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 de Agricultura*** (GFAR), y el ***Instituto de los Estudios del Desarrollo*** (IDS), ***The 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 3: Creando datos abiertos

Lección 3.1: La gestion conjuntos de datos Dinamicos

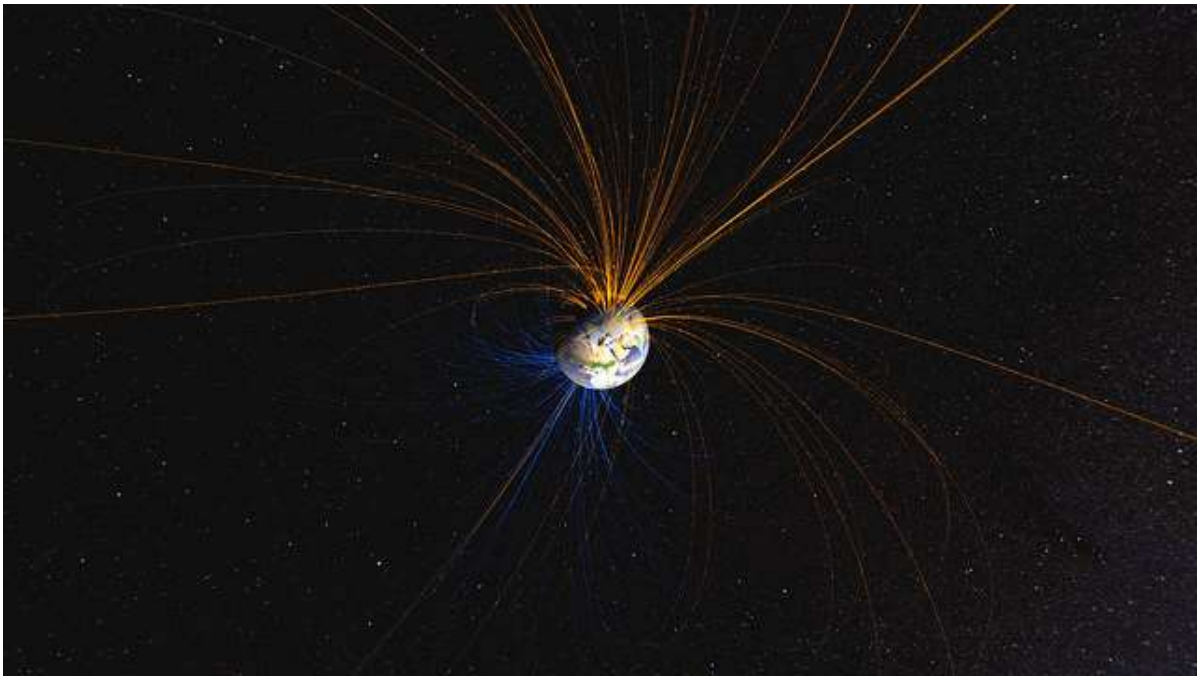


Foto por Nasa Goddard Space Flight Center licenciado bajo CC BY 2.0

Objetivos y metas de aprendizaje

Se puede utilizar a datos abiertos cuando

Esta lección tiene el objetivo de:

- Proveer un bosquejo de procesos automáticos para publicar datos y la entrada de datos y proveer ejemplos de ello.

Después de estudiar esta lección debes saber cómo:

- Entender a los procesos que juegan el papel de mantener y exponer a conjuntos de datos dinámicos
- Estar consciente de ejemplos existentes de conjuntos de datos dinámicos
- Estar consciente de métodos para evaluar la consistencia de datos en conjuntos de datos dinámicos

Contenido

Unidad 3: Creando datos abiertos

Lección 3.2: Gestión de Conjuntos de Datos Dinámicos

Objetivos y metas de aprendizaje

Lista de tablas

Lista de gráficas

Lista de tablas

1. Introducción

En la lección previa, miramos a los principios de gestión de datos y el rango grande de actividades desde los aspectos administrativos a los técnicos. Se puede publicar a datos de una gran variedad de maneras y esto depende en si los datos están recopilados por mano y agregado o recopilados por procesos automáticos.

Es importante entender la diferencia entre los dos tipos de conjuntos de datos que será discutido en esta unidad – estáticos y dinámicos – en conjunto al proceso para manejarlos. Las mejoras que vemos en la tecnología hoy han conducido al desarrollo de herramientas que podemos usar para proveer acceso a los datos, y daremos ejemplos para las herramientas que se puede usar para exponer a datos dinámicos. Esta unidad también describirá rutinas sencillas de consistencia para usar en revisar y verificar conjuntos de datos dinámicos.

2. Conjuntos de datos dinámicos y estáticos

Datos dinámicos son los que cambian de manera asincrónico cuando actualizaciones están disponibles[1]. El opuesto de esto son datos estáticos, también conocidos como datos persistentes, que no son accedidos con frecuencia y rara vez será modificado. Datos dinámicos son diferentes de datos fluyendo en que no hay un flujo constante de información; en cambio, las actualizaciones pueden venir a cualquier hora, con periodos de inactividad en el medio.

En la ciencia informática, las estructuras de datos estáticos tienen un área fija de memoria donde se pueden operar. No es posible expandir este tamaño fijo en el tiempo de correr. Por lo tanto, la ubicación de cada elemento es fijo y conocido por el programa. En cambio, las estructuras de datos dinámicos tienen un área flexible donde se pueden operar. Es posible expandir o contraer esta área como se requiere, para agregar o remover elementos de la estructura de datos. Así será inefectivo usar estructuras dinámicas para almacenar conjuntos de datos que no cambiaran. Usando estructuras de datos estáticos en tal caso ahorraría recursos del sistema y proveer acceso más rápido a elementos. Usuarios o desarrolladores están responsables para usar estructuras adecuadas para datos, según la situación.

