Diseño de un módulo web automatizado para la recuperación de metadatos de referencias de artículos

Alma Delia Apale Zitzihua, Ignacio López Martínez,
Giner Alor Hernández, José Luis Sánchez Cervantes,
Luis Ángel Reyes Hernández

Tecnológico Nacional de México,
Maestría en Sistemas Computacionales
México

almazitzihua@gmail.com,
{ilopez, galor}@ito-depi.edu.mx,
jlsanchez@conacyt.mx, lreyesh@orizaba.tecnm.mx

Resumen. En la actualidad surge la necesidad por parte de estudiantes e investigadores de extraer e integrar la información bibliográfica de los artículos científicos que han sido publicados en distintas revistas o libros de divulgación científica que se concentran en internet con el fin de visualizar, extraer, almacenar y gestionar estos resultados en los repositorios de sus instituciones que servirán para llevar un control así como el seguimiento del trabajo y producción de trabajos de investigación, los metadatos son indispensables al momento de realizar una investigación de la localización de un artículo digital. Entre los metadatos que contienen las referencias de un artículo uno de los más relevantes que poseen es el Identificador de Objeto Digital (DOI); el cual facilita la localización dentro de las bases de datos en las que son almacenados, ya que se compone de un código alfanumérico que es único para cada objeto digital. La cantidad de información concentrada en Internet es enorme y conforme pasa el tiempo se van desarrollando nuevas y mejores herramientas así también métodos de recuperación de esta información. En este trabajo se presenta el diseño de un módulo Web automatizado para la recuperación de metadatos de los artículos científicos de distintas bases de datos y repositorios de gran relevancia. Adicionalmente se presenta y se describe una arquitectura de los componentes y su funcionamiento.

Palabras clave: Internet, divulgación científica, repositorios, Metadatos, DOI.

Design of an Automated Web Module for Article Reference Metadata Retrieval

Abstract. Currently, there is a need for students and researchers to extract and integrate the bibliographic information of scientific articles that have been published in different popular science magazines or books that are concentrated on the Internet in order to visualize, extract, store and manage these results in

the repositories of their institutions that will serve to control and monitor the work and production of research work, metadata is essential when conducting an investigation of the location of a digital article. Among the metadata contained in the references of an article, one of the most relevant they have is the Digital Object Identifier (DOI), which facilitates the location within the databases in which they are stored, since it is made up of an alphanumeric code that is unique for each digital object. The amount of information concentrated on the Internet is enormous and as time goes by, new and better tools are developed, as well as methods for retrieving this information. This paper presents the design of an automated Web module for retrieving metadata of scientific articles from different highly relevant databases and repositories. Additionally, an architecture of the components and their operation is presented and described.

Keywords: Internet, popular science, repositories, Metadata, DOI.

1. Introducción

La innovación, así como los nuevos descubrimientos científicos requieren largos periodos de investigación que resultan en nuevas herramientas y métodos que ayudan a mejorar la vida diaria, este tipo de resultados en su mayoría se reflejan actualmente en productos digitales como lo son artículos. Es importante recordar que estos productos de investigación se clasifican e identifican mediante su referencia bibliográfica; la cual se compone de un conjunto de datos como: autor, título del artículo, fecha de publicación, título de la revista, así como un identificador único, estos datos sirven para buscar, clasificar y almacenar los artículos.

Existen herramientas como los gestores de referencias [1] que se encargan de gestionar grandes cantidades de productos digitales de investigación, además permiten buscar y almacenar de manera organizada dichos productos. Mendeley [2], Zotero [3], EndNote [4] son tres de los gestores más conocidos y utilizados, entre sus características se destacan que Mendeley provee 2 GB de almacenamiento gratuito y funciona mediante una aplicación de escritorio que cuenta con cinco maneras de extracción de artículos los cuales están incorporados a las bases de datos pertenecientes a la editorial Holandesa Elsevier [5], estas herramientas utilizan *plugins* (programa informático de inserción) [6] en páginas Web y bases de datos pero para la búsqueda se requiere ingresar la mayor parte de los datos del trabajo de investigación en algunos casos no se presentan problemas de compatibilidad.

