ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2023.1.1.987

Automation of the transformation process of publication formats in scientific journals through a Python script

Automatización del proceso de transformación de formatos de publicaciones en revistas científicas a través de script en Phyton

Danny Murillo-Gonzalez, Master[1], Sucel López, Doctor[2]
[1]Universidad Tecnológica de Panamá, Panamá, danny.murillo@utp.ac.pa
[2]Universidad Tecnológica de Panamá, Panamá, sucel.lopez@utp.ac.pa

Abstract– ***Scientific disclosure and diffusion is the way to inform society and other scientists about the results of research and the generation of new knowledge. In recent years, scientific journals in digital format have become the most widely used medium to demonstrate these results, but when evaluating a journal for publication, it is necessary to evaluate: its presentation, form of distribution, quality of its content, and impact of publication. magazine. Of these elements, the form of distribution is of great relevance since it is linked to the visibility of the journal, if it is not found, it is not read or cited, but if the publication formats are not diverse, we will not be able to improve the digital reach either. of those who use this content.***

According to data from the Scholastica report, a payment platform that includes more than 900 academic journal publishers, indicates that the most used formats are pdf and html. In some studies carried out in Central America on more than 185 journals, specifically from Costa Rica and Panama, the most used scientific journal formats are pdf, html, ePub, xml-jats, audio and Flipbook. Of the evaluated journals, only 50% use two formats and only 15% use more than three formats, the most common being html and pdf. However, the limitation is not only the format but the software to transform from pdf to html due to the limitations, but according to the publishers they do not use other formats because they do not know the software used for this process. In the case of Panamanian journals, out of 30 journals evaluated, 100% use pdf, only six use html, and only four journals use more than three formats, so we can say that there is a deficiency in the number of formats and the amount of time the transformation process may take publishers.

The objective of this work is to generate a script using Python as a programming language to automate the process of transforming scientific article formats in docx, to other formats such as pdf, html, ePub, txt and audio, minimizing the use of software and reducing the processing time of these documents. In the tests carried out, it was necessary to generate document character style formats to achieve good results, where it was possible to transform 24 articles from two magazines into five formats, also the transformation time was reduced from 15 hours to 15 minutes to transform them.

Keywords— automation, formats, journals, Python, visibility

Resumen– ***La divulgación y difusión científica es la forma de dar a conocer a la sociedad ya otros científicos los resultados de la investigación y la generación de nuevo conocimiento. En los últimos años, las revistas científicas en formato digital se han convertido en el medio más utilizado para demostrar estos resultados, pero al evaluar una revista para publicar es necesario evaluar: su presentación, forma de distribución, calidad de su contenido e impacto de la revista. De estos elementos la forma de distribución es de gran relevancia ya que está ligada a la visibilidad de la revista, si no se encuentra, no se lee ni se cita, pero si los formatos de publicación no son diversos, tampoco podremos mejorar el alcance digital de quienes usan este contenido.***

Según datos del informe Scholastica, una plataforma de pago que incluye a más de 900 editores de revistas académicas indica que los formatos más utilizados son pdf y html. En algunos estudios realizados en Centroamérica a más de 185 revistas, específicamente de Costa Rica y Panamá, los formatos de revistas científicas más utilizados son pdf, html, ePub, xml-jats, audio y Flipbook. De las revistas evaluadas, solo el 50% utiliza dos formatos y apenas el 15% utiliza más de tres formatos, siendo los más comunes html y pdf. Sin embargo, la limitación no es solo el formato sino el software para transformar de pdf a html debido a las limitaciones, sino que según los editores no utilizan otros formatos porque desconocen el software utilizado para este proceso. En el caso de las revistas panameñas, de 30 revistas evaluadas, el 100% usa pdf, solo seis usan html, y solo cuatro revistas usan más de tres formatos, por lo que podemos decir que hay una deficiencia en la cantidad de formatos y el tiempo que el proceso de transformación puede llevar a los editores.

El objetivo de este trabajo es generar un script usando Python como lenguaje de programación para automatizar el proceso de transformación de formatos de artículos científicos en docx, a otros formatos como pdf, html, ePub, txt y audio, minimizando el uso de software y reduciendo el tiempo de tramitación de estos documentos. En las pruebas realizadas fue necesario generar formatos de estilo de caracter de documento para lograr buenos resultados, donde se logró transformar 24 artículos de dos revistas en cinco formatos, también el tiempo de transformación se redujo de 15 horas a 15 minutos para transformarlos.

Palabras claves— automation, formats, journals, Python, visibility

I. INTRODUCCIÓN

La divulgación y difusión científica es un proceso de comunicación para dar a conocer a la sociedad y los pares los resultados obtenidos y conocimiento generado de la labor de investigación a través de diversos medios de comunicación.

Uno de los principales medios de comunicación científica han sido las revistas de edición científica tradicional o impresa que en los últimos años ha generado mayo alcance por la aparición de la edición electrónica, el cual supone nuevas formas de comunicación de contenidos, nuevos formatos, nuevos servicios de valor añadido, nuevos estándares [1] y nuevos enfoques entorno a mejorar el alcance de la revista a nivel de acceso a través de la web.

