Semantically Identifying Regional-Indexed Publications, a Web-Exploring Approach

José Ortiz*, Xavier Sumba†, José Segarra‡, and Víctor Saquicela‖
Departamento de Ciencias de la Computación, Universidad de Cuenca,
Cuenca, Ecuador.
{*jose.ortizv, †xavier.sumba93, ‡jose.segarra,
‖victor.saquicela}@ucuenca.edu.ec

***Abstract*—The indexing services are an important element of researching process because they make publicly available research results and articles for the community. Furthermore, regional indices such Latindex have contributed to spreading scientific works and incentivizing research in the Latin-American context. However, the uncentralized publication approach that its member journals follow has made impossible to form a unified view of Latindex-indexed articles and its corresponding journals. This drawback has limited activities such bibliometric studies and integration from the perspective of information systems. In this paper, a linking mechanism between journals and publications is outlined, which aims to identify explicitly whether or not an arbitrary article belongs to Latindex. The proposed approach leverages on the Linked Data principles and takes advantage of web search engines to validate its results. This proposal has successfully been evaluated on an Ecuadorian publications dataset obtaining a 0.91 f-score respect to a manually classified sample.**

***Index Terms*—Latindex, Linked Data, Web Correlation, Indexing Services**

I. Introducción

El nivel de producción científica y tecnológica de un país se evalúa generalmente a través del número de trabajos científicos que se producen y el nivel de impacto que estos presentan. Para encontrar el impacto de estos trabajos es común el uso de estudios bibliométricos que se encargan de obtener valores cuantitativos, los cuales son comparables en función de un criterio para la evaluación de un trabajo, revista o investigador. La evaluación de este tipo de trabajos son de especial importancia para el ámbito científico pues muchas veces son utilizados en labores de acreditación, aprobación de proyectos, asignación de fondos y en múltiples otros escenarios donde se requiere evaluar aspectos relacionados con la producción de investigación.

Muchas de las métricas empleadas en procesos de evaluación tienen un especial énfasis a la revista en la que el trabajo es publicado. Esto debido a que las revistas se encargan de validar los trabajos previo a su publicación, y por lo tanto, la rigidez científica con la que cuentan. A nivel internacional este criterio es ampliamente usado y aceptado principalmente por organismos de larga trayectoria en producción científica [1]. Un ejemplo de este factor es el *índice de impacto*, el cual provee de un medio para evaluar las revistas a través del número de citas y el número de artículos [2]. Sin embargo, a nivel latinoamericano y del Caribe este tipo de métricas no han sido favorables, frente a la realidad particular en la que se vive, dejando a los trabajos producidos en la región en segundo plano. Consientes de este hecho, en función de fortalecer los trabajos desarrollados a nivel regional nace Latindex como un proyecto conjunto de varias instituciones con el afán de elevar y destacar la producción científica producida en la región.

Latindex actualmente cuenta con gran presencia a nivel latinoamericano, del Caribe e incluso se han anexado a la propuesta países como España y Portugal, pues ofrece una alternativa válida para la diseminación de la producción científica. Para esto, Latindex cuenta con varios productos que incluyen bases de datos con información de las revistas y buscadores sobre este tipo de información [3]. Sin embargo a pesar de todo esto, Latindex actualmente cuenta con una gran problemática, pues aunque la determinación de si una revista es Latindex es relativamente sencilla (a partir de un título o ISSN[1]), el proceso de determinar si un trabajo que pertenece a un revista indexada en muchos casos no lo es.

El portal de portales[2] brinda acceso a publicaciones que pertenecen a Latindex, pero cuenta con un número de recursos limitado, que son insuficiente cuando se quiere identificar a todas las publicaciones que pertenecen a Latindex. Por otro lado, la labor de determinar la revista a la que pertenece una publicación no resulta sencilla puesto que la información con la que cuentan las publicaciones es mínima y está sujeta a ambigüedad.

Para mejorar el proceso de determinación de publicaciones pertenecientes a Latindex, se ha presentado una propuesta para abordar esta problemática en [4], en el cual se esboza un proceso para la identificación automática de revistas Latindex. En el presente trabajo se ha mejorado dicha propuesta para abordar más escenarios y conseguir determinar si una publicación con metadatos completos o parciales pertenecen o no a Latindex. Este enfoque ha demostrado proveer resultados satisfactorios que se han obtenido a partir de su evaluación con datos de publicaciones de autores ecuatorianos. Además, se hace público un conjunto de datos con información pertinente a revistas Latindex en una estructura de datos enlazados.

Las siguientes secciones de este trabajo han sido organizadas como sigue: en la sección II se provee información

[1]International Standard Serial Number
[2]http://www.latindex.ppl.unam.mx/

de antecedentes y trabajos relacionados; en la sección III se detalla el proceso planteado y su funcionamiento; en la sección IV se detalla el proceso que se ha seguido para validar la propuesta y los resultados obtenidos; finalmente, en la sección V se detalla las conclusiones de la propuesta y posibles trabajos futuros.

