

Preservación Digital de Páginas Web

Rondineli Gama Saad
SciELO Brasil

25/09/2024

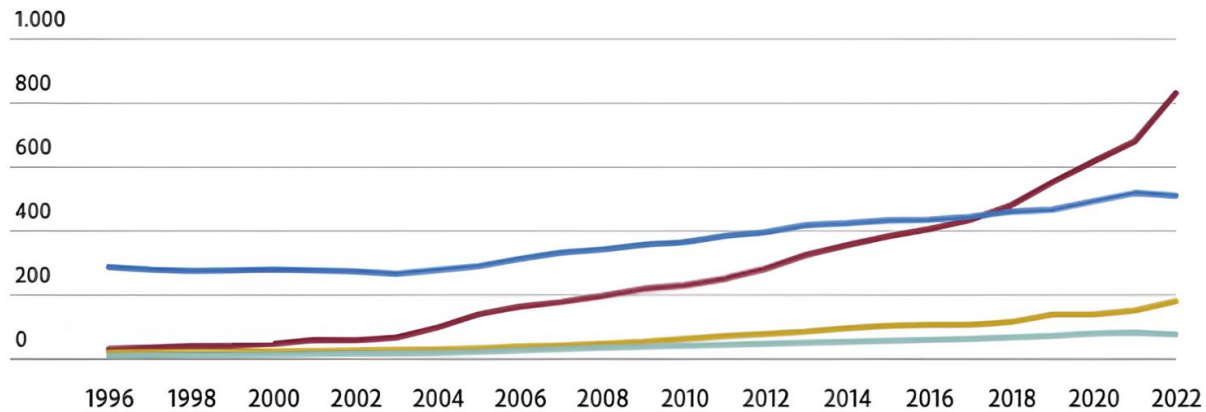


Publicación Acelerada

China, EUA e Índia são os que mais produzem artigos, enquanto Brasil é o 14º colocado; veja evolução dos 4 países

Quantidade de artigos, em milhares

- China
- EUA
- Índia
- Brasil



Fuente: Levantamento da Revista Bori e Elsevier (2023)

Producción de artículos de Acceso Abierto

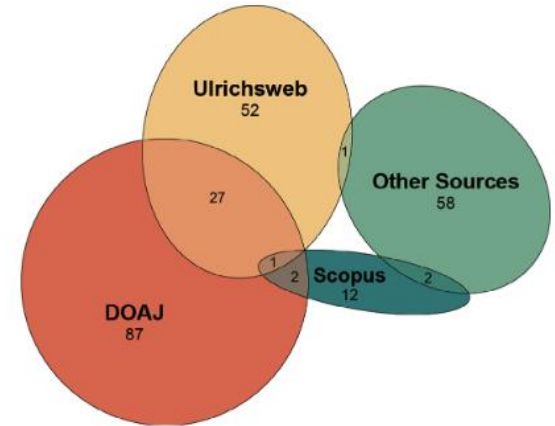
- 2,7 millones de artículos publicados por miembros de la OASPA (Open Access Scholarly Publishing Association) entre 2000 y 2022;
- El número de artículos publicados cada año creció aproximadamente 13 veces de 2011 a 2020.
- Más de 579 mil de ellos fueron publicados en 2020, lo que representa cerca del 28% en comparación con el año anterior;
- Alrededor del 82% de los artículos en acceso abierto se publican en revistas completamente de acceso abierto.

La importancia de preservar

COSTA, Miguel; GOMES, Daniel; SILVA, Mário J. **The evolution of web archiving.** *International Journal on Digital Libraries*, p. 1-15, 2016. Disponible en: <https://doi.org/10.1007/s00799-016-0171-9>

Laakso M, Matthias L, Jahn N. **Open is not forever: A study of vanished open access journals.** *J Assoc Inf Sci Technol.* 2021; 1–14. Disponible en: <https://doi.org/10.1002/asi.24460>

- Entre los años 2000 y 2019, 174 revistas de acceso abierto desaparecieron, de las cuales:
 - 154 revistas desaparecieron completamente;
 - Los autores detectaron que 900 estaban inactivas y en riesgo de desaparecer.



