Manejo de Datos Abiertos en la Agricultura y Nutrición

Este curso de aprendizaje digital (e-learning) es el resultado de una colaboración entre socios de GODAN Action, incluyendo a ***Investigaciones Ambientales Wageininen (WUR)***, ***AgroKnow***, ***AidData***, la ***Organización de las Naciones Unidas para la Alimentación y la Agricultura*** (FAO por sus siglas en Ingles), ***El Foro Global sobre Investigaciones Agricultures*** (GFAR), y el ***Instituto de los Estudios del Desarrollo*** (IDS), T***he Land Portal***, ***el Instituto de Datos Abiertos*** (IDI) y el ***Centro Técnico de Agricultura y cooperación Rural*** (CTA).

GODAN Action es un proyecto de tres años [por] el Departamento del Desarrollo Internacional del Reino Unido para capacitar a los que usan, producen, e intermediarios de datos para conectarse efectivamente con datos abiertos y maximizar la potencial por su impacto en los sectores de agricultura y nutrición. En particular, trabajamos para mejorar la capacitación, promover estándares comunes y mejores prácticas para medir el impacto. [www.godan.info]

GODAN ACTION
Global Open Data for Agriculture & Nutrition

UKaid
from the British people

Unidad 2: Usando los datos abiertos

Lección 2.1: Descubriendo a datos abiertos

[Imagen]

Foto por Niel Palmer (CIAT) licenciado bajo CC BY-SA 2.0

Objetivos y metas de aprendizaje

Esta lección propone proveer un fundamento en cómo descubrir y acceder datos que están disponibles en la red. Desde descargas de datos y proveedores de servicios de datos, a publicadores de datos abiertos relacionados, esta lección cubrirá el kit de herramientas completo para obtener los datos correctos de la red más rápido.

Después de estudiar esta lección debes saber como:

- Enumerar diferentes tipos de servicios que proveen acceso a datos abiertos
- Enumerar diferentes tipos de métodos por lo cual se puede acceder datos de estos servicios
- Explicar la diferencia entre estos tipos de servicios
- Usar los diferentes tipos de servicios para acceder datos abiertos
- Descubrir datos descargables
- Descubrir datos escondidos
- Identificar si una fuente de datos es abierta para estos tipos de servicios
- Describe las ventajas y desventajas de estos tipos de servicios.

Contenido

1. Introducción

Como la web ha evolucionado, así también su uso continuo para compartir recursos y datos más complejos, pero reta a paradigmas existentes. La web mundial (*www* o *World Wide Web* en inglés) era central para datos e información y es visto como un espacio donde se accede otros documentos u otros recursos en la web (World Wide Web).

Viendo a la web en esta manera ha resultado en el desarrollo de una red o "web" (telarañas) de documentos o páginas web como se conoce comúnmente, diseñado para las personas para acceder y leer. Existen alrededor de 4.75 mil millones de ellos [http://www.worldwebsize.com]. Muchos de los documentos están relacionados o vinculados, y esas conexiones agregan valor. En un blog, artículo del periódico o artículo académico, podemos usar estos vínculos o enlaces para añadir discusiones previas o apuntar a recursos factuales. Esto nos ayudan a explorar la red o "web" de documentos.

2. Las generaciones de la web[1]

Había mucho menos documentos en los días tempranos de la web, pero la gente todavía necesitaba hallar cosas. Los primeros esfuerzos para mantener un índice por mano[2] fueron realizados por Sir Tim Berners-Lee, presidente y co-fundador de la ODI (Instituto de Datos Abiertos)[3]. Gente podía ir a la lista y saltar a páginas que le parecían interesantes.

Entonces crearon portales como DMOZ[4] y el antiguo Yahoo![5] Estas eran listas ordenadas de sitios web y páginas organizadas por tema. Cuando crecía la web, estos portales ya no eran viables y la gente transicionaron a los buscadores de metadatos, como Altavista[6] y Lycos[7], que usaron metadatos que fueron definidos en el sitio web que proveía información sobre el documento.

El método de búsquedas era más flexible porque las páginas se podrían hallar automáticamente, pero los resultados no eran confiables y se podían manipular fácilmente. La próxima generación de búsquedas en la web vino con el estilo de búsqueda de *PageRank*[8] (rango de página), como Google, que usó muchas más pistas para realizar una búsqueda, incluso al entendimiento del contenido, utilización y uso de enlaces. Esta tercera generación aprendió como buscar dentro de la red de documentos para descubrir cuan relevante cada documento seria para usuarios.