El auge de tecnologías de precisión en la agricultura está cambiando la manera en que agricultores manejan su tierra y ganado, como con sistemas de posicionamiento geográfico generado por satélite y sensores que detectan nutrientes y agua en tierras. Estas tecnologías por último resultan en la colección de más datos dinámicos. Que son procesados automáticamente.

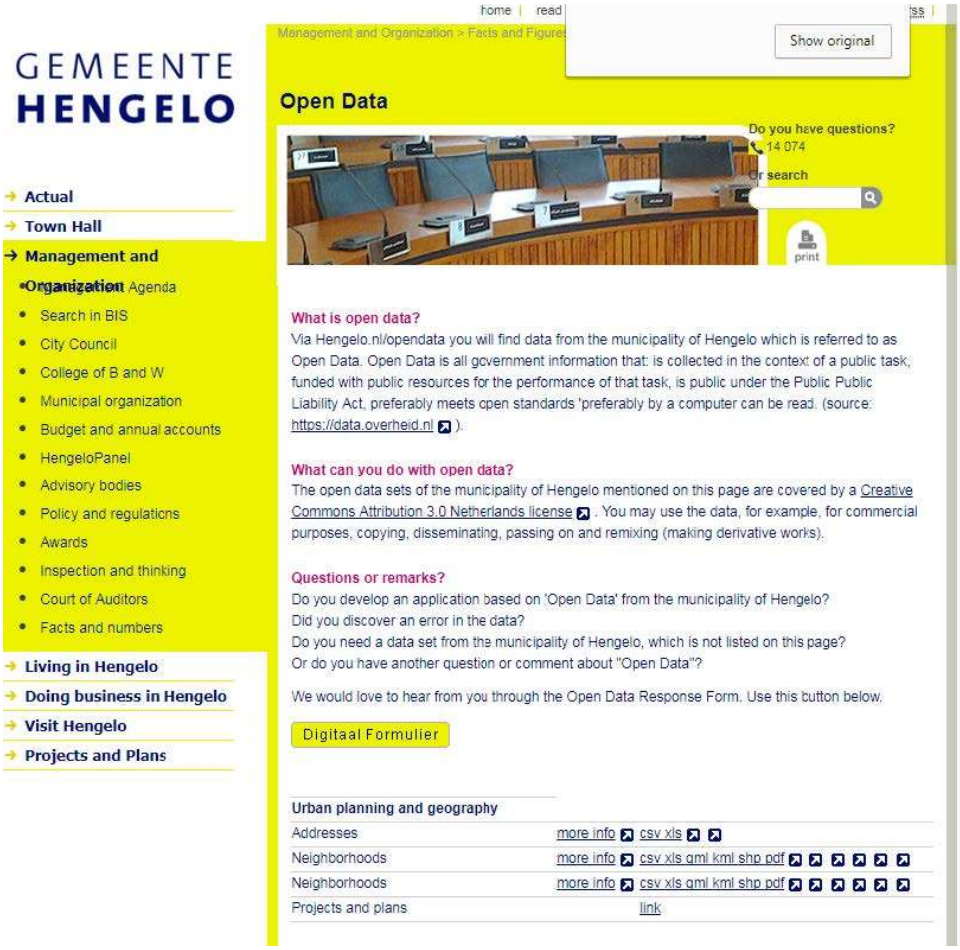
Próximamente, miraremos los varios métodos y procesos usados para publicar datos – de mano o automáticamente.

[1] https://en.wikipedia.org/wiki/Dynamic_data

2.1. Publicando datos – entrada manual

2.1.1. Publicando datos en un sitio web

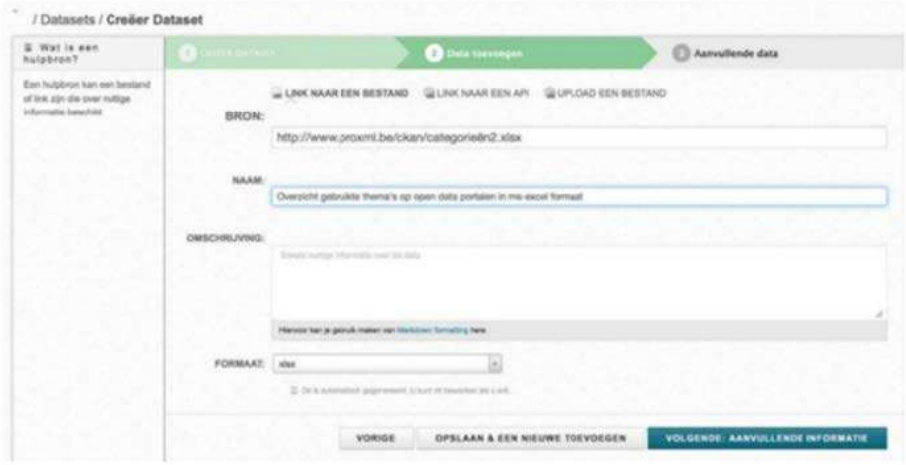
Se puede compartir los conjuntos de datos estáticos por medios estáticos como sitios web. Algunas organizaciones solamente tienen pocos conjuntos de datos para compartir, que hacen por publicar los archivos en su sitio web. La solución de software usado podría ser cualquier *Content Management System* (CMS, o Sistema de Gestión de Contenido), como Drupal o Wordpress en combinación con una base de datos interno. Por ejemplo, la ciudad Holandés de Hengelo hizo una página web para compartir datos en su sitio web de la ciudad, como en Gráfica 1.



