La Gestión de Datos de Investigación

Marisa Pérez Aliende
Universidad Autónoma de Madrid
mp.aliende@uam.es
Sevilla, 13 de Junio 2017

UNIVERSIDAD D SEVILLA
Biblioteca

Compartir y citar los datos de investigación

- Beneficios y desafíos de compartir
- Confidencialidad
  - Anonimización
- Citación de los datos

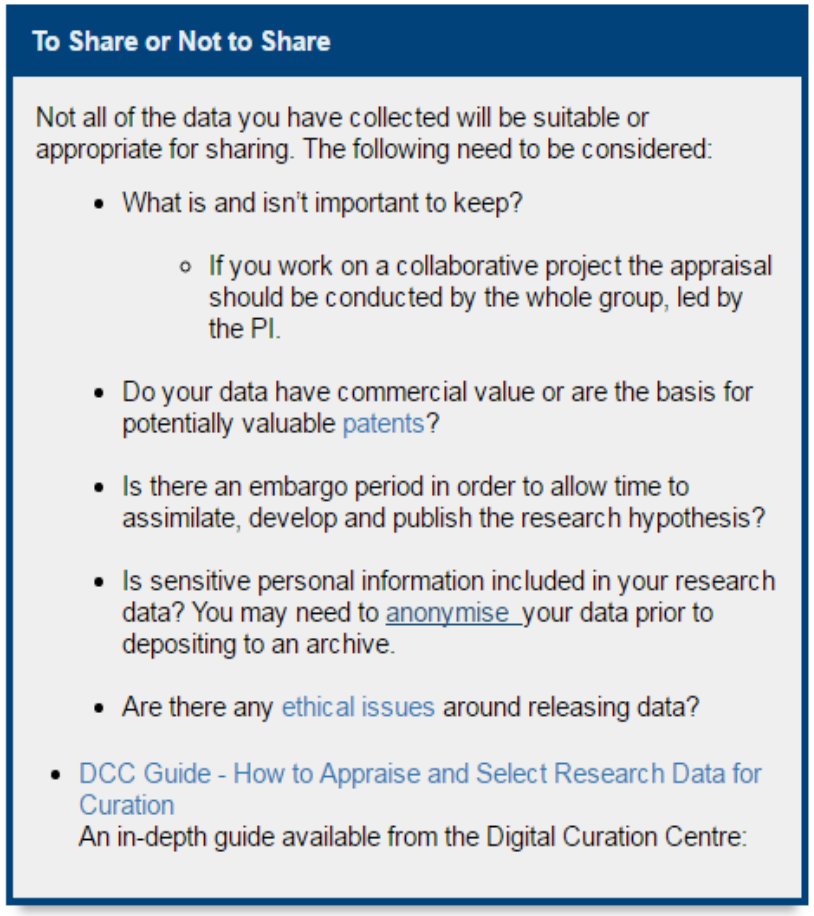
Compartir los datos de investigación: Beneficios

- Refuerza la investigación científica abierta
- Apoya la verificación y replicación de los resultados originales
- Promueve nuevas investigaciones y permite la prueba de métodos nuevos o alternativos
- Fomenta la colaboración
- Proporciona importantes recursos didácticos
- Reduce costes evitando duplicar los esfuerzos
- Protege contra la facultad o fecha fraudulenta
- Mejora la visibilidad y el impacto general de los proyectos de investigación
- Conserva los datos para uso futuro
- Ayuda a una comunidad más amplia y a investigadores individuales a hacer una mejor investigación

MANTRA_Sharing, Preservation and Licensing.

