Ecuador were classified in both knowledge areas and research areas. This procedure groups authors within the classifications of the UNESCO nomenclature and detects specific research topics with DBpedia concepts. These results are extremely useful for the stakeholders of REDI and allow researchers and organisations to navigate and interpret the research status in Ecuador. For example, it is possible to answer questions such as: what are the largest research areas in terms of the number of researchers?, which researchers work on a certain subject? or who researcher has most influence or collaboration with other researchers?

Table 10. Features of the REDI's classification.

Feature	Value
Knowledge areas detected	199 areas
Research areas detected	807 topics
Average size of a knowledge area	142 researchers
Classified researchers	28,323 researchers
Unclassified researchers	1,818 researchers

Table 11. Classification of authors by areas – REDI versus Scopus.

Repository	Biggest knowledge area	Biggest research area
REDI	Computer Sciences	Computer Network
Scopus	Computer Sciences	–

In Table 10, we present some details of the classification results. Notice that, some authors could not be classified due to the lack of keywords from their publications. This issue is caused because some academic sources do not provide this metadata.

Scopus, like REDI, has mechanisms for classifying authors within knowledge areas through its taxonomy that has a single level of granularity. However, REDI offers two levels of classification one based on the UNESCO nomenclature, and a second based on DBpedia concepts. However, as it can be seen in Table 11, these classification methods usually coincide on their results, for example, the knowledge area with the largest number of researchers is Computer Sciences in both systems. However, REDI can identify a second level of classification, in this case, the research topic Computer Network was found as the most common research area within Computer Science.

Another important feature that has to be compared between Scopus and REDI is the visualisation of collaboration networks. As shown in Figure 6, REDI presents collaboration networks as visual graphs that include components such as author profiles' photos, co-authorship linkage, and filters for knowledge areas and research areas. However, Scopus lacks such visualisations and only allows the exploration of clusters of authors through the use of links on the website.

All these resources are publicly available for the academic community benefit through the webpage of the project https://redi.cedia.edu.ec/. Furthermore, the compliance of semantic web standards (RDF/SPARQL) guarantees that developers and other researchers can easily access the REDI's knowledge base. So, REDI is a valuable service for students, researchers, universities and government agencies, which now have a centralised endpoint for accessing academic assets and research networking in Ecuador.

4.2. Use cases B and C

In this section, we discuss two additional implementations of the REDI framework within the Latin American context. Those implementations follow the same pattern as the Ecuadorian instance. Therefore, some technical details have been omitted for the sake of brevity.

First, the information of five member universities of KIMUK[8] in Costa Rica was centralised in an REDI instance as part of a concept-proof implementation. This node has approximately 20,000 authors' profiles with a similar number of publications, and represents a large proportion of scientific production of Costa Rica. The information of authors was extracted from the institutional repositories of the members of KIMUK. Specifically, the universities that were harvested are KERWA, TEC, UNA, UNED and UTN. Those researchers were processed in order to obtain their publications and they were classified in a similar manner to their Ecuadorian counterparts. Moreover, all the visualisations and search functionalities were implemented for this node, so it could be used for diverse purposes by the community. This implementation of REDI can be accessed on the web page https://rediclon.cedia.edu.ec/kimuk.

Second, an integration prototype was developed in order to unify both the REDI instances of Ecuador and Costa Rica into a single central repository. This prototype is a slightly modified version of REDI that adds multi-country support for visualisations and other data import utilities. The prototype combines the information from both Ecuador and Costa Rica in a single knowledge base. Therefore, it is possible to discover some interesting facts about collaboration networks in specific topics among universities from both countries. Similarly, this prototype can be found on https://rediclon.cedia.edu.ec/central.

5. Conclusion and future work

Scholarly data management and research networking are fairly complex tasks in academics. In addition, the evolution and growing of the Web have introduced new challenges and opportunities on the subject. Although numerous tools and research works have emerged in recent years, there are still plenty of limitations regarding integration, standardisation and completeness of the information.

This work presents a new perspective for managing academic information based on the construction of a knowledge base that facilitates the development of advanced functionalities such as the generation of research networks. The REDI framework implements an integration process based upon Linked Data principles that unify data from several information sources into a knowledge base. Moreover, that knowledge is analysed, enriched, enhanced and published on a single endpoint, so that the community can use it for diverse purposes. The approach followed by REDI has been tested on real scenarios for covering the scholarly information needs in countries like Ecuador and Costa Rica. Soon, it could even become an official reference database in Ecuador due to its complete catalogue of researchers.