Las revistas científicas como medio de difusión de los avances de la investigación, son uno de los ámbitos de la comunicación que más han evolucionado en la última década, especialmente en los aspectos de su estructura, medios técnicos, normalización y evaluación [2], logrando integrar un grupo de expertos que no solo trabajan en el proceso de dirigir la revista sino de diagramación, indexación , divulgación y también de medición de diferentes indicadores científicos y webmétricos de la revista.

El análisis de la producción científica es una manera de comprobar el auge expansivo de una revista científica, conocer si su contenido es útil y pertinente, o si hay académicos, investigadores o estudiantes que consultan estos contenidos para elaborar o sustentar otros resultados, si se cumple con los niveles de calidad o si su contenido es visible en un contexto nacional e internacional [3].

Cabe destacar que el proceso de consulta, no existe, sin la divulgación o difusión de los contenidos de la revista, esta debe entenderse como la capacidad que esta tiene de ser visible para la comunidad científica a la que se dirige, siendo la finalidad de una revista científica como medio de comunicación, alcanzar al público al que se dirige [4]. generando un mayor impacto del conocimiento proveniente del nuevo aporte o innovación en los resultados plasmados.

Es importante señalar que en el ámbitos de la revistas existe un ecosistema de publicación científica que contiene varias etapas del ciclo de vida del documento, relacionado son los tipos de formatos utilizados, formato de adquisición, formatos de producción, formatos de difusión, formato de preservación [5]. Entre los formatos de difusión tradicional están el pdf y html que son altamente recomendados en revistas académicas de Acceso Abierto [6] porque permiten la accesibilidad de sus contenidos, sin embargo otros especialistas recomiendan nuevos formatos de difusión del contenido como, videos, audio [7], y los formatos electrónicos ePub e ePub3 para incrustar contenido multimedia [8] además del formato XML-JATS, un formato de marcaje semántico del texto de una publicación [9].

La difusión de los contenidos de la revista está ligado a los formatos utilizados en la divulgación que son necesarios que mejoran la calidad estructural del contenido [5] si se utilizan los estándares adecuados, pero también es necesario que este formato contribuya a dar mayor visibilidad y accesibilidad al documento. Es por ello que dentro de los aspectos de visibilidad las revistas deben ser de acceso abierto (OA) al conocimiento sin restricciones económicas, tecnológico o de índole legal [5].

Si bien el solo hecho que una revista sea publicada en OA no necesariamente implica calidad o sea un factor para que tenga más lecturas, descargas y citaciones [10], sí permite una mayor difusión de sus contenidos [11], ya que existen plataformas OA como indexadores, directorios, repositorios y bases de datos bibliográficas multidisciplinares, que integrar los diverso formatos de estas revistas OA como material de consulta.

Hay que mencionar también que al realizar la búsqueda de una revista antes enviar un artículo a está, es necesario evaluar algunos aspectos, entre los que se destacan, su acceso online [12], forma de distribución, calidad y tipo de contenido [13] e impacto de la revista. Si bien todos estos elementos son de interés, la forma de distribución toma gran relevancia ya que está vinculada a dar mayor visibilidad de la revista, en términos de lograr encontrarla en las diversas plataformas y en que formatos de publicación se encuentra para su consulta.

Según datos encontrado de estudios realizados en Centroamérica sobre tipos de formatos de revistas científicas en Costa Rica, en el año 2017 el Tecnológico de Costa Rica evaluó 98 de sus revistas, de la cuales, PDF (98), html (22), ePub(10) y xml (2) [14]. En el año 2022 la Universidad Nacional de Costa Rica (UNA) evaluó 27 revistas, donde sus usos fueron pdf (27), html(24), ePub (19) , audio (3), xml (2) y FlipBook colo uno [15]. En el caso de Panamá en el año 2019, el Consejo de Rectores de Panamá (CRP) y la Secretaría Nacional de Ciencia y Tecnología (SENACYT) realizó el primer catálogo de revistas académicas-científicas de Panamá donde se evaluaron 37 revistas, los formatos más comunes fueron, pdf (25), html (4), sin identificar otros formatos [16].

Según datos del informe de Scholastica 2020, plataforma web de pago que incluye más de 900 editores de revistas académicas, los formatos más empleados son el pdf (98%), html (48%), papel (43%), epub (14%), xml (3%) [17]. Estos datos dan un indicativo de la importancia de los formatos, siendo los más utilizados el pdf y el html donde en promedio solo el 48.5% de estas revistas utiliza solo dos formatos, lo que minimiza el nivel de visibilidad, por otro lado, estos datos no indican el tiempo que lleva el transformar de estos artículos de revistas en los diferentes formatos.

En el año 2022 el Consejo de rectores de Panamá (CRP) y la SENACYT de Panamá realizaron diversos talleres para el mejoramiento de las revistas científicas panameñas encuestando a diverso participantes y responsables de 30 revistas científicas identificando que todas utilizaban el formato pdf(30), FlipBook (6), html (6), ePub (4) y una mhtml, solo cuatro revistas utilizaban tres formatos. Se mencionó que ocho de estas revoisas utilizaban XML-JATS, sin embargo, no lo utilizaban en sus portales de revistas, sino que se generaban en otra plataforma donde estaban indexadas.