Gráfica 1: ***El sitio web de Datos Abiertos para Hengelo*** ***http://www.hengelo.nl/opendata***

2.1.2. Subiendo a portales de web

El proceso de subir a un portal es uno de los medios más usados para publicar datos abiertos. Gráfica 2 abajo muestra una captura de pantalla del portal CKAN donde se sube a los datos por mano2. Este es el Inferfaz manual de CKAN, que tiene algunas restricciones, como no poder reutilizar marcos similares y sin apoyo para otros idiomas.



Gráfica 2: ***Ejemplo de un conjunto de datos agregado por mano al portal de datos Flamenco usando el interfaz CKAN3***

Un portal provee, por medio de un catálogo de metadatos, un punto de acceso singular a los datos. Abajo hay ejemplos de portales de datos abiertos con datos agrícolas:

- Unión Europea: https://data.europa.eu/euodp/en/data
- Estados Unidos de América: https://www.data.gov/
- Reino Unido: https://data.gov.uk/
- FAO (Organización de las Naciones Unidas para la Alimentación y la Agricultura): http://www.fao.org/data/en/
- Banco Mundial: http://data.worldbank.org/
- Ending Rural Hunger (ERH, Acabando con el Hambre Rural): https://endingruralhunger.org/
- CGIAR: http://www.cgiar.org/resources/open/data-managementsystem/
- Ayuda Alimentaria – WFP: http://www.wfp.org/fais/
- Nutrición para Animales – FAO: http://www.feedipedia.org/
- Pescadería – FAO: http://www.fao.org/fishery/statistics/en
- Alimentos y nutrición: para enlaces para alimentos y nutrición en esta página web, haz click aquí

[2] https://www.europeandataportal.eu/en/providing-data/goldbook/publishing-data
[3] https://www.europeandataportal.eu/en/providing-data/goldbook/publishing-data

Todos los portales (2500+) se encuentran aquí:

https://www.opendatasoft.com/a-comprehensive-list-of-all-open-dataportals-around-the-world/

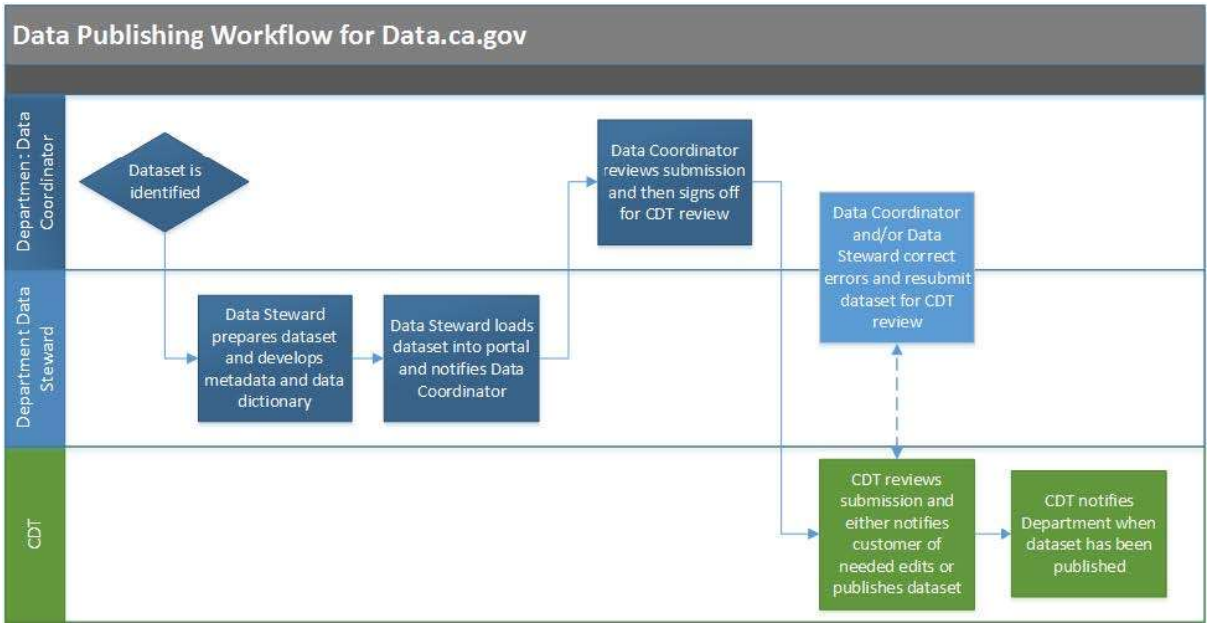
Datos Abiertos en California – un ejemplo de un flujo de trabajo manual para publicar datos en un portal de datos abiertos.

La publicación de datos es controlada por un flujo de trabajo por secuencia. Un coordinador de datos o mayordomo es asignado a uno de estos dos roles:

Creador de Contenido, Contribuidor al Flujo de Trabajo – este rol se asigna a especialistas al nivel departamental quienes crean y suben estos conjuntos de datos.

Editor, Moderador de Flujo de Trabajo – este rol se asigna a un mayordomo de datos al nivel departamental que revisa y aprueba a datos antes de publicar.