Es vital que aprendamos como usar estos diferentes tipos de búsqueda. La red de documentos no podía escalar hasta que la búsqueda[9] se convirtió en la técnica principal de descubrimiento. Todos estos métodos todavía existen, y cumplen varios roles, uno de ellos siendo poner datos abiertos en la web.

---

[1] https://theodi.org/blog/we-need-to-learn-how-to-search-the-web-of-data
[2] https://www.w3.org/History/19921103-hypertext/hypertext/DataSources/WWW/Servers.html
[3] https://theodi.org/team/timbl
[4] https://www.dmoz.org
[5] https://en.wikipedia.org/wiki/Yahoo!_Directory
[6] https://en.wikipedia.org/wiki/AltaVista
[7] https://en.wikipedia.org/wiki/Lycos
[8] https://en.wikipedia.org/wiki/PageRank
[9] http://onlinelibrary.wiley.com/doi/10.1002/1097-4571(2000)9999:9999%3C::AIDASI1607%3E3.0.CO%3B2-F/full

3. Las generaciones de datos abiertos en la web

Las técnicas tempranas de publicación de datos reflejan las mismas generaciones que los de documentos en la web. Primero, creamos portales, como dat.gov.uk, opendata.go.tz y data.sncf.com. Estos son catálogos ordenados de conjuntos de datos preparados por temas particulares o estructuras organizacionales.

Mientras que la cantidad de datos continúa creciendo, agregadores de datos empiezan a aparecer para proveer servicios para facilitar el descubrimiento de datos relacionados a temas particulares o regiones. Por ejemplo, hay enigma.io[10], el portal de datos de Europa[11] y el API transport[12]. Tales servicios dependen de la disponibilidad de metadatos de otros portales, páginas web y proveedores de servicios para proveer la información sobre los datos y como accederlos.

La tercera generación de buscadores de la web para datos todavía está muy centrada en la web como un espacio de información, así que tenemos que depender en los mismos buscadores para encontrar datos, al igual que información. El desarrollo de buscadores para datos todavía está en su infancia y es en gran parte relacionado a los métodos usados para publicar datos en (o dentro de) la web.

4. Datos son solamente uno de varios recursos en la web

Mientras que la web evolucionaba, empezó a convertirse en un lugar donde compartir recursos de multimedia. La inclusión de imágenes, audio y video abrió nuevo potencial para proveer servicios como los de *streaming* (flujos o corrientes de datos). Audio era el pionero, donde paginas como last.fm utilizaba metadatos sobre pistas de música para crear recomendaciones personalizadas para usuarios. Esta tecnología precede y forma la fundación para mucho de los sistemas de recomendaciones utilizados hoy en día. Last.fm daba recomendaciones, pero no permitía que el usuario escuchara la música recomendada. Esta funcionalidad no salió hasta 3 años después en 2005 con el lanzamiento de Pandora, una aplicación de emisora de radio personalizado. 3 años después se lanzó Spotify, que era el primer servicio de *streaming* que usó tecnologías de la web para proveer una plataforma dedicada fuera de un buscador de la web.

Avanza unos años y la web e internet en actualidad son la plataforma de servicio para una cantidad gigantesca de recursos diferentes. Dentro del buscador de la web, servicios como Google han agregado la capacidad de buscar por multimedia específicamente para encontrar imágenes y videos, mientras que portales específicos como Yahoo y YouTube ahora proveen acceso por la web en la página web o por aplicaciones o televisores conectadas al internet.

[10] https://www.enigma.com
[11] https://www.europeandataportal.eu
[12] https://www.transportapi.com

Lo mismo no es cierto para datos. Buscadores todavía no tienen búsquedas específicas para datos, y tal vez nunca lo tendrán. Pero encontrar datos en la web puede ser difícil, y empieza con la definición de datos en sí.

¿Qué son datos?

Una imagen es una representación visual de algo (una foto). Un archivo de audio hace un sonido cuando se abre. Un video combina múltiples imágenes con audio para hacer una foto moviendo. Datos son… difícil de definir y por eso realizar una búsqueda por ellos.

Los datos son en el nivel más básico de abstracción de donde se deriva la información y conocimiento. Estos son términos abstractos y, por lo tanto, datos podrían ser una imagen, una hoja de cálculo, o un archivo de audio. Adicionalmente, si la web es un espacio para información, ¿son los datos algo subyacente o más fundamental?

Tradicionalmente, datos son entendidos como una hoja de cálculo o conjunto de números que se puede analizar de alguna forma. En la web, tales datos se suelen compartir a través de portales de datos simplemente en forma de un archivo que se puede descargar. Algunos portales proveen una funcionalidad como de YouTube donde los datos se pueden explorar sin descargar, mientras que los datos en sí todavía son un recurso estático, subido y disponible para ser descargado por otro.