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

La encuesta también consultaba a los editores de revistas que herramienta utilizaban para transformar los artículos en diversos formatos, donde todos indicaron que utilizan Microsoft Word para transformar los artículos en formato PDF y html; para el formato ePub utilizan el software Calibre y para el formato de audio, herramientas online para transformar texto a audio. De las seis revistas de Panamá con formato html, solo tres mostraban el contenido del artículo formateado por lo que se infiere que el MS Word no había exportado lo estilos del documento de forma adecuada. En el caso de las cuatro revistas con tres formatos incluyendo PDF, html y FlipBook, es un trabajo realizado por una empresa externa a la universidad por lo que no es posible conocer el tiempo que demoraban en esa transformación.

De las instituciones panameñas que utilizan en sus revista el formato html, la Universidad Tecnológica Panamá muestra un contenido limpio y personalizado, esto se debe a que realiza un proceso que implica exportar el documento .docx a html y posteriormente realiza una limpieza del código html y CSS que garantiza eliminar el código basura html y minimizar el tiempo de carga del documento en el navegador, también permite personalizar visualmente el articulo y una mejor presentación, sin embargo, el tiempo que se utiliza en este proceso de transformación por cada artículo varía entre cinco a ocho días (8 horas) en función del número de páginas del artículo y demás componentes del artículo, a este tiempo no se le incluye crear otros formatos como ePub o audio.

Para las revistas, es de suma importancia que sus publicaciones puedan tener diversos formatos para mejorar su indexación en buscadores y bases de datos para mejora su visibilidad [18], pero no todas cuentas con recurso técnico, económicos y humano [21], para minimizar el tiempo y costes de transformación de los formatos que les ayude en difusión de la revista [22].

El objetivo de este trabajo es mostrar los resultados al crear un script para automatizar el proceso de transformación de formatos de artículos científicos de docx, a otros formatos como pdf, html, ePub, txt y audio, utilizando como lenguaje de programación, Phyton, buscando minimizae el uso de softwares y reduciendo el tiempo de transformación de estos documentos.

II. Metodología

A. Herramientas

Los softwares utilizados para transformar los artículos fueron: Microsoft Word, Adobe Dreamweaver 2021, Sublime Text 3.2, Calibre 5.44, página web texttomp3 para transformar de texto a mp3, typeset.io plataforma online de pago para transformar .docx a pdf, html, epub, xml-jats, aunque este último formato no se contempla en las pruebas.

Para crear los estilos de caracteres en MS Word se creó una plantilla con los estilos de cada uno de los componentes del artículo como títulos, palabras claves, resumen encabezados, tablas, figuras, referencias, entre otros,

Se utilizó como editor de código para programar en Phyton, Jupiter Notebook con la versión de Phyton: 3.8.8.

Las librerías integradas al código Phyton fueron: numpy, pandas, scv, os, glob, openpyxl, docx2pdf, zipfile, html_from_epub.

Phyton es un lenguaje dinámico de script multiplataforma, lo que permite que el código generado pueda realizarse en diversos entornos de programación y su código final pueda migrarse a plataforma online o de escritorio para facilitar su uso [19], [20].

A. Datos para las pruebas.

Para las pruebas de transformación de formatos de publicaciones se utilizaron los artículos de dos revistas, la revistas I+D Tecnológico y RIC de la Universidad Tecnológica de Panamá. Se seleccionó para la primera prueba dos artículos de cada revista, dos de cuatro páginas y dos de 10 páginas o más que incluyeran en su contenido, imágenes, tablas, fórmulas, referencias enumeradas, más de tres autores con su afiliación,

En la prueba por volumen de la revista y tiempo de procesamiento se seleccionaron 12 artículos de cada revista sin limitaciones de páginas para el script y 12 artículos de cuatro páginas para la transformación por software.

B. Procesos utilizado para transformar artículos en diversos formatos utilizando herramientas de software.

El proceso utilizado actualmente en la Universidad Tecnológica de Panamá para transformar los artículos de sus revistas en otros formatos estaba basado enteramente en el uso de softwares, como se muestra en la figura 1 como Esquema (a). El proceso de transformación inicia con los artículos en formato .docx con la plantilla correspondientes de cada revista y aprobados previamente por los responsables de la revista. Cada artículo de forma individual es transformado desde Microsoft Word a PDF y a HTML. En el caso del formato HTML este se depura para eliminar el código basura utilizando el software Dreamweaver de Adobe, este proceso se elimina etiquetas HTML y el código CSS generado por Word para que el archivo quede solo con las etiquetas HTML básica. Luego se integra los códigos CSS creados de forma personalizada en un archivo externo al HTML para lograr un código limpio y fácil de mantener.