II. Antecedentes y trabajos relacionados

En el ámbito científico, día a día se produce una gran cantidad de documentación científica proveniente de investigaciones en diferentes áreas de conocimiento alrededor del mundo. Un inmenso esfuerzo que ha llevado al desarrollo tecnológico, social y económico que actualmente conocemos. Frente a esta situación los países de América Latina y el Caribe, no han querido quedarse indiferentes por lo que se ha trabajado en mejorar el aporte que se realiza a la comunidad científica desde su propia perspectiva. Este aporte aunque no es nuevo, ha mejorado a través del tiempo, no solo en función del número de trabajos científicos que se realizan en cada uno de los países, si no que se ha tratado de mejorar su calidad para poder encarar y estar a la altura de grandes potencias en esta área. Este proceso no ha sido inmediato y una de las iniciativas que ha actuado en función de recopilar y reflejar estos esfuerzos a nivel regional ha sido Latindex.

Latindex ha aparecido como un espacio para acoger el desarrollo científico producido en la región, centrándose en difundir, hacer accesible y elevar la calidad de las revistas académicas editadas dentro de este contexto. Este suceso ha sido de gran importancia para los países a los que se han acogido a esta propuesta porque la falta de estandarización sumada a otros factores como la falta de mecanismos para la difusión de este tipo información provocaba una limitada visibilidad e impacto tanto a nivel local como externo. Latindex nació como propuesta concebida en el año 1994, y es el resultado del esfuerzo aunado de varias instituciones interesadas en mejorar el manejo, difusión y visibilidad y de las revistas seriadas en América Latina, el Caribe, España y Portugal [5]. Actualmente forman parte de este proyecto la mayoría de países de América Latina que han visto una oportunidad para mejorar la calidad y visibilidad de su producción científica a nivel regional [6]. Latindex actualmente cuenta con varios productos, como son:

Catálogo: Contiene datos referentes a las distintas revistas registradas en Latindex, tanto en sus formatos digital y físico.

Directorio: Cuenta con información de un número selecto de revistas que cumplen con determinados criterios de calidad.

Revistas en línea: Proporciona acceso a textos completos de revistas digitales disponibles en línea y disponibles en hemerotecas digitales de América Latina, el Caribe, España y Portugal.

La relevancia que ha alcanzado Latindex y el gran aporte de sus criterios para la evaluación de las revistas ha servido como referencia para realizar un análisis del estado investigativo que viven las revistas de muchos países acogidos dentro de la cobertura de Latindex. Es así, por ejemplo, que se ha utilizado los criterios de Latindex para evaluar el estado de las revistas de ciencias sociales y humanidades [7] y para el ámbito de salud [8] en España. Por otro lado, en [9] se analizan las características editoriales de las revistas de Ecuador con respecto a los criterios Latindex. En el trabajo [10] se realiza un análisis para las revistas de ciencias sociales en Perú. En este último, se pone en evidencia las dificultades presentadas con la recopilación de datos desde Latindex al momento de manejar publicaciones, provocando que en muchos casos se tenga que realizar una gran labor manual. Para casos como estos–relacionados con la generación de estudios bibliométricos a partir de Latindex–el presente trabajo puede llegar a ser de gran utilidad.

Otro escenario donde la asociación de revistas con Latindex puede ser de utilidad es en el campo de desambiguación de autores. En este campo de estudio se investigan técnicas para identificar a autores plenamente y distinguirlos de otros que pueden llegar a estar asociados por ser homónimos o autores que disponen una representación heterogénea que al final resultan ser el mismo autor. En estos casos al igual que en el presente trabajo, se recopila la mayor cantidad de información de las fuentes disponibles siendo la más común la Web, permitiendo enriquecer la información existente y poseer mayores criterios de comparación [11]. También es común usar el criterio de correlación Web para validar dichos resultados [12]. Para casos como los mencionados, disponer de información de la revistas a las que pertenecen las publicaciones de un autor o incluso saber si el autor suele publicar en Latindex podría servir como un valioso recurso de información.

Por otro lado, la propuesta de datos enlazados ha emergido en los últimos años como un medio vanguardista en la representación y compartición de la información. Esta iniciativa está enfocada principalmente en que los datos de la Web puedan ser conectados y que dicha información pueda ser aprovechada tanto por las personas como por las máquinas [13]. Los datos bibliográficos también han adoptado este tipo de estándares por los beneficios que ofrecen en términos de visibilidad y estandarización. Es por esta razón que en el presente proyecto se ha decidido emplear este tipo de tecnología dentro de la propuesta facilitando de esta manera su reutilización en otros proyectos que también se basan en *Linked Data*. Para homogenizar el manejo de datos se ha empleado dicho medio de representación tanto para manejar los datos de las revistas extraídas de los repositorios Latindex, como los datos de las publicaciones utilizados el proceso de identificación expuesto a continuación.