Se describe en detalle el proceso de publicar en https://office-of-digitalinnovation.github.io/data-training/publishing/



Coordinador de Datos Departamental

- *[Se identifica un Conjunto de Datos]*
- *[El Coordinador de Datos revisa el material subido y lo somete para revisión por CDT]*
- *[El Coordinador de datos y/o el Mayordomo corrigen errores y reenvían al conjunto de datos para revisión por CDT]*

Mayordomo de Datos Departamental

- *[El Mayordomo de Datos prepara el conjunto de datos y desarrolla a los metadatos y diccionario de datos]*

- *[El Mayordomo sube al conjunto de datos al portal y notifica al Coordinador]*

[CDT]

- *[CDT revisa el material y o notifica al cliente de cambios necesarios o publica el conjunto de datos]*
- *[CDT notifica al Departamento cuando se publica el conjunto de datos]*

Gráfica 3: ***Datos Abiertos de California provee un ejemplo de un flujo de trabajo manual para publicar datos en un portal de datos abiertos***

2.2. Publicando datos – procesos automáticos

Fuentes de datos automáticos. Mientras que el *Internet of Things* (IoT, el Internet de Cosas) expande, particularmente con el desarrollo de sensores baratos, la posible atomización de colección de datos expande. Con sensores de precisión, agricultores pueden colectar de manera automática datos sobre el clima, tierra, calidad de aire, y madurez de cosechas, habilitando decisiones más inteligentes. Dispositivos como teléfonos móviles también se usa para colectar datos pasivos: por ejemplo, la aplicación WeatherSignal[4] utiliza sensores ya incluidos en el teléfono para medir las condiciones atmosféricas locales, que se muestran en la pantalla en su mapa de clima que se auto-actualiza. Ahora hay potencial para sensores en ciudades, casas, carros, torres de teléfonos celulares y hasta teléfonos móviles para contribuir a datos de observaciones que también puede ser usado para modelos de pronóstico. Mientras que conjuntos de datos continúan de crecer exponencialmente, se han introducido.

Automatizando Datos es el proceso de actualizar datos en su portal de datos abiertos por la programación en vez de hacerse manual. Automatizando el proceso de subir a los datos es importante para la sostenibilidad a largo plazo de su programa de datos abiertos. Cualquier dato subido manual corre el riesgo de ser atrasado porque es una cosa más que hacer como parte del resto de su carga de trabajo.[5]

Automatizando al proceso de publicar datos provee para los usuarios actualizaciones de datos rutinarias y predecibles, y hacer que el proceso de publicar sea más eficiente. Hay 3 elementos comunes para la automatización de datos -Extraer, Transformar, y Subir:

Extraer: El proceso de extraer a los datos de uno a varios sistemas de fuente

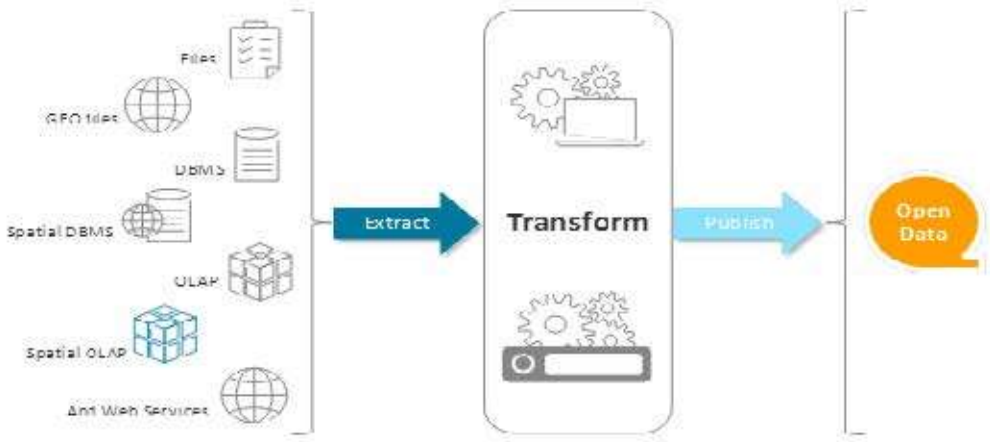
Transformar: El proceso de transformar a los datos al formato necesario, como un archivo CSV; eso puede incluir a cosas como cambiar a abreviaturas a nombres completos

Subir: El proceso de subir a los datos al sistema final, en este caso el portal de datos abiertos.

El proceso de publicación automática antes era altamente personalizado, con publicadores haciendo un programa de *Extract-Transform-Load* (ETL, Extraer-Transformar-Subir) de la nada, que requiere mucho esfuerzo. Los procesos ETL hacen esas tres cosas: extrayendo a los datos brutos de la fuente original, transformándolo a un formato más útil, y subiéndolo al portal de datos abiertos. Cada uno de estos

[4] https://play.google.com/store/apps/details?id=com.opensignal.weathersignal&hl=en
[5] https://support.socrata.com/hc/en-us/articles/212871018-Data-Automation-Overview

procesos es crítico para completamente automatizar al proceso de subir datos, y para hacerlo exitosamente. La Gráfica 4 es una diagrama de un proceso estándar de ETL.