Compartir los datos de investigación: Desafíos

- Se necesita tiempo y esfuerzo para que los datos se puedan compartir
  - 54% tiempo insuficiente
  - 40% falta de financiación
- Riesgos percibidos por la pérdida de control de los datos
- Proteger los datos sensibles
- La propiedad de los datos puede no estar clara o ser problemática
- Faltan incentivos para compartir los datos
- Falta de experiencia en la gestión de datos:
  - Se requieren profesionales de la información que preparen los datos con herramientas para compartirlos de forma eficiente y efectiva



To Share or Not to Share

Not all of the data you have collected will be suitable or appropriate for sharing. The following need to be considered:

- What is and isn't important to keep?
  - If you work on a collaborative project the appraisal should be conducted by the whole group, led by the PI.
- Do your data have commercial value or are the basis for potentially valuable patents?
- Is there an embargo period in order to allow time to assimilate, develop and publish the research hypothesis?
- Is sensitive personal information included in your research data? You may need to anonymise your data prior to depositing to an archive.
- Are there any ethical issues around releasing data?

- DCC Guide - How to Appraise and Select Research Data for Curation
  An in-depth guide available from the Digital Curation Centre:

UCD Library. Research Data Management: Data Sharing

Compartir los datos de investigación: Barreras

No todos los datos se pueden compartir.
Las principales barreras están relacionadas con temas:

- **Financieros**
  - Sobre todo en el ámbito anglosajón, las universidades buscan aprovechar el valor comercial de su propiedad intelectual. Sus datos pueden tener valor financiero
- **Confidencialidad**
  - En estudios con personas

Proteger la confidencialidad recurriendo a la ley:

Los investigadores deben estar familiarizados con el tema de la protección de datos, conocer la normativa y las técnicas de anonimización y aplicar métodos estadísticos para “desidentificar” el dataset.

Información que debe ser protegida:

- Información de identificación personal

Cualquier representación de la información que permita identificar a un individuo, a través de medios:

- Directos: Nombre, nº de SS, teléfono…
- Indirectos (Cuando se enlaza con otra información pública permite identificar al individuo): Raza y etnia, ingresos, profesión…
- Información sanitaria protegida
- Información sensible: Vida sexual, religión, registro criminal .

Según el contabilizan hasta 18 tipos de identificadores (US NIH):

1. Nombre
2. Geografía
3. Fechas

4. Telephone numbers.
5. Facsimile numbers.
6. Electronic mail addresses.
7. Social security numbers.
8. Medical record numbers.
9. Health plan beneficiary numbers.
10. Account numbers.
11. Certificate/license numbers.
12. Vehicle identifiers and serial numbers, including license plate numbers.
13. Device identifiers and serial numbers.
14. Web universal resource locators (URLs).
15. Internet protocol (IP) address numbers.
16. Biometric identifiers, including fingerprints and voiceprints.
17. Full-face photographic images and any comparable images.
18. Any other unique identifying number, characteristic, or code, unless otherwise permitted by the Privacy Rule for re-identification.

Compartir los datos de investigación: Confidencialidad I

Proteger la confidencialidad de los datos personales o sensibles antes de almacenar, compartir o usar. Para ello:

- **ANONIMIZACIÓN**: Procesar los datos personales de forma que se prevenga de manera irreversible la identificación. Puede ser por razones éticas, como proteger la identidad de las personas participantes en la investigación, por razones legales, o comerciales.

  Se pueden aplicar a:

  - **Datos cuantitativos**: Eliminar o agregar variables o reducir la precisión o el significado textual detallado de una variable. Cuidado al relacionar variables en datasets, o con los datos georeferenciados se puede revelar una identidad.
  - **Datos cualitativos:** como transcripción de entrevistas, audios…, se debe usar seudónimos o descriptores más vagos.

  Técnicas de anonimización:

  - Eliminación de **identificadores directos** (nombres, iniciales, dirección fotos, fechas personales…)
  - Eliminación de **identificadores indirectos** (profesión, sexo, enfermedad, lugar de nacimiento…)
  - Modificando los datos para limitar la identificación: eliminando líneas de información, representación parcial de los datos.