Por otro lado, Zotero lleva a cabo la misma función, pero a diferencia del gestor anterior, solo provee de forma gratuita 300 MB de almacenamiento y presenta problemas de compatibilidad con algunas versiones de Word. Finalmente, el gestor EndNote permite almacenar distintos datos de los trabajos de investigación como imágenes y referencias, pero es muy poco conocido por la comunidad investigadora. Los tres gestores descritos anteriormente son de gran utilidad al momento de realizar una investigación, pero existen varios motivos que hacen de ellos muy complejos al momento de utilizarlos como dificultades al momento de la instalación en los dispositivos que en ocasiones presentan incompatibilidad en las versiones de software, el uso de complementos en los distintos navegadores, entre otros. A causa de esto es

necesario el desarrollo de una herramienta más intuitiva, sencilla y fácil de utilizar para la extracción de referencias con base en las necesidades del usuario final.

Tomando en cuenta lo anterior este artículo presenta una propuesta de solución que consiste en el desarrollo de un módulo Web automatizado que permita recuperar información de referencias de artículos a nivel de metadatos para asegurar una información más precisa y sin duplicidad, es decir, una búsqueda más simple ingresando la menor cantidad de datos del trabajo de investigación y así localizarlo en las principales bases de datos; la sección 2 presenta trabajos relacionados sobre métodos de extracción de metadatos con diferentes tipos de tecnologías; la sección 3 explica el diseño y funcionalidad del módulo Web; por último en la sección 4 se muestran las conclusiones del trabajo que se está realizando y los trabajos a futuro.

2. Trabajos relacionados

La producción en el ámbito de divulgación científica crece día a día y genera que los investigadores, así como las instituciones donde llevan a cabo este proceso se vean en la necesidad de llevar el control de manera organizada, así como de cada uno de sus productos generados; se han desarrollado herramientas que ponen a disposición para buscar, citar y almacenar estos trabajos de investigación. A continuación, se presenta una revisión del estado del arte de los trabajos existentes que han utilizado herramientas y/o métodos de búsqueda y extracción de datos de artículos científicos.

Klein y Van de Sompel [7] afirmaron que en las últimas dos décadas, la comunicación de la investigación ha pasado de ser un esfuerzo basado en el papel a una empresa digital basada en la Web. Con el paso del tiempo el proceso de investigación ha comenzado a evolucionar, ya que pasó a ser una actividad abierta y visible a nivel mundial. Con el fin de apoyar a los distintos grupos de investigadores surgió una gran variedad de repositorios, bases de datos, plataformas virtuales como una herramienta de divulgación científica. Sin embargo, algo que ha pasado rara vez pero que significa mucho para los usuarios es que las plataformas pueden desaparecer sin dejar rastro causando la pérdida de información importante.

Algunas de ellas no ofrecen garantías del contenido publicado, es por eso que se propuso un nuevo paradigma de archivo que se centra en el descubrimiento de entidades y artefactos Web en el ámbito del enfoque de archivo Web. Dado que el ORCID [8] surgió como una infraestructura Web académica de alto potencial que asigna identificadores únicas digitales a los académicos, permite listar identidades Web adicionales, así como artefactos. Sin embargo, también se descubrió que los crecimientos de ORCID tienen una tasa muy significativa que supera el crecimiento de los investigadores y eso ha causado también que varias plataformas últimamente opten por añadir esta función a sus tecnologías. Por lo tanto, existe una posibilidad real de que ORCID alcance un nivel de cobertura en un futuro próximo más adecuado para resolver las necesidades de los investigadores.

Hozlmann y Runnwerth [9] mencionaron que existen muchos tipos de identificadores para referirse a los trabajos académicos principalmente hicieron notar la existencia de DOI (*Digital Object Identifier*, Identificador de Objetos Digitales) Sistema que se inició en el año 1998[10] por Crossref una organización de miembros sin ánimos de lucro[11] , estos identificadores están constituidos por datos en conjunto,

referentes a la obra y que son llamados metadatos. Como problemática presentada en este artículo se argumentó que las referencias provenientes de sitios Web como lo son los blogs o programas informáticos, son muy vagas, aunque esta contenga la URL (*Uniform Resource Locator*, Identificador de recursos uniforme) [12] y la fecha en que se visitó, ya que, los sitios están en constante modificación de su contenido. Por lo anterior se presentaron los Micro Archivos que son colecciones microscópicas como su nombre lo dice, de los recursos archivados en la Web, utilizados para describir los objetos o entidades mediante el prototipo Micrawler (*Micro Crawler*, micro rastreador) que es una implementación y prueba de referencia encargado de la gestión de estos conjuntos de micro información. Micrawler tiene como objetivo principal que los archivos consultados en la Web archiven todos los recursos relacionados incluyendo el código fuente resultando en un micro Archivo para una consulta permanente mediante una URL corta o un mediante DOI.