[archivos]

[Archivos geográficos]

[Base de Datos]

[Base de datos espaciales] *-[Extraer]- -[Transformar] –[Publicar]-[Datos Abiertos]*

[OLAP]

[OLAP espacial]

[servicios web]

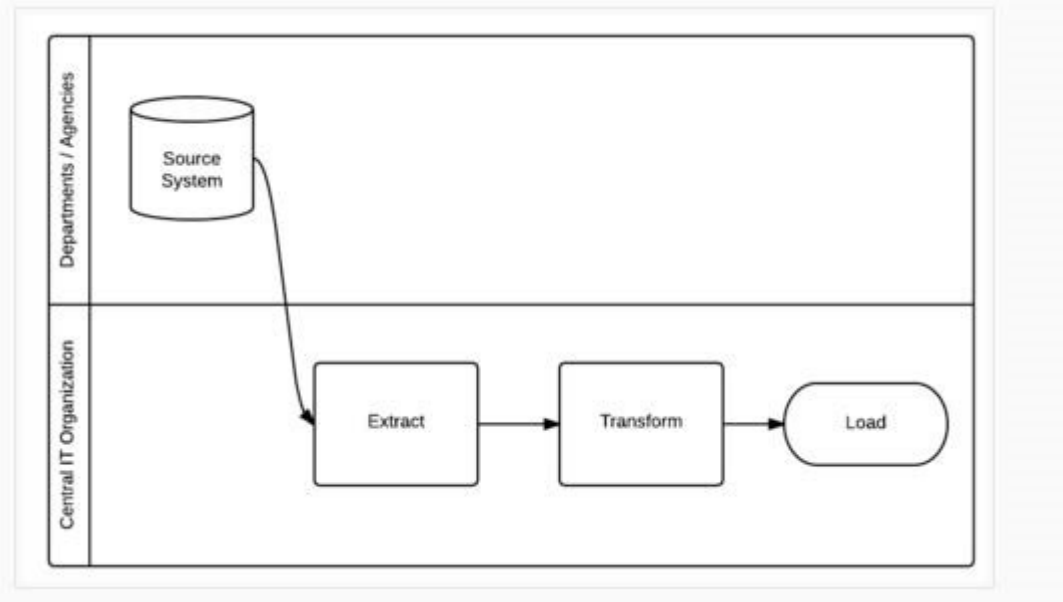
Gráfica 4: ***Un proceso de ETL: extrae datos brutos de la fuente original, lo transforma en un formato más útil, y lo sube/publica en el portal de datos abiertos***[6]

Estrategia de automatización de datos. Primeramente, es importante determinar una estrategia general de automatización para su organización. Tener una estrategia desde antemano te ayudará conectar con la gente correcta, las herramientas correctas, al tiempo correcto dentro de su organización.

Identifica a quien es el dueño de la automatización de datos en su organización. Diferentes grupos tendrán partes diferentes en el proceso ETL:

Centralizado: El departamento central de Informática es responsable para el proceso completo de ETL y automatización de datos (vea gráfica 5).

[6] https://www.europeandataportal.eu/en/providing-data/goldbook/technical-preparationand-Implementation



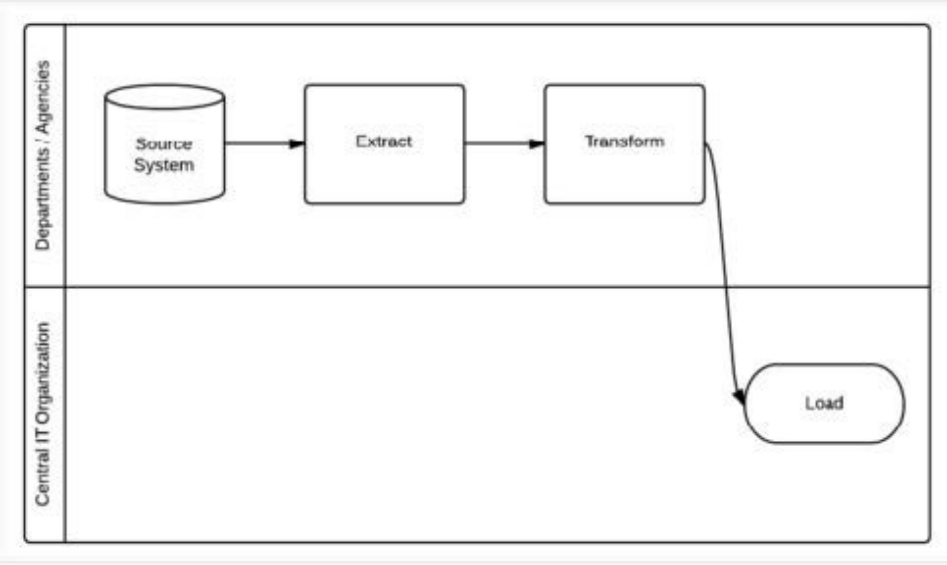
[Departamentos y Agencias]

[Fuente de Sistema]

[Organización Central de Informática]

[Extraer] – [Transformar] – [Subir]

Gráfica 5: ***Un proceso ETL centralizado***



[Departamentos y Agencias]

[Fuente de Sistema] – [Extraer] – [Transformar]