El documento HTML generado de Word se transforma al formato ePUB utilizando la herramienta Calibre, en este proceso no se realiza integración de metadatos, índices, ni portada a las publicaciones. El proceso de transformación a audio, solo se ha realizado en un volumen, donde el documento HTML exportado se extrae el número del volumen, año, título, resumen, autores y enlace web, los cuales se guardan en un archivo en formato txt. El texto generado se copia en una web que permite transformar el texto a audio en diversos idiomas y tipos de voz (hombre o mujer), la limitante del esquema (a) como se ha indicado es que cada artículo debe pasar por el mismo proceso y el tiempo de procesamiento de la limpieza del código html, depende del

número de páginas del documento, cantidad de imágenes y tablas.

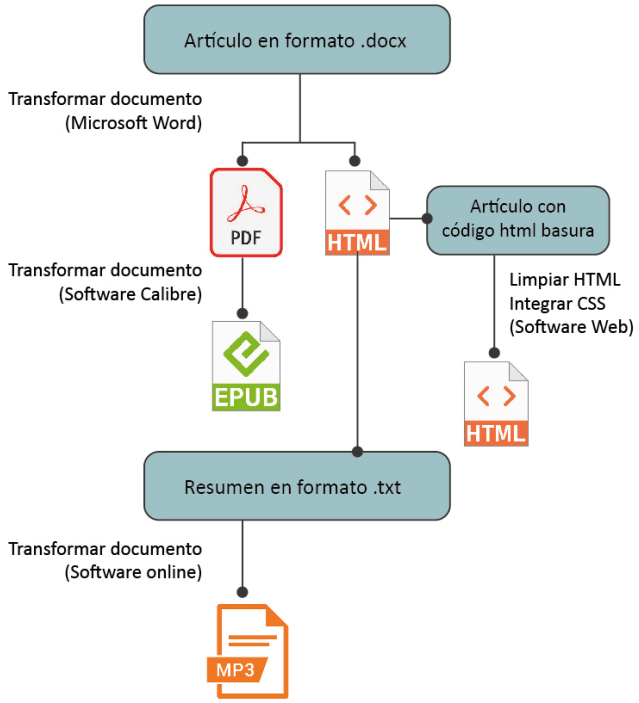

Figura 1. Esquema de los procesos para transformar artículos de revistas en cuatros formatos utilizando softwares

C. Esquema de script en Phyton para transformar artículos en diversos formatos utilizando funciones

Con el objetivo de no utilizar herramientas de Software en el proceso de transformación de los artículos y minimizar la intervención humana se creó un Script en Phyton con diversas funciones que contienen el código programado utilizando librerías propias del lenguaje, este esquema se muestra en la figura 2. El proceso inicia con los artículos en formato .docx que por razones de simplificar la programación del script se evaluó transformar el documento de dos columnas a una columna. Se realiza un proceso de normalización del contenido del documento formateando el texto utilizando estilos creados en una plantilla en Word antes de ser guardados en la carpeta de trabajo que se utilizará en el script.

Al iniciar el script en Phyton, este tiene un listado de los estilos creados en Word que fueron interpretados en Phyton integrados a través de la función estilos_phyton(). El nombre de estos estilos fue vinculados y creados en la función estilos_css(), para generar un estilo CSS de cada uno de los nombres de estilos creados desde Word, esto con el objetivo de no utilizar los estilos CSS predeterminados de MS Word sino con código independiente, personalizado que pueda ser manipulable posteriormente por programación.

En el siguiente paso se cargan las librerías de Phyton que son utilizadas por las funciones, el script lee la ruta de la carpeta de trabajo donde se encuentran las publicaciones en formato .docx, esta ruta es personalizable y el script leerá solo archivos en el formato .docx o .doc.

El componente principal del script contiene tres funciones y tres cicles de repetición. El primer ciclo permite cargar la función doc_html() que realiza la lectura del primero documento en formato Word, lo transforma a PDF y luego lo transforma a HTM. Una vez termina de transformar el HTM, selecciona el archivó y lo transforma a formato ePUB. Este proceso se realiza hasta terminar el número de documentos .docx en la carpeta de trabajo.

El segundo ciclo es un proceso para extraer del archivo html los textos relacionados con el volumen de la revista, título del artículo, resumen, autores y url donde está el enlace de la publicación, estos datos son almacenadas en variables que luego son integradas en un archivo .txt el cual debe guardarse con el encoding= utf8, esto fue necesario para asegurar los caracteres latinos en el documento y evitar caracteres extraños en el texto

El tercer ciclo de repetición integra la función txt_mp3(), que lee los archivos .txt generados de cada publicación y los transforma en formato de audio latino utilizando la librería gTTS de Phyton.