III. Proceso de identificación y validación de revistas indexadas a través de la Web

El presente trabajo aborda el problema de la identificación de publicaciones indexadas desde una perspectiva de integración de la información, debido a que tanto los repositorios de publicaciones como catálogos de revistas indexadas actualmente proveen su información en la Web, pero de forma aislada. La falta de acuerdos respecto a estándares, o la ausencia de un repositorio central ha originado problemas

de heterogeneidad y falta de entrelazamiento explicito entre artículos y revistas desde la perspectiva de sistemas de información. Es por esta razón, que la presente propuesta intenta solventar esta necesidad, integrando estas fuentes a través de la utilización de recursos y tecnologías de la Web.

En la figura 1 se presentan los principales componentes desarrollados, los cuales se explican a más detalle en esta sección. El proceso propuesto se ha divido en dos fases con el fin de lidiar con las dos problemáticas mencionadas anteriormente. La primera fase consiste la homogenización y estandarización de la información, para esto se han seguido principios de *datos enlazados*. La segunda fase propuesta es el propio algoritmo de detección de publicaciones indexadas que aprovecha la información de la Web para crear los vínculos entre artículos y revistas. La ejecución de las dos fases no solo permite la identificación del servicio de indexación de una publicación, sino que además garantiza acceso homogéneo a la información.

III-A. Fuentes de información

En el contexto planteado, las fuentes de información a ser consideradas principalmente son los catálogos de revistas indexadas y repositorios de publicaciones. Estos dos tipos de fuentes de información deberán proveer listados de revistas y publicaciones respectivamente, los cuales incluyan los metadatos asociados con cada recurso. Adicionalmente, la inclusión de un tercer tipo de fuente se ha considerado para mejorar la calidad y cantidad de metadatos con los cuales el algoritmo de identificación opera. Específicamente se ha incluido información del registro de identificadores de colecciones ISSN, el cual es ampliamente usado para la identificación de revistas en diversos ámbitos.

En referencia a las fuentes de revistas indexadas se ha considerado al catálogo de Latindex como objetivo del proceso de identificación, este catálogo provee una lista de las revistas que han cumplido con los estándares de calidad establecidos por esta iniciativa. Revistas que cubren principalmente el ámbito latinoamericano y que contienen un gran parte de los resultados académicos y científicos desarrollado en la región. Latindex provee acceso programático a su catálogo a través del buscador de su página oficial[3], del cual se obtuvo un listado de 25180 revistas con sus correspondientes metadatos. Los campos de información provistos por este catalogo se presentan en la tabla I.

Como se puede notar, en la revista de ejemplo presentada en la tabla I, ciertos campos como ISSN-L[4] son opcionales, además la información obtenida del catalogo no esta necesariamente actualizada. Por esta razón en esta propuesta se adiciona nuevas fuentes de información como lo son ISSN.org[5] y ROAD[6], los cuales permitirán enriquecer las revistas con más metadatos. Para esto se utiliza como partida el ISSN de cada una de las revistas de catálogo Latindex para realizar consultas

[3]http://latindex.org/latindex/bAvanzada
[4]ISSN Linking
[5]The ISSN International Centre
[6]Directory of Open Access scholarly Resources

Cuadro I
METADATOS DE REVISTAS LATINDEX

Campo	**Ejemplo**
Título	Revista Politécnica (Quito)
Año de Inicio	1961
ISSN	1390-0129
ISSN-L	*No disponible
País	Ecuador
Editorial	Escuela Politécnica Nacional
Idioma	Español, inglés
Temas	Ciencias de la Ingeniería, Ciencias Exactas y Naturales
Folio	12020

sobre la API de ISSN.org[7] y sobre los datos de ROAD [8]. Los metadatos que los servicios antes mencionados adicionan a la proveniente del catálogo de Latindex, se presentan en la tabla II.

Cuadro II
METADATOS PROVENIENTES DE ISSN.ORG

Campo	**Ejemplo**
ISSN-L	2477-8990
Web	http://www.revistapolitecnica.epn.edu.ec/
Email	epnjournal@epn.edu.ec

Con el objetivo de mejorar la estandarización y visibilidad de los datos cosechados se aplicó un proceso de generación de *datos enlazados*. La información proveniente tanto del catálogo de Latindex como de los servicios de ISSN.org y ROAD es procesada y consolidada en un *dataset* único (grafo), usando el lenguaje de representación RDF[9]. Para el modelamiento de los datos se utilizó una ontología *Ad-Hoc* basada en URIs simples diseñadas a partir del dominio de la Universidad de Cuenca[10]. La información de publicaciones Latindex puede ser descargada[11] o accedida a partir de un SPARQL endpoint[12]. Un ejemplo del RDF resultante de "Revista Politécnica" se presenta en el segmento de código 1.

```
@prefix ns0: <http://www.ucuenca.edu.ec/ontology/>.
@prefix ns1: <http://www.ucuenca.edu.ec/resources/>.

ns1:12020 a ns0:Journal ;
  ns0:pais "Ecuador" ;
  ns0:folio "12020" ;
  ns0:nombre "Revista Polite'cnica (Quito)" ;
  ns0:temas "Ciencias de la ingenieri'a..." ;
  ns0:editorial "Escuela Polite'cnica Nacional" ;
  ns0:issn "1390-0129" ;
  ns0:issnl "2477-8990" ;
  ns0:email "epnjournal@epn.edu.ec" ;
  ns0:inicio "1961" ;
  ns0:folio "12020" .
```