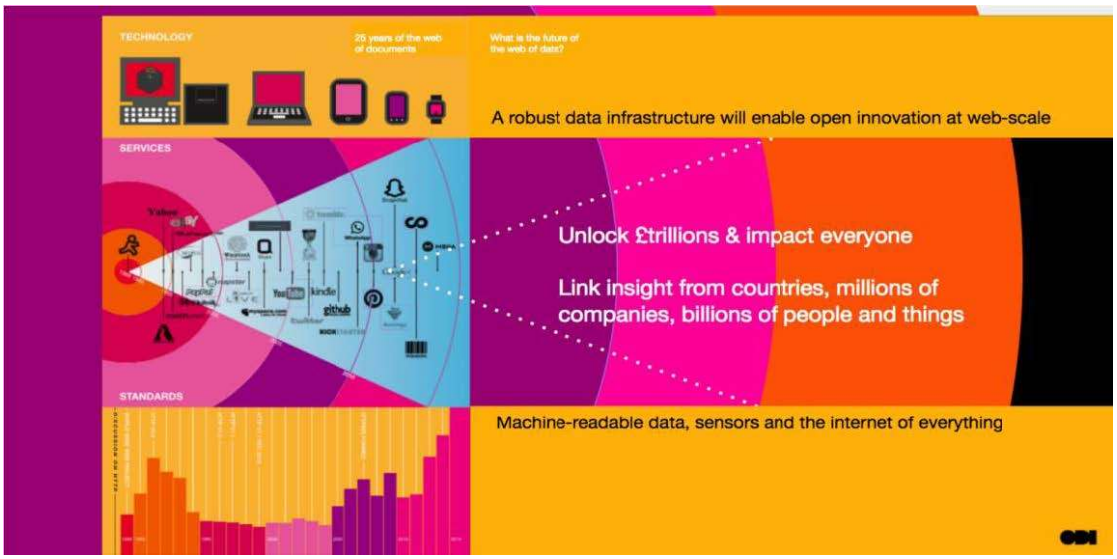
Mientras que los datos siguen siendo un recurso estático, servicios de la web de segunda y tercera generación son perfectamente adecuados para extraer metadatos sobre estos recursos estáticos y proveer entradas en los resultados de una búsqueda. Se puede conectar a estos recursos estáticos por enlaces y entonces un algoritmo como *PageRank* sigue siendo relevante.

Este método es adecuado para la red de documentos, donde los metadatos siguen siendo la manera clave de encontrar datos. Esto significa que se puede encontrar datos usando a buscadores si haces la consulta correctamente.

Pero no todos los datos son estáticos.

Cuando visitas a una página de mercado, de viajes, o del clima, el contenido probablemente sea diferente cada vez dependiendo en los datos en bruto. En una página de mercado, ciertos productos pueden variar en precio y nivel de oferta y de inventario. En una página de viajes, opciones varían en disponibilidad y precio dependiendo en los criterios de búsqueda mientras que el clima cambia constantemente.

Los datos que hacen correr esas páginas son vastos y escondidos en la web. Datos legibles-por-máquina están creando una red de datos completamente nueva que los expertos dicen tienen el potencial de abrir una era de los datos. Tal vez la red entonces se convierte de un espacio de información a un espacio de datos e información.



Tecnología | 25 años de la red de documentos | ¿Cuál será el futuro de la red de documentos?
Servicios

Abrir $Billones e Impactar a todos

Conectar ideas de países, millones de compañías, miles de millones de personas y cosas

Estándares

Datos legibles por máquina, sensores y el internet de cosas

]

Grafica 1: *El alcance expansivo de la web*

Aplicaciones que usan tales datos que ya son prevalentes, de planificadores de viaje, aplicaciones del clima y de comercio hasta juegos, tales aplicaciones intercambian datos para ayudarles a funcionar, pero estos datos están escondidos, haciendo que sea difícil que otros lo utilicen.

5. Portales del gobierno

La historia de datos abiertos está estrechamente conectada con leyes que gobiernan el acceso a información pública. Tales leyes aseguran que el público tiene el derecho de acceder a la información que proviene de esos servicios públicos. El movimiento para datos abiertos intenta una inversión completa de esta lógica. En vez de tener que solicitar datos, los datos deben de estar 'abiertos por defecto'. Para cerrar un conjunto de datos requiere una buena razón en vez del opuesto. Los gobiernos y proveedores de servicios deben trabajar abierta y proactivamente.

