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 2: Usando los datos abiertos

Lección 2.2: Calidad y procedencia

Objetivos y metas de aprendizaje

Se puede utilizar datos abiertos cuando:

Después de estudiar esta lección debes saber cómo:

- Enumerar, describir, y aplicar los factores que afectan la utilidad de datos abiertos
- Usar herramientas que te ayuden a evaluar la utilidad de datos abiertos
- Verificar sus datos contra otras fuentes de datos
- Identificar la procedencia de un conjunto de datos abiertos (para un proceso legible por máquina/legible por humanos
- Describe cuáles son datos usables desde diferentes puntos de vista

Contenido

Lista de imágenes

Lista de tablas

1. Introducción

La calidad y procedencia son dos aspectos importantes que determinan la utilidad de un conjunto de datos. Esta lección hace una vista general a los diferentes aspectos que hacen que un conjunto de datos de buena calidad y una variedad de las guías de buenas prácticas pueden ayudar a la publicación de datos útiles de alta calidad.

Parte de la calidad de un conjunto de datos depende de la historia, o procedencia, de ese conjunto de datos. Saber que los datos provienen de una fuente confiable y que fue recopilado utilizando métodos confiables (o por métodos con limites conocidos) puede muchas veces ser más importante que tener un conjunto de datos con un vocabulario o esquema bien controlada

Aunque los datos son de carácter técnico por su naturaleza, no todas las medidas de calidad son técnicos; asegurar que un conjunto de datos sea legible por máquina no siempre quiere decir que se puede utilizar; no todos los requisitos son técnicos. La evolución de guías de buena práctica refleja esto y en esta lección miramos a los aspectos técnicos y no técnicos que hacen que un conjuntos de datos sea útil y de alta calidad.

2. Marcas y medidas de calidad

Evaluando la calidad de datos abiertos no se puede hacer rápidamente. Hay una gran cantidad de estándares que provienen de la comunidad y marcas de calidad y le pueden ayudar en evaluar la calidad y utilidad de datos.

Una de las primeras marcas que salió para datos abiertos es la de “las 5 estrellas de datos abiertos vinculados”. Premiando cada estrella es parte de una secuencia y empieza con el requisito de aplicar una licencia de datos abiertos. Las siguientes estrellas están divididas en dos: se enfocan en que los datos están disponibles “en la web” para descargar y dos que se enfocan en datos que están “dentro de la web” para usar con un API que puede conseguir los recursos al instante. Aparte del aspecto que tiene que ver con las licencias, la guía de 5-estrellas se enfoca en la disponibilidad técnica de datos y carece de los aspectos no técnicos en que consisten los datos útiles de alta calidad.

De forma similar, los principios de *FAIR*[1] proveen una guía similar para publicadores. Según estos principios los datos deben ser: **Fáciles de encontrar** (*Findable*), **Accesibles** (*Accessible*), **Interoperables y Reusables** (**FAIR**). Mientras que se enfocan en gran parte otra vez en los aspectos técnicos, los principios sin embargo si establecen algunos términos con relación a la procedencia, diciendo que “objetos de datos publicados deben referirse a sus fuentes con metadatos y procedencia amplios para *fomentar citación adecuada*”.

El esquema de 5-estrellas y los principios FAIR están descritos con más detalle en la Unidad 4: Intercambiando Datos Abiertos. Esta lección examinará otros aspectos de la utilidad como los definido por los Certificados de Datos Abiertos del Instituto de Datos Abiertos (ODI).

[1] Force11. Guiding Principles For Findable, Accessible, Interoperable And Reusable Data Publishing Version B1.0. https://www.force11.org/fairprinciples

3. Los certificados de datos abiertos

El Certificado de Datos Abiertos es una herramienta gratis que fue desarrollado y es mantenido por el Instituto de Datos Abiertos para evaluar y reconocer la publicación sostenible de datos abiertos de alta calidad. Se aborda los aspectos legales, prácticos, técnicos y sociales de publicar datos abiertos utilizando una guía de buenas prácticas.