Listing 1. Representación RDF-Turtle de una revista

[7]https://portal.issn.org/services
[8]http://road.issn.org/en/contenu/download-road-records
[9]Resource Description Framework
[10]http://www.ucuenca.edu.ec/
[11]https://goo.gl/vLTZVp
[12]https://redi.cedia.edu.ec/sparql/admin/squebi.html

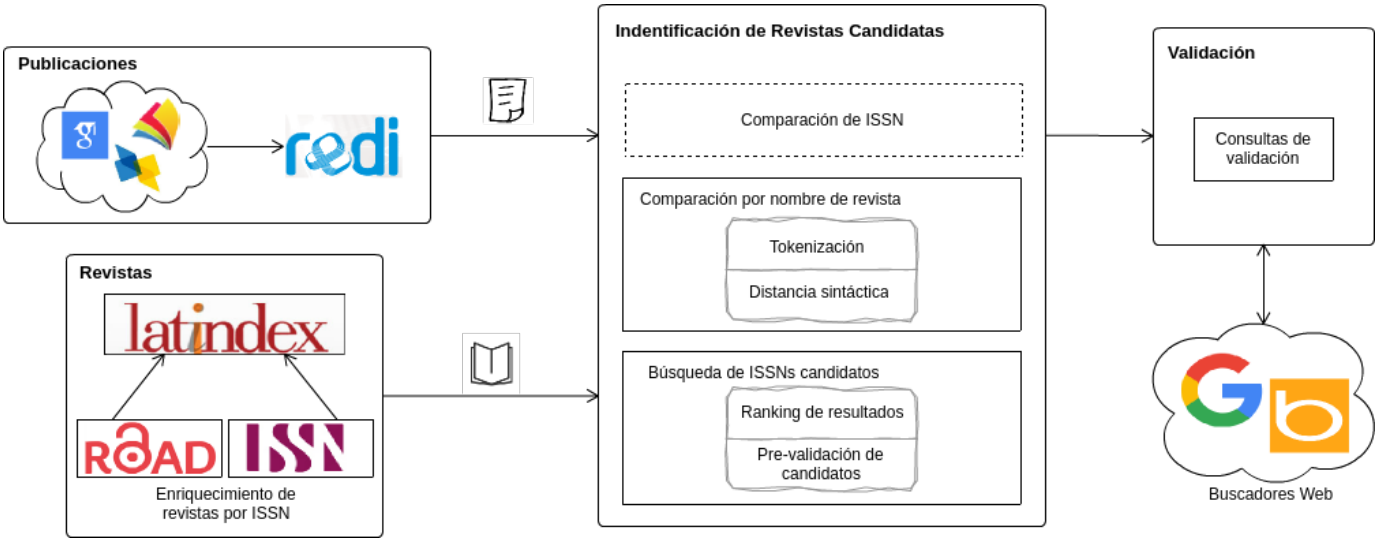

Figura 1. Proceso de identificación de publicaciones indexadas

Con respecto a la fuente de información encargada de retornar las publicaciones a ser identificadas se decidió reutilizar los datos disponibles en el *Repositorio Ecuatoriano de Investigadores - REDI* [14] con el fin de simplificar esta etapa. Este repositorio cosecha e integra publicaciones de autores ecuatorianos proveniente desde varios repositorios en línea como Google Scholar[13], Scopus[14], Scielo[15], etc. Adicionalmente, REDI[16] ofrece sus contenidos como *datos enlazados* a través de los estándares RDF/SPARQL[17] lo que elimina la necesidad de procesamiento adicional. Sin embargo, es necesario destacar que cualquier otra fuente de información podría ser utilizada en esta fase, por ejemplo, la información proveniente desde repositorios institucionales de universidades podría ser procesada y usada como origen de las publicaciones.

Los datos provenientes de REDI son una versión consolidada de todas las fuentes de información que integra, por tanto, el modelo de datos es común para todas las publicaciones. El modelo empleado en REDI está principalmente basado en las ontologías *BIBO* y *DCTERMS*. En la tabla III se presentan los principales metadatos provistos por REDI y que por tanto serán utilizados para la ejecución del algoritmo de identificación.

III-B. Algoritmo de identificación

El algoritmo de identificación utiliza los metadatos enriquecidos de publicaciones y revistas para encontrar enlaces explícitos entre estos recursos. El esquema propuesto sigue un patrón de reducción del campo de búsqueda (*blocking*) mediante la obtención de revistas candidatas para cada publicación y una etapa validación que filtra los resultados eliminando problemas de ambigüedad. La selección de revistas candidatas se consigue mediante una comparación de los atributos de las publicaciones con los metadatos de las revistas Latindex. Por otro lado, la validación de enlaces publicación-revista es realizada a través de búsquedas en la Web. Los detalles y razonamientos que guiaron a este esquema se presentan a continuación.