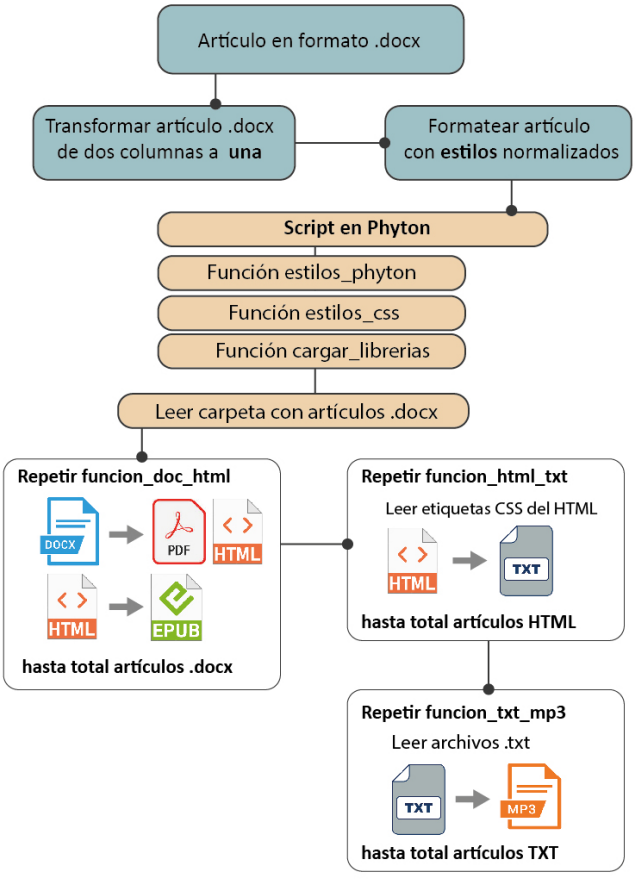

Figura 2.. Estructura de las funciones para transformar artículos de revistas en cuatros formatos utilizando Script en Phyton

III. RESULTADOS

Para la prueba de transformación de formatos, la tabla 1 muestra los diversos procesos utilizados en los cuatro artículos seleccionados de las dos revistas utilizando softwares, plataforma online y el script en Phyton.

Los resultados indican que utilizando los softwares se logró transformar los cuatro documentos de prueba en diversos formatos, con la plataforma (typeset.io) se logró realizar la transformación a pdf, html no estructurado, el formato epub se logró transformar, pero no estructurado visualmente. Para el formato XML-JATS, es necesario hacer un proceso de marcaje de elementos del documento Word como título, autores, subtítulos, palabras claves, resumen, imágenes y tablas, antes de hacer la transformación y que este pueda interpretar la estructura del documento, antes de transformarlo. Aunque la interface de la plataforma es simple, no es muy intuitiva y los resultados en html e ePub no fueron del todo favorables en su apariencia. Con el script solo se pudo hacer la transformación de pdf y html sin formato, ni estructura, el resto de los formatos no se pudieron transformar porque el script no pudo identificar algunos elementos en los artículos .docx, como número de columnas en el documento, viñetas, imágenes, tablas, títulos ya que loa artículos tenían los estilos predefinidos de la plantilla de la revista, por los que creemos que los autores no habían seguido las directrices de los formatos.

TABLA I
NÚMERO DE DOCUMENTOS TRANSFORMADOS POR CADA PROCESO CON LOS SOFTWARES, HERRAMIENTA ONLINE Y EL SCRIPT EN PHYTON

Procesos	Software	typeset.io	Script Phyton
word a pdf	4	4	4
word a html	4	4	4
limpiar html/css	4	0	0
html a epub	4	4	0
html a txt	4	0	0
txt a mp3	4	0	0

Entre los elementos en los artículos .docx que impidieron la transformación a otros formatos se identificaron 11 los cuales se muestran en la figura 3. Se destaca que todos los artículos tenían dos columnas, por lo que en el formato html, la visualización no era la correcta como si sucedía en el pdf. Los artículos no tenían estilos de caracteres parametrizados para poder identificarlos en el script en Phyton, sino que estilos generados por los autores, además los subtítulos, imágenes, tablas, fórmulas y referencias auto numerados en Ms Word no se visibilizan en el html.

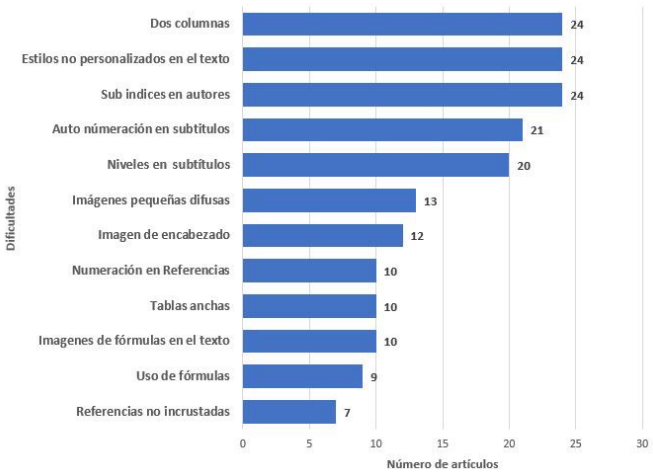

Figura 3. Estilos de carácter creados en Word para formatear el documento.

Entre las correcciones realizadas al artículo .docx antes de hacer la evaluación del script creado en Phyton fue necesario formatear el documento con diversos estilos de carácter personalizados y normalizados creados desde MS Word para las diferentes secciones del documento. Como se muestra en la figura 4 donde se selecciona un estilo llamado RevistaTituloPrincipal, para formatear el título en español del artículo. A cada artículo en formato .docx se le eliminó el formato de párrafo y carácter y se le asigno uno de los estilos creados para cada sección, titulo español, título en inglés, autores, afiliación, resumen, abstract, palabras claves en español e inglés, tablas, imágenes, párrafos, referencias, títulos de tablas e imágenes.