Como FAIR, el proceso de los certificados de datos abiertos[2] toman una mirada alternativa pero complementaria al esquema de 5-estrellas. Un certificado mide cuan efectivamente alguien se comparte un conjunto de datos por la facilidad del rehuso. El alcance cubre más que asuntos técnicos[3], incluso a derechos y licenciamiento, documentación y garantías sobre la disponibilidad. Un certificado por lo tanto ofrece una evaluación más completa de la calidad de publicación de un conjunto de datos.

Para los publicadores de datos el proceso de evaluar un conjunto de datos provee una idea en cómo podrían mejorar su proceso de publicar. El proceso de evaluación por lo tanto tiene valor en sí, pero el certificado que se produce es también de valor a los reutilizadores.

3.1. Siendo un reutilizador de datos

Para los usuarios de datos, la calidad técnica en el momento de uso puede ser suficiente para su uso particular. Sin embargo, para los reutilizadores es probable que la sostenibilidad y el apoyo sean más dominantes en el proceso de tomar decisiones que el formato de archivo.

A Los reutilizadores se necesita ofrecer la seguridad que los datos serán consistentes y confiables. Un certificado de datos abiertos reta a los publicadores pensar más allá de los datos y abordar consideraciones claves de consideraciones de política en apoyo de los datos. Estas consideraciones pueden estar desglosados en 4 categorías cruciales para reutilizadores.

3.2. Una lista de chequeo para un reutilizador de datos

Utilizando los certificados de datos abiertos como guía, lo siguiente presenta una lista de chequeo para datos abiertos. La lista de chequeo se divide entre 4 categorías que reflejan las secciones del Certificado de Datos Abiertos.

Legal

- ¿Están los datos licenciados y usables legales?
- ¿Están el modelo, formato, o estructura también licenciados abiertamente y usable legalmente?
- ¿Están claras las declaraciones del derecho del autor?
- ¿Hay datos o algún parte de los datos que no están licenciados descritos?
- ¿Hay algunas potenciales limitaciones éticas o de privacidad en los datos descritos?

[2] https://certificates.theodi.org
[3] https://certificates.theodi.org/about

Prácticos

- Están los datos bien descritos
- ¿Es la razón por que se colectan los datos clara?
- ¿Es el uso del publicador de los datos claro?
- ¿Hay otros usos actuales de los datos descritos?
- ¿Están los datos accesibles?
- ¿Tienen los datos una marca de tiempo o están actualizados?
- ¿Estarán los datos disponibles por lo menos por un año?
- ¿Estarán los datos actualizados con regularidad?
- ¿Hay un proceso de control de calidad?

Técnico

- ¿Están los datos disponibles en un formato adecuado para el contenido?
- ¿Están los datos disponibles de un lugar consistente?
- ¿Son los datos bien estructurados y legible por datos?
- ¿Están términos complejos y siglas en los datos definidos?
- ¿Los datos utilizan una esquema o estándar de datos?
- ¿Hay un API disponible para acceder a los datos?

Social

- ¿Hay una comunidad existente de usuarios de los datos?
- ¿Ya una gran cantidad de personas depende del uso de los datos?
- ¿Están los datos apoyados oficialmente?
- ¿Hay acuerdos al nivel de servicio disponibles para los datos?
- ¿Es claro quien mantiene y puede ser contactado sobre los datos?

Si un publicador ha completado un certificado de datos abiertos y aplicado la marca de calidad a sus datos entonces un reutilizador puede rápidamente encontrar las respuestas a todas las preguntas de la lista de chequeo más arriba. Alternativamente una lista completa de todos los certificados está disponible en el registro de conjuntos de datos certificados[4].