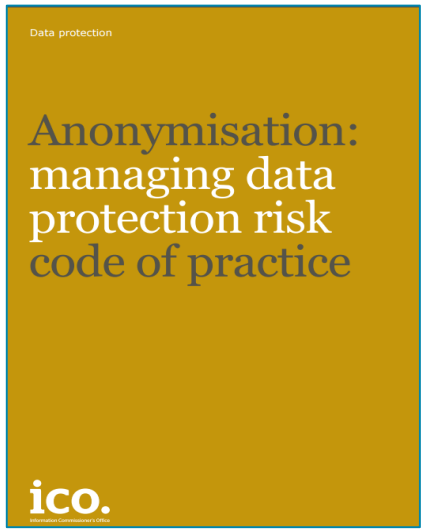
(UK . Data-Archive. Create & manage data. Anonymisation)

Compartir los datos de investigación: Confidencialidad II

Ejemplos de anonimización:

Un nombre propio se sustituye por otra cosa que describe aspectos demográficos con el individuo o la relación personal que existe con el estudiado:

- [F/34/Hispanic] – María Sabater
  [husband] - John



Interview and page number	Original	Changed to
Int1		
p1	Age 27	Age range 20-30
p1	Spain	European country
p3	Manchester	Northern metropolitan city or English provincial city
p2	20th June	June
p2	Amy (real name)	Moira (pseudonym)
Int2		
p1	Francis	my friend
p8	Station Road primary school	a primary school
p10	Head Buyer, Produce, Sainsburys	Senior Executive with leading supermarket chain

(UK. Data-Archive. Create & manage data. Anonymisation)

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income	Expenses/month
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	M	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

Income & Expenses Individual-level dataset

Age	Sex	Postcode	Income (low if <25,000; medium if between 25,000 to 45,000; high if >45,000)	Expenses/month (low if ,1,800; medium if between 1,800 to 2,400; high if >2,400)
20-24	F	SO17-19	low	low
25-29	M	SO17-19	low	low
30-34	M	SO14-16	medium	medium
35-39	F	SO17-19	medium	medium
40-44	F	SO14-16	high	high
50-54	M	SO14-16	medium	low

Restrict the upper or lower ranges of a continuous variable to hide outliers if the values for certain individuals are unusual or atypical within the wider group researched. In such circumstances the unusually large or small values might be collapsed into a single code, even if the other responses are kept as actual quantities, or one might code all responses.

example: *Annual salary could be 'top-coded' to avoid identifying highly paid individuals. A top code of £100,000 or more could be applied, even if lower incomes are not coded into groups.*

(ICO. Anonymisation: managing data protection risk code of practice. 2012)

Compartir los datos de investigación: Confidencialidad III

Recursos de ayuda en la anonimización:

- UKAN: UK Anonymisation network
  http://ukanon.net/external-resources/
- ICO. Information Commissioner´s Office
  https://ico.org.uk/for-organisations/guide-to-data-protection/anonymisation/
- Dictamen del Data Protection Working Party
  http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

(Jarque, P. “Taller sobre gestión de datos para bibliotecarios: Aspectos legales de los datos de investigación”. 2016)

Compartir los datos de investigación: Citación I

La Australian National Data Service (ANDS) describe la citación de los datos como:

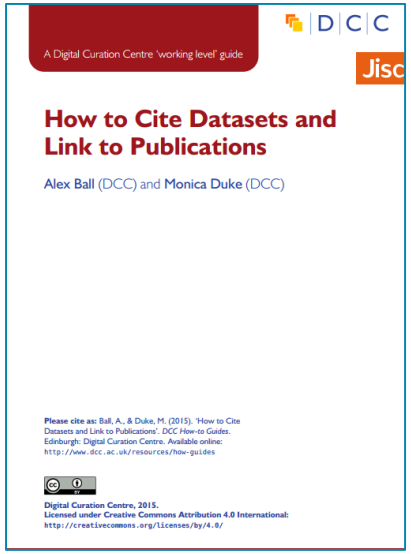
> "La práctica de proporcionar una referencia a un dato, del mismo modo que los investigadores de manera rutinaria proporcionan la referencia bibliográfica de un recurso"

