

Seguridad, privacidad y aspectos legales

Álvaro López García

Grupo de Computación Avanzada y e-Ciencia
Instituto de Física de Cantabria (IFCA) - CSIC-UC

Máster universitario en ciencia de datos / Master in Data Science



CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

CSIC



UIMP
Universidad Internacional
Menéndez Pelayo

Parte I

Aplicación en el entorno Open Science

1. ¿Cómo publicar un dataset en abierto?
2. Licencias de datos
3. Minimizar los riesgos

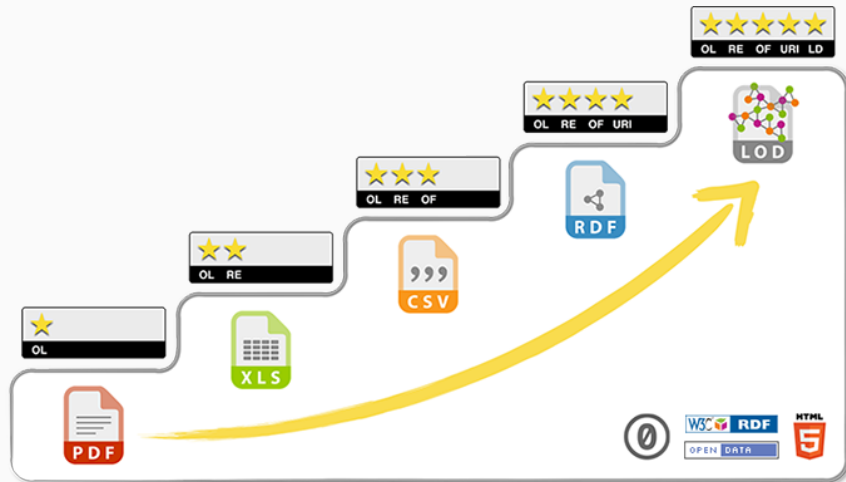
¿Cómo publicar un dataset en abierto?

Como publicar un dataset *open data*? I

Tim Berners-Lee 5* data: <http://5stardata.info/en/>

- * Make your stuff available on the Web (whatever format) under an open license.
- ** Make it available as structured data (e.g., Excel instead of image scan of a table).
- *** Make it available in a non-proprietary open format (e.g., CSV instead of Excel).
- **** Use URIs to denote things, so that people can point at your stuff.
- ***** Link your data to other data to provide context.

Como publicar un dataset *open data*? II



Pasos:

- Escoger los datos que se quieran publicar en abierto.
- Escoger un portal de datos abiertos.
- **Escoger una licencia.**
- Para cada dataset:
 - Identificar un estándar y/o formato aplicable.
 - **Aplicar técnicas para eliminar los riesgos de publicar datos en abierto.**
 - Exportar los datos al formato elegido.
 - Publicar el dataset.
- Actualizar, curar y mantener los datos.

Licencias de datos

Licencia

Contrato o permiso oficial que se concede a alguien para utilizar, copiar, modificar, estudiar, distribuir un bien, normalmente no tangible.

- Es **necesario** establecer que queremos que se pueda hacer con nuestros datos.
- Los usuarios de los datos necesitan saber que se puede hacer con unos datos.
- Preservar la visibilidad de la organización.
- La EU PSI Directive⁵ establece: «Conditions for re-use shall be non-discriminatory for comparable categories of re-use»
- Diferentes licencias: datos, contenido, código, etc.

- Respecto a la propiedad, se puede.
 - Transferir la propiedad y todos los derechos.
 - Renunciar a los derechos (dominio público)
- La licencia da un permiso, pero:
 - El *copyright* sobre los trabajos/contenidos generados es del creador.
 - El *database-right* sobre las colecciones de datos recopiladas es de la persona que la generó.

⁵ <https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information>

- Una licencia abierta puede permitir:
 - Republicar el contenido o los datos (gratis o no).
 - Crear contenido derivado.
 - Hacer dinero con el contenido.
- Únicas restricciones aceptables, según la Open Definition⁶:
 - Atribución** (*attribution*) hay que decir explícitamente quien es la fuente.
 - Compartir-igual** (*Share-alike*) hay que compartir los datos o cualquier trabajo derivado de la misma manera.
- Se puede aplicar ninguna, una o las dos.
- Tres grandes grupos:
 - Dominio público** No hay restricciones.
 - Atribución** Hay que decir quien originó los datos.
 - Atribución y compartir-igual** Hay que decir quién originó los datos y, cualquier trabajo derivado, hay que distribuirlo bajo la misma licencia.

