ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2023.1.1.1206

Diagnosis of the data used to measure scientific production. Case Study: CvLAC (Colombia)

Juan-Sebastián González-Sanabria, M. Sc. [1], Alexander Castro-Romero, M. Sc. [2], Esteban Novoa-Quiñones, Eng. (c) [3] and Samuel-Felipe Ruiz Duarte, Eng. (c) [4]
[1,2,,3,4] Universidad Pedagógica y Tecnológica de Colombia, Colombia, juansebastian.gonzalez@uptc.edu.co, alexander.castro01@uptc.edu.co, esteban.novoa@uptc.edu.co, samuel.ruiza@uptc.edu.co

Abstract–Each country uses different technological mechanisms to inventory its investigative capacities, however, in a large part of Latin America, platforms that are not very interoperable are used and that require the information to be entered manually by researchers, either due to bad practices or due to involuntary errors, they make continuous failures that affect having a real knowledge of the scientific production of the countries. For the case study, the extraction of information was carried out through an own development for the CvLAC platform of the Ministry of Science of Colombia, in order to obtain the data reported in the profiles of researchers. After the extraction, the data obtained was analyzed, where a significant flaw in the data recording process is evident, obtaining that 65.3% of the records studied present errors in the reported data. Due to the above, the need for an interoperable platform is evident, which is integrated with systems such as ORCID or CrossRef to automatically select or associate the production of researchers directly from publishers.

Keywords--data quality; science measurement; research platforms: institutional repositories.

I. INTRODUCCIÓN

La representación y difusión del conocimiento es una de las tareas de continuo interés de la comunidad científica, en este sentido, los investigadores se valen de diferentes mecanismos como la realización de ponencias o trabajo social, la publicación de artículos científicos en revistas especializadas, entre otros, siendo este último medio, uno de los que más tiene impacto en la comunidad, pues se suelen usar mecanismos de validación como revisión por pares, procesos de correcciones de estilo y realimentaciones, que hacen que los conocimientos que allí se presenten tengan mejor receptividad por la sociedad en general [1].

Es por lo anterior, que la mayoría de los sistemas de ciencia y tecnología de los diferentes países usen una serie de métricas, en gran parte asociada a la producción de artículos científicos, para la generación de sus indicadores de investigación [2], tal es el caso del Ministerio de Ciencia Tecnología e Innovación - Minciencias, en Colombia, que mediante su "Modelo de Medición de Grupos de Investigación, Desarrollo Tecnológico o de Innovación y de Reconocimiento de Investigadores del Sistema Nacional de Ciencia, Tecnología e Innovación" incluye los mismos como uno de los factores esenciales para la categorización (ranqueo) de los investigadores y de los grupos de investigación, así como para presentar las estadísticas y capacidad investigativa del país.

Profundizando en el caso colombiano, la medición se hace a través de la información reportada por los investigadores en su hoja de vida en la plataforma CvLAC (Currículum Vitae para Latinoamérica y el Caribe), en dicho sistema cada persona registra la información de los productos generados, para el caso de los artículos la información corresponde principalmente al título, revista en la que se publicó, autores, tipo (impreso o electrónico) volumen y numero, DOI, entre otros; sin embargo, no es posible validar de manera automática la información allí reportada, particularmente porque esta plataforma actúa de manera independiente y no permite la recuperación de la información, puesto que no cumple con estándares de interoperabilidad existentes para el manejo de la información en la Web.

Evidencia de lo anterior es que en esta era de datos enlazados (principio básico de la Web) en el currículo de un investigador no se puede usar enlaces y mucho menos compartir o exportar la información registrada. Adicionalmente, es importante mencionar que es desgastante para un investigador registrar manualmente sus productos en plataformas donde no hay coherencia en los datos y formatos solicitados, y también difícil para las instituciones verificar la veracidad de la información consignada, dado que, si bien se establece un procedimiento de verificación a través del aplicativo InstituLAC, allí solo se puede validar para el caso de los artículos el título del artículo, el volumen, el número, las páginas y el año de publicación.

Por lo anteriormente planteado, es notorio que esto contrae problemas tanto para investigadores como para instituciones, e incluso para el gobierno y los indicadores presentados, pues al querer presentar y soportar cifras reales de la productividad científica (bibliometría) del país y generar estrategias para la visibilidad de esta, no se cuenta con los mecanismos y fuentes necesarias.