In future work, we plan to expand the REDI scope to the Latin American ambit through the implementation of new instances in countries such as Colombia and Mexico. In this context, it is expected that REDI can scale up through national nodes, which satisfy the information needs of a particular country, and also can be accessed through a central instance. However, the integration of multiple instances of REDI into a centralised repository is an open problem that should be tackled shortly. Besides, we are exploring Graph Neural Networks (GNNs) as a means for improving the effectiveness of the collaboration network generation process. Finally, we aim to continuously improve internal processes and better exploit all the advantages of Linked Data.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This work was supported by RED CEDIA[9] in the project 'Ecuadorian Repository of Researchers'.

ORCID iD

José Ortiz Vivar https://orcid.org/0000-0002-2313-8526

Notes

1. Spanish acronym for Repositorio Ecuatoriano de Investigadores.
2. Open Archive Initiative – Protocol for Metadata Harvesting.
3. Extract, Transform and Load.
4. RDB to RDF Mapping Language.
5. Software or service for converting non-RDF data sources into one or more of the RDF data model serialisations.
6. Current Research Information System.
7. Vector Space Model.
8. Kímuk is the National Repository of open access scientific publications of Costa Rica maintained by CONARE.
9. https://www.cedia.edu.ec/

References

[1] Bizer C, Heath T and Berners-Lee T. Linked data – the story so far. *Int J Semant Web Inform Syst* 2009; 5(3): 1–22.
[2] Bäcker A, Pietsch C, Summann F et al. Base (bielefeld academic search engine): Eine suchmaschinenlösung zur indexierung wissenschaftlicher metadaten. *Datenbank-Spektrum* 2017; 17(1): 5–13.
[3] Knoth P and Zdrahal Z. Core: three access levels to underpin open access. *D-Lib Mag* 2012; 18(11/12), 4, http://oro.open.ac.uk/35755/
[4] Ammar W, Groeneveld D, Bhagavatula C et al. Construction of the literature graph in semantic scholar. In: Srinivas B, Jennifer C-C and Yunyao L (eds) *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: human language technologies*, vol. 3 (industry papers). New Orleans, LA: Association for Computational Linguistics, pp. 84–91.

[5] Chacón Candia FI. *Desarrollo de un repositorio de artículos científicos*. Master's Thesis, Department of Computer Sciences, University of Chile, Chile, 2012.

[6] Osborne F, Motta E and Mulholland P. Exploring scholarly data with rexplore. In: Alani H, Kagal L, Fokoue A et al. (eds) *The semantic web – ISWC 2013*. Berlin: Springer, pp. 460–477.

[7] Atanassova I and Bertin M. Faceted semantic search for scientific papers. In: *The semantic publishing challenge, task 3: in-use task, 11th European Semantic Web Conference (ESWC-2014)*, Crete, Greece, May 2014, pp. 1–5. Springer International Publishing.

[8] Fokou Pelap G, Faron Zucker C, Gandon F et al. Web semantic technologies in web based educational system integration. In: Escalona MJ, Domínguez Mayo F, Majchrzak TA et al. (eds) *Web information systems and technologies*. Cham: Springer International Publishing. pp. 170–194.

[9] Conlon M, Woods A, Triggs G et al. Vivo: a system for research discovery. *J Open Source Softw* 2019; 4(39): 1182.

[10] Patrício HS, Cordeiro MI and Ramos PN. Formalizing enrichment mechanisms for bibliographic ontologies in the semantic web. In: Garoufallou E, Sartori F, Siatri R et al. (eds) *Metadata and semantic research*. Cham: Springer International Publishing, pp. 147–158.

[11] Piedra N, Chicaiza J, Quichimbo P et al. Marco de Trabajo para la Integración de Recursos Digitales Basado en un Enfoque de Web Semántica. *RISTI – Revista Ibérica de Sistemas e Tecnologias de Informação*, 2015, pp. 55–70, http://www.scielo.mec.pt/scielo.php?script=sci_arttext&pid=S1646-98952015000100006&nrm=iso

[12] Segarra J, Ortiz J, Espinoza M et al. Integration of digital repositories through federated queries using semantic technologies. In: *2016 XLII Latin American computing conference (CLEI)*, Valparaiso, 10–14 October 2016, pp. 1–9. New York: IEEE.