Aunque los datos se compartían en el pasado, raramente se citaban del mismo modo que los artículos u otras publicaciones. Cuando se cita un dataset se puede:

Localizar
Reusar
Volver a citar por otros

Los datos son una parte fundamental del proceso de investigación y una correcta citación debiera ser una característica importante de las publicaciones, permitiendo:

- Conocer las fuentes de autor
- Facilitar la identificación de los datos
- Favorecer la reproducción de los resultados de investigación
- Encontrar los datos
- Rastrear el impacto de los datos
- Reconocer y recompensar a los creadores de los datos



(UK Data Service)

Compartir los datos de investigación: Citación II

Los Data Citation Principles reconocen la doble necesidad de crear prácticas de citación que sean a la vez comprensibles por el hombre y procesadas por la máquina. Define 8 principios:

1. **Importancia**: Los datos debieran considerarse legítimos, citables. Su cita al mismo nivel que el de las publicaciones.
2. **Crédito y atribución**: A todos los creadores de los datos.
3. **Evidencia**: Cuando la demanda se basa en datos, los datos se deben citar.
4. **Identificación única:** La cita debiera incluir un método de identificación persistente que sea accionable de manera automática, único y ampliamente usado.
5. **Acceso:** Las citas debieran facilitar el acceso a los datos y a los metadatos asociados, documentación, códigos y otros materiales necesarios tanto para las personas como para las máquinas de manera que puedan usarse los datos referenciados.
6. **Persistencia:** Los identificadores únicos y los metadatos debieran persistir.
7. **Especificidad y verificabilidad:** Incluir información sobre la procedencia para facilitar la verificación de que los datos recuperados sean los citados.
8. **Interoperabilidad y flexibilidad:** Acomodarse a las prácticas de diferentes comunidades pero sin que interfiera en la interoperabilidad.

(Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014)[

Compartir los datos de investigación: Citación III

Dataverse normaliza la generación de citas, creadas automáticamente al depositar un dataset

Principle 2: Credit and Attribution (e.g. authors, repositories or other distributors and contributors)

Principle 4: Unique Identifier (e.g. DOI, Handle.). **Principle 5, 6 Access, Persistence:** A persistent identifier that provides access and metadata

Author(s), Year, Dataset Title, Data Repository or Archive, Version, Global Persistent Identifier

Principle 7: Specificity and verification
(e.g. the specific version used).
Versioning or timeslice information should be supplied with any updated or dynamic dataset.

Following are two authentic examples of replication data citations:

From International Studies Quarterly, King and Zeng, 2006, p. 209:

> Gary King; Langche Zeng, 2006, "Replication data for: When Can History be Our Guide? The Pitfalls of Counterfactual Inference", Harvard Dataverse, V2, http://hdl.handle.net/1902.1/DXRXCFAWPK UNF:3:DaYIT6QSX9r0D50ye+tXpA==

From Political Analysis, Hanmer, Banks, and White, 2013:

> Hanmer, Michael J.; Banks, Antoine J., White, Ismail K., 2013, "Replication data for: Experiments to Reduce the Over-reporting of Voting: A Pipeline to the Truth", Harvard Dataverse, V1, http://dx.doi.org/10.7910/DVN/22893 UNF:5:eJOVAjDU0E0jzSQ2bRCg9g==

(Dataverse Project. Data Citation)

UNF (Universal Numerical Fingerprints) es un *hash* criptográfico de los datos que asegurará que no se han producido cambios. Sus datos son idénticos a los usados en las publicaciones en épocas anteriores, incluso aunque haya cambiado el hardware, almacenamiento.