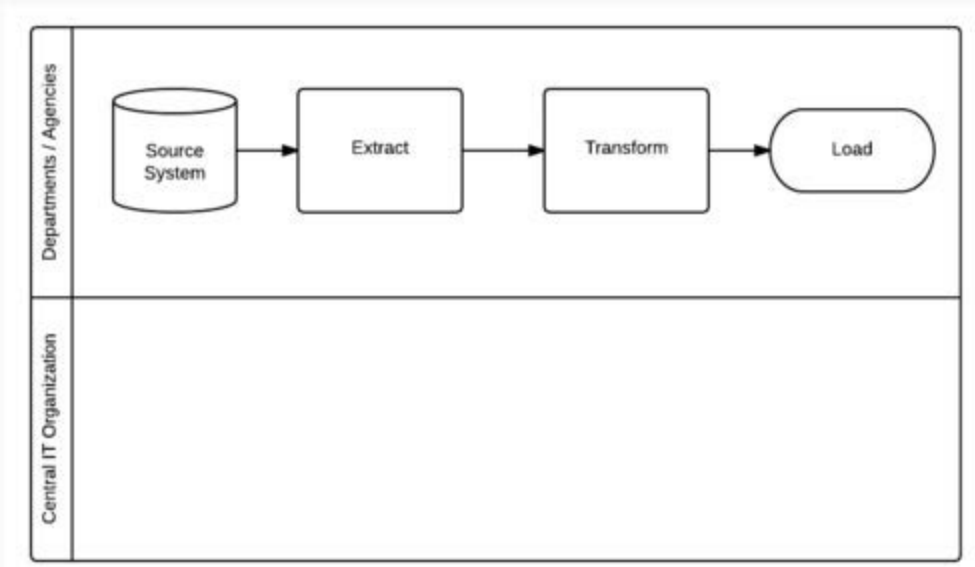
[Organización Central de Informática]

[Subir]

Gráfica 6: ***Un proceso híbrido ETL***

Híbrido: Este modelo puede variar, pero muchas veces las agencias/departamentos individuales serán responsables para los procesos de extracción y transformación, y el departamento central de Informática será responsable para el proceso de subir (vea Imagen 6).

Descentralizado: los departamentos/agencias son responsables por sus propios procesos de ETL (vea gráfica 7).



[Departamentos y Agencias]

[Fuente de Sistema] – [Extraer] – [Transformar] – [Subir]

[Organización Central de Informática]

Gráfica 7: ***Un proceso descentralizado de ETL***

Cuando has definido donde las etapas de ETL ocurrirán dentro de la organización, tendrás que determinar donde las etapas encajarían en el flujo de trabajo de publicar. Mientras que creas al diagrama de flujo para publicar, este seguro de clarificar a los pasos para Extraer, Transformar, y Subir, en conjunto con los que serán responsables para cada uno de ellos.

Seleccionando a los datos que serán automatizados. Lo más que adoptas a un enfoque de “automatizar donde sea posible” a subir datos, lo menos recursos tendrás que dedicar en el largo plazo para

mantener datos de alta calidad. Aquí están algunos tips/consejos para encontrar conjuntos de datos candidatos para procesos automáticos para subir:

- ¿Es el conjunto de datos actualizado cada 3 meses o más frecuente?
- ¿Hay algunas transformaciones u otro tipo de manipulación que se necesita hacer con ese conjunto de datos antes de subirlo?
- ¿El conjunto de datos es grande (más que 250MB)?
- ¿Puedes solamente obtener a los renglones cambiados para cada actualización siguiente (en vez del archivo completo)?
- ¿Es posible obtener a los datos del sistema fuente, en vez de un individuo?

Los conjuntos que reciben un fuerte "si" como respuesta a cualquiera de las preguntas anteriores son muy buenos candidatos para automatizar a las actualizaciones porque la automatización puede quitar el riesgo de una falta de tiempo y recursos para preparar manualmente a los conjuntos de datos.

Pasos para la automatización. Una vez que entiendes al entorno de la automatización dentro de tu organización, puedes empezar a poner a emplear a su estrategia de automatización. Aquí están algunos pasos para empezar:

1. Identificar a datos: seleccione a uno o dos conjuntos de datos de alto valor donde acceder a los sistemas de la fuente será fácil (empezando con algo fácil que sería "como un pan comido")
2. Revisa al inventario del sistema de la fuente para determinar a cuáles sistemas ya tienes acceso.
3. Determina el acceso: determina como el departamento o agencia relevante obtendrá a los datos. ¿Será por una consulta de SQL, descargar un CSV, etc.? El custodio de datos a lo mejor sería el mejor recurso para acceder el sistema de fuente para un conjunto de datos.
4. Definir transformaciones: resaltar cualquieras transformaciones necesarias para el conjunto de datos. Esto puede ser tan simple como cambiar acrónimos complejos a nombres completos, o tan complicado como transformar un base de datos relacional a un archivo CSV.
5. Trabajar con ambos el mayordomo y custodio
6. Desarrollar y probar el proceso ETL: basado en los requisitos definidos en los pasos 2 y 3, selecciona una herramienta para publicar con ETL, y publicar el conjunto de datos con el portal de datos abiertos. Confirmar que el conjunto de datos ha sido subido o actualizado con éxito sin errores.
7. Horario: programa a tu conjunto de datos para ser actualizado en momentos oportunos.
8. Refiere a los campos de metadatos sobre su colección, la frecuencia de reiniciar y actualizar.

3. Consistencia de datos

Datos consistentes son los que técnicamente están correctos y aptos para análisis estadísticos. Estos son datos que han sido revisados para valores faltantes, valores especiales, errores o valores atípicos (obvios), que son o borrados, corregidos, o imputados. Los datos están consistentes con límites basados en un conocimiento del mundo real sobre el tema que los datos describen[7].