[13] Suma F, Ortíz J, Segarra J et al. Integración de fuentes de datos bibliográficas utilizando tecnologías de linked data – caso de uso: Biblioteca de la Universidad de cuenca. *Maskana* 2017; 8: 189–203, https://publicaciones.ucuenca.edu.ec/ojs/index.php/maskana/article/view/1462

[14] Piedra N, Chicaiza J, Cadme E et al. Una aproximación basada en linked data para la detección de potenciales redes de colaboración científica a partir de la anotación semántica de producción científica: Piloto aplicado con producción científica de investigadores ecuatorianos. *Maskana* 2016; 5(Ed. Esp.), https://publicaciones.ucuenca.edu.ec/ojs/index.php/maskana/article/view/727

[15] Chicaiza J, Piedra N, Lopez-Vargas J et al. Discovery of potential collaboration networks from open knowledge sources. In: *2018 IEEE global engineering education conference (EDUCON)*, Tenerife, 17–20 April 2018, pp. 1320–1325. New York: IEEE.

[16] Sumba X, Segarra J, Ortiz J et al. Redi: a linked data-powered research networking platform. In: Gangemi A, Gentile AL, Nuzzolese AG et al. (eds) *The semantic web: ESWC 2018 satellite events*. Cham: Springer International Publishing, pp. 121–125.

[17] Saquicela V, Segarra J, Ortiz J et al. LOD-GF: an integral linked open data generation framework. *Adv Intell Syst* 2019; 884: 283–300.

[18] Bizer C and Schultz A. The R2R framework: publishing and discovering mappings on the web. In: Olaf H, Andreas H and Juan S (eds) *Proceedings of the first international conference on consuming linked data (COLD'10)*, vol. 665. Aachen: CEUR-WS.org, pp. 97–108.

[19] Sumba X, Sumba F, Tello A et al. Detecting similar areas of knowledge using semantic and data mining technologies. *Electron Notes Theor Comput Sci* 2016; 329: 149–167.

[20] Ortiz J, Segarra J, Sumba X et al. Authors semantic disambiguation on heterogeneous bibliographic sources. In: *2017 43rd Latin American computer conference, CLEI 2017*, Cordoba, Argentina, 4–8 September 2017, pp. 1–9. New York: IEEE.

[21] Scimago. SJR – SCImago Journal Country Rank [portal], n.d., http://www.scimagojr.com (accessed 8 September 2019).

[22] Ortiz J, Sumba X, Segarra J et al. Semantically identifying regional-indexed publications, a web-exploring approach. In: *2018 XLIV Latin American computer conference (CLEI)*, pp. 519–526, https://www2.scopus.com/inward/record.uri?eid=2-s2.0-85071088417&doi=10.1109%2fCLEI.2018.00068&partnerID=40&md5=f943364f3e2086a7ab5fd0070c978aee

[23] Pastor-Sánchez JA, Rodriguez-Muñoz JV and Lopez-Carreño R. Publicación como linked open data de la nomenclatura internacional de ciencia y tecnología y del tesauro unesco. In: *I Congresso ISKO Espanha e Portugal*, Oporto, 7–9 November 2013, pp. 1022–1043. Faculdade de Letras da Universidade do Porto.

[24] Auer S, Bizer C, Kobilarov G et al. Dbpedia: a nucleus for a web of open data. In: *Proceedings of the 6th international the semantic web and 2nd Asian conference on Asian semantic web conference (ISWC'07/ASWC'07)*. Berlin: Springer-Verlag, pp. 722–735, http://dl.acm.org/citation.cfm?id=1785162.1785216

[25] Daiber J, Jakob M, Hokamp C et al. Improving efficiency and accuracy in multilingual entity extraction. In: *Proceedings of the 9th international conference on semantic systems (ISEMANTICS)*, pp. 121–124, https://dl.acm.org/doi/10.1145/2506182.2506198

[26] Miles A and Bechhofer S. SKOS simple knowledge organization system reference. *Working draft, W3C*, 2008, http://www.w3.org/TR/skos-reference

[27] Segarra J, Sumba X, Ortiz J et al. Author-topic classification based on semantic knowledge. In: Villazón-Terrazas B and Hidalgo-Delgado Y (eds) *Knowledge graphs and semantic web*. Cham: Springer International Publishing, pp. 56–71.