Para validar la calidad de los datos reportados en CvLAC, en el presente trabajo se realiza un análisis de la información registrada por los investigadores de una facultad de una universidad, con el fin de determinar su veracidad e integridad. Enfatizando en que, si se quieren obtener datos más aproximados a la realidad, organismos como Minciencias deben establecer estrategias de datos abiertos e interoperables en sus plataformas, dado que estas no cumplen con estándares de calidad de datos en tres frentes: de indización Web, datos abiertos y formatos de información científica.

Con el fin de determinar las falencias es necesario conocer aspectos relacionados tanto con la producción científica, especialmente de artículos científicos, y lo relacionado con las fallas a nivel de bases de datos, los cuales se presentan a continuación.

A. *Identificadores Persistentes Digitales*

Organizaciones diseñaron estándares de identificación única, abarcando diferentes características. Los identificadores Persistentes se dividen en: identificadores de autores o colaboradores, identificadores de objetos e identificadores de organizaciones. Para el alcance de este trabajo, se hablará principalmente de los dos primeros.

1) Identificador de autores. La variación del nombre de un mismo autor en trabajos o citaciones debido a factores como el contexto cultural, error de escritura, omisión de partes o reglas de publicación, propende a problemas de ambigüedad, donde se puede adjudicar la autoría de un trabajo a la persona incorrecta. Esto conlleva a disminuir la visibilidad de trabajos de los autores que tengan un nombre similar, afectando los resultados de indicadores a nivel de autor y revista, junto con la pérdida de transparencia [3].

La identificación apropiada de autores y de sus trabajos es crucial para todos los involucrados en el entorno de la comunicación científica. Es por esto que se busca identificar a investigadores, autores y científicos, evitando la ambigüedad entre colaboradores y la producción asociada a estos [4]. Existen identificadores con alcances locales que se centran en campos específicos a nivel de país, institución o área de estudio. Sin embargo, se vuelve más necesaria una solución con un enfoque global que permita establecer una relación autor-producto-institución.

Pese a que existen diversas iniciativas como ORCID, ScopusID, ResearcherID, se encuentra que la primera es la más utilizada. El Open Research and Contributor Identifier (ORCID) resuelve el problema de ambigüedad de nombres de manera más precisa, así mismo, permite acceder se encuentra la hoja de vida del autor, reuniendo todos los trabajos, resultados de investigación e instituciones a las que ha estado vinculado. Asimismo, brinda garantía de la atribución correcta de trabajos, interoperabilidad y enlace con identificadores externos, simplificación de las búsquedas y la globalidad del identificador [5].

2) Identificador de objetos. Los objetos pueden corresponder a libros, artículos, borradores, capítulos, conjuntos de datos, figuras, videos o cualquier dato significativo. Los identificadores de objetos son URL's persistentes en el tiempo que direccionan a servicios de resolución, los cuales mantienen información sobre la ubicación actual del recurso; permitiendo así, mantener el acceso y trazabilidad de los objetos, independientemente a su ubicación.

Respecto a los identificadores de objetos, el Digital Object Identifier (DOI) es mayormente usado para identificar artículos científicos, revistas o cualquier tipo de objeto digital. Consiste en una URL que indica el servicio que resolverá el DOI, un prefijo que identifica a una institución y un sufijo que identifica el objeto digital, cuya estructura es definida libremente por los editores. El uso del DOI garantiza el acceso a los recursos y conlleva a un aumento en la visibilidad e impacto de una publicación científica. Además, es interoperable con otras plataformas o repositorios de contenido [6].

Digital Object Identifier: (only for full papers, inserted by LACCEI).
ISSN, ISBN: (to be inserted by LACCEI).
DO NOT REMOVE

B. *Cosechadores de metadatos*

El ecosistema de investigación académica tiene principalmente dos tipos de participantes [7-8]: i) los proveedores de datos: incluyen los trabajos realizados para su visualización y descarga para el lector y sus metadatos para ser distribuidos a través de los canales de comunicación entre sistemas, y, ii) los proveedores de servicio: encargados de cosechar los metadatos y ofrecer interfaces de integración y búsqueda. Asegurando su uso, los proveedores de datos ceden la gestión de sus metadatos a las agencias especializadas que centralizan los registros y proporcionan una infraestructura para su recuperación a través de múltiples fuentes y dominios [9].

