Todo lo que siempre quisieron saber de la CDPedia y nunca se atrevieron a preguntar

Facundo Batista

¿Qué es?

La CDPedia es la Wikipedia Offline. O sea, la Wikipedia, lo más fiel posible a su formato y contenido original, pero armada (construida, compactada) de una manera que no se necesita nada de Internet para acceder a toda la info de la misma.

Se llama CDPedia porque la idea original era tenerla en un CD. Hoy por hoy se generaron cuatro imágenes de cada liberación de CDPedia: un CD, un DVD, y dos archivos comprimidos (uno mediano y otro grande) que se pueden poner en un pendrive o en cualquier disco rígido.

La CDPedia es multiplataforma: el mismo CD, DVD o archivo comprimido se puede usar en Linux, Windows, o Mac, sin necesitar nada instalado previamente por fuera de lo que cada sistema trae normalmente.

¿Cómo surgió?

El proyecto arrancó en el sprint posterior al primer PyDay de Santa Fé, en Junio del 2006, con la idea base de poder distribuir la Wikipedia a aquellos lugares que no tenían o tienen acceso a Internet (en particular teníamos en mente a escuelas de frontera o de ciudades chicas, bibliotecas de barrio, centros culturales de pueblos pequeños, etc.).

El proyecto continuó siempre, y aunque no siempre se le pudo dedicar tiempo. Las mejoras en el proyecto fueron paulatinas. Se destaca que fueron casi 30 personas quienes colaboraron en el proyecto a lo largo de los años.

Se trabajó mucho en este proyecto durante los PyCamps (los dos en Los Cocos, el de Verónica, y el de La Falda), donde muchas personas le dedicaron un buen tiempo, y también se realizó bastante durante otras reuniones, especialmente durante el 2010 y 2011.

A modo de ejemplo, dos sprints: uno fue en un incipiente hacklab, donde se experimentó mucho sobre el índice para las búsquedas, y también durante la fundación de Wikimedia Argentina, donde se presentó por primera vez el proyecto y se realizó un gran avance en la primera parte del procesamiento de datos.

Una gran característica de la CDPedia, indiscutiblemente el proyecto más grande y más largo de Python Argentina, es que siempre se mantuvo orientado a los mismos objetivos: tener una Wikipedia offline con fines sociales (distribuir en escuelas sin conexión a Internet, que el conocimiento sea libre, etcétera), que sea divertido de hacer (es decir, hacerlo en Python), y mantenerlo libre (no sólo el producto final, que recomendamos copiarlo y repartirlo, sino el código en sí).

¿Cómo se usa? ¿Cómo se ve?

Una decisión estratégica de la CDPedia es tomar el HTML generado por los servers de Wikipedia y usarlos casi directamente. Exploramos en algún momento tomar la info de la base de datos directamente, pero no logramos generar una página web igual a la de Wikipedia online.

Y eso es una fortaleza de la CDPedia: por la manera en que se armaron las páginas, la forma de ver y usar las páginas, de explorar y acceder a la información, es igual a la Wikipedia

online, de manera que el usuario no tiene un costo cognitivo en pasar de la versión online a offline. Es más, también se puede considerar a la CDPedia como el paso previo de consumo de contenido a la Wikipedia: una persona se puede acostumbrar a explorar las páginas, leer, cruzar y criticar la información, etc, y recién cuando tiene todo armado va a la Wikipedia Online y al resto de Internet para completar su investigación.

Más allá de la página a nivel contenido, lo que sí se modificó mucho es la barra de la izquierda. No tiene la original de Wikipedia, porque no tiene sentido al ser todo offline, así que se reemplazaron los botones y enlaces por otros: hay un botón para ver una página al azar, un campo de texto de búsqueda, el logo de CDPedia, el logo de PyAr, enlace a una página de ayuda, etc.

Algo que también se modificó bastante es como se señalizaron los enlaces en la página misma, en el contenido de Wikipedia. Hay principalmente tres tipos, distinguibles en cómo se decoró el texto convertido en enlace:

- Azul: un link normal, apunta a otra página de Wikipedia que se incluyó dentro de CDPedia.
- Rojo, subrayado con guiones: un enlace a otra página de Wikipedia pero que no fue incluida en CDPedia por razones de espacio.
- Azul, subrayado con guiones: un link que los sacaría de CDPedia, ya que apunta a recursos online (útiles solamente si tenés Internet, claramente).