A su vez Bangert y Frances [13] afirmaron que nos encontramos en un panorama en el que los productos académicos aumentan exponencialmente, debido a esto; los Identificadores Persistentes (PID) son una tecnología clave para permitir el acceso y la interoperabilidad entre los sistemas involucrados en la comunicación académica, estos identificadores permiten que un trabajo de investigación sea rastreado, visualizado, pero debido al surgimiento de nuevas plataformas que almacenan estos documentos de investigación surgió la necesidad de implementar mejoras en esta tecnología, es así, que varias organizaciones han trabajado para mejorar la integración en la infraestructura de investigación internacional y se propuso que sea a través de proyectos de colaboración con plataformas como ORCID que provee una forma de identificación al autor y la Red de Interoperabilidad de DataCite.

Este proyecto consistió en el diseño y la entrega de los servicios de PID en la Biblioteca de la UNSW que está guiado por las características de los identificadores como es el caso de DOI.

Los identificadores asignados a los resultados de la investigación son interoperables, basados en las fuentes verídicas, y contienen metadatos legibles por humanos y máquinas. DOI permite tener un acceso persistente al recurso está garantizado por la biblioteca como custodio de los identificadores y del contenido del repositorio asociado. Los resultados fueron que a medida que los identificadores pasan a formar parte de cada etapa del ciclo de vida de la investigación, el desafío para las instituciones será seguir rigiendo su asignación de manera efectiva, seguir las normas de la comunidad, y optimizar su uso para y por los investigadores.

Además Yang y Zhang [14] dedujeron que debido a las diversas plataformas que albergan miles de trabajos de investigación científica los investigadores tienen un acceso ilimitado a estas herramientas digitales, existen distintas formas de realizar la búsqueda de información y una de las más utilizadas es empleando las palabras claves para una recopilación precisa y necesaria de lo que se busca, esto se lleva a cabo por medio de la utilización de algoritmos de búsqueda o métodos de rastreo, en algunos casos estas formas muestran artículos más citados recientemente o su reputación dependiendo el autor, esto en algunos casos puede no ser lo que se requiere para el trabajo de investigación que se está desarrollando. Teniendo esta problemática se propuso la utilización del algoritmo de integración de texto llamado "TextRank" que es utilizado para ayudar a los investigadores en la realización de la revisión de la literatura. El objetivo principal de TextRank es resumir una obra determinada de documento que

Tabla 1. Comparativa de trabajos que muestran la falta de un sistema Web que se complemente con estas tecnologías, en el que se incluyan identificadores de autores, así como de artículos para recopilar y gestionar información precisa de obras de investigación que se encuentren almacenadas en distintas bases de datos y de diferentes editoriales.

Artículos	Extracción por ID	Extracción de Metadatos	Base de Datos	Integración de ORCID	Integración de Crossref
Klein y Van de Sompel [7]	✓	✓	✓	X	**X**
Holzmann y Runnwerth [9]	X	X	✓	✓	**X**
Bangert y Frances [13]	✓	**X**	✓	✓	**X**
Yang y Zhang [14]	**X**	✓	**X**	**X**	**X**
Módulo Web Propuesto	✓	✓	✓	✓	✓

Tabla 2. Tipos de metadatos a utilizar para la extracción de referencias de artículos.

Tipo de metadato	Contenido
Descriptivos	– Titulo – Autor
Estructurales	– Volumen de revista – Número de página

se encuentre contenido en alguna base de datos, por lo que se puede decir de acuerdo a su funcionamiento que este algoritmo se basa en la minería de texto y al ser compleja se requiere un gran esfuerzo humano.

A continuación, se presenta una tabla comparativa de trabajos relacionados y el alcance que cada uno tiene respecto a los métodos y tecnologías utilizadas.