Agencias, como Crossref y DataCite, recopilan la información de las instituciones adscritas para ofrecer un valor agregado a los datos para cualquier interesado. Lo que permite la recolección de datos para la construcción de un dataset de un proyecto de investigación, su uso en la generación de métricas por parte de la revista evaluada, entre otros. Generalmente los cosechadores establecen diferentes protocolos de comunicación (OAI-PMH) y estándares de metadatos (Dublin Core).

Sin embargo, una de las características en común entre los cosechadores es la recuperación de los metadatos, esto corresponde a información de trabajos, autores y entidades, permitiendo hacer búsquedas especializadas por identificadores que estén indexados sobre el recurso. Así mismo, el registro del contenido, por el cual las agencias permiten la serialización de un recurso digital para ser identificado a través de la red, ejemplo de esto es el DOI, que es registrado por la agencia para ser usado al momento de la publicación de un trabajo en una revista y así tener una trazabilidad a través de las citaciones en artículos, mención en redes sociales y aparición en blogs [10].

II. METADATOS EN LA CIENCIA

Bajo el protocolo OAI, para la recolección de metadatos, las herramientas de gestión editorial exponen sus metadatos a disposición de servicios bajo el estándar HTTP (Hypertext Transport Protocol), encargado de la comunicación con el repositorio y XML (Extensible Markup Language), como estándar para la codificación de documentos que envíen como respuesta. La petición se realiza con unos argumentos propios del estándar OAI, habilitando una lista de peticiones con argumentos específicos, dependiendo del recurso que se desea recuperar, junto con un modelo estructurado para proporcionar un nivel básico de interoperabilidad.

Un modelo bastante usado para describir los recursos, y facilitar la recuperación de sus metadatos, es el estándar Dublin Core. En la Tabla I se visualizan los atributos que

describen a los recursos, y un ejemplo de los metadatos que contienen; ninguno de los atributos es obligatorio, pero se recomienda su completitud para asegurar su correcta identificación.

TABLA I
EJEMPLIFICACIÓN DE LOS ELEMENTOS DENTRO DE DUBLIN CORE. ADAPTADO DE [11]

Campo	Descripción	Ejemplo
Title	Nombre asignado al trabajo	Modeling and simulation of a pervaporator coupled to a simultaneous saccharification-fermentation process for the ethanol production
Author or Creator	Persona u organización principalmente responsable de la creación del trabajo	Cubillos-Lobo, Jairo Antonio. Bustamante-Londoño, Felipe Acosta-Cárdenas, Alejandro
Subject and Keywords	Palabras clave o frases que describen el trabajo	bioetanol, pervaporation, silicalite
Description	Descripción textual del contenido del recurso; incluye resúmenes o descripciones de contenido	Process integration is now days considered a viable option for reducing ethanol production costs from biomass
Publisher	Editorial responsable que el recurso esté disponible	Universidad Pedagógica y Tecnológica de Colombia
Date	Fecha asociada a la creación o disponibilidad del recurso con la norma ISO 8601	2015-09-10
Resource Type	Categoría del recurso	article
Format	El formato del recurso para identificar el software necesario para visualizar o hacer funcionar	application/pdf text/html
Resource Identifier	Cadena o número utilizado para identificar de forma exclusiva del recurso	10.19053/01211129.3848
Source	Información sobre un segundo recurso del que se deriva el presente recurso	Revista Facultad de Ingeniería; Vol 24 No 40 (2015); 49-66
Language	La lengua del contenido intelectual del recurso	spa (Español)
Relation	Este elemento se utiliza para expresar vínculos entre recursos relacionados	/*ref*/R. Wooley et al., "Process design and costing of bioethanol technology: a tool for determining the status and direction of research and development". Biotechnol. Prog Vol. 15, pp. 794-803, 1999.

II. METODOLOGÍA

Para analizar la calidad del manejo de la información científica que se realiza en la plataforma CvLAC es necesario un proceso de Web Scraping, es decir extraer la información directamente de cada página HTML (Lenguaje de Marcas de Hipertexto) de los investigadores, esto debido a que la organización no cuenta con módulos de consulta o reportes, ni un API (Interfaz de Programación de Aplicaciones) pública que pueda ser consultada y de la que se pueda extraer la información para ser analizada de mejor manera. Para ello se definieron las siguientes etapas:

1) Identificación y caracterización de fuente de datos. Inicialmente se analizó el archivo HTML de cada perfil de investigador, encontrándose que el sitio Web no sigue ningún estándar de metadatos, lo cual dificultó la tarea de extracción de la información.