Figura 4. Estilos de carácter creados en Word para formatear el documento.

Para que Phyton interpretará los estilos de carácter creados en MS wordAntes y poder manipularlos desde entono de Jupyter Notebook y el lenguaje de programación estos fueron asignados a diversas variables en Phyton. El estilo RevistaTituloPrincipal de MS Word en Phyton se llamaría h1. tituloEs como se muestra en la figura 5, que posteriormente se utilizará como el nombre de un estilo en CSS.

```
custom_styles = """ p.Revistatituloprincipal => h1.tituloEs
                    p.Revistasubtitulo => h1.tituloEn
                    p.Revistaautores => p.autores
                    p.Revistaafilia => p.afiliacion
                    p.Revistacuerpo => p.cuerpo
                    p.Revistaresumen => p.resumen
                    p.Revistapalabraclave => h1.palabrasC
                    p.Revistatitulo1 => h2.subtitulo
                    p.Revistatitulo2 => h3.subtitulo2
                    p.Revistatitulo3 => h4.subtitulo3
                    p.Revistatabla => p.titulotabla
                    p.Revistapie => p.pieTexto
                    p.RevistaReferencias => p.referencias
```

Figura 5. Nombre de los estilos de carácter creados en Word asignados a estilos en Phyton

Los estilos de los artículos creados en Word son asignados al exportar el html, sin embargo, estos son eliminados por el script de Phyton ya que contienen elementos no relevantes y no personalizados. Los nuevos estilos CSS son creados en Phyton como se muestra en la figura 6 y estos luego son integrados al html limpio como formato de salida del documento final, los cuales pueden personalizarse desde el código de Phyton según el formato de la revista.

```
.autores{
     text-align: center;
     font-size: 13px;
     font-weight:bold;

}

  .afiliacion{
    text-align: center;
    font-size: 12px;
    display:block;
}

h1.tituloEs{
    color:#074666;
    font-family:Arial, sans-serif;
    font-size:23px;
    font-weight:bold;
    line-height:0.957;
    margin-bottom:25px;
    margin-top:25px;
    text-align:center;
    color:#2B3690;
}
```

Figura 6. Estilos CSS creados en Phyton con los nombres de los estilos formateados en Word

Para que el script logrará el proceso de transformación de los documentos se tuvo que cambiar los artículos de formato de dos columnas a una columna, luego se realizó l parametrización y formateo del documento .docx de los contenidos con los estilos personalizados en cada bloque de contenido.

Al iniciar el ciclo de repetición para transformar varios artículos, el proceso inicia con leer la carpeta de trabajo para identificar el número de documentos en formato .docx y el nombre de cada archivo. El ciclo de repetición contiene tres funciones que realizan la transformación de .docx a html, pdf e ePUB como se muestra en la figura 7, una vez se transforman los documentos se imprimen los nombres de los documentos generados y una barra de carga que muestra que los archivos han sido transformados sin ningún inconveniente. Si hay un error, el script no muestra el nombre del documento transformado. Es importante resaltar que el MS Word no debe estar abierto durante el uso del script porque Phyton envía un error de código que no está relacionado con MS Word peo si con el proceso de escritura del documento. El formato txt y audio en formato .mp3 no se incluye en este ciclo ya que es necesario que todos los documentos html se hayan generados previamente para hacer la transformación.

```
for archivo in archivos_texto:
    file= archivo
    file_html =  file.replace(".docx",".html")
    #crear html
    word_to_html(file, file_html)
    file_pdf =  file.replace(".docx",".pdf")
    file_pdf_ = carpeta  + file_pdf
    file_ = carpeta  + file
    #crear pdf
    convert(file_,file_pdf_) # barra de carga
    # funciona este
    file_epub =  file.replace(".docx",".epub")
    file_epub = carpeta  + file_epub
    #crear epub
    html_epub(file_html,file_epub)
    print(file_, file_html)
    print(file_, file_pdf_)
    print(file_, file_epub)
```

```
100% 1/1 [00:03<00:00, 3.11s/it]

../docs2/articulo_1.docx articulo_1.html
../docs2/articulo_1.docx ../docs2/articulo_1.pdf
../docs2/articulo_1.docx ../docs2/articulo_1.epub
```

Figura 7. Código en phyton que integra las funciones de transformación de docx a html, pedf e ePub

Al realizar el proceso de transformación de formatos, en la figura 7 se muestra la vista de los tres formatos transformados con la vista del artículo original en formato .docx. El script permitió hacer la transformación y la personalización utilizando las hojas de estilos el CSS creadas en Phyton manteniendo el estilo original del documento en MS Word, por lo que se emula la apariencia en los formatos, pdf, html e ePub, siendo no solo legibles, sino accesible, por lo tanto, su contenido se puede copiar. Estos documentos se lograron acceder a través de navegadores web, Adobe Acrobat, y el software Calibre para verificar su contenido.