Cuadro III
METADATOS DE PUBLICACIONES DE REDI

Propiedad	**Ejemplo**
Titulo	Infraestructura basada en Globus Toolkit para dar soporte a repositorios..
Abstract	Grid computing is a recent and innovative technology...
Keywords*	Medical imaging, distributed systems, grid computing...
Nombre de la revista*	Maskana
ISSN*	No disponible

*Campos opcionales. Dependiendo de la fuente de datos esta información puede o no estar disponible.

III-B1. Identificación de revistas candidatas: El proceso de encontrar revistas potenciales en las que un artículo fue publicado parte de un análisis de los metadatos de las publicaciones debido a que los campos disponibles de dichos recursos son variables. Sin embargo, el objetivo general es encontrar similitud sintáctica entre los campos comunes de publicaciones y revistas. Es por esto que se establecieron tres posibles casos para encontrar revistas candidatas de una publicación en base a los metadatos disponibles.

Comparación por ISSN: Uno de los métodos más simples de referenciar unívocamente a una revista desde los metadatos de una publicación es la utilización de un código ISSN. La búsqueda directa del ISSN de una publicación dentro del catálogo Latindex en ciertos casos es suficiente para identificar a dicha publicación como indexada o no. Sin embargo, las fuentes de información de publicaciones no suelen proveer este metadato lo que reduce el número de publicaciones que pueden

[13]http://scholar.google.com/
[14]https://www.scopus.com/
[15]http://www.scielo.org/
[16]https://redi.cedia.edu.ec/
[17]https://www.w3.org/TR/rdf-sparql-query/

ser identificadas con este método. Además, la desactualización de ciertas revistas del catálogo Latindex y el uso de ISSNs alternativos (ISSN, ISSN-L) presenta una dificultad adicional.

La detección de revistas candidatas implementada para este caso utiliza una comparación exacta del campo ISSN de las publicaciones con sus pares en el catálogo Latindex. Es necesario recordar que el catalogo Latindex fue previamente enriquecido mediante información de ISSN.org y ROAD como se menciona en la sección III-A. Esto con el objetivo de evitar los problemas de desactualización del catálogo antes mencionado.

Debido a la naturaleza del identificador ISSN la lista de revistas candidatas retornada de este proceso es potencialmente una lista vacía o con un solo elemento. Esto implica que las publicaciones que posean información de ISSN puedan ser identificadas como indexadas o no de forma directa. Por esta razón, estas publicaciones no se procesan nuevamente a pesar de la posibilidad de que cuenten con otros metadatos como el nombre de revista.

Comparación por nombre de revista: Un metadato común en las publicaciones del repositorio REDI es el nombre de la revista. Este metadato puede ser fácilmente comparado con la información del catálogo Latindex con el objetivo de encontrar revistas candidatas que tengan nombre similar al registrado en las publicaciones. Sin embargo, hay que destacar que debido a la falta de estandarización en este campo es común encontrar grandes diferencias sintácticas en nombres que se refieren a la misma revista, por ejemplo es común encontrar publicaciones con revista "Revista Científica Maskana (Online)" o "Revista MASKANA" que hacen referencia al registro del catálogo Latindex de la revista "Maskana". Para proporcionar mayor flexibilidad en la generación del listado de revistas candidatas para una publicación a partir del nombre de una revista se aplicaron varios filtros de texto. Entre las principales transformaciones realizadas a los nombres de revistas se tiene, la eliminación de textos encerrados en paréntesis, por ejemplo "Maskana (Online)" es transformada en "Maskana". Adicionalmente, mediante una revisión manual se identificó una lista de palabras vacías (*stopwords*) como "revista" que pueden introducir ambigüedad en la búsqueda, por ejemplo "Revista Estoa" es transfomada en "Estoa".

Respecto a la comparación de los nombres de revista filtrados, se aplicó una comparación basada en *tokens* que mejore el emparejamiento de nombres de revistas potencialmente equivalentes. El algoritmo especifico que se utilizó para esto es *Jaccard* [15], el cual fue seleccionado tras un análisis empírico con varias revistas de ejemplo. La implementación de este tipo de comparación facilita que nombres de revistas diferentes desde un punto de vista estrictamente sintáctico como "Maskana. Edición especial" y "Maskana" pueda emparejarse.

El resultado final de esta estrategia de extracción es un listado de revistas candidatas del catálogo Latindex que compartan similitud con el nombre de la revista especificado en la publicación. Al tratarse de una comparación de texto no se puede garantizar la exactitud de los emparejamientos publicación-revista por lo que una etapa de validación es requerida, la cual es detallada en secciones siguientes. Sin embargo, la aplicación de esta comparación de nombre de revista reduce el campo de búsqueda enormemente.

Obtención de candidatas por exploración Web: La detección de revistas candidatas para una publicación trae consigo un grado adicional de dificultad cuando no se posee ningún metadato que permita encontrar algún tipo de similitud con el catalogo Latindex. Este caso es común en el repositorio REDI debido a que buena parte de las fuentes de información de donde los artículos son extraídos proporcionan únicamente información básica como: autores, título, abstract y palabras clave. Para estos casos, es necesario implementar mecanismo de búsqueda indirecta que permitan encontrar posibles revistas relacionadas con el artículo.