fue desde entonces tomado de manera generalizada para denominar al período.
3. ↑ La cifra exacta está sujeta a debate: la CONADEP registró 8961 casos, mientras que otros organismos de derechos humanos elevan la cifra a 30 000. El número de las indemnizaciones otorgadas por el Estado a familiares de desaparecidos llega a 13 000.
4. ↑ La Ciudad de Buenos Aires es una entidad de segundo grado constitucional, pero no organizada como provincia sino según un régimen especial (Ciudad Autónoma), similar y equiparable al propio de provincia.
5. ↑ El dato incluye la parte del campo de hielo Patagónico Sur que Argentina disputa con Chile.
6. ↑ [a] [b] No incluyen 980 874 km² de la Antártida Argentina y las Islas del Atlántico Sur, que están bajo administración del Reino Unido, por los que Argentina reclama soberanía, totalizando una superficie de 3 761 274 km².
7. ↑ De acuerdo al informe de la WWF, Argentina solo es antecedida en riqueza natural y biodiversidad por los países continentes de Brasil, China, EE.UU. Rusia, India, Australia y por la, mediana en extensión, Indonesia.
8. ↑ Entre 2003 y 2013 las siguientes boxeadoras argentinas obtuvieron títulos mundiales: Marcela Acuña, Mónica Acosta, Fernanda Alegre, Yésica Bopp, Carolina Duer, Érica Farías, Carolina Gutiérrez Gaite, Yésica Marcos, Alejandra Oliveras, Patricia Quirico y Claudia Andrea López.

Referencias

1. ↑ [a] [b] [c] «Superficie de la República Argentina» (XLS). *Territorio/geografía*. Instituto Nacional de Estadística y Censos. Consultado el 19 de junio de 2008.
2. ↑ «Estimaciones y proyecciones de población 2010-2040. Total del país». *INDEC*.
3. ↑ [a] [b] «Total del país. Población total y variación intercensal absoluta y relativa por provincia o jurisdicción. Años 2001-2010» (XLS). *INDEC*. Consultado el 1 de septiembre de 2011.
4. ↑ [a] [b] «Report for Selected Countries and Subjects» (en inglés). FMI. Consultado el 5 de noviembre de 2014.
5. ↑ «Data» (en inglés). Consultado el 5 de agosto.
6. ↑ «GDP per capita (current US$)» (en inglés). Consultado el 5 de agosto.
7. ↑ PNUD, ed. (25 de julio de 2014). «Informe sobre Desarrollo Humano 2014» (PDF) (en ingles). Washington, Estados Unidos. Consultado el 25 de julio de 2014.
8. ↑ UNDP, ONU, ed. (24 de julio de 2014). «Table 2:

177. ↑ «La deuda pública es la más baja respecto del PBI en los últimos diez años». diariobae.com. (enlace roto disponible en Internet Archive; véase el historial y la última versión).
178. ↑ Ministerio de Economía. Serie histórica correspondiente al PBI y la deuda externa.
179. ↑ «La recaudación en provincias en el primer semestre». Consultado el 18 de mayo de 2014.
180. ↑ «Presión tributaria uruguaya es la tercera más alta de la región». elobservador.com.uy. Consultado el 18 de mayo de 2014.
181. ↑ Ministerio de Economía. Informe Trimestral Nº 67, "Finanzas Públicas", 2008.
182. ↑ «El FMI presenta una 'moción de censura' a Argentina por fallos en sus estadísticas». Consultado el 2 de febrero de 2013.
183. ↑ «Argentina acusa al FMI de 'trato desigual' y anuncia un nuevo índice para la inflación». Consultado el 2 de febrero de 2013.

Otra sección que se modificó es el pie de cada página: se pone un enlace a la misma página pero online, en Wikipedia misma, por si el usuario necesita la información actualizada.

Cabe mencionar que la CDPedia funciona también en Modo Servidor. De esta manera, se puede instalar la CDPedia en el servidor de una escuela, y que todas las computadoras del establecimiento puedan usar la información desde allí. Así logramos el efecto deseado de que los chicos puedan tener acceso al contenido de Wikipedia sin realmente tener que tener Internet, pero sin la complicación o el incordio de tener que instalar CDPedia en cada una de las computadoras.

¿Y qué contenido tiene?

El contenido de la CDPedia está fuertemente determinado por dos características intrínsecas del proyecto: la CDPedia es estática y fácilmente distribuible en un disco o pendrive.

Decimos que la CDPedia es estática porque una vez armada, no se actualiza. Por eso, como "fotografía de un momento de Wikipedia", por definición siempre va a estar desactualizada.

Cuando se comienza a generar una nueva versión de la CDPedia, se baja todo el contenido de Wikipedia y se empieza a procesar. Este procesamiento puede llevar varias semanas, incluso un par de meses. Entonces, cuando se libera una nueva versión de CDPedia, no incluye todos los cambios desde que se empezó a procesar.