Compartir los datos de investigación: Citación IV

Reconoce que los desafíos asociados con la publicación de datos varían entre disciplinas

Recomienda el siguiente formato:

Autor (Año de publicación). Título. Editor. Identificador

Es recomendable incluir información sobre el tipo de recurso y la versión:

Autor (Año de publicación). Título. Versión. Editor. Tipo de recurso. Identificador

Donde:

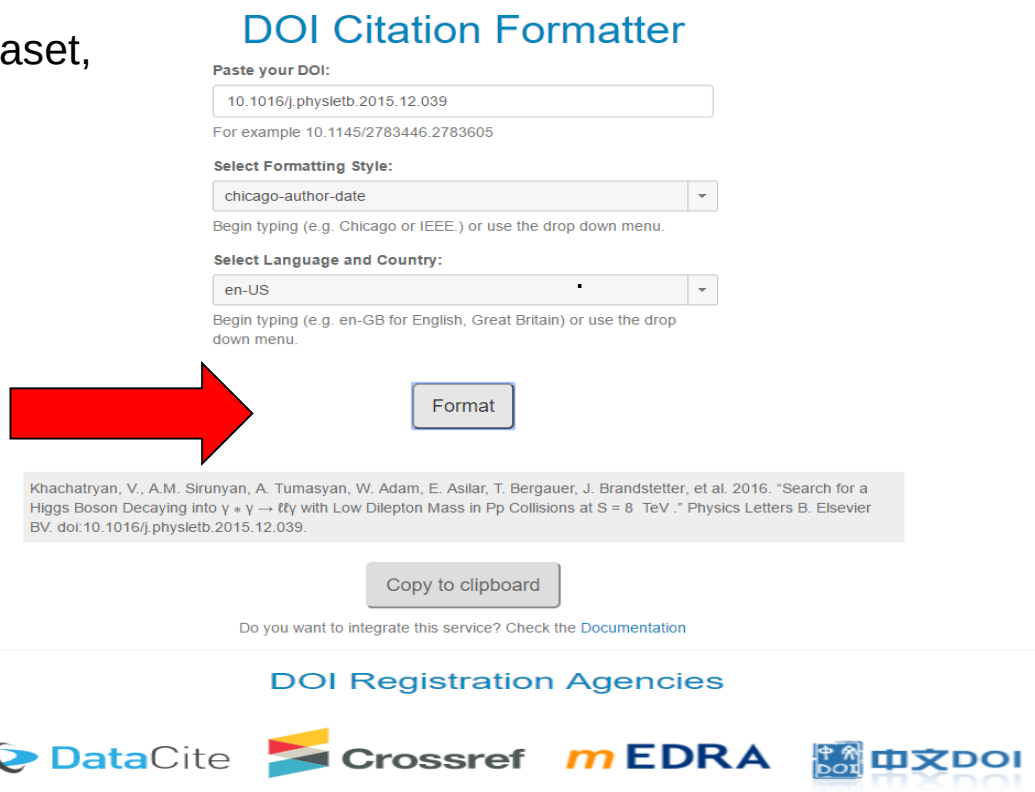
Editor: Es la organización que proporciona acceso al dataset, ej.: Dryad, Zenodo

Tipo de recurso: Por ejemplo, base de datos, dataset

DataCite en colaboración con Crossref, mEDRA e ISTIC:

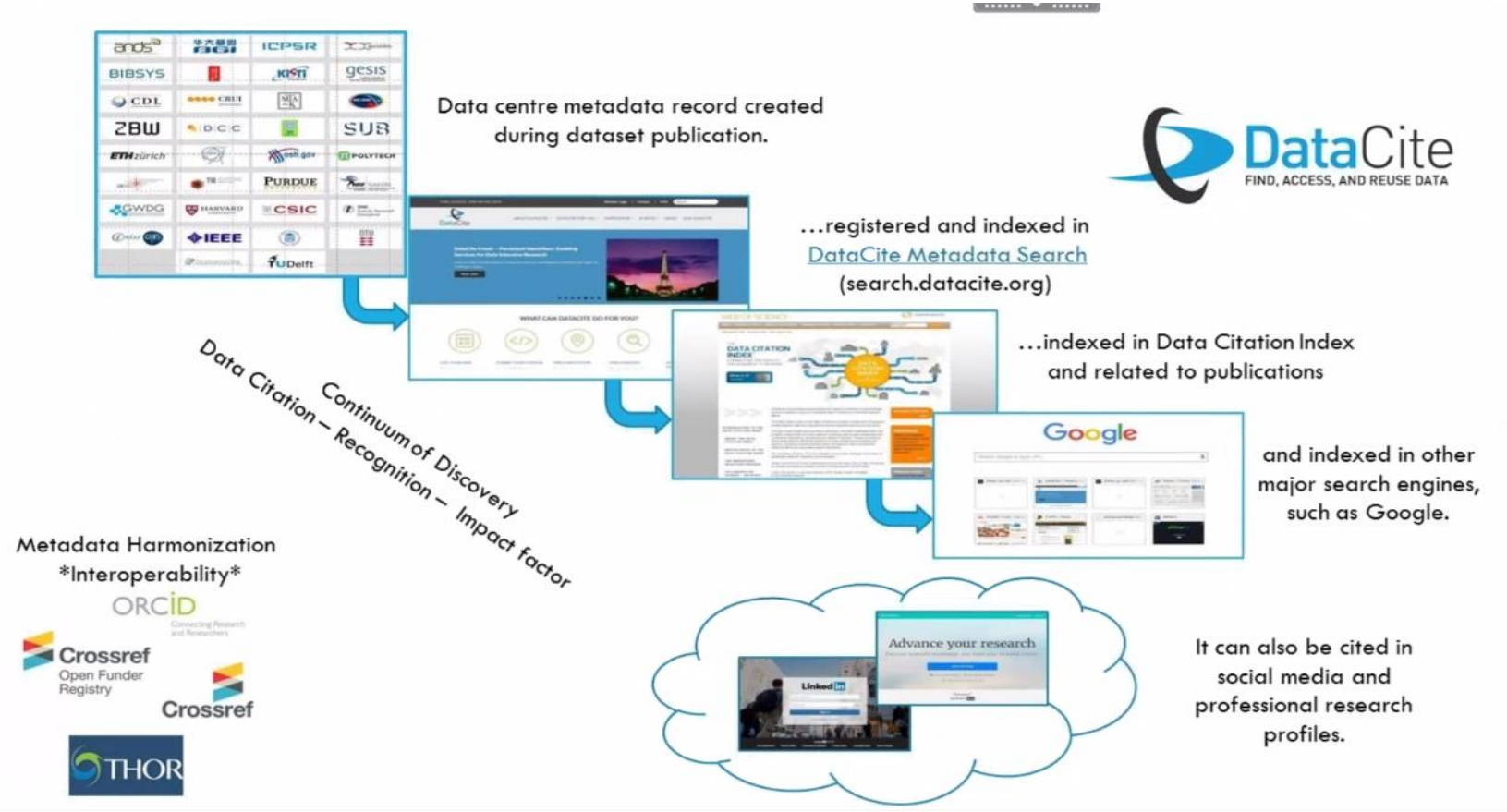
The DOI Citation Formatter
(http://citation.crosscite.org/)

A partir del DOI contruye una cita completa en múltiples estilos de citación.