En trabajos anteriores se exploró la idea de utilizar métricas semánticas [4] para comparar las palabras clave de los artículos con los temas de las revistas y así generar una lista de revistas candidatas. Sin embargo, dicha estrategia tiene serias limitaciones con revistas de ámbito general como la revista Maskana, que aborda temas tan amplios como *"Arte, Tecnología, Ciencias Sociales..."*. Estos temas difícilmente pueden ser relacionados con las palabras clave específicas de un artículo como *"DICOM, Semantic Web, "*.

En este contexto el presente trabajo presenta un nuevo enfoque para lidiar con este tipo de publicaciones. La estrategia ideada, intenta identificar las páginas Web descriptivas de los artículos a través de búsquedas Web, estrategia que es también usada en la validación de revistas candidatas y que se describe a más detalle en la siguiente sección. El objetivo de la búsqueda es analizar las páginas descriptivas de los artículos para encontrar potenciales identificadores ISSN que son usados para detectar revistas Latindex candidatas.

El método de búsqueda realiza una consulta exacta con el título y resúmen de la publicación siendo analizada, usando motores de búsqueda como *Google Search* o *Bing*. Los enlaces encontrados a su vez son consultados (HTML[18]) y analizados para encontrar patrones numéricos similares a un ISSN tanto en el cuerpo como en los metadatos de la página. Los resultados son filtrados comprobando que los potenciales ISSNs pertenezcan a Latindex. Adicionalmente, una variante inversa de *TF-IDF* es aplicada con el objetivo de eliminar la influencia de páginas Web que recompilen listados de publicaciones o secciones de referencias que pueden incluir ruido. Esto por cuanto se busca priorizar los ISSNs encontrados en páginas con un número de resultados potenciales reducido y que a su vez estén presentes en múltiples páginas diferentes.

$$w_i = \frac{1}{N} \sum_{j}^{N} \frac{tf_{i,j} \cdot df_i}{nt_j} \tag{1}$$

$tf_{i,j}$: Número de apariciones del ISSN i en la página j .
df_i: Número de páginas que contienen el ISSN i .
nt_j: Número de apariciones de ISSNs en la página j.
N: Número total de páginas.

[18]HyperText Markup Language

En la formula 1 se presenta el cálculo del peso w_i de un ISSN candidato (i) respecto a los demás resultados. Una vez estimado el peso de cada ISSN candidato se procede a filtrarlos a través de un umbral obtenido empíricamente. De está forma se obtiene los ISSNs con mayor aparición en una búsqueda Web exacta de términos. Entonces, el resultado de esta etapa es un conjunto de revistas candidatas Latindex obtenidas a partir de los ISSNs que pasen el filtro.

III-B2. Validación: Una vez realizada la identificación de revistas candidatas para las publicaciones se requiere de un método efectivo para comprobar si los pares publicación-revista son realmente válidos. El enfoque propuesto en este trabajo para dicha tarea es la comprobación a través de búsquedas Web de páginas que contengan tanto lo metadatos de las publicaciones así como de la revista candidata. El razonamiento tras este planteamiento es que la gran mayoría de revistas indexadas en Latindex poseen una página Web donde se publican fichas informativas de sus artículos (o texto completo) o bien publican sus contenidos dentro de repositorios de libre acceso. Esto debido a que dentro de los lineamientos para formar parte de Latindex se establece el requerimiento de contar con servicios de información virtuales y estar registrado en servicios de búsqueda[19].

Para ejecutar la validación propuesta se generan un conjunto de consultas exactas con los metadatos de las publicaciones y las revistas. Específicamente se toma el título y resumen de las publicaciones, mientras que de las revistas se toma su ISSN y nombre. Previamente a la validación se ejecuta un último filtro, cuando el metadato de fecha de publicación está disponible en el artículo científico debe estar en rango de fechas de las ediciones de una revista candidata. Es decir, se filtran las revistas candidatas comprobando que la fecha de el artículo científico sea posterior a la fecha de inicio de la revista candidata registrada en Latindex. En caso de no cumplir está restricción las revistas son filtradas dado que no se puede cumplir esa condición en un escenario real.

En la tabla IV se presentan las combinaciones de consultas que son ejecutadas. Si estas consultas retornan resultados dentro de buscadores Web como Google o Bing se asume que el par publicación-revista es válido y la publicación es presentada como Latindex, si por el contrario ningún par produce resultados se asume que la publicación no es Latindex. Finalmente, las publicaciones validadas son enlazadas a través de enlaces apropiados a nivel de datos mediante las propiedades *isPartOf* de DCTERMS o *sameAs* de OWL. En otras palabras, se define si una publicación pertenece o no a Latindex y además se asigna dicha revista a los metadatos de la publicación.