3. Diseño y funcionalidad del módulo Web

La propuesta de solución surge con el fin de proveer a los investigadores y académicos una herramienta automatizada para recopilar las referencias de las obras publicadas en distintas bases de datos y que permita extraer metadatos para llevar el control de referencias de los artículos, así como almacenarlas además de monitorear el estatus de dicha información que se incluye en otras investigaciones en donde han sido citadas.

Se muestra una descripción general de la propuesta de solución y su arquitectura preliminar para el funcionamiento del módulo Web presentando las partes más importantes que son la búsqueda, recolección de información, selección y

Fig. 1 Diagrama del funcionamiento de las herramientas seleccionadas.

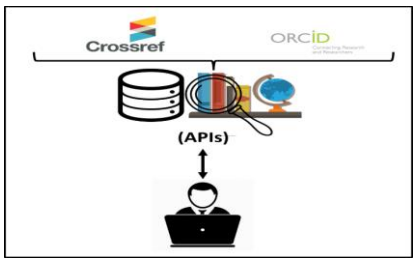

Fig. 2 Funcionamiento de un segmento de la arquitectura preliminar.

funcionamiento. Para llevar a cabo esta extracción es importante mencionar que se realizará la extracción de metadatos que son datos que describen otros datos [15] para el caso de las referencias de artículos se utilizaran los siguientes dos tipos de metadatos:

Búsqueda y recolección de metadatos. Se describe el módulo Web que interactúe con las bases de datos científicas que pertenezcan a trabajos de investigación en sus distintos formatos se logren recuperar metadatos de referencias bibliográficas de artículos utilizando principalmente los identificadores internacionales y el ingreso de la menor cantidad de datos para el rastreo de los trabajos de investigación, para la búsqueda y extracción.

Selección de herramientas. Para la selección de las herramientas a implementar en el módulo se basó en el siguiente diagrama que muestra el funcionamiento de recolección de productos de investigación:

Funcionamiento. En primer lugar, las bases de datos como: Scopus, Web of Science, entre otros; pertenecen a la mayor editorial de contenido científico que es Elsevier la cual posteriormente se encarga de depositar sus metadatos a la agencia de identificadores que en este caso es Crossref [10] ; de igual forma los investigadores depositan sus metadatos mediante el uso de ORCID. En base a este funcionamiento se eligió utilizar ambas herramientas para realizar la búsqueda y extracción requerida para el módulo tomando en cuenta la forma en que ellos recolectan la información que se necesita para alimentar el módulo y aprovechando la oportunidad de alcance a esta concentración de datos al público en general, así como desarrolladores y este caso resolviendo los problemas anteriormente planteados.

En el siguiente diagrama se muestra el funcionamiento entre usuario y las herramientas seleccionadas. En primera instancia el usuario realiza una petición de

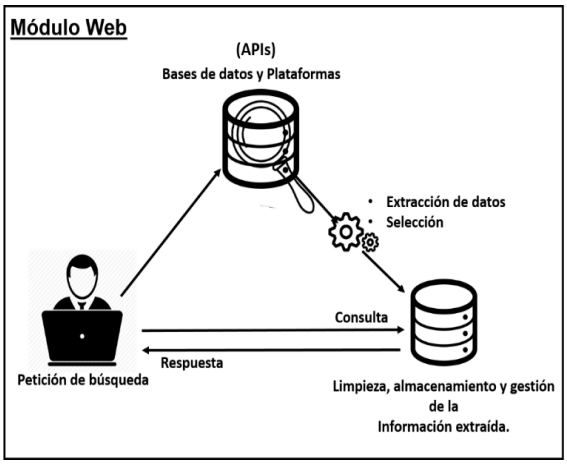

Fig. 3. Arquitectura preliminar basado en el modelo orientado a servicios.

