# WikiTeam, un proyecto para la preservación digital de wikis

Emilio J. Rodríguez-Posada
Wikis.cc
emijrp@gmail.com

29 de junio de 2017

## Resumen

Los internautas juegan hoy un papel decisivo en la generación de contenido. Existen soluciones para la preservación digital de la web, siendo la más destacada Internet Archive, pero se vuelven ineficaces a la hora de archivar wikis. En este artículo exploramos tanto los problemas que surgen como la falta de herramientas para preservar wikis y presentamos nuestra solución, el proyecto WikiTeam. Desde su creación más de 26.000 wikis han sido preservados, abriéndose nuevas posibilidades de investigación sobre estas comunidades en línea.

**Keywords:** preservación digital, wikis, mediawiki, archivos digitales, internet archive

## 1. Introducción

En 2016, Internet Archive[1] cumplió su 20 aniversario, el proyecto más ambicioso de preservación de la web. Cuenta con un acervo de 298.000 millones de páginas web, una colección que supera los 20 petabytes y que crece cada día. Internet Archive ha logrado generar conciencia entre los internautas de la importancia de la preservación digital de la web. Cualquier persona que navegue durante unos minutos por Internet es muy probable que se encuentre con lo que popularmente se conoce como enlace roto, o más técnicamente como error 404. Ya nadie duda de la fugacidad de los contenidos en la red, pues se estima que la vida media de una página en línea es de solamente 75 días,[1] y de la importancia de archivar la web para las generaciones futuras.

## 2. Preservación de wikis

## 3. Planteamiento

Los wikis son un caso especial de contenido web ya que además de textos e imágenes, también disponen de historiales con todas las versiones anteriores de cada página. Este histórico y sus metadatos (autores, fechas, comentarios, etc) son de suma relevancia, no solo para mantener la información acerca de la autoría de los contenidos, sino de cara al estudio de su evolución y del comportamiento de la comunidad. Asimismo, los textos están escritos usando una sintaxis que varía según cada motor wiki y que incluye una rica información sobre enlaces entre páginas, inserción de imágenes y estilo.

Todo este contenido se encuentra en continuo riesgo de desaparición. Wikis que son abandonados por sus autores, administradores que descuidan el mantenimiento del servidor, dominios que caducan, ataques de vándalos y spammers o fallos de software y hardware, son algunas de las causas que hacen que peligre la integridad de los datos. Con la excepción de Wikipedia[2] y algún caso aislado como WikiTravel,[3] decenas de miles

[1] https://archive.org

[2] https://dumps.wikimedia.org

[3] https://code.google.com/archive/p/oxygenpump/

de wikis no ofrecen copias de seguridad completas y públicas a sus usuarios. El copiar manualmente los textos e imágenes de un wiki, incluso si es pequeño, es una tarea ardua; en el caso de wikis de tamaño medio o grande es una tarea impracticable.

No es la intención de este documento recopilar un listado de wikis desaparecidos ni adentrarse en cada caso, eso podría trabajo de futuras investigaciones, pero podemos mencionar un ejemplo para comprobar la importancia de disponer de backups. La *wikifarm* ScribbleWiki perdió todos sus wikis por un problema técnico con el servidor y el sistema de copias de seguridad. A pesar de las esperanzadoras primeras palabras por parte de los administradores, que aseguraban que el servicio volvería a estar pronto en línea, los usuarios jamás volvieron a saber de sus wikis.[4]

### 3.1. Problemas y soluciones

Las iniciativas de preservación web existentes como Internet Archive o WebCitation[5] no son capaces de extraer completamente y almacenar los historiales, metadatos y sintaxis, ya que tratan las páginas del wiki como páginas web normales, guardando simplemente el código HTML mostrado por el sitio en vez del contenido original a partir del cual el servidor genera dicho HTML. A consecuencia de esto, la preservación de wikis se venía realizando con muchas dificultades y severas omisiones que daban lugar a archivos muy incompletos y poco usables (véase Cuadro 1).

Pero no todo iban a ser problemas, a diferencia del resto de sitios web, los wikis suelen publicarse con algún tipo de licencia libre como GFDL o Creative Commons y sus variantes, por lo que no existe ningún obstáculo legal a la hora de preservar los contenidos y redistribuir las copias. Solo era necesario que alguien desarrollara el software adecuado.

Hubo algún intento de resolver ese vacío de herramientas para archivar wikis, como el proyecto Urobe[2] que no pasó de ser un prototipo al quedarse sin financiación. La creciente cantidad de contenido wiki disponible en Internet convertía la preservación de wikis en un problema abierto y con bastantes particularidades dentro del área de la preservación web, requiriendo de soluciones específicas y eficaces. Para dar solución a ello se fundó el proyecto WikiTeam.

Cuadro 1: Comparación de los datos extraídos de wikis según la herramienta utilizada

| Datos | Int. Archive | WikiTeam |
|---|---|---|
| **Historial** | No / Parcial | Completo |
| **Imágenes** | Miniaturas | Máx. resolución |
| **Metadatos** | No / Parcial | Completo |
| **Sintaxis** | No / HTML | Sí |
| **Formato** | HTML | XML |
| **Importable** | No | Sí |

## 4. WikiTeam

WikiTeam[6] es un proyecto para la preservación digital de wikis. Sus miembros desarrollan software libre que permite exportar los contenidos de los wikis (textos, historiales, metadatos e imágenes) y almacenarlos en formatos estándares y estructurados como XML. Hasta el momento sus esfuerzos se han concentrado en MediaWiki,[7] el motor wiki más extendido, aunque se planea añadir soporte para otros motores.