⁶ <https://opendefinition.org/>

- Licencias para fotos, textos, etc.
- Licencias recomendadas: Creative Commons (CC).
- Última versión: 4.0, validez internacionales.
- Hay varias licencias CC, pero no todas se consideran abiertas.

Dominio público CC0

















Atribución CC-by

Atribución y compartir-igual CC-by-sa



Creative Commons

Compatibilidad entre licencias

	 PUBLIC DOMAIN	 PUBLIC DOMAIN	 BY	 BY SA	 BY NC	 BY ND	 BY NC SA	 BY NC ND
 PUBLIC DOMAIN	✓	✓	✓	✓	✓	✗	✓	✗
 PUBLIC DOMAIN	✓	✓	✓	✓	✓	✗	✓	✗
 BY	✓	✓	✓	✓	✓	✗	✓	✗
 BY SA	✓	✓	✓	✓	✗	✗	✗	✗
 BY NC	✓	✓	✓	✗	✓	✗	✓	✗
 BY ND	✗	✗	✗	✗	✗	✗	✗	✗
 BY NC SA	✓	✓	✓	✗	✓	✗	✓	✗
 BY NC ND	✗	✗	✗	✗	✗	✗	✗	✗

- Licencias recomendadas: CC 4.0 y Open Data Commons.
- Se puede diferenciar entre la base de datos y el contenido (diferentes licencias).

Dominio público CC0 PDDL

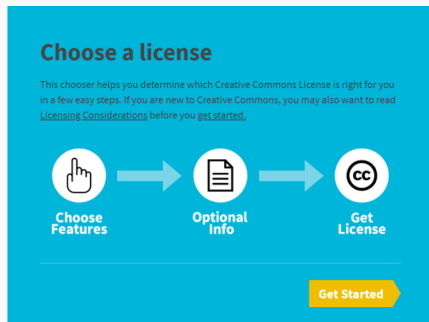
Atribución CC-by ODC-by

Atribución y compartir-igual CC-by-sa ODBL

- Depende del modelo de negocio, si es que hay alguno.
- Establecer el tipo de atribución que se requiere.
- Establacer cómo se requiere la atribución.
- Licencia dual, datos bajo dos licencias: una open y otra no. Menos restricciones.
- Si el contenido es republicado o derivado de otro contenido (y la licencia nos ha dejado), hay que publicarlo con la misma licencia (share-alike).

Publicador de datos II

¿Qué licencia escoger?



License chooser: <https://creativecommons.org/share-your-work/>

Utilizador o consumidor de datos

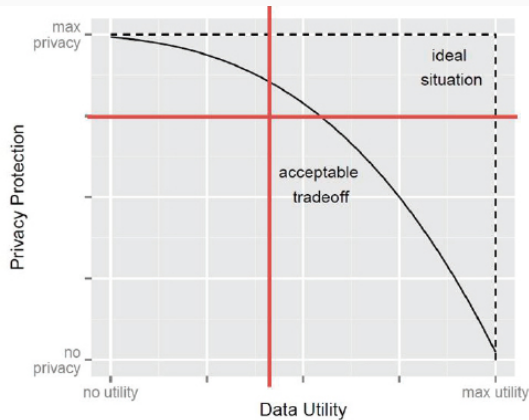
Qué puedo hacer?

- A tener en cuenta, no siempre se puede...
 - Republicar.
 - Aportar valor añadido.
 - Publicar extractos.
 - Publicar contenido derivado.
- Hay que comprobar que nos permite hacer una licencia.
- CC0: <https://www.kaggle.com/donorschoose/io>
- CC-BY-SA 4.0:
<https://www.kaggle.com/ardamavi/sign-language-digits-dataset>
- ODbL, DbCL: <https://www.kaggle.com/nickhould/craft-cans>
- Otros: <https://www.kaggle.com/unitednations/global-commodity-trade-statistics>

Minimizar los riesgos

¿Riesgos? ¿Qué riesgos?

- Como hemos visto publicar datos en abierto puede tener un riesgo.
- Aunque no existan datos personales existen riesgos.
- Compromiso entre utilidad de los datos y privacidad.
- Es necesario evaluar los riesgos de publicar un dataset.



Activo Elementos que pueden generar un riesgo o un beneficio

Evento Situaciones generadas por un activo, que llevan consigo una resultado positivo (beneficio) o negativo (riesgo)

Fuente Generadores de eventos

Probabilidad Certidumbre de que algo suceda o no.

Impacto Efecto negativo o positivo del evento. La magnitud del impacto depende de la escala y de la gravedad.