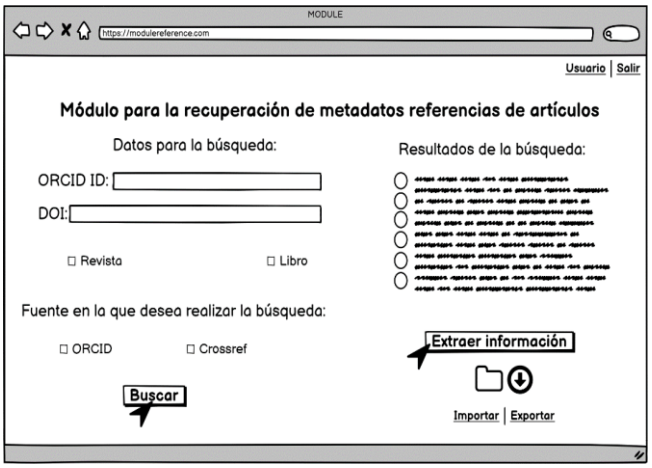

Fig. 4. Mockup de la interfaz del Módulo Web propuesto.

búsqueda de metadatos mediante el módulo que con ayuda de las API´s (Interfaces de Programación de Aplicaciones del Navegador, *Application Programming Interface*) [16] realizará la entrega de la respuesta de la información requerida.

Agregando la otra parte del funcionamiento de la arquitectura propuesta contempla la utilización de una base de datos local que se encargará de reunir la información para su posterior selección de la información requerida y envío para la concentración en algún repositorio dependiendo la decisión de uso del usuario.

La figura que se presenta a continuación muestra la interfaz que va interactuar con el usuario (estudiante o investigador) que tiene como objetivo realizar la búsqueda automatizada minimizando el ingreso de datos de manera que ingresando el Identificador de Objeto o Identificador de autor puede ser buscado también en libros y revistas.

Tabla 3. Requerimientos solicitados por ORCID para el cumplimiento de políticas de seguridad de datos y sus funciones principales a realizar dentro del módulo Web.

Requerimientos	**Función dentro del módulo**
Almacenar Identificadores	Cada usuario con cuenta en ORCID maneja un identificador único que ayudará a clasificar el contenido dentro del módulo de forma más segura.
Usar Tokens de acceso persistentes y actualizados	Los Tokens de acceso que funcionan como llaves digitales dentro del módulo serán utilizados para el tratamiento seguro de cada uno de los productos.
Registro de interacciones con la API	Se realizará el registro de todas las llamadas y respuestas recibidas.
Contacto de soporte	Se debe incluir una función que sirva como medio de contacto con soporte técnico cuando ocurra alguna interacción inesperada.

La búsqueda entre las bases de datos será opcional entre utilizar una en específico o ambas simultáneamente. Una vez finalizada la búsqueda el módulo mostrará los resultados obtenidos para la selección de interés por parte del usuario y con las opciones de importación y exportación de sus datos contemplando las políticas de seguridad que pongan en vigor ambas organizaciones que se incluyen en el módulo.

Las organizaciones en la elección de búsqueda ponen a disposición las siguientes Apis:

API REST Pública Crossref: La cual expone los metadatos que son depositados por los miembros en la plataforma de Crossref que son importantes para localizar publicaciones por autor, cabe mencionar que dichos datos públicos se pueden utilizar sin restricción [17].

Funcionalidades de los componentes de búsqueda:

- Trabajos utilizando el DOI,
- Miembros,
- Tipos de publicaciones,
- Revistas.

Clasificación:

- Relevancia,
- Fecha,
- Hora.

API RESTful Pública ORCID: API que proporciona el acceso a la base de datos de los datos de registros ORCID ID públicos función que es elección del miembro[18].

Funcionalidades de los componentes de búsqueda:

- Obtener ORCID ID Autenticado,
- Buscar,

Tabla 4. Propuesta de tecnologías para el desarrollo del módulo, se muestra la solución propuesta la cual considera la utilización del lenguaje de programación Node.js esto debido a que es adecuado para el desarrollo Web, como entorno de desarrollo se contempló el uso de NetBeans ya que provee soporte para las aplicaciones orientadas a servicios. En cuanto al sistema gestor de base de datos MariaDB para almacenar los datos extraídos, y finalmente la metodología Scrum la cual maneja un entorno colaborativo, organizado por iteraciones entres otras cualidades.

Lenguaje de programación	IDE	SGBD	Metodología
Node.js	NetBeans	MariaDB	Scrum

– Recuperar datos públicos.

Para una implementación óptima y segura de la API de ORCID es indispensable cumplir con los siguientes requerimientos como políticas de seguridad por parte de ORCID en su documentación de uso de su API [19]:

El desarrollo de este módulo contempla la utilización de tecnologías que permitan una interacción entre la base de datos con el servidor y el usuario con el módulo. Para ello se requiere de un entorno de desarrollo que soporte aplicaciones orientadas a servicios. Finalmente se propone una base de datos local para un mejor control y gestión de la información extraída.