### 4.1. Backups individuales

El software está desarrollado en lenguaje Python y funciona a través de consola de comandos, siendo compatible con sistemas operativos Windows y GNU/Linux. El programa recibe la URL del wiki a preservar, ya sea su página principal o la dirección de la API, y tras extraer el listado de todas las páginas del wiki e imágenes, se dispone a exportarlas en XML utilizando la función `Special:Export` de MediaWiki.

[4] http://wikiindex.org/ScribbleWiki
[5] https://www.webcitation.org
[6] https://github.com/WikiTeam
[7] https://www.mediawiki.org

Existen opciones para seleccionar conjuntos de páginas, por si el usuario solo desea descargar aquellas que se encuentren en cierto espacio de nombres, así como la posibilidad de extraer solamente la versión actual de cada página o el historial completo. Sea como fuere, el resultado es un único XML en el que se encuentran fusionados los historiales de las páginas seleccionadas. En el caso de las imágenes, el software las extrae a máxima resolución y también obtiene la página de descripción que suele incluir información acerca de la autoría y licencia.

Una orden típica para archivar un wiki al completo, tanto páginas como imágenes con historiales completos, sería la siguiente:

```
python dumpgenerator.py
http://wiki.domain.org --xml --images
```

Existe un tutorial completo con todas las opciones disponibles.[8]

### 4.2. Backups por lotes

Dado que muchos wikis estaban desapareciendo de Internet sin que quedaran copias de seguridad de sus contenidos, los miembros de WikiTeam pensaron en generar listados de wikis y descargarlos periódicamente para su preservación. Así surgió la necesidad de desarrollar un módulo que permitiera lanzar el software sobre un lote de wikis.

Generando previamente la lista en un fichero de texto, la orden para lanzar el backup por lotes sería la siguiente:

```
python launcher.py lista-de-wikis.txt
```

El software recorrerá todos los wikis de la lista por orden, archivándolos al completo (XML e imágenes) y realizando algunas tareas de comprobación de integridad para verificar que los datos se han bajado correctamente. Cualquier error encontrado será mostrado por pantalla.

Cuadro 2: Distribución de idiomas sobre el total de 26.000 wikis archivados

| **Idioma** | **Wikis** | **% del total** |
|---|---|---|
| Inglés | 16.286 | 62 % |
| Alemán | 1.644 | 6 % |
| Español | 955 | 3 % |
| Ruso | 882 | 3 % |
| Francés | 697 | 2 % |
| Holandés | 246 | <1 % |
| Chino | 237 | <1 % |
| Italiano | 214 | <1 % |
| Japonés | 205 | <1 % |
| Portugués | 192 | <1 % |
| Polaco | 169 | <1 % |
| Checo | 130 | <1 % |
| Finlandés | 116 | <1 % |
| Sueco | 104 | <1 % |
| Otros | 3.923 | 15 % |

## 5. Resultados

Las herramientas creadas por WikiTeam cubren un importante hueco existente en el área de la preservación digital de este tipo de sitios web. Lo hacen maximizando el contenido y los metadatos recuperados para cada wiki y de una manera escalable que permite el archivado por lotes de miles de sitios.

Prueba de ello son los más de 26.000 wikis preservados, así como varias *wikifarms* y 34 terabytes de imágenes de Wikimedia Commons, que han sido publicados en una colección específica en Internet Archive.[9] Un análisis de la distribución de idiomas de los wikis archivados arroja los siguientes resultados (véase Cuadro 2).

El contenido preservado representa un enorme conjunto de datos de la *wikiesfera*, con un incalculable valor histórico y un gran potencial para la investigación. Como prueba de ello, el estudio más amplio del que se tiene noticia tuvo en cuenta tan solo 151 wikis distintos.[3] En la actualidad, WikiTeam ha puesto a disposición de la comunidad investigadora más de 100 veces ese

[8] https://github.com/WikiTeam/wikiteam/wiki/Tutorial [9] https://archive.org/details/wikiteam

número.

## 6. Trabajo futuro

Entre las líneas de trabajo que se presentan destacan la expansión a otros motores wiki (quizás DokuWiki) y mejorar la cobertura de las listas de wikis además de mantenerlas actualizadas. Asímismo es necesario seguir produciendo backups periódicos dado que el contenido de los wikis siguen creciendo.

Como efecto colateral, los backups generados pueden ser utilizados a su vez como conjuntos de datos para investigar el comportamiento de estas comunidades en línea. Prácticamente todas las investigaciones realizadas se han centrado en Wikipedia, Wiktionary y más reciéntemente Wikidata. Poco se ha estudiado el resto de wikis que componen la *wikiesfera*, seguramente en gran medida por la dificultad de acceder a los datos de manera estructurada. Ahora es posible.

## Agradecimientos

Agradecemos el trabajo realizado por los voluntarios de WikiTeam, desde los más activos hasta los esporádicos, que han ayudado reportando errores, enviando sugerencias, mejorando la documentación, haciendo pruebas y ejecutando los scripts para generar los miles de backups desde sus hogares o servidores.

## Licencia



## Referencias

[1] Steve Lawrence, David M. Pennock, Gary William Flake, Robert Krovetz, Frans M. Coetzee, Eric Glover, Finn Årup Nielsen, Andries Kruger, and C. Lee Giles. Persistence of web references in scientific research. *Computer*, 34, February 2001.

[2] Niko Popitsch, Robert Mosser, and Wolfgang Philipp. Urobe: a prototype for wiki preservation. In *IPRES 2010*, September 2010.

[3] Jeff Stuckman and James Purtilo. Measuring the wikisphere. In *WikiSym '09*, 2009.