[7] E de Jonge and M van der Loo, 2013, *An introduction to data cleaning with R* Statistics

La calidad de datos se evalúa para aplicar métodos de verificación y validación como parte del proceso de control de calidad. La verificación y validación son componentes importantes de la gestión de datos que puede asegurar que los datos son válidos y confiables. La Agencia de la Protección del Ambiente en los Estados Unidos define a la verificación de datos como el proceso de evaluar lo completo, correcto, y la conformidad de un conjunto de datos con los procesos requeridos para asegurar que los datos son lo que deben representar. La validación de datos sigue a la verificación de datos, y tiene que ver con evaluar a datos verificados para determinar si las metas de la calidad de datos han sido cumplidos y las razones para cualquier desviación.

Se necesita aplicar a los principios de la calidad de datos en todas las etapas del proceso de gestión de datos (captura, digitalización, almacenamiento, análisis, presentación, y utilización). Hay dos elementos claves para la mejora de la calidad de datos – la prevención y corrección. La prevención es estrechamente relacionada con la colección de los datos y de entrarlos en una base de datos. Aunque se debe y puede dedicar un gran esfuerzo a la prevención de errores, el hecho es que habrán errores en conjuntos de datos grandes y, la validación y corrección de datos no se puede ignorar.

3.1. Listas de chequeo en consideración a la consistencia de datos

1. ¿Está completo tu conjunto de datos? Cada conjunto de datos debe:
   a) Tener una fila de encabezado con una descripción única para lo que está allí. Esto significa que una vez que una estructura para un conjunto de datos está establecida, no se debe cambiar una vez que se agrega fuentes. En los metadatos, se debe describir el encabezado.
   b) Estar etiquetado con un numero de versión. Una vez que se realiza una actualización, el conjunto de datos debe recibir un número nuevo de versión para que la audiencia pueda seguir a los cambios
2. Tener información sobre su origen. ¿De qué se tratan los datos, de dónde viene, y para qué propósito se publicó?
   a) Tener un estado: borrador, validado, final
   b) ¿Están los datos limpios? Revisa a los siguientes aspectos:
   c) Campos vacíos
   d) ¿Datos “dummy” y valores estándares – están correctos?
   e) Valores incorrectos
   f) Entradas duplicadas
   g) Información sensitiva y privada

Hay varios ejemplos de errores e inconsistencias en los datos y como se resuelvan, con la validación de esquema y otras herramientas avanzadas como Open Refine[8] en la Unidad 2, Lección 2.2

1. ¿Están los datos correctos? ¿Está el conjunto de datos correcto? Los aspectos más importantes en cuestión de la precisión son:
   a) ¿Están los datos lo suficiente de precisos para su propósito potencial?

Netherlands, The Hague, Netherlands. Available at: https://cran.rproject.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf

[8] http://openrefine.org

b) ¿La precisión afecta a la confianza que tienes en ellos?
c) ¿Se describe las opciones que tiene que ver con el intervalo?
d) ¿Los datos necesitan desagregación o agregación?

La documentación es clave para datos de buena calidad. Sin buena documentación, es difícil que los usuarios determinen lo apto del uso de los datos y difícil para los custodios para saber qué y por quien las revisiones de calidad de datos han sido realizadas. La documentación es por lo general de dos tipos, y se deben instalar dentro del diseño de una base de datos. Lo primero tiene que ver con cada registro, y registrar cuales revisiones se ha realizado y cuales cambios han sido hechos y por quien. La segunda son los metadatos que registran a la información al nivel del conjunto de datos.

4. Exponiendo a Conjuntos da Datos Dinámicos

4.1. Los API y datos abiertos

Un API (*Application Programming Interface*, o Interfaz de Programación para Aplicaciones) es un nombre colectivo para vincular a sistemas por un interfaz de programación. Un API se puede usar para hacer disponible a datos abiertos para proveer un usuario con acceso directo a los datos abiertos de un proveedor.

Por lo general, los datos tienen un conjunto específico de campos o columnas y códigos, que los usuarios tienen que entender para trabajar de manera efectiva con datos. Si se consiguen por un API o servicio, los usuarios también necesitarán entender cómo funciona los API.[9]

El Open Data Institute (Instituto de Datos Abiertos) sugiere que la documentación técnica tiene que proveer con un API que debe incluir:

- Documentación de formato sobre los formatos de datos que estas proveyendo, posiblemente incluso a esquemas para cualquier vocabulario se utilice;
- Lista de códigos que provee más detalles sobre cada uno de los códigos que se usa dentro de sus datos; una manera de proveer esta información es tener un URL que provee documentación sobre cada código e incluir un enlace para ese URL dentro de los datos;
- Documentación de servicio que describe como un API que provees funciona; esto puede incluir enlaces a descripciones de servicio legibles-por-máquina si aplican.

Equipados con esta información, usuarios deben entender a los datos que se publica y como crear aplicaciones que lo utilizan.

4.1.1. Ejemplo de un API en agricultura

Food Security Portal (el Portal de Seguridad Alimenticia)[10]: Este sitio web del IFPRI[11] contiene más de 40 indicadores relacionadas con la seguridad alimenticia, precios de productos básicos, la economía, y bienestar humano. Muchos de estos datos están disponibles para cada país del mundo y se remontan a

[9] https://theodi.org/guides/engaging-reusers
[10] http://www.foodsecurityportal.org
[11] http://www.ifpri.org

más de 50 años. Provienen de fuentes públicos y autoritarias como el Banco mundial, el FAO, UNICEF, y otros, en conjunto a datos prioritarias.

Para hacer que estos datos en el sitio web sean lo más útiles posibles, están disponibles para descargar gratuitamente por el API y para agregar, mezclar, y compartir. El portal está diseñado a agrupar tal información de maneras estructuradas y para revisar la calidad y relevancia de los datos.