Fonte: Laakso M, Matthias L, Jahn N (2021)

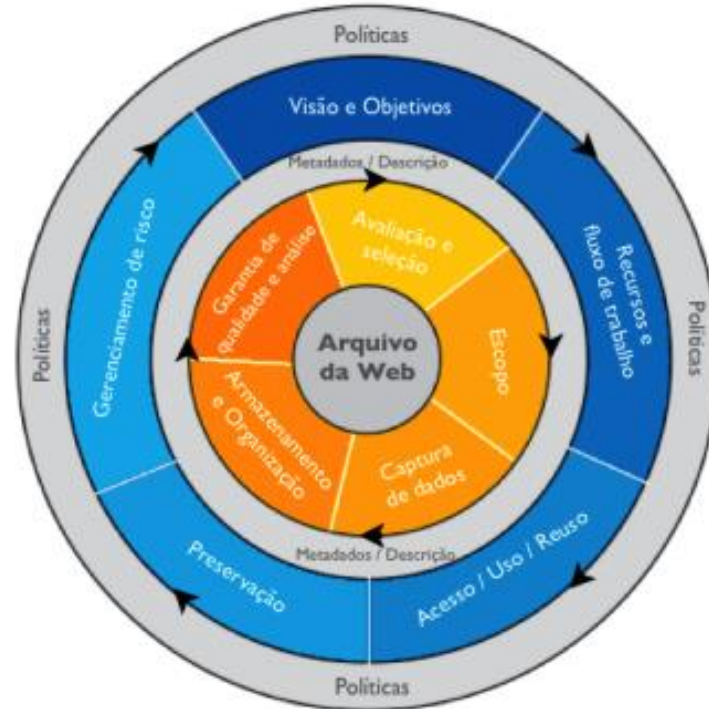
Archivado Web

El archivado web puede verse como un flujo de trabajo a través del cual los recursos son seleccionados, capturados, preservados y puestos a disposición del usuario final. Para Rockembach (2018), la selección y captura se realizan de forma continua, y esto puede tener en cuenta una serie de factores, como el contenido a ser capturado, si también se capturaron los enlaces externos al sitio seleccionado y la frecuencia de la recopilación. Para el autor, considerar los factores mencionados tiene como objetivo garantizar la calidad de la captura para asegurar un mejor archivado y recuperación de esa información.

Formato de Archivado

- El formato de archivo WARC (Web ARChive) ofrece una convención para concatenar varios registros de recursos (objetos de datos), cada uno compuesto por un conjunto de encabezados de texto simple y un bloque de datos arbitrario en un archivo largo.
- ISO 28500:2017 (WARC File Format)
- Extensiones abiertas .warc y .warcz
- Reproducido por software específico (wayback)

Modelo de ciclo de vida del archivado web



Flujo de archivado web

- Mapear y seleccionar los sitios web;
- Analizar sus contenidos para medir la "*archivabilidad*";
- Archivar los sitios web seleccionados, generando archivos del tipo WARC;
- Indexar los archivos WARC y usar software de reproducción de archivos de sitios web archivados;
- Comparar el contenido archivado con el sitio web seleccionado;
- Presentar las posibilidades de archivado web con énfasis en la preservación digital.

Flujo de archivado web



Selección

- Sitio web <https://antigo.ibict.br>;
- "Archivabilidad" del sitio medida por ArchiveReady:
 - Apariencia (layout de presentación)
 - Comportamiento

The screenshot displays the ArchiveReady website archiving tool interface. At the top, it says "ArchiveReady website archiving evaluation tool". Below that, it indicates "Checking website: http://antigo.ibict.br". There are navigation tabs for "Summary", "HTML and CSS 15", "HTTP 4", and "Media 18". The main content area shows an "Overall Rating" of 77%. To the right, there is a "Results" table with the following data:

Category	Rating
Web Archiving Facet	Rating
Accessibility	43%
Cohesion	93%
Metadata	100%
Standards Compliance	72%

Additional options visible include "One page printable HTML" and "EARL XML results".

Captura

- Heritrix
- Pywb
- wget (comando de Unix)
- [ArchiveWeb.Page](#) (extensión de Chrome)

Hazlo tú mismo



Webrecorder

Web archiving for all!

[Blog](#)

[Tools](#)

[Community](#)

[About](#)

[Contact](#)

[FAQ](#)

[Jobs](#)

The Webrecorder Project has developed many tools to help with web archive capture and replay.

There are tools for everyone who is interested in creating or replaying web archives on their own, and more advanced tools for developers, which provide tools for software developers to integrate into other workflows and require a bit more technical expertise.