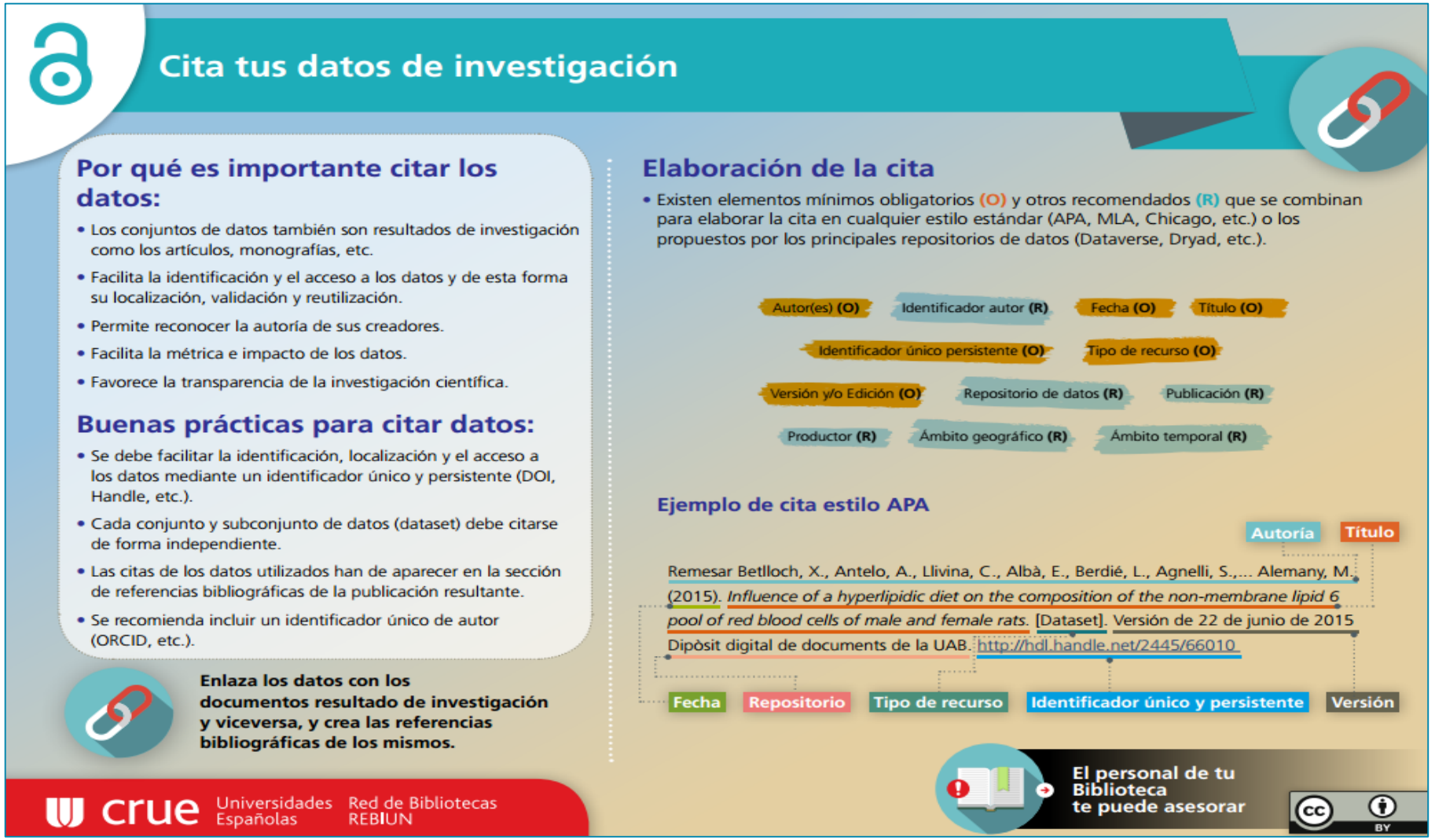
Compartir los datos de investigación: Citación V

En Septiembre 2016 DataCite lanza el Metadata Schema 4.0 (http://doi.org/10.5438/0012) :
Campos obligatorios: Identificador, autor, título, editor, años de publicación y tipo de recurso
Campos opcionales: Idioma, formato, versión, etc.



(Starr. Joan (2016). DataCite Metadata Schema. https://player.vimeo.com/video/172929697)

Compartir los datos de investigación: Citación VI



Compartir los datos de investigación: Citación VI

Publicar ideas cientificas en vez de resultados | Lucas Sanchez | TEDxValladolid

La Gestión de Datos