Figura 8. Vista de los formatos de artículos de revistas transformado con el Script en Phyton

Para la prueba del tiempo en minutos de transformación de los diferentes formatos (pdf, html, epub, txt, audio) en la Tabla II se muestra el resultado de un solo artículo utilizando los softwares el cual en su totalidad conlleva 73 minutos para generarlo. La mayor cantidad de tiempo se utiliza en limpiar el html y CSS basura creado por Word al transformar y luego integrar una hoja de estilo CSS personalizada el documento html. En cambio, utilizando el script, el tiempo fue de siete minutos.

En la prueba de un volumen por cada revista, el tiempo de transformación de los 12 artículos utilizando los softwares fue de 894 minutos, aproximadamente 15 horas o dos días laborables de ocho horas, sin embargo, en esta prueba solo se transformaron documento de cuatro páginas por lo que el tiempo promedio para transformar un volumen completo de 12 artículos de más de 10 páginas es de 15 a 20 días. Utilizando el script en Phyton y los 12 artículos sin límites de páginas (12 páginas como promedio) el tiempo de transformación fue de 15 minutos, lo que muestra una mejora sustancial en minimizar el tiempo de procesamiento de estos formatos.

TABLA II
TIEMPO EN MINUTOS DE LA TRANSFORMACIÓN DE ARTICULO EN CUATRO FORMATOS UTILIZANDO SOFTWARES Y EL SCRIPT EN PHYTON

	un articulo		**12 artículos**	
Procesos	**Software**	**Script Phyton**	**Software**	**Script Phyton**
word a pdf	1	1	12	2
word a html	1	1	12	2
limpiar html/css	60	1	800	3
html a epub	3	1	20	2
html a txt	6	2	35	4
txt a mp3	2	1	15	2
	73	**7**	**894**	**15**

IV. CONCLUSIONES

Según datos mostrados de las revistas de Panamá y Centroamérica, es recomendable que las revistas científicas utilicen otros formatos apartes de los comunes pdf y html, como son el formato de audio (mp3), epub y el formato semántico XML-JATS.

El uso de software de escritorio y online en la transformación de artículos en diversos formatos de divulgación puede resultar más fácil para el editor, pero según los resultados el formato de salida no fue visualmente parecido al docx, aunque fue necesario hacer un proceso de formateo de los componentes del documento, no fue posible personalizar su apariencia por lo que se pierde el control de como se ve el artículo final.

La generación de esta iniciativa dio algunos elementos de apoyo en cuanto a la posible simplicidad que debe tener el artículo .docx para que este pueda ser transformado en otros formatos, como el hecho de tener que eliminar los encabezados visuales de las revistas, el cual puede ir en un Texto, pasar de dos columnas a una, transformar formulas a imágenes y normalizar los textos del contenido con estilos de carácter.

El uso de procesos automatizados utilizando lenguajes de programación como Phyton, no solo permite minimizar el tiempo de procesamiento sino también integrar códigos que personalicen la visualización de los formatos con poca intervención del rol humano, además es posible evaluar otras librerías del lenguaje para crear funciones que permitan generar nuevos formatos.

El tiempo de procesamiento del script en Phyton versus el software es sustancial según resultados ha sido menor y con una gran diferencia de 800 minutos, lo que reduce el tiempo de transformación y por ende el de divulgación de los artículos de revista, dando una visibilidad en menor tiempo a las artículos y evidencia a los autores del trabajo publicado.

Entre las limitantes del script están que es necesario hacer un proceso de formateo de la plantilla docx de la revista, que

puede no ser aceptado por los editores ya que es necesario formatear el documento con los nombres de los estilos predefinidos en Word antes de la transformación, sin este proceso el script no puede identificar un título de texto y un título de tabla. Si existen fórmulas como imagen en el texto, no es posible dejarlas incluidas en el texto por lo tanto las fórmulas deben estar separadas como imágenes. El formato html utilizado hasta ahora es de una sola página, es decir las imágenes están integradas en el documento y no como archivos independientes por lo que las imágenes no serían integradas en buscadores.

En cuanto a las limitantes técnicas es que el script debe ejecutarse en un entono de Phyton por lo que su uso es implica que quienes lo utilicen deban conocer este lenguaje.

Otra limitante es que las pruebas solo fueron realizadas a 24 artículos de dos revistas institucionales de la Universidad Tecnológica de Panamá y aunque su formato era diferente es una muestra poco representativa.

TRABAJOS FUTUROS

Se está trabajando en utilizar html con las imágenes independientes del texto, sin embargo, algunas imágenes no se pudieron transformar porque no fueron integradas en el documento original con los formatos (jpg, png, gif).

Se evalúa algunas librerías en Phyton para poder transformar los artículos de las revistas del formato HTML en formato en XML JATS, pero requiere que las etiquetas en este formato sean estudiadas con más detalle ya que no son tan genéricas como el HTML.

Para mejorar su uso se investiga como utilizar el código de Phyton en un entorno de escritorio con una interfaz más amigables de utilizar para los editores de revistas a través de una aplicación ejecutable y no tener que utilizar Jupyter Notebook.