4. Conclusiones y trabajos futuros

Existen gestores de referencias y componentes que ayudan a extraer información referente a trabajos de investigación en bases de datos científicas pero en particular con el módulo Web propuesto realizará esta función de manera más sencilla y eficaz con una búsqueda ingresando la menor cantidad de datos además de que sea compatible con los buscadores resolviendo los problemas que presentan los actuales gestores de referencias anteriormente mencionados para esto se utilizaran las normas y estándares de calidad para recopilar dicha información , esto se logrará utilizando los identificadores usados en los distintos formatos para mayor confiabilidad tomando en cuenta sus políticas de seguridad así como las recomendaciones de mejores prácticas.

El seguimiento, almacenamiento y control de productos científicos es relevante e importante al momento de llevar a cabo el desarrollo de una nueva tecnología, método o herramienta que resuelva una problemática actual, ya que, como en este caso se toman en cuenta los resultados obtenidos para tomarlos en cuenta en todo momento del proceso de desarrollo.

El objetivo principal y más importante de este trabajo es beneficiar no solo a personas que participen en algún tipo de investigación como lo hacen los estudiantes al momento de concluir un grado de estudio, sino también a instituciones que requieran llevar un control de las obras realizadas por sus académicos en las distintas áreas.

Como trabajo a futuro se desarrollará el módulo Web automatizado propuesto y se validará mediante un caso de estudio implementándolo en una institución para gestionar los productos sus estudiantes o investigadores autores además de agregar un apartado extra que obtenga el informe sobre el estatus de producción para encontrar los puntos de oportunidad en las cuales trabajar y mejorar. Finalmente se realizarán las pruebas

necesarias con el equipo de ORCID para la validación, aprobación y publicación del módulo desarrollado para llevar a cabo la transición a la API de producción.

Referencias

1. PoliScience: Gestores de referencias. https://poliscience.blogs.upv.es/investiga dores-2/mis-citas/gestores-de-citas/ (2020)
2. Mendeley: Software de gestión de referencias y red de investigadores. https://mendeley.com/?interaction_required=true (2020)
3. Zotero: Tu asistente de investigación personal. https://zotero.org/ (2020)
4. EndNote: Clarivate Analytics. https://endnote.com/ (2020)
5. Elsevier: Zona de Lectura. https://elsevier.es/es (2020)
6. Neo Wiki: ¿Qué es un Plugin y para que sirve?. https://neoattack.com/ neowiki/plugin/ (2020)
7. Klein, M., van de Sompel, H.: Discovering scholarly orphans using ORCID. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 1–10 (2017)
8. ORCID: Un sistema para identificar de manera única a los investigadores. HAAK, Learned Publishing (2012)
9. Holzmann, H., Runnwerth, M.: Micro archives as rich digital object representations. In: Proceedings of the 10th ACM Conference on Web Science, pp. 353–357 (2018)
10. Doi.org: Agencias de registro de DOI. https://doi.org/registration_agencies.html (2020)
11. Wilkinson, L.J.: You are Crossref, https://crossref.org/ (2020)
12. Significados.com: Significado de URL (Qué es, Concepto y Definición) Significados. https://significados.com/url/ (2020)
13. Bangert, D., Frances, M.: PIDs to support discovery and citation: Persistent identifier service design and delivery at UNSW library. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 1–2 (2017)
14. Yang, D., Zhang, A.N.: Performing literature review using text mining, Part III: Summarizing articles using TextRank. In: IEEE International Conference on Big Data (Big Data), Seattle, pp. 3186−3190 (2018)
15. PowerData: ¿Qué son los metadatos y cuál es su utilidad? https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/que-son-los-metadatos-y-cual-es-su-utilidad (2020)
16. Hipertextual.com: Qué es una API. https://hipertextual.com/archivo /2014/05/que-es-api/ (2020)
17. Crossref: Rest Api. https://crossref.org/education/retrieve-metadata/rest-api/ (2020)
18. Krznarich, L.: Orcid Api. https://orcid.org/organizations/integrators/API (2014)
19. Orcid Members: Getting started with your ORCID integration. https://members.orcid.org/api/getting-started (2020)