4.1.2. Ejemplo de un API para datos del clima

OpenWeatherMap[12]: esta página publica datos abiertos del clima por un API para desarrolladores que se hace fácil exportar información del clima en una variedad de aplicaciones, incluyendo a aplicaciones web y móvil, y soluciones para los seguros, la agricultura, deportes, y muchas áreas más.

El servicio *OpenWeatherMap* recopila datos de estaciones de clima profesionales y privadas. Hoy, tienen más de 40,000 estaciones de clima; la mayoridad son estaciones profesionales que están instaladas en aeropuertos, ciudades grandes, etc.

4.1.3. La Tecnología detrás de los API

Los API corren por un conjunto de tecnologías específicas, haciendo que sea fácil para que desarrolladores los entiendan. Este tipo de enfoque quiere decir que los API pueden funcionar con cualquier lenguaje de programación, y la manera más popular de crear API de la web siendo *REST* (*REpresentational Estate Transfer*, Transferencia Representacional de Estado).

REST usa los mismos mecanismos del internet que se usa para ver a las páginas web normal, creando muchas ventajas que resultan en implementaciones más rápidas y fáciles para que los desarrolladores puedan entender y usar[13]. Las implementaciones de los API REST permiten que uno tome datos y funciones que ya están disponibles en su página web y hacerlos disponibles por un API para programación que aplicaciones en la web y por móvil pueden utilizar. Entonces, en vez de regresar a HTML para representar información como haría un sitio web, un API puede producir una respuesta en uno de dos formatos"

- *Extensible Markup Language* (XML) o
- *JavaScript Object Notation* (JSON).

Los desarrolladores pueden tomar estos datos y usarlos en aplicaciones de la web o por móvil. No Obstante, XML y JSON pueden ser traducidos fácilmente para hojas de cálculo y otras herramientas que gente que no son desarrolladores pueden usar también, que hace que los API sean útiles para todos.

[12] http://www.openweathermap.org
[13] https://project-open-data.cio.gov/api-basics/

Resumen

Se puede publicar datos de varias maneras y esto depende de si los datos están recopilados y agregados manualmente o por un proceso automático.

Los datos dinámicos son datos que son cambiados de manera asincrónica mientras que sea posible actualizarlos

Los conjuntos de datos se pueden publicar por páginas web, portales, o los API.

La automatización de datos es el proceso de actualizar datos en su portal de datos abiertos por programación, en vez de ser manual. Automatizar el proceso de subir datos es importante para la sostenibilidad de largo plazo de su programa de datos abiertos.

Las tres etapas comunes en la automatización de datos son **Extraer**, **Transformar**, y **Subir** (ETL por sus siglas en Ingles):

- **Extraer**: El proceso de extraer los datos de uno o varios sistemas de fuente
- **Transformar**: El proceso de transformar los datos a la estructura necesaria, como de un archivo CSV.
- **Subir**: El proceso de subir los datos al sistema final.

Datos consistentes son los que técnicamente son correctos y aptos para análisis estadísticos. Los datos han sido revisados por si valores faltantes, valores especiales, errores (obvios) y valores atípicos están borrados, corregidos, o imputados. Los principios de calidad de datos se tienen que aplicar a todas las etapas del proceso de gestión de datos (captura, digitalizar, almacenamiento, análisis, presentación, e utilización). En consideración de la consistencia de sus datos tienes que asegurar que: los datos están completos, limpios, y precisos.

Un API (Application Programming Interface) es un nombre colectivo para conectar sistemas por un interfaz de programación. Un API se puede usar para hacer que datos abiertos estén disponibles para proveer a un usuario el acceso directo a los datos abiertos de un proveedor.

En la próxima lección 3.3 Creando y Gestionando Repositorios de Datos Abiertos, exploraremos las varias opciones disponibles para el alojamiento de un repositorio de datos abiertos y cómo manejarlo.

Lecturas Adicionales

Arms, C. R., Fleischhauer, C. and Murray, K. (2013). Sustainability of digital formats: planning for Library of Congress collections. Library of Congress, Washington DC, USA. Available at: www.digitalpreservation.gov/formats

Beagrie, N. and Houghton, J. (2014). *The value and impact of data sharing and curation - synthesis of three recent UK studies*. Jisc. Available at: https://repository.jisc.ac.uk/5568/

Charles Beagrie Ltd (2013). *Keeping research data safe: cost / benefit studies, tools, and methodologies focussing on long-lived data*. Available at: http://www.beagrie.com/krds.php (accessed 4 August 2014)

Digital Curation Centre (DCC). (2010). *Data management plans*. Available at: http://www.dcc.ac.uk/resources/data-management-plans (accessed 4 August 2014)

Drummond, C.G. (2009). Replicability is not reproducibility: nor is it good science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop*, 26th ICML, Montreal, Canada. Available at: http://cogprints.org/7691

European Commission (EC). (2017). *Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*. Available at: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf (Version 3.2, 21 March 2017)

GODAN 2016 *A Global Data Ecosystem for Agriculture and Food*. GODAN, Wallingford, UK. Available at: http://www.godan.info/documents/dataecosystem-agriculture-and-food

Jones, S. (2011). *How to Develop a Data Management and Sharing Plan*. DCC How-to Guides, Digital Curation Centre, Edinburgh, UK. Available at: www.dcc.ac.uk/resources/how-guides/develop-data-plan

UK Data Archive (UKDA). *Plan to Share*. Available at: www.dataarchive.ac.uk/create-manage/planning-for-sharing (accessed 4 August 2014)