REFERENCIAS

[1] P. Lifante Alonso, "Las revistas científicas en los repositorios Dialnet, E-Revistas, Infomine, Latindex, REDALYC y SCOPUS," *Tejuelo Rev. ANABAD Murcia*, no. 9, pp. 46–63, 2009, [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=3309666.

[2] A. D. Fernández, M. Ruiz-Corbella, and A. Galán, "Calidad editorial y científica en las revistas de educación. endencias y oportunidades en el contexto 2.0," *Rev. Investig. Educ.*, vol. 35, no. 1, pp. 235–250, 2017, doi: 10.6018/rie.35.1.244761.

[3] L. F. Morales Morante, "Producción e impacto de las revistas peruanas del ámbito de las Ciencias Sociales en el catálogo Latindex," *Investig. Bibl.*, vol. 30, no. 69, pp. 179–204, 2016, doi: 10.1016/j.ibbai.2016.04.017.

[4] E. Delgado López-Cózar, "Evaluar Revistas Científicas: Un Afán Con Mucho Presente Y Pasado E Incierto Futuro," *Rev. científicas situación actual y retos Futur.*, pp. 73–103, 2017, [Online]. Available: http://eprints.rclis.org/32132/.

[5] M. da L. Antunes, T. Sanches, C. Lopes, and J. Alonso-Arévalo, "Publicar en el ecosistema de la ciencia abierta," *Cuad. Doc. Multimed.*, vol. 31, p. e71449, 2020, doi: 10.5209/cdmu.71449.

[6] M. A. Rojas V. and S. Rivera Mena, "Guía de Buenas Prácticas para Revistas Académicas de Acceso Abierto," no. August 2011, p. 26, 2011, [Online]. Available: http://www.revistasabiertas.com/wp-content/uploads/Manual-Buenas_Practica_Revistas_Academicas.pdf.

[7] E. Vázquez, "El videoartículo: nuevo formato de divulgación en revistas científicas y su integración en MOOCs," *Comunicar*, vol. 41, no. 21, pp. 83–91, 2013.

[8] M. R. De Giusti, A. J. Lira, J. P. Rodríguez Vuan, and G. L. Villarreal, "Accesibilidad de los contenidos en un repositorio institucional," *e-Ciencias la Inf. ISSN-e 1659-4142, Vol. 6, Nº. 2, 2016, pág. 23*, vol. 6, no. 2, p. 23, 2016, [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=5575967%0Afile://C:/Users/cabal/Downloads/Dialnet-AccesibilidadDeLosContenidosEnUnRepositorioInstitu-5575967.pdf.

[9] E. X. Un and J. Z. Patiño, "Estandarización XML-JATS. Un modelo para la publicación en acceso abierto."

[10] L. F. M. Morante, "Visibility and impact of peruvian social science journals in open access," *Biblios*, vol. 65, no. 65, pp. 29–51, 2016, doi: 10.5195/biblios.2016.320.

[11] J. Veiga de Cabo and H. Martín-Rodero, "Open access: New models of scientific publishing in web 2.0 environments," *Salud Colect.*, vol. 7, no. SUPPL, pp. 19–27, 2011, doi: 10.18294/sc.2011.387.

[12] P. P. Palma, J. Benavides, and L. M. Saltos, "Los formatos bibliográficos en la redacción de textos científicos," vol. 5, pp. 62–71, 2020.

[13] M. josé G. José Manuel Barrueco Cruz, Cristina García Testal, "Una aproximación a las revistas científicas en Formato Electrónico," pp. 121–129, 2001, doi: 10.3989/egeogr.2001.i245.267.

[14] A. Ramírez Vega and F. Abarca Fedullo, "Publicación científica de revistas electrónicas en formato EPUB," *Taller Calidad, interoperabilidad, evaluación y certificación Repos. – el Present. y el Futur.*, pp. 297–306, 2017, [Online]. Available: http://sedici.unlp.edu.ar/handle/10915/63601.

[15] M. A. Penabad-Camacho, "Caracterización de las revistas de acceso abierto en la Universidad Nacional, Costa Rica," p. 2022, 2022.

[16] J. Estrella, D. Murillo, M. Férnandez, and H. Calderón, "Catálogo de Revistas de Panamá 2019," *Cons. Rectores Panamá*, 2019.

[17] A. A. R. G. Catalina Naumis Peña, "La Investigación Bibliotecología y del ainformación hacia el 2030: Desarollo Sostenoble," 2022.

[18] M. GarcÍA-GonzÁLez and L. B. Villalobos, "Experiencia De La Revista Saber En Su Transformación Del Formato Físico Al Digital," *Exp. J. Saber Its Transform. From Phys. To Digit. Format.*, no. 29, pp. 809–820, 2017, [Online]. Available: https://0-search.ebscohost.com.biblioteca-ils.tec.mx/login.aspx?direct=true&db=asn&AN=138073707&lang=es&site=ehost-live.

[19] J. J. M. Torres and R. M. A. Chinea, "Obtencion automatica de indices de calidad de sitios web." p. 2018.

[20] J. Tusons, "Introduccion al lenguaje Phyton," no. September 2018, p. 185, 2017, [Online]. Available: https://elibro.net/es/lc/uladech/titulos/56308.