3.3. Una guía para un reutilizador para la procedencia

Todas las guías de mejores prácticas presentadas hasta aquí están de una forma enfocados en los productores de datos, quien ya son dueños y manejan los datos fuentes. Sin embargo, no todos los datos abiertos provienen del publicador; una gran cantidad de datos se derivan de otros datos. Por ejemplo, un pronóstico del tiempo se deriva después de aplicar modelos complejos de datos meteorológicos. Habiendo muchas fuentes de datos meteorológicos y muchas organizaciones que usan diferentes modelos para crear un pronóstico, esto puede conducir a situaciones donde hasta

[4] https://certificates.theodi.org/en/datasets

pronósticos basados en los mismos aportes de datos pueden variar grandemente, con consecuencias potencialmente devastadores[5].

3.3.1. Una lista de chequeo de procedencia

La siguiente lista de chequeos ayuda establecer la proveniencia de un conjunto de datos y ayuda establecer el nivel de confianza en el conjunto de datos.

- ¿Están los datos producidos la propiedad completa del proveedor de los datos?
- ¿Alguien más produce datos comparables para comprobar a los datos?
- ¿Está claro si los datos son derivados de otras fuentes de datos?
- ¿Están las otras fuentes de datos claros?
- ¿Están las otras fuentes de datos confiables y comparables con otros proveedores de datos?
- ¿Es claro si cualquier dato ha cambiado (de cualquier fuente), y como, antes de estar disponibles como datos abiertos en su punto de acceso?

Seguir estos puntos ayudará determinar cuán confiable sean las fuentes de datos abiertos. Hasta puede revelar la posibilidad de evitar la fuente de datos y seguir el camino de procedencia hasta la fuente original que podría ser más confiable u ofrecer un servicio más completo y/o apoyado.

4. Chequeo de calidad post-acceso

Al establecer la confianza en un proveedor de datos, potencialmente obtener un acuerdo de servicio y después acceder a los datos, no significa el fin del proceso de evaluar a la calidad de ellos. Una vez que se obtienen los datos es esencial verificar a los datos, especialmente con su propio uso en mente.

Esta etapa de verificar y preparar datos para estar listos para usar tiene muchos aspectos, la mayoría basados en procesos técnicos. Es más que probable que se encuentre problemas o inconsistencias en los datos durante esta etapa. Aquí es cuando es esencial estar conectado con la comunidad o proveedores y/o tener acceso a procedimientos de control de calidad para ayudar con su entendimiento de los datos o para arreglar estos problemas.

4.1. Validación de datos

Una de las primeras cosas que hacer con datos es verificarlos comparándolos con cualquier esquema de datos y una descripción de la estructura de datos disponibles. Esto revelará potenciales omisiones de la documentación para términos que han surgido en los datos. Verificar los datos con el esquema también ayudará verificar que datos correctos existen dentro del conjunto de datos.

Si tus datos no tienen esquema, puede que sea necesario transformar a los datos a un formato donde sea fácil de aplicar o que haga su propio esquema para el conjunto de datos.

[5] http://www.bbc.co.uk/news/av/world-24713504/michael-fish-s-denial-of-ahurricane-in-1987

4.1.1. Diseñando esquemas

Los esquemas juegan un papel en permitir la reutilización amplia y automática de datos. Un esquema es un cianotipo para datos que define un conjunto de límites de integridad y reglas relacionados con la estructura y el contenido de un recurso de datos. Un esquema define tres cosas que están descritas abajo.

Títulos de columna y claves en los datos: definir un grupo consistente de títulos de columna (o claves) para un conjunto de datos es esencial para asegurar que los conjuntos de datos a la misma vez se pueden combinar y analizar fácilmente. Frecuentemente, los títulos de columna cambian o pueden ser abreviados para ahorrar tiempo, pero esto causa muchos problemas cuando se analizan los datos con periodos largos de tiempo. Agregar títulos de columna es un poco menos problemático, pero se tiene que tomar en cuenta cuando analizan datos.