IV. Evaluación y resultados

Para probar la validez de la propuesta se ha ejecutado el proceso anteriormente descrito con los datos de investigadores y sus publicaciones disponibles en el proyecto REDI. De esta manera se pretende determinar el porcentaje de aportes que los investigadores ecuatorianos realizan dentro de Latindex. REDI actualmente dispone de datos de investigadores de más de 27 instituciones académicas del país, de los cuales se han recopilado alrededor de 4000 trabajos investigativos en total a través de varias fuentes como Google Schoolar, Scopus, Microsoft Academics, Scielo entre otras. Para evaluar los resultados obtenidos de este proceso se empleó como criterio de comparación un *Gold Standard*. Dicho *Gold Standard* se realizo mediante un grupo de personas a los cuales se les proporciono una muestra de 25 % de las publicaciones disponibles en el repositorio REDI. Los expertos se dieron la tarea de determinar manualmente y con la ayuda de un buscador Web, en primer lugar, la revista a la que pertenecía dichas publicaciones para posteriormente determinar si esta pertenece al catalogo Latindex. Los datos con los que contaban los expertos consistían en el *título de la publicación, autores e ISSN* en caso de estar disponible. Los resultados obtenidos por los expertos a través del *Gold Standard* determinaron que el 33.2 % pertenecían a Latindex tal como se presenta en la tabla V. Cabe recalcar que si a una publicación no pudo determinarse su revista, se considero como que no pertenecía a Latindex.

Cuadro IV
Búsquedas de validación

Combinación	**Ejemplo**
Título, resumen, ISSN	"Infraestructura basada en Globus Toolkit ..." "Grid computing is a recent and innovative..." "1390-6143"
Título, resumen, nombre de la revista	"Infraestructura basada en Globus Toolkit ..." "Grid computing is a recent and innovative..." "Maskana"

Cuadro V
Resultados del Gold Standard

Gold Standard		
Num. Latindex	Num. No Latindex	Total
332	668	1000

Con el *Gold Standard* mencionado se realizó una comparación respecto a los resultados obtenidos con la propuesta. Esta comparativa se resumen en la tabla VI donde se identifican los valores para verdaderos positivos (*VP*), falsos positivos (*FP*), falsos negativos (*FN*) y verdaderos negativos (*VN*) .

Cuadro VI
Resultados obtenidos del algoritmo en comparación con Gold Standard

	Gold Standard	
Algoritmo	Condición Verdadera	Condición Falsa
	VP: 286	FN: 37
	FP: 18	VN: 659

Los datos recopilados sirvieron para el cálculo de *precisión* y *exahustividad* [16], tal como se presenta en la ecuación 2 y

[19]http://www.latindex.org/latindex/revistaselec

3. Finalmente con los valores anteriores, se calcula el *valor-F* en la ecuación 4.

$$Precisión = \frac{VP}{VP + FP} = \frac{286}{286 + 18} = 0{,}94 \quad (2)$$

$$Exahustividad = \frac{VP}{VP + FN} = \frac{286}{286 + 37} = 0{,}88 \quad (3)$$

$$valor - f = 2 * \frac{precisión * exahustividad}{precisión + exahustividad}$$

$$= 2 * \frac{0{,}94 * 0{,}88}{0{,}94 + 0{,}88} = 0{,}91 \quad (4)$$

Los resultados obtenidos demuestran un alto nivel de precisión alrededor del 94 % de la muestra, por lo que se puede concluir que la mayoría de recursos que son Latindex son identificados como tal. En el caso de la exahustividad se dispone de un valor ligeramente menor llegando a ser del 88 %. Era predecible obtener un menor nivel de exahustividad en la evaluación debido a que muchas veces las publicaciones carecen de la información necesaria para identificar la revista a la que pertenecen y por lo tanto identificarlos como publicaciones que pertenecen a revistas Latindex.

Finalmente, podemos concluir que la propuesta de identificación tiene un número reducido de errores cuando clasifica una publicación como Latindex (precisión), pero que sin embargo tiende a ignorar con más frecuencia otras publicaciones que también son Latindex (exhaustividad). El problema de exhaustividad en la detección se debe principalmente a la falta de metadatos de ciertas publicaciones, problema que es difícil de solucionar inclusive para un humano.

V. Conclusión y trabajos futuros

El esquema de publicación de artículos que sigue una gran parte de las revistas pertenecientes a catálogos o servicios de indexación regionales como Latindex trae consigo serias limitaciones en cuanto a estandarización e integración de la información. La falta de un repositorio central que unifique explícitamente los artículos publicados y el catálogo de revistas dificultan la ejecución de estudios bibliométricos y difusión efectiva de la información. En este contexto surge la necesidad de mecanismos de integración flexibles que permitan obtener enlaces explícitos entre los artículos publicados en revistas y el catalogo Latindex, considerando la naturaleza descentralizada de las fuentes.