Dentro del perfil de investigador se tienen las siguientes secciones: Datos generales, Actividades formación, Actividades evaluador, Apropiación social, Producción bibliográfica, Producción Técnica, Más información y Producción en arte. Estas secciones no están identificadas claramente dentro del HTML, la única forma de identificar donde inician y terminan es por comentarios dentro del código de la página. De estas la idea es centrarse en Datos generales y del menú de producción bibliográfica en la opción de artículos científicos.

Ahora bien, la información de los artículos se muestra en una tabla de HTML, donde cada artículo es una fila, solo hay una columna que es una etiqueta tipo "blockquote" donde la información del artículo se guarda en texto plano. La única ventaja es que se tiene un orden en los datos y estos están separados en ocasiones por caracteres como puntos y comas. Los datos no siempre están completos y se detectan bastantes problemas con la calidad de estos, pero en general se pudieron recuperar datos como: autores, título del artículo, país, nombre de la revista, ISSN (Número Internacional Normalizado de Publicaciones Seriadas), volumen, numero, páginas y DOI (Digital Object Identificator).

En esta etapa también se identifica que la pagina funciona bajo una conexión cifrada HTTPS (Protocolo de transferencia de hipertexto seguro) y que no requiere autenticación o tiene un límite de peticiones.

2) Extracción de datos. Teniendo en cuenta la caracterización de la etapa anterior fue necesario utilizar varias técnicas para extraer la información:

• Trabajo por lotes usando colas e hilos: se estableció una arquitectura de procesamiento que consistía en que la entrada era un archivo plano con una lista de URL (Localizador Uniforme de Recursos) de los perfiles de los perfiles de los investigadores. El sistema toma cada URL y descarga a memoria RAM el archivo HTML de cada perfil. El sistema maneja dos hilos para aprovechar las capacidades de cómputo

del servidor y si falla en algún perfil, pasa al siguiente. Esto se logra usando un programa hecho Java usando OpenJDK.

• Luego de que se obtiene el HTML y por medio de la librería jsoup que es un parser para trabajar con el HMTL se pueden extraer las partes de la página que interesa analizar.

•Después fue necesario usar expresiones regulares para extraer cada parte de información de cada artículo, también fue necesario la creación de funciones específicas para encontrar donde terminaban metadatos como el ISSN usando funciones de Java para el manejo de vectores y cadenas de caracteres.

3) Validación y limpieza de datos. Durante el proceso anterior fue importante tomar decisiones respecto a cómo trabajar con datos faltantes. Los registros que definitivamente no contenían información suficiente fueron descartados. También se removieron caracteres especiales (la página maneja una codificación ISO-8859-1) y espacios que podrían ocasionar errores en procesos posteriores. Adicionalmente fue necesario una verificación manual de registros para identificar aspectos como errores de ortografía o de deletreo en metadatos como nombre de las revistas.

4) Análisis de datos. Posteriormente se realiza un análisis de tipo exploratorio y otro de tipo critico que permita interpretar los resultados de la exploración. Para ello los resultados del procesamiento se exportaron en un archivo tipo CVS (Valores Separados por Comas).

5) Propuesta de solución. Dentro de las alternativas, se propone el uso de expresiones regulares y bases de datos inteligentes, complementadas con estrategias que permitan que la base de datos "aprenda" y alimente su algoritmo de búsqueda e inclusión de la información.

III. Resultados

Para el caso de prueba se trabajó con una lista de 70 perfiles y como resultado se obtuvo un archivo de 15 columnas con 545 filas o registros en un periodo 2011-2021. En total se pudieron extraer un total de 545 registros de artículos científicos que se obtuvieron del proceso de extracción de cada uno de los perfiles de los 70 investigadores tomados en la muestra, en la Tabla II se presentan los resultados obtenidos de analizar "manualmente" cada registro; se hace referencia a una validación manual dado que al no aplicarse estándares de metadatos en el sistema CvLAC, el cotejamiento de la información debió hacerse personalizando el sistema de extracción de información.