ArchiveWeb.page



ReplayWeb.page



pywb



Browsertrix Crawler



Browsertrix Cloud



All Tools



[Tutorial ArchiveWeb.page](https://webrecorder.net/tools)

<https://webrecorder.net/tools>

Evaluación de los atributos del sitio web

- . Layout del sitio web
- . Audio o radio web
- . Herramienta de búsqueda
- . Videos de eventos, conferencias y reuniones
- . Imágenes ilustrativas para íconos/enlaces
- . Mapa del sitio
- . Interactividad con la agenda
- . Interoperabilidad con redes sociales
- . Menú de navegación
- . Presentación del feed de noticias
- . Banner rotativo

Layout del sitio web

The screenshot shows the top section of the website. At the top, there is a browser address bar with the URL `localhost:8080/ibict/20220117232052/https://antigo.ibict.br/`. Below the browser bar is a dark navigation bar with the text "Instituto Brasileiro de Informação em Ciência e Tecnologia - Página inicial" and the date "Mon, 17 Jan 2022 23:20:52 GMT". A secondary navigation bar contains "Portal do Governo Brasileiro" and "Atualize sua Barra de Governo". The main header area is blue and features the IBICT logo, the full name of the institute, and the acronym "MCTI". A search bar is located on the right side of the header. Below the header is a horizontal menu with links: "Página inicial", "Sobre o IBICT", "Cooperação Técnico-Científica", "Editais", "Oportunidades", "Carta de Serviços", and "Sala de Imprensa".



ASSUNTOS

- Informação para a Sociedade
- Informação para a Pesquisa
- Resúmenes de B...



Últimas notícias

Sugestão de Leitura: "Direitos das mulheres e a encontrabilidade da informação no portal da Câmara dos Deputados: perspectivas brasileiras rumo à Agenda 2030 das Nações Unidas"

Sugestão de Leitura: "Direitos das mulheres e a encontrabilidade..."

Plataformas de Archivado Web

- [Conifer Archive](#)
- [Arquivo.PT](#)
- [Arquiweb IBICT](#)
- [Archive-it](#)
- [NUAWEB](#)
- [Graúna](#)

Plan de Archivado Web

- 1. Defina los objetivos:** Determine por qué desea archivar el contenido web. Esto puede incluir la preservación de información histórica, garantizar el cumplimiento legal o capturar datos para investigaciones futuras.
- 2. Identifique el alcance:** Decida qué tipos de contenido desea archivar, como sitios completos, páginas específicas, redes sociales o blogs. También considere si desea archivar solo texto o también capturar imágenes, videos y otros tipos de medios.
- 3. Elija una herramienta de archivado:** Existen varias herramientas disponibles para ayudar en el archivado web, como Webrecorder. Investigue y seleccione la herramienta más adecuada para sus necesidades.
- 4. Establezca una estrategia de captura:** Determine la frecuencia con la que desea archivar el contenido web. Puede ser diariamente, semanalmente, mensualmente o en intervalos personalizados, dependiendo de la naturaleza del contenido y sus necesidades.
- 5. Defina un sistema de organización:** Desarrolle un sistema de categorización y metadatos para ayudar en la organización de los archivos. Esto puede incluir información como la fecha de archivado, la fuente original, el autor y otros datos relevantes.

Plan de Archivado Web

- 6. Implemente una estrategia de almacenamiento:** Elija un método adecuado para almacenar sus archivos. Esto puede involucrar el uso de servidores locales, servicios de almacenamiento en la nube o una combinación de ambos. Asegúrese de tener copias de seguridad adecuadas y medidas de seguridad para proteger los archivos.
- 7. Establezca políticas de acceso y preservación:** Defina quién tendrá acceso a los archivos archivados y durante cuánto tiempo se mantendrán. Considere los requisitos legales y éticos relacionados con la privacidad y la protección de datos.
- 8. Realice pruebas regulares:** Verifique periódicamente si el proceso de archivado está funcionando correctamente. Realice pruebas de recuperación para garantizar que los archivos puedan ser accesibles cuando sea necesario.
- 9. Monitoree los cambios en la web:** Tenga en cuenta que la web está en constante evolución, con sitios que se actualizan o se eliminan. Manténgase al tanto de los cambios y ajuste su plan de archivado según sea necesario.
- 10. Documente el proceso:** Registre todas las etapas de su plan de archivado, incluidas las políticas, herramientas utilizadas, estrategias de recolección y almacenamiento. Esto ayudará a mantener el plan consistente y facilitará futuras referencias.