Tipos de valores: con los títulos de columnas y claves definidos es importante definir el tipo de dato válido para los valores, por ejemplo, número, texto, fecha, coordenada, etc.

Límites de valores: con el tipo de valor definido, límites válidos como ser requeridos, únicos, ser una unidad particular (por ejemplo, galones (UK)) o dentro de un rango particular deben de estar definidos. Por ejemplo, una columna puede tener el título de “costo ($m)”; por lo tanto, cualquier valor deben ser números (sin coma o punto). Establecer un rango válido también ayuda evitar y explicar errores en los datos. Por ejemplo, establecer un rango de .001-100 en el “costo ($m)” (si se conoce que el costo no puede ser mas que $100m) la validación de rango previene que alguien por accidente lea la columna “titulo/unidades” y registrar 100000000 en vez de 100 para $100m.

4.1.2. Usando esquemas

Todo paquete de hoja de cálculo permite que uno crea reglas de validación de datos relacionados con cada título de columna. No obstante, muy pocos paquetes permite la exportación de esquema también con los datos para que otros la utilicen sin agregar extensiones de desarrolladores complejas.

Una de las razones principales para esto es la conexión entre los esquemas y bases de datos jerárquicos relacionados con los diseños antiguos de bases de datos por EF Codd (1970). Los esquemas son una parte clave de una base de datos donde múltiples tablas están relacionadas con relaciones predefinidas. Desde este punto las implementaciones han sido guiados por este modelo teorético y aplicados técnicamente en paquetes de bases de datos relacionales como MySQL.

El desarrollo del *eXtensible Markup Language* (*XML*, Lenguaje de Marca Expandible) como mecanismo para compartir datos emergió mucho más tarde en 1996. Cinco años después en el año 2001, el actualmente-popular esquema de la especificación XML fue lanzado para ayudar de formalizar el compartir de datos consistentes y verificables. Durante la mitad de la década de los 90, a finales de la década 2000 hasta hoy de un sentido, XML era el formato estándar para la representación de datos intercambiables; diseñados para ser legible por máquina y legible por humanos.

Tabla 1: Datos ejemplares

FirstName (Nombre)	LastName (Apellido)	Instrument (Instrumento)	DateOfBirth (Fecha De Nacimiento)
John	Lennon	Vocal	1940-10-09
Paul	McCartney	Bass Guitar	1942-06-18
George	Harrison	Guitar	1943-02-25
Ringo	Starr	Drums	1940-07-07

XML

```
<People>
   <Person>
      <FirstName>John</FirstName>
      <LastName>Lennon</LastName>
      <Instrument>Vocal</Instrument>
      <DateOfBirth>1940-10-09</DateOfBirth>
   </Person>
   <Person>
      <FirstName>Paul</FirstName>
      <LastName>McCartney</LastName>
      <Instrument>Bass Guitar</Instrument>
      <DateOfBirth>1942-06-18</DateOfBirth>
   </Person>
   <Person>
      <FirstName>George</FirstName>
      <LastName>Harrison</LastName>
      <Instrument>Guitar</Instrument>
      <DateOfBirth>1943-02-25</DateOfBirth>
   </Person>
   <Person>
      <FirstName>Ringo</FirstName>
      <LastName>Starr</LastName>
      <Instrument>Drums</Instrument>
      <DateOfBirth>1940-07-07</DateOfBirth>
   </Person>
</People>
```