TABLA II
CANTIDAD DE REGISTROS SEGÚN ESTADO

Estado	Cantidad
Artículo registrado dos veces por el mismo autor	16
Artículo con mal registro de información (Faltan campos, están incompletos o la información registrada es errónea)	309
Artículo no tiene como autor el investigador que lo registró	3
Artículo no es publicado en una revista científica	18
Artículo no existe, no se encuentra con los datos reportados	10
Artículo registrado correctamente	189
Total	545

En la Tabla II, llama la atención que aproximadamente 356 registros (65.3%) son incorrectos, considerándose una masa crítica los errores humanos dado un mal registro de información, correspondiente al 56.7% de los datos, dados principalmente por los errores relacionados en la Tabla III.

TABLA III
ERRORES POR MAL REGISTRO DE INFORMACIÓN

Error	Cantidad
Año de publicación	1
Aparecen una cantidad diferente de autores de los que realmente son	155
Nombres incorrectos de los autores registrados	38
DOI mal digitado o no valido	173
País donde se edita la Revista	10
Título del artículo	7

En cuanto a la coautoría es interesante destacar dos aspectos, el primero es que solo 36 de los 501 registros (Eliminando los duplicados, los no publicados en revistas científicas y los inexistentes) corresponden a artículos trabajados por un solo autor, los artículos restantes son trabajados en colaboración, según distribución de la Tabla III.

TABLA IV
CANTIDAD DE ARTÍCULOS SEGÚN NÚMERO DE AUTORES

Cantidad de autores	Número de artículos
Uno	36
Dos	98
Tres	242
Cuatro	65
Cinco	34
Seis	8
Siete	8
Ocho	3
Nueve	4
Diez o más	3

Al analizar lo presentado en las Tablas I y II, se encuentran serias dificultades de recuperación e interoperabilidad de los datos, especialmente a no propender por la integridad referencial de los datos, así mismo evidencia una ausencia de aplicación de los principios de bases de datos, al manejar registros redundantes y sin validación mínima de coincidencia.

Por ejemplo, para validar la información sería suficiente definir una serie de datos a extraer de para integración de la información a los perfiles de cada uno de los autores, validando adicionalmente que los demás datos si sean correctos, por ejemplo, en los artículos y libros, mediante la interacción con otrs bases de datos como la disponible por CrossRef. En la Figura 1, se exponen, como ejemplo, las

expresiones regulares utilizadas para capturar el dato que se desea extraer.

FIG. 1. Expresiones regulares para extracción de información.

Al ingresar dichas expresiones regular en proceso automatizados en bases de datos, se empiezan a incorporar elementos de aprendizaje requeridos para la implementación de bases de datos inteligentes, capaces de determina y "alimentarse" de los registros acorde a patrones, características y comportamientos rutinarios. Así mismo, el poder integrarse y aprende de otras bases de datos, permite llevar a cabo una implementación de lo que se considera una base de datos inteligente. Por ejemplo, un ejemplo de integración de diversas bases de datos que, para el caso de artículos científicos, permitirían alimentar unaa base de datos de productividad de los investigadores en todos los países, sin tener que pedir registro de la información, solo usando un dato de comparación como el nombre como lo hace Google Scholar, o el ORCID, para una mejor recuperación de la información.

Por otra parte, el uso de un servidor web, desarrollado en Python, realiza la petición al repositorio institucional, captura los datos correspondientes, los cuales dependerán del archivo de la Revista. De los datos recopilados se obtienen los DOI; estos identificadores son usados para hacer peticiones a los datos almacenados en diversas bases de datos mediante las APIs de Crossref (https://www.crossref.org/), Datacite (https://datacite.org/) y Freya (https://www.project-freya.eu/en). De allí se puede solicitar información como: las visualizaciones, las citaciones, el perfil del investigador, los cambios de versiones que tenga el trabajo y las menciones que hayan tenido los trabajos en redes sociales y académicas. Estos datos son capturados y transformados para ser incluidos en la base de datos propia, que se sugiere sea la del Ministerio (CvLAC).

Esta arquitectura permite el uso de múltiples usuarios en aplicación, así como desacoplar las funcionalidades de captura de datos y transformación con su visualización, haciéndola adaptable a cambios de lógica como de presentación y brindando la posibilidad de realizar cambios sin afectar la codificación ya realizada. Como a nivel de servidor solo se hacen los procesos de extracción y transformación, si es necesario se puede hacer de un balance de carga, teniendo múltiples instancias del servidor para ofrecer una mayor concurrencia.