Resultado Riesgo o beneficio. Síntesis de la probabilidad e impacto, con respecto a un peligro (riesgo) o a una oportunidad (beneficio).

Calculando el riesgo/beneficio

1. Identificar los activos.

Filas, columnas, entradas, conjuntos de entradas que contribuyen al beneficio.

Filas, columnas, entradas, conjuntos de entradas que contribuyen al riesgo.

2. Identificar los eventos (¿voy a publicar datos individuales o agregados).

¿De qué forma es beneficioso este dataset? ¿Cómo se va a usar?

¿De qué forma comporta un riesgo este dataset? ¿Cómo se va a explotar?

3. Identificar las fuentes.

¿Quién puede usar este dataset?

¿Quién puede explotar este dataset?

4. Identificar beneficio/riesgo

Probabilidad e impacto.

Dataset de recogidas taxis NYC

Evaluación riesgo/beneficio

Activos

Trayectos, recogidas y entregas de taxis de NYC

Eventos

- Entradas individuales

Fuentes

Beneficio-Riesgo / Probabilidad

Dataset de recogidas taxis NYC

Evaluación riesgo/beneficio

Activos

Trayectos, recogidas y entregas de taxis de NYC

Beneficios

Eventos

- Entradas individuales
- Entender patrones de tráfico
- Estudiar ubicación de paradas
- Estudiar condiciones de trabajo de conductores

Fuentes

- Ciudadanos
- Periodistas
- Investigadores
- Empresas innovadoras

Beneficio-Riesgo / Probabilidad

	B	M	A
B	B	B	M
M	B	M	A
A	M	A	A

Dataset de recogidas taxis NYC

Evaluación riesgo/beneficio

Activos

Trayectos, recogidas y entregas de taxis de NYC

Eventos

- Entradas individuales

Fuentes

Beneficio-Riesgo / Probabilidad

Beneficios

- Entender patrones de tráfico
- Estudiar ubicación de paradas
- Estudiar condiciones de trabajo de conductores

- Ciudadanos
- Periodistas
- Investigadores
- Empresas innovadoras

	B	M	A
B	B	B	M
M	B	M	A
A	M	A	A

Riesgos

- Identificación de conductores
- Identificación de pasajeros y trayectos
- Uso por otras compañías

- Otras compañías (comptentencia)
- Ciudadanos
- Periodistas
- Investigadores

	B	M	A
B	B	B	M
M	B	M	A
A	M	A	A

Una vez calculado el riesgo, tenemos que aplicar una mitigación y evaluar la privacidad y utilidad del nuevo dataset. Ejemplos:

- Eliminar campos que contengan información sensible.

Privacidad Alta, eliminar el riesgo.

Utilidad Se elimina la utilidad que puedan aportar esos campos.

- Eliminar registros que sean particularmente sensibles.

Privacidad Alta, elimina el riesgo de los registros eliminados.

Utilidad Alta, solo se elimina un subconjunto de datos.

- Agregar datos, producir datos terciarios, estadísticas, etc.

Privacidad Alta, no se hacen públicos datos individuales.

Utilidad Baja, no se pueden analizar entradas individuales, solo estadísticas

- Generalizar datos, reducir precisión de los datos.

Privacidad Depende de la generalización. A mayor generalización, mayor privacidad

Utilidad Depende de la generalización. A mayor generalización, menor utilidad.

- k-anonimato.

Privacidad Similar a la generalización.

Utilidad Similar a la generalización.

- Añadir ruido a los datos originales.

Privacidad A mayor ruido, mayor privacidad, pero dependiendo de la densidad de la población.

Utilidad A mayor ruido, menor utilidad.

- Identificadores anónimos, eliminando atributos individuales y sustituyéndolos por un identificador sin relación con los datos.

Privacidad Si es aleatorio, alta. Sin embargo, no protegen contra la re-identificación. Una vez identificado un individuo en una entrada se puede generalizar (para ese individuo).

Utilidad Alta, no hay impacto en la utilidad.

- Privacidad diferencial. Permite analizar una población sin acceder a las entradas individuales.

Privacidad Alta, protección robusta.

Utilidad Alta, no se modifican los datos.

1. Ratio riesgo/beneficio.

Con el riesgo y el beneficio ya calculados, generando una nueva tabla

2. Obtener posibles mitigaciones.

Establecer qué se puede hacer para controlar el riesgo.