XML Schema extract

```
.
.
.
      <xs:simpleType
    name="birthsDate">
       <xs:restriction
      base="xs:date">
      <xs:minInclusive value="1800-01-01"/>
      <xs:maxInclusive value="2017-07-31"/>
      <xs:pattern value=".{10}"/>
   </xs:restriction>
</xs:simpleType>
<xs:simpleType name="instrument">
  <xs:restriction base="xs:token">
    <xs:enumeration value="Vocal"/>
    <xs:enumeration value="Guitar"/>
    <xs:enumeration value="Bass Guitar"/>
    <xs:enumeration value="Drums"/>
   </xs:restriction>
</xs:simpleType>
...
<xs:complexType>
   <xs:sequence>
      <xs:element name="FirstName"
      type="xs:string"/>
      <xs:element name="LastName"
      type="xs:string"/>
      <xs:element name="Instrument"
      type="instrument"/>
      <xs:element name="DateOfBirth"
      type="birthsDate"/>
</xs:sequ
    ence>
</xs:compl
  exType>
...
```

Grafica 1: *Datos ejemplares en formato XML con extractos asociados del esquema*

Los siguientes datos ejemplares (mostrados aquí en formato tabular) son un ejemplo de datos tomados de un conjunto de datos.

El mismo esquema se ve abajo (Figura 1) en formato XML con extractos asociados con el esquema.

Se puede utilizar tales esquemas con máquinas para automáticamente validar la estructura y el contenido de un conjunto de datos; también hay una herramienta de validación en línea[6] que se puede usar con los dos ejemplos arriba.

Mientras que XML ha sido adoptado ampliamente, el año 2005 fue significante para datos en la web cuando salió *Ajax* (Garret 2005). Con su significado de *Asyncronous JavaScript and XML (Javascript y XML Asyncronizados*), el Ajax es un conjunto de técnicas del desarrollo de la web que usa JavaScript para dinámicamente cargar datos en aplicaciones de web. La meta inicial era para permitir el uso dinámico de XML en aplicaciones de la web, fomentando y animando la emisión de datos. En práctica, aplicaciones modernas comúnmente sustituyan el JSON (*JavaScript Object Notation*, Notación de Objetos en JavaScript) en vez del XML por las ventajas de que JSON es nativo a JavaScript, 21% más rápido para usar y significativamente menos verbosa.

Similar a XML, JSON provee una especificación de esquema JSON[7]; pero todavía no es tan desarrollado como la especificación de esquema XML y no tiene la habilidad de validar un rango de input (por ejemplo valores mínimo y máximo permitidos). La figura 2.2.2 abajo muestra los mismos datos como del ejemplo XML, esta vez en JSON con el equivalente en el esquema JSON.

Aquí se muestra el esquema JSON para nuestro archivo de datos tabulares. Probado por sí mismo para descargar los datos tabulares[8] (en CSV) y esquema[9] (en JSON) y entonces usar csvlint[10] (para validar a datos tabulares) para validar que los datos conforman al esquema.

Uno de los retos con ambos formatos es el requisito de tener dos archivos: de los datos y del esquema. Esto crea la situación predecible donde los datos están mantenidos y compartidos, pero el esquema se olvida y es perdida. El uso de *namespaces* (código que sirve como referencia a un objeto en la programación) y datos vinculados como una solución técnica ayuda un poco a proveer una solución, pero el problema principal permanece en que la falta de integración de tales estándares en paquetes de mantenimiento en su forma original.

Paquetes de hoja de cálculo como Excel tienden a hacer el proceso de definir el formato de datos y establecer reglas de validación demasiado complejo, comparado con la facilidad de crearlos en herramientas como *airtable*[11]. Pero de nuevo, actualmente exportando el esquema no es posible, ni tampoco el uso de esquemas vinculados o con *namespace*.

Lo mismo es el caso para software de bases de datos y el movimiento hacia las estructuras de bases de datos *noSQL* que no están controlados por un esquema relacional rígidamente definido. Aun cuando el movimiento hacia nuevas estructuras de bases de datos es bueno para la rapidez cuando se mantiene conjuntos de datos grandes, el uso de los *namespace* es todavía faltante.