La selección de tecnologías en la arquitectura fue por el rendimiento que estás brindan, en el servidor se necesitaba un lenguaje robusto que permitiera el manejo de grandes volúmenes de datos sin consumir muchos recursos a nivel de hardware y sin llegar a cuellos de botella al realizar los procesos de transformación de los datos.

V. CONCLUSIONES

A hoy en día el uso de bases de datos es común en todas las organizaciones, incluyendo las gubernamentales, sin embargo, sus diseños son desactualizados o incorrectos, llevando a tener errores de almacenamiento, como redundancia, duplicidad de los datos, entre otros. Lo anterior, en los entes del estado se torna de mayor critica, puesto que con base en esto se toman decisiones del país o se miden las capacidades de este, implicando no contar con fuentes de datos confiables y actualizadas.

Las bases de datos inteligentes, pese a que hoy aun sigue siendo complejo implementarlas en el pleno de su definición, ayudan a validar, mejorar y automatizar procesos mediante aprendizaje de maquina y el uso de elementos bases, como expresiones regulares, haciendo que la base de datos pueda de manera automática extraer información de otra bases de datos, acorde a una serie de reglas definidas, cada vez con mejores resultados pro el perfeccionamiento y aprendizaje que se tendría desde los algoritmos.

REFERENCIAS

[1] M. C. García-Cepero, "El estudio de productividad académica de profesores universitarios a través de análisis factorial confirmatorio; el caso de psicología en Estados Unidos de América," *Universitas Psychologica*, vol. 9, no. 1, pp. 13–26, 2010. https://doi.org/10.11144/Javeriana.upsy9-1.epap

[2] M. González-Zabala, E. Galvis-Lista, G. Angulo-Cuentas, "Análisis de indicadores de ciencia, tecnología e innovación (CTI) propuestos por organizaciones nacionales de CTI en América Latina," *Revista Virtual Universidad Católica del Norte*, vol. 52, pp. 23-40, 2017.

[3] M. K. McNutt, *et al.*, "Transparency in authors' contributions and responsibilities to promote integrity in scientific publication," *Proceedings of the National Academy of Sciences*, vol. 115, no. 11, pp. 2557-2560, 2018. https://doi.org/10.1073/pnas.1715374115

[4] A. Raoof, M. Ehab, "Open Researcher and Contributor Identifier and other author identifiers: Perspective from Pakistan," *Journal Of Pakistan Medical Association*, vol. 69, no. 6, pp. 888-891, 2019.

[5] A. Subbiah, M. Muthu, "Adopting ORCID as a unique identifier will benefit all involved in scholarly communication," *The National Medical Journal of India*, vol. 29, no. 4, pp. 1-8, 2016.

[6] N. Paskin, "Digital Object Identifier (DOI) System," en *Encyclopedia of Library and Information Sciences*, 2015, pp. 1-7. https://doi.org/10.1081/E-ELIS3-120044418

[7] T. Habermann, "Metadata and Reuse: Antidotes to Information Entropy," *Patterns*, vol. 1, no. 1, e100004, 2020. https://doi.org/10.1016/j.patter.2020.100004

[8] D. A. Allison, "OAI-PMH Harvested Collections and User Engagement," *Journal of Web Librarianship*, vol. 10, no. 1, pp. 14-27, 2016. https://doi.org/10.1080/19322909.2015.1128867

[9] L. L. Haak, A. Meadows, J. Brown, "Using ORCID, DOI, and Other Open Identifiers in Research Evaluation," *Frontiers in Research Metrics and Analytics*, vol. 3, pp. 1-7, 2018. https://doi.org/10.3389/frma.2018.00028

[10]R. Zhao, X. Wang, Z. Liu, Y. Qi, Z. Zhang, R. Chang, "Research on the impact evaluation of academic journals based on altmetrics and citation indicators," *Proceedings of the Association for Information Science and Technology*, vol. 56, no. 1, pp. 336-345, 2019. https://doi.org/10.1002/pra2.27

[11]S. Weibel, J. Kunze, C. Lagoze, M. Wolf, "Dublin Core Metadata for Resource Discovery," *Internet Engineering Task Force*, RFC 2413, 1998.