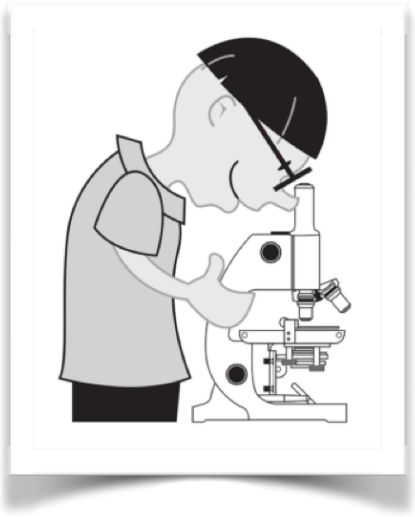
Es por esto que se trata de liberar CDPedias al menos una vez por año, para que contenga todo lo último.

La CDPedia se puede distribuir fácilmente: sólo hace falta grabar un CD o DVD, o incluso pasarse los archivos mediante un pendrive. En casi todas las versiones (menos la más grande), por una cuestión de formato, no entra todo el contenido de la Wikipedia. Por ejemplo, para la versión 0.8.3, tenemos lo siguiente:

- CD (693 MB): 54 mil páginas y 5% de las imágenes
- Tarball medio (3.6 GB): 400 mil páginas y 20% de las imágenes
- DVD (4.3 GB): Todas las páginas y 8% de las imágenes
- Tarball grande (8.7 GB): Todas las páginas y todas las imágenes

Entonces, a menos que se arme un tarball grande, es evidente que tenemos que decidir cuáles páginas e imágenes van a entrar, y cuales van a quedar afuera.

Esa decisión se toma ordenando todas las páginas por un determinado puntaje (que explico abajo), y se eligen las primeras N páginas (para el ejemplo anterior, las primeras 54 mil para el CD, las primeras 400 mil para el tarball medio, etc). Esas páginas tienen a su vez imágenes, que naturalmente también quedan ordenadas por el puntaje de las páginas: se toma un primer porcentaje de imágenes que se incluyen al 100%, otro porcentaje de imágenes que se escalan al 75%, otro porcentaje de imágenes que se escalan al 50%, y el resto no se incluye.



La selección de las páginas

La selección de la páginas que estén incluídas en CDpedia es un tema clave. Por lo que se trata de darle un puntaje a las mismas. Este puntaje está formado (hoy por hoy) en base a dos factores: levemente por el largo de la página (una página larga tiene más puntaje que una corta), y fuertemente por lo que llamamos "peishranc", que es la cantidad de otras páginas que enlazan a la que estamos evaluando (entonces, si a una página se la menciona en otras mil páginas es mucho más importante que una página que casi no se la menciona en el resto de la Wikipedia).

¿Qué hacemos para evitar el vandalismo?

Cuando comienza el proceso de generar una nueva versión de la CDPedia, se bajan todas las páginas de Wikipedia, ¡pero no siempre bajamos la última versión! Lo que se hace es revisar cuándo fue modificada y por quién: si fue modificada por un usuario normal, perfecto; pero si fue modificada por un usuario anónimo (como sucede en la mayoría de las vandalizaciones) se fijan cuando fue modificada: si fue hace más de varios días, se incluye (se asume que los bibliotecarios de Wikipedia ya tuvo tiempo de verificar el cambio), pero si es muy reciente se evita la última versión de la página, y se utiliza la versión anterior (y se aplican nuevamente todos esto.

Estado actual del proyecto

El proyecto avanza, pero lento. Uno de los objetivos del proyecto es lograr la internacionalización de la CDPedia. Cuando esté terminado, se van a poder crear CDPedias no sólo a partir de la Wikipedia en español, sino también de la Wikipedia en otros idiomas: portugués, aymara, guaraní, alemán, ruso, etc.

El otro cambio es más bien la construcción de una infrastructura en particular. La idea es tener una generación continuas de CDPedias, que se arme la CDPedia en español, y automáticamente luego se arme la de otro idioma, y otro, y otro, y otro, y luego de varios meses, vuelva a arrancar con la de español.

Pero hay muchas cosas para hacer.
Unos chicos en un PyCamp hicieron una app para Android que, luego de copiar los datos a mano, correría la CDPedia en cualquier teléfono o tablet.

¿Que necesita CDPedia?

Programadores con ganas de trabajar y aprender, tiempo de programador para continuar llevando este proyecto tan interesante y valioso por buen camino. Si tienen ganas de participar de cualquier manera, lo principal es que se pongan en contacto con el grupo en general, a través de esta lista de correo o del foro asociado (son espejo uno del otro, usen el que sientan más cómodo):

- La dirección de mail de la lista es: cdpedia@googlegroups.com
- Y la URL del foro es: https://groups.google.com/forum/#!forum/cdpedia