3. Calcular el riesgo beneficio después de las mitigaciones.

4. Decisión final. Resultado.

Dataset de recogidas taxis NYC

XXX

Activos

Riesgo/beneficio

Mitigaciones

Riesgo/Beneficio

- Trayectos, recogidas y entregas de taxis de NYC

Dataset de recogidas taxis NYC

XXX

Activos

- Trayectos, recogidas y entregas de taxis de NYC

Riesgo/beneficio

- Beneficio: Alto
- Riesgo: Alto

Mitigaciones

Riesgo/Beneficio

	B	M	A
B	M	B	B
M	A	M	B
A	A	A	M

Dataset de recogidas taxis NYC

XXX

Activos

- Trayectos, recogidas y entregas de taxis de NYC

Riesgo/beneficio

- Beneficio: Alto
- Riesgo: Alto

Mitigaciones

Riesgo/Beneficio

	B	M	A
B	M	B	B
M	A	M	B
A	A	A	M

Dataset de recogidas taxis NYC

XXX

Activos

- Trayectos, recogidas y entregas de taxis de NYC

Riesgo/beneficio

- Beneficio: Alto
- Riesgo: Alto

	B	M	A
B	M	B	B
M	A	M	B
A	A	A	M

Mitigaciones

- Ratio riesgo/beneficio: Medio
- Eliminación de atributos.
- Eliminación de entradas.
- Agregar datos.
- Generalizar datos.
- Anonimato.
- etc.

Riesgo/Beneficio

Dataset de recogidas taxis NYC

XXX

Activos

- Trayectos, recogidas y entregas de taxis de NYC

Riesgo/beneficio

- Beneficio: Alto
- Riesgo: Alto

	B	M	A
B	M	B	B
M	A	M	B
A	A	A	M

Mitigaciones

- Ratio riesgo/beneficio: Medio

- Eliminación de atributos.
- Eliminación de entradas.
- Agregar datos.
- Generalizar datos.
- Anonimato.
- etc.

Riesgo/Beneficio

- Nuevo beneficio: Alto
- Nuevo riesgo: Medio

	B	M	A
B	M	B	B
M	A	M	B
A	A	A	M

Dataset de recogidas taxis NYC

XXX

Activos

- Trayectos, recogidas y entregas de taxis de NYC

Riesgo/beneficio

- Beneficio: Alto
- Riesgo: Alto

	B	M	A
B	M	B	B
M	A	M	B
A	A	A	M

Mitigaciones

- Ratio riesgo/beneficio: Medio

- Eliminación de atributos.
- Eliminación de entradas.
- Agregar datos.
- Generalizar datos.
- Anonimato.
- etc.

Riesgo/Beneficio

- Nuevo beneficio: Alto
- Nuevo riesgo: Medio

	B	M	A
B	M	B	B
M	A	M	B
A	A	A	M

Dataset de recogidas taxis NYC

XXX

Activos

- Trayectos, recogidas y entregas de taxis de NYC

Riesgo/beneficio

- Beneficio: Alto
- Riesgo: Alto

	B	M	A
B	M	B	B
M	A	M	B
A	A	A	M

Mitigaciones

- Ratio riesgo/beneficio: Medio

- Eliminación de atributos.
- Eliminación de entradas.
- Agregar datos.
- Generalizar datos.
- Anonimato.
- etc.

Riesgo/Beneficio

- Nuevo beneficio: Alto
- Nuevo riesgo: Medio

	B	M	A
B	M	B	B
M	A	M	B
A	A	A	M

Dataset de recogidas taxis NYC

XXX

Activos

- Trayectos, recogidas y entregas de taxis de NYC

Riesgo/beneficio

- Beneficio: Alto
- Riesgo: Alto

	B	M	A
B	M	B	B
M	A	M	B
A	A	A	M

Mitigaciones

- Ratio riesgo/beneficio: Medio
- Eliminación de atributos.
- Eliminación de entradas.
- Agregar datos.
- Generalizar datos.
- Anonimato.
- etc.

Riesgo/Beneficio

- Nuevo beneficio: Alto
- Nuevo riesgo: Medio

	B	M	A
B	M	B	B
M	A	M	B
A	A	A	M

Resultado Publicación de dataset con datos de recogida y entrega generalizados.

- San Francisco «Open Data Release Toolkit»
 - <https://datasf.org/resources/open-data-release-toolkit/>
 - <https://docs.google.com/document/d/1ZGknij29YfoYtYZFn6uGqpqGTXwmZNkdUNaNhK1-6pc>
- Open Data Commons <https://opendatacommons.org/>
- Australia «Open Council Data»
<https://opencouncildata.org/how-to-publish-data/>
- NYU Databrary
<http://databrary.org/resources/policies/best-practices.html>
- European Data Portal «Open Data Licensing»
<https://www.europeandataportal.eu/en/resources/training-companion/open-data-licensing>

Preguntas?