---

[6] http://www.utilities-online.info/xsdvalidation/#.WhK94rZ0cUs
[7] http://json-schema.org
[8] http://odedu.learndata.info/D5/course/en/assets/d466f522c9b280bdf8e0050e0899efe7cb184c9d.csv
[9] http://odedu.learndata.info/D5/course/en/assets/e367089a2d91deea7449638787755946d8b5a91e.json
[10] http://csvlint.io
[11] https://airtable.com

Los esquemas claramente tienen ventajas y se deben de adoptar cuando una organización depende mucho en datos de alta calidad.

JSON	JSON Schema (Download)

[
  {
  "FirstName": "John",
  "LastName": "Lennon",
  "Instrument": "Vocal",
  "DateOfBirth": "1940-10-09"
  },
  {
  "FirstName": "Paul",
  "LastName": "McCartney",
  "Instrument": "Bass Guitar",
  "DateOfBirth": "1942-06-18"
  },
  {
  "FirstName": "George",
  "LastName": "Harrison",
  "Instrument": "Guitar",
  "DateOfBirth": "1943-02-25"
  },
  {
  "FirstName": "Ringo",
  "LastName": "Starr",
  "Instrument": "Drums",
  "DateOfBirth": "1940-07-07"
  }
]
	{
	  "fields": [
	  {
	  "name": "FirstName",
	  "type": "string",
	  "constraints": {
	  "required": true
	  }
	  },
	  {
	  "name": "LastName",
	  "type": "string",
	  "constraints": {
	  "required": true
	  }
	  },
	  {
	  "name": "Instrument",
	  "enum": ["Vocal", "Guitar", "Bass Guitar", "Drums"],
	  "constraints": {
	  "required": true
	  }
	  },
	  {
	  "name": "DateOfBirth",
	  "type": "string",
	  "format": "date-time"
	  }
	  ]
	}

Grafica 2: *Datos ejemplares en formato JSON con extractos asociados del esquema*

4.2. Limpiando datos

Uno de los retos más grandes cuando uno trabaja con datos es lidiar con errores. Muchas veces los publicadores de datos ni se dan cuenta de errores porque los datos pueden cambiar mucho a través de los años. En otros casos, los errores pueden ser el resultado de errores en el proceso de registrar los datos, como errores de tipografías o abreviaturas incorrectas.

Aun cuando un esquema está disponible, errores e inconsistencias pueden existir en los datos. Cuando uno trabaja con datos, es importante saber cómo encontrar errores y corregirlos para hacer que los datos sean más útiles.

En esta sección se introduce un número de ejemplos diferentes de errores e inconsistencias en los datos, marcos o bosquejos que se puede arreglar con validación de esquema y que necesitan herramientas más avanzadas como Open Refine[12]

[12] http://openrefine.org

4.2.1. Formatos de fechas equivocados

Las fechas pueden estar escritas de forma inconsistente y según diferentes estándares. Una de las confusiones más grandes existe entre la forma en que se escribe una fecha en el Reino Unido y los Estados Unidos. En los Estados Unidos el mes se escribe primero, y después el día (por ej. 12/30/2017), mientras que en el Reino Unido es al revés (30/12/2017). Esto es fácil de ver donde el día es mas grande que el 12, pero puede causar confusión en otro caso (por ej. 6/7/2017).

El estándar *ISO 8601* especifica una serie de reglas para escribir fechas y horas. Para resolver esto y otros problemas *el ISO 8601* especifica que las fechas se deben escribir con el año primer (por ej. AAAA-MM-DD HH:mm:ss). Esto no solamente es fácil de leer, pero también funciona como una manera de ordenar la fecha con el factor más significante primero.