Un paso más allá, los gobiernos de varios países firmaron a la Sociedad de Gobiernos Abiertos (*Open Government Partnership, OGP)*. Esta OGP se lanzó en 2011 para proveer una plataforma internacional por reformadores domésticos con el compromiso de hacer que sus gobiernos sean más abiertos,

responsables, y responsivos a sus ciudadanos. Un parte clave del OGP es el compromiso de ser “abierto por defecto” a datos abiertos. Esto condujo al lanzamiento de varios portales gubernamentales de datos abiertos que fueron construidos para almacenar datos del gobierno y hacerlo disponible para el público.

Tras los años, actividades relacionadas con datos abiertos han evolucionado a lo extenso, donde los gobiernos están calificados por cuan buenos siguen y cuan sostenible sean.

El Barómetro de Datos Abiertos (Open Data Barometer, ODB) por la Fundación del *World Wide Web* (World Wide Web Foundation) califica a gobiernos en tres aspectos: la preparación, implementación, e impacto. La calificación para implementación se mide por evaluar si el conjunto de datos claves están presentes, accesibles, y actualizadas. En el *Barómetro* del 2016 la calificación para implementación se consideraba por la disponibilidad de 15 tipos de datos.

- Datos de mapas
- Datos de propiedad de terrenos
- Datos detallados del censo
- Datos detallados del presupuesto gubernamental
- Datos detallados de gastos gubernamentales
- Registro de compañías
- Legislación
- Horarios de transporte público
- Datos del intercambio internacional
- Desempeño del sector de salud
- Desempeño de la educación primaria y/o secundaria
- Estadísticas de crimen
- Estadísticas del ambiente nacional
- Resultados de elecciones nacionales
- Contratos públicos

Conjuntos de datos como los que son para los resultados de elecciones, gastos gubernamentales, y desempeño en educación son registros de importancia histórica. Tales conjuntos de datos son muy estáticos por su naturaleza y fácilmente se puede hacer disponible en forma de un archivo de hoja de cálculo para descargarlo vía un portal. A la misma vez, son estos mismos registros que están más vinculados con acceso a leyes de informáticos y tienen menos potencial económico para el uso amplio.

Al contrario, los conjuntos de datos como de mapas, compañías, y datos sobre el intercambio son mucho más dinámicos y así son más adecuados para otro nivel de servicio que un simple descargo de archivos. Esto es un caso en particular porque datos para un mapa toman la forma de varios formatos complejos y los datos de intercambios puede exceder el tamaño sensible para archivos para descarga.

La agricultura es un área donde se cruza con muchos de estas otras áreas, y de esa forma muchos conjuntos de datos existen en portales. Por ejemplo, datos de mapas son muy importantes para informar a la agricultura, más allá de asuntos de ser dueño de terrenos, como la captación o escurrimiento de agua, el uso de tierras y áreas protegidas o restringidas, que pueden informar el uso de tierras.

Como mencionamos previamente, hallando datos en estos portales depende mucho en búsquedas por metadatos. Así, el título y la descripción de cada conjunto de datos son críticos para hacer posible encontrarlo. Sin embargo, como el portal a lo mejor estaría organizada por un departamento o actividad, encontrar un conjunto de datos requiere o un conocimiento especial del dominio de éstos y/o conocimientos como el del gobierno local que organiza y describe sus datos en el portal. Un buen punto de partida es hacer una vista general por el portal por sí mismo para encontrar pistas en cómo se organiza y descubrir cómo se describen los datos antes de especificar su búsqueda con este nuevo conocimiento.

6. Obtener datos "en la web"

Como mencionamos previamente, mucho de los datos abiertos disponibles allí afuera es solamente disponible "en la web"; o por un botón de descargar o dentro de las páginas en sí. Esta sección revisa las técnicas para empezar a descubrir y abrir estos datos listos para usar.

a. Encontrando archivos descargables

Muchos de los proveedores son de mucha ayuda y te dan enlaces de descargar en formato legible por las personas para obtener sus datos. La mayoría de los conjuntos de datos en portales gubernamentales para datos funcionan así; explora a los enlaces para descubrir más.

- Estadísticas Nacionales del Reino Unido – datos sobre las reservas de cereales actualizados
- Censo agrícola de Tanzania[13]

Muchos buscadores, como Google, le permite a uno usar búsquedas y prefijos avanzados para poder hallar datos de fuentes como los de arriba. Búsquedas avanzadas utilizan filtros y prefijos para limitar el tipo de búsqueda y resultados. Se puede ver una lista de prefijos y ejemplos con enlaces en la Tabla 1.