En este trabajo se presenta un enfoque de identificación de revistas indexadas basado en la Web acorde con las problemáticas presentadas. La propuesta parte por unificar e enriquecer varias fuentes de información bibliográfica en línea para obtener los metadatos de los recursos que necesitan ser enlazados. La información obtenida es procesada mediante varias métricas sintácticas para encontrar potenciales enlaces entre publicaciones y revistas. Finalmente, los potenciales enlaces encontrados son validados usando búsquedas en la Web que tienen por propósito inferir relaciones de pertenecía ente publicaciones y revistas a través de páginas descriptivas o textos completos disponibles en la Web.

Con el objetivo de cuantificar el potencial de la propuesta se realizó una evaluación en base a un Gold standard anotado por expertos. Los algoritmos fueron probados sobre una muestra de mil publicaciones disponibles del proyecto REDI de Ecuador. Los resultados obtenidos demostraron la validez del enfoque planteado especialmente en términos de precisión.

El trabajo futuro se enfocará en optimizar la presente propuesta para su aplicación a una mayor escala en términos de número de publicaciones a ser identificadas. Esta mejora estará enfocada en cubrir fuentes de publicaciones Latinoamericanas que requieran la identificación de artículos indexados en Latindex. Adicionalmente, se desarrollaran mecanismos más elaborados para la extracción de revistas candidatas a partir de la exploración en la Web con el objetivo de reducir el número de errores.

Agradecimientos

Al Departamento de Ciencias de la Computación de la Universidad de Cuenca. Adicionalmente, a la Corporación Ecuatoriana para el Desarrollo de la Investigación y la Academia (RED-CEDIA), por el financiamiento brindado a esta investigación mediante el proyecto "Repositorio Semántico de Investigadores del Ecuador".

Referencias

[1] L. Feetham, "Can you measure the impact of your research?" *Veterinary Record*, vol. 176, no. 21, pp. 542–543, 2015. [Online]. Available: http://veterinaryrecord.bmj.com/content/176/21/542

[2] E. Garfield, "Journal impact factor: a brief review," *CMAJ*, vol. 161, no. 8, pp. 979–980, 1999. [Online]. Available: http://www.cmaj.ca/content/161/8/979

[3] J. Alonso Gamboa, "Bases de datos y calidad de las revistas científicas: la aportación de latindex," *Espacio I+ D Innovación y Desarrollo*, vol. 6, no. 13, 2017. [Online]. Available: http://eprints.rclis.org/31531/

[4] J. Cullcay, J. Ortiz, X. Sumba, F. Sumba, and V. Saquicela, "Identificación automática de artículos indexados en latindex," *Maskana*, vol. 8, pp. 103–111, 2017.

[5] A. Cetto, "Ciencia y producción científica en américa latina. el proyecto latindex." *International Microbiology*, vol. 1, no. 3, pp. 181–182, 1998.

[6] M. C. Ratto and A. B. Dellamea, "Difusión acceso y visibilidad de publicaciones científicas seriadas de iberoamérica. el sistema latindex," *Dominguezia*, vol. 17, no. 1, pp. 51–57, 2001.

[7] A. R. Román, M. V. Valero, and C. U. Caminos, "Los criterios de calidad editorial latindex en el marco de la evaluación de las revistas españolas de humanidades y ciencias sociales," *Revista española de documentación científica*, vol. 25, no. 3, pp. 286–307, 2002.

[8] M. V. Valero, C. U. Caminos, and A. R. Román, "Las revistas españolas de ciencias de la salud frente a los criterios de calidad editorial latindex," *Revista española de documentación científica*, vol. 26, no. 4, pp. 418–432, 2003.

[9] V. F. Andrade, A. V. G. Sierra, and A. R. G. García, "Características editoriales de las revistas electrónicas ecuatorianas indexadas en catálogo de latindex," *Revista Publicando*, vol. 4, no. 10 (1), pp. 118–130, 2017.

[10] L. F. M. Morante, "Producción e impacto de las revistas peruanas del ámbito de las ciencias sociales en el catálogo latindex," *Investigación Bibliotecológica: archivonomía, bibliotecología e información*, vol. 30, no. 69, pp. 179–204, 2016.

[11] D. A. Pereira, B. Ribeiro-Neto, N. Ziviani, A. H. Laender, M. A. Gonçalves, and A. A. Ferreira, "Using web information for author name disambiguation," in *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2009, pp. 49–58.

[12] K.-H. Yang, H.-T. Peng, J.-Y. Jiang, H.-M. Lee, and J.-M. Ho, "Author name disambiguation for citations using topic and web correlation," in *International Conference on Theory and Practice of Digital Libraries*. Springer, 2008, pp. 185–196.

[13] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *International journal on semantic web and information systems*, vol. 5, no. 3, pp. 1–22, 2009.

[14] X. Sumba, F. Sumba, A. Tello, F. Baculima, M. Espinoza, and V. Saquicela, "Detecting similar areas of knowledge using semantic and data mining technologies," *Electronic Notes in Theoretical Computer Science*, vol. 329, pp. 149–167, 2016.

[15] S. Niwattanakul, J. Singthongchai, E. Naenudorn, and S. Wanapu, "Using of jaccard coefficient for keywords similarity," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, no. 6, 2013.

[16] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011.