Como las fechas son complicadas, se ha hecho un esfuerzo de esconder el mantenimiento de datos de usuarios. Por ejemplo, si uno escribe '8-7' en una celda de Excel automáticamente se lo convierte en una fecha como '08-Jul' en un CSV, que no es un estándar ISO. Al cambiar el formato de las celdas en fecha permite un formato específico, pero mezclar las fechas de EEUU y RU todavía es posible y solamente visible cuando el contenido está alineado a la izquierda (incorrecto) o derecha (correcto) en la celda. Mas preocupante es que pasa cuando uno importa datos en archivo CSV a Excel, por ejemplo una fecha en formato del estándar ISO (YYYY-MM-DD HH:mm:ss) se traduce a un formato personalizado (DD/MM/YYYY HH:mm:ss) y este formato se guarda al archivo cuando terminas de editar en Excel, a menos que el formato se guarde antes de guardar el archivo. La mayoría de los usuarios no sabrán que Excel ha hecho esta conversión antes de mostrar los datos.

Las fechas son difíciles para mantener, especialmente cuando el software hace cambios que el usuario no ve. Los esquemas pueden ayudar a proveer que los datos coincidan con el formato requerido y las traducciones no pasen en el medio de todo.

4.2.2. Representaciones múltiples

La gente suele de tratar de ahorrar tiempo registrando datos por abreviar a términos. Si estas abreviaturas no están consistentes, puede causar errores en el conjunto de datos. Los esquemas que usan listas predefinidas y enumeradas de términos adecuadas pueden ayudar aquí, ya que los usuarios no pueden fácilmente añadir a la lista y así recrear el mismo error.

Otros errores que existen en esta categoría incluyen las diferencias en el uso de mayúsculas, espacios, género y plural de adjetivos (ej. "juez" o "jueces"), que pueden causar problemas interesantes.

4.2.3. Registros duplicados

Un registro duplicado es donde el mismo dato se ha registrado mas de una vez. Los registros duplicados ocurren cuando los conjuntos de datos se han combinado o porque no se sabe que ya fue registrado. Adicionalmente un registro duplicado puede ocurrir cuando se refiere a una persona por dos o más nombres (ej. "José" o "pepe" o "cheo"). Esto puede crear instantes donde se tiene que investigar si los

autores de las publicaciones son los mismos o diferentes. Este tipo de error no se puede captar con una validación de esquema.

4.2.4. Datos combinados o redundantes

Datos redundantes son cualquiera que no sean relevantes a su uso con el conjunto de datos. Muchas veces un conjunto de datos se ha creado para un propósito específico que requiere detalles que tal vez no necesitaras. Ocurrencias comunes de datos redundantes puede ser excluir a filas que representan cantidades totales. Estas muchas veces aparecen cuando se exporta un conjunto de datos en Excel sin antes haber removido la fila "total." En otros instantes se combina algunas columnas para ayudar en su legibilidad para humanos.

4.2.5. El uso mezclado de escalas numéricas

Valores numéricos en los conjuntos de datos muchas veces usan escalas diferentes para hacerlo más fácil para que una persona lo lea. En los conjuntos de datos para un presupuesto, por ejemplo, las unidades muchas veces están expresadas en millones. $1,200,000 se convierte en $1.2m. Pero, cantidades más pequeñas como $800,000 todavía se expresa completas. Para una máquina, esto significa que lea a las cifras grandes como $1.2, que causa errores. Alternativamente, si la columna debe de estar en millones, la segunda cifra se convierte en $800,000,000,000.

Desafortunadamente los esquemas no son buenos en identificar este tipo de error. Esto es porque todos los valores pueden ser de formato valor y los errores se puede causar en cualquier nivel. Establecer unos límites puede ayudar, pero tal vez no resuelvan el problema. Hacer que las unidades de medida sean claros al punto de colección o uso es imprescindible aquí para garantizar la consistencia de los datos y asegurar que no ocasione un desastre[13].

4.2.6. Rangos numéricos

Los datos a veces se miden en rangos, como un rango de edades o de salario. Para que una máquina los entienda es importante separar los valores altos de bajos para analizar mejor. Tal vez puede ser necesario crear grupos diferentes si han cambiado a través de los años (por ej. Las edades de las personas o las edades de jubilación si han cambiado).