Tabla 1. *Prefijos para realizar una búsqueda avanzada*

Prefijo	descripción	Ejemplo de búsqueda
"filetype:"	Búsqueda para tipos específicos de archivos	"filetype: xls reservas de cereales"
"site:"	Buscar con un dominio o sitio solamente	"site: opendata.go.tz agricultura"
"related:"	Buscar por contenido relacionado con una página conocida	"related: https://www.gov.uk/government/statistics/cereal-stocks"

[13] http://opendata.go.tz/dataset/tanzania-utafiti-wa-sampluli-sensa-ya-kilimo-2007-2008

"link:"	Lista solamente páginas que hacen enlace a las especificadas	"link: https://www.gov.uk/government/statistics/cereal-stocks"

Cada uno de estos prefijos puede ayudar en refinar una búsqueda por datos. Mientras que los primeros dos ayudan filtrar la búsqueda, los dos de abajo lo hacen más amplio una vez que encuentras un recurso relevante. Mientras que "*related*" puede ayudar encontrar otro tipo de contenido relevante, los que hacen enlace a su conjunto de datos podría haber utilizado a los datos y ayudar proveer contexto del uso actual. Dado que la mayoría de los conjuntos de datos con licencia abierta requiere una atribución, una búsqueda con el prefijo "*link:*" debe dar un resultado de por lo menos un número de resultados si el conjunto de datos se ha utilizado.

b. Agregadores de datos

Uno de los desafíos principales cuando se trata de datos "en la web" es la falta de la capacidad de realizar una búsqueda dentro de los datos en sí. Los buscadores en existencia solamente permiten una búsqueda de metadatos, sin importar el formato de archivo (lo mismo es el caso para audio y video, pero Google si permite una búsqueda utilizando una imagen existente).

Muchos portales de datos sirven como agregadores y permiten un poco de exploración de datos, o de un solo proveedor o un conjunto. En relación con la agricultura, el Banco Mundial agrega datos claves de estadísticas de varios países y permite la exploración y descarga de estos datos.

Un ejemplo del Banco Mundial son los indicadores del Desarrollo Rural y agrícola[14]. Agregando los datos juntos permite la exploración de indicadores como área de tierra agrícola contra la población rural. El ejemplo vinculado arriba examina a la comparación entre una gran variedad de países en el este y oeste de África. El servicio *DataBank* (banco de datos) del Banco Mundial provee acceso fácil para explorar y después descargar los datos disponibles para usar.

Tales agregadores pueden ser una fuente crucial de datos abiertos cuando portales de países son o no actualizados o simplemente no están disponibles.

Enigma.io, ganador de *Techcrunch Disrupt* en el año 2013[15], junta a datos de una multitud de fuentes de datos abiertos y hace posible una búsqueda muy precisa dentro de los datos en sí. Esto es en efecto una búsqueda al revés; en vez de buscar por los metadatos para encontrar los datos, enigma.io busca en los datos y te enseña en cuales conjuntos de datos se encuentra su consulta. Por ejemplo, una búsqueda por "Monsanto"[16] te da todo tipo de conjunto de datos interesante, desde contribuciones de campaña política en 2016 a reportes de clima desde el año de 1949.

[14] https://data.worldbank.org/topic/agriculture-and-ruraldevelopment?locations=KE-TZ-RW-GH-NG-ML-BF
[15] http://www.businessinsider.com/techcrunch-disrupt-winner-enigma-2013-5?IR=T
[16] https://public.enigma.com/search/Monsanto

c. Los scrapers

A veces los datos no estarán disponibles para descargar en un formato usable. A veces los datos estarán disponibles sólo desde una página web como una tabla o lista. En otros casos los datos pueden ser disponibles en un formato de documento (como PDF) en vez de un formato de datos. En ambos casos el uso de *scrapers* (extraidores) de datos pueden ayudar a extraer los datos visibles.

d. Los *scrapers* de datos de la web

Los scrapers de la web permiten la extracción automática de datos estructurados en una página web. Herramientas como grepsr[17] permiten la extracción automática de datos de páginas estructuradas en segundos, incluso a la habilidad para manejar la paginación y lo que resulta en un sin fin de resultados de desplazamiento. En la actualidad tales herramientas requieren un costo por cada registro para una extracción con un número limitado de créditos gratis por mes.

e. Los *scrapers* de PDF

Otro formato donde se encuentra datos útiles empotrados es dentro de reportes de PDF producidos por agencias de estadísticas. Estos suelen tener apéndices largos de datos en tablas que se puede extraer con herramientas como PDFTables.com. Pruébalo por sí mismo con algunas estadísticas de agricultura[18] de Tanzania. Los datos empiezan en página 124 de este reporte y es recomendable reducir al PDF a las páginas exactas que tienen los datos deseados antes de subirlo a un extractor de PDF como PDFTables.com.

7. Obtener datos “de la web”

La evolución de la web ha conducido al requerimiento de separar la infraestructura de fondo y datos del nivel de presentación como sitios web y aplicaciones móviles que usan los mismos datos. Aplicaciones de comercio, del clima y de viajes ofrecen varias opciones para que usuarios interactúen con esencialmente los mismos datos.

Estas aplicaciones están usando servicios dedicados de datos para acceder y hacer consultas a los datos. Muchos de estos servicios están documentados para que cualquier persona los utilice mientras que muchos quedan escondidos por razones comerciales o de presupuesto. Esta sección examina las diferentes técnicas que se podría intentar para acceder a datos “en la web” que ayuda a que funcionen estas aplicaciones.

[17] https://www.grepsr.com
[18] http://harvestchoice.org/sites/default/files/downloads/publications/Tanzania_2007-8_Vol_5g.pdf

a. Extensiones de tipo de archivos

Algunos sitios web están construidos para ofrecer una manera de extraer datos por agregar una extensión de archivo al *URL* de la página web que se ve. Para tales sitios de web, comúnmente mantenido por organizaciones que publican datos abiertos descargables, agregar la extensión correcta resulta en un descargo de esa página en formato de datos, en vez de un formato de documento.

Un buen ejemplo de esto es el sitio web del Gobierno del Reino Unido (www.gov.uk), que provee cualquier página en formato de datos por simplemente agregar la extensión relevante como '.json', por ejemplo www.gov.uk/browse/business.json. Para ver a los datos en un formato más entendible por las personas, copialos en jsonlint.io.

La página de Aranceles del Intercambio del Reino Unido[19] tiene la misma funcionalidad y contiene detalles en los códigos del intercambio internacional que se puede vincular con los datos disponibles de Entradas y Aduanas[20].

Desafortunadamente, no muchos sitios de web hacen claro a las personas que formatos alternos (como de JSON) están disponibles. Un buen indicador es para encontrar un sitio de web que parece ser moderno y de donde claramente las páginas contienen datos, como registros sobre compañías individuales[21], de donde se puede probar tales extensiones. La tabla 2.1.2 enumera formatos de datos comunes disponibles "en la web".

Extension	**Descripcion**
.csv	*Comma Separated Values* (valores separados por coma). Datos en tablas como de Excel, pero reducidos para sólo almacenar datos en una estructura básica.
.json	*JavaScript Object Notation* (notación de objetos en JavaScript). Datos en formato jerárquico nativo al lenguaje *JavaScript* que se usa comúnmente en la web, ya que forma parte de la especificación de HTML5.
.xml	eXtensible Markup Language (Lenguaje de anotaciones extensibles). Una especificación de anotaciones que tiene una variedad grande de usos. Ha sido criticado por su complejidad y verbosidad en comparación a JSON.
.rdf	Aunque el RDF no debe ser un formato de datos (no cubierto aquí), RDF define una estructura formal de datos que se puede aplicar en formatos de xml, json, y csv. El

[19] https://www.trade-tariff.service.gov.uk/trade-tariff/sections
[20] https://www.uktradeinfo.com/Statistics/BuildYourOwnTables/Pages/Home.aspx
[21] https://opencorporates.com/companies/ch/471356

	uso de la extensión implica que se usa esa estructura y que probablemente los datos en sí están en formato de XML/
.rss	Otra estructura especifica de XML que se usa comúnmente de fuentes de datos que se actualizan frecuentemente como noticias y clima

b. Interfaz de Programación de Aplicaciones (API)

Los API son uno de las mejores maneras de acceder a datos. Los API son un servicio mejor descrito como una "promesa" por un sistema constantemente y consistentemente de proveer un servicio a otro que permite que los dos interactúen. Por esta razón, los API tiene muchas ventajas sobre cualquier otra manera de acceder a datos, como descrito abajo.

1. *Acuerdos de servicio.* Como un API es un servicio, garantiza acceso a datos y muchas veces se acompañan con acuerdos al nivel del servicio para los que los quieren usar.
2. *Acceso en vivo.* Los API provee un mecanismo donde los datos pueden estar incluidos en vivo con una aplicación. El ejemplo más común de un API de datos es en tiempos de tránsito en vivo. En base a un solo API, se puede crear cientos de aplicaciones.
3. *Diseñados para datos.* Tal vez la ventaja más grande de un API es que son desinados para datos y máquinas en vez de humanos. Esto quiere decir que la disponibilidad de datos no se limita por los paradigmas de como los humanos usan a la web, aunque si crea dificultades cuando se realiza una búsqueda por datos que podrían estar dentro de un API.

La desventaja principal de los API es que los datos no son fáciles de acceder para descargar y usar de una vez. Algunas aplicaciones de terceros, como enigma.io ya usan los API para acceder datos de otros servicios para dar acceso fácil, y otros como *OpenCorporates* permiten descargos por la extensión como parte de su API.

Ejemplos de servicios que tienen API incluyen a: *OpenCorporates*[22], *OpenStreetMap*[23], *Twitter*[24], *Flickr*[25], y *LinkedIn*[26]. Estos API proveen acceso directo a los datos en bruto o consultas amplias para permitir búsquedas multidimensionales.

Muchas de las plataformas de datos abiertos proveen un API para acceder datos incluso a *Socrata* y *OpenDataSoft*. Tales plataformas son utilizadas por varios gobiernos y departamentos; *Socrata* esta principalmente en los EEUU y *OpenDataSoft* en Europa. *CKAN*, una alternativa de código abierto también tiene un API, aunque este API solamente da acceso a los registros de metadatos en muchos instantes.

[22] https://api.opencorporates.com
[23] http://wiki.openstreetmap.org/wiki/API
[24] https://developer.twitter.com/en/docs
[25] https://www.flickr.com/services/api/
[26] https://developer.linkedin.com/docs/rest-api

La tabla 3 contiene algunos ejemplos de cada plataforma y algunos conjuntos de datos de agricultura disponibles, que son algunos que mencionamos anteriormente.

Tabla 3: *Ejemplos de plataformas de datos abiertos y conjuntos de datos de agricultura disponibles allí*

CKAN	*Página web:* https://data.gov.uk/dataset/cereal_stocks_england_and_wales *API:* https://data.gov.uk/api/3/action/package_show?id=cereal_stocks_england_and_wales
Socrata	*Página web:* https://data.code4sa.org/dataset/List-of-Registered-Dams-2014/iety-gmha *API:* https://data.code4sa.org/resource/cig6-sz38.csv
OpenData Soft	*Ningún ejemplo en agricultura encontrado*

c. Usando los API

Muchos API de la web tienen el formato de un *REST API*. Representational State Transfer (REST, o *Transferencia del Estado Representacional*) es un API diseñado específicamente para la web. Tiene un conjunto específico de pautas y reglas que controlan algo si es un *API RESTful*.

Hablando en lo general, un API requiere el uso de identificadores de recursos con que se puede interactuar para subir/descargar ese recurso deseado. En el caso de *Socrata* de la sección anterior, el API está ubicado en el identificador de la web del recurso (https://data.code4sa.org/resource/cig6-sz38). Al hacer click a este recurso serás transferido a la página web, no porque eso es lo que hace el enlace, sino porque eso es lo que fue consultado cuando se hizo click en un buscador de la web. El buscador va a una solicitud GET para una representación de página web (texto/html) de este recurso como mostrado aquí.

```
GET /resource/cig6-sz38 HTTP/1.1
HOST: data.code4sa.org
ACCEPT: text/html
```

Un *API REST* específica que una máquina debe poder cambiar la solicitud para pedir otras representaciones diferentes del mismo recurso. Esto es un poco como agregar una extensión de formato de archivo, excepto cuando el recurso pedido no cambia ninguna parte de su ubicación en la web (como agregar '.csv' en efecto cambia el URL). El siguiente ejemplo muestra dos pedidos modelos para una versión JSON y CSV del mismo recurso utilizando un API REST.

JSON	CSV
GET /resource/cig6-sz38 HTTP/1.1 HOST: data.code4sa.org **ACCEPT: text/csv**	GET /resource/cig6-sz38 HTTP/1.1 HOST: data.code4sa.org **ACCEPT: application/json**

Los API REST simplemente son extensiones de la HTTP (*HyperText Transfer Protocol*, Protocolo de Transferencia en HyperText) existente en la web utilizados paras los datos. Así, es posible cambiar el tipo de la solicitud de un GET (conseguir) a un PUT (colocar) y enviar datos estructurados al servidor para reemplazar los datos existentes con datos nuevos (obviamente autorizado). La Ciudad de Chicago usa el método *POST* para subir estadísticas de crimen actualizadas[27] a su portal de datos y lo han hecho todos los días desde el año 2001.

Los API no solamente proveen acceso para usuarios de datos, sino también forman una parte clave de la infraestructura de datos del proveedor, permitiendo que los datos se mantengan y actualicen.

[27] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzpq8t2/data

d. Los API escondidos

No todos los sitios web que cargan a datos de manera dinámica hacen que su API sea público, aun si uno existe. No obstante, es posible descubrirlo. Hacerlo requiere un cierto conocimiento técnico; no obstante, una búsqueda buena por Google puede revelar comunidades de personas que tal vez ya han construido algo para el servicio en particular de donde quieres extraer los datos

Como muchos API están basados en el modelo del *API REST*, en muchos casos es relativamente fácil para alguien que conoce a los *API REST* encuentre si un servicio tiene uno y como funciona. Esto se puede hacer probando unas solicitudes *REST* con extensiones de buscador como *Postman para Google Chrome*[28].

La herramienta experimental de la ODI, el extractor de datos escondidos[29] fue construido para automáticamente buscar a API REST que intercambian datos de JSON cuando carga una página web.

8. Revisando los derechos de usar los datos

Hay muchas maneras de obtener datos en la web, sea visible claramente por un botón de descargar o por un API público o escondido. Sin importar el método de adquirir los datos es crítico revisar los derechos de ambos para usar ese método y los datos asociados.

Tal como los datos en sí, algunas declaraciones de derechos serán únicamente legible por las personas, algunas legibles por máquina y otras una combinación de los dos. No obstante, comúnmente proveedores de servicio tendrán una versión legible por humanos de los términos de uso y/o una licencia que cubrirá ambos los términos del uso del servicio y derechos de usar los datos una vez adquiridos.

Muchos portales de datos gubernamentales tendrán una licencia de datos incluida como una parte de metadatos para el archivo visto. Por ejemplo, en data.gov.uk todas las licencias están enumeradas directamente debajo del conjunto de datos en forma de un enlace. La plataforma CKAN (de lo cual data.gov.uk es una versión) es particularmente buena en exponer declaraciones de derechos. Eso ayuda que los usuarios aseguren que los datos que usan son datos abiertos.

Servicios como Flickr también tienen licencias asociadas con cada foto. Cada usuario de Flickr puede especificar la licencia para sus propias fotos. Flickr hasta provee una búsqueda que les permite encontrar fotos con licencias específicas.

Si las licencias también están legibles por máquina (como en el caso de CKAN y Flickr), entonces los buscadores pueden usar esto como pedazo de metadatos, que significa que los resultados de búsqueda se pueden filtrar instantáneamente para contener solamente contenido con licencia abierta (prueba una búsqueda avanzada de Google[30])

[28] https://chrome.google.com/webstore/detail/postman/fhbjgbiflinjbdggehcddcbncdddomop?hl=en
[29] http://odinprac.theodi.org/hidden-data-extractor/
[30] https://www.google.co.uk/advanced_search

Si se usa un API REST, la declaración de derechos podría ser incluido como parte de un enlace encabezado[31]. Esto separa la declaración de derechos del contenido, permitiendo que una respuesta a la solicitud sea sólo datos, por ejemplo un archivo CSV.

Si ninguna de estas opciones existe, podría ser necesario leer los términos y condiciones de los proveedores para asegurar que su método de acceso y derechos de usar los datos están permitidos. Sólo por poder acceder a algo en la web no le da el derecho para que todos lo usen.

Resumen

En esta lección se ha introducido muchos de los métodos y limitaciones de descubrir datos abiertos en la web. Todavía está en la evolución de la 'era de los datos' siguiendo a la 'era de información' y servicios que especializan en proveer acceso rápido de datos continúan de evolucionar.

A la misma vez el número de servicios que proveen datos está creciendo, imitando a los días tempranas de la web. Todavía hay lecciones para aprender, pero los métodos que se usa para acceder a los datos se están estabilizando con el surgimiento de los API comunes como de REST.

Formatos de datos también han evolucionado y así también los métodos de hallar y acceder datos. Los buscadores se están poniendo más inteligentes y se pueden personalizar para realizar consultas con mucha precisión. A la misma vez las herramientas para extraer y trabajar con datos han evolucionado tanto que es muy fácil empezar con datos sin importar el formato.

Esta evolución de aplicaciones móviles que demandan acceso instantáneo a datos ha incrementado también el número de API disponible, aun si algunos siguen estando escondidos.

Es claro que vivimos en la era de datos, no obstante, necesitamos tener cuidado con nuestros derechos de usar tales datos. Tener licencias claras de datos abiertos es esencial para el futuro de nuestra infraestructura de datos.

[31] https://theodi.org/guides/publishers-guide-to-the-open-data-rights-statementvocabulary#linkingtorightsstatementsfromwebapis