4.2.7. Errores de ortografía

Si tiene mucho texto libre en los datos, es importante revisar para que la ortografía sea consistente para asegurar que se puede realizar un análisis con otros conjuntos de datos. La ortografía puede no estar "correcto", (por ejemplo "gandules" versus "guandules") pero debe ser consistente para permitir que sea comparable e interoperable.

---

[13] http://www.nytimes.com/1983/07/30/us/jet-s-fuel-ran-out-after-metric-conversionerrors.html?mcubz=1

4.3. Kit de herramientas de limpieza de datos

Cuando estas buscando errores en los datos, puede ser necesario descargar y subir conjuntos de datos en varias herramientas para limpiar y procesar. También es importante hacer notas de cuales cambios se han hecho y compartir estos abiertamente con otros para que todos puedan beneficiarse de su trabajo, particularmente si los datos que estas limpiando son datos abiertos existentes que se han publicados.

Ya hemos mirado a un número de herramientas de validación de esquema, pero hay un numero de otras herramientas que pueden ayudar a limpiar datos desordenados.

OpenRefine

OpenRefine es una herramienta de software diseñada para lidiar con datos desordenados. La herramienta es un buscador que usa columnas que permite que uno repara errores por todo el conjunto de datos en una sola acción. Los errores que se puede arreglar incluyen a:

- Formatos de fechas
- Representaciones múltiples
- Registros duplicados
- Datos redundantes
- Escalas de números mezcladas
- Rangos mezclados.

Programas de hoja de cálculo

Open Refine es una herramienta clave para la limpieza de datos. No obstante, a veces es más fácil arreglar algunos errores en un programa de hoja de cálculo como:

- Errores ortográficos
- Datos redundantes
- Verificación numérica
- Arreglando datos convertidos.

Otras herramientas

- Drake[14]
- Data Wrangler[15] (una herramienta desarrollada en conjunto de las universidades Stanford y Berkeley en los Estados Unidos)
- Data Cleaner[16]
- WinPure[17]

[14] https://github.com/Factual/drake
[15] http://vis.stanford.edu/wrangler/
[16] http://datacleaner.org
[17] http://www.winpure.com/article-datacleaningtool.html

Resumen

Para que los datos sean útiles requiere mucho más que sean técnicamente buenos datos. Muchas guías tratan de simplificar los requisitos para que los datos sean realmente útiles y cado uno tiene sus méritos. La lista creada de los Certificados de Datos Abiertos (Open Data Certificates) demuestra cuan desafiante puede ser producir datos útiles de alta calidad. El Certificado de Datos Abiertos esta destinado a evaluar los aspectos legales, prácticas, técnicas, y sociales de publicar datos abiertos usando la guía de buenas prácticas. No obstante, aun si esto no considera la revisión en comparación con otras fuentes de datos para asegurar que se utiliza los datos adecuados. La procedencia de datos también es clave, pero de nuevo difícil de seguir completamente en un conjunto de datos.

Una vez que se ha establecido la confianza en un conjunto de datos, verificar su contenido es el próximo desafío en espera. La creación y adopción de esquemas y validación de datos ayuda en un sentido, pero la necesidad de limpiar y validar datos (potencialmente a mano) no se va a desaparecer pronto. De hecho, mientras que exploramos en la próxima lección el proceso de preparar a los datos para análisis puede tomar hasta 80% de su tiempo!

Referencias

Dodds, L. 2015. Comparing the 5-star scheme with Open Data Certificates. Available at: https://theodi.org/blog/5-star-open-data-certificates-timberners-lee

Codd, E. F. 1970. A relational model of data for large shared data banks. *Communications of the ACM* **13 (**6), 377–387.

Garrett, J. J. 2005. Ajax: A new approach to web applications. Available at: http://adaptivepath.org/ideas/ajax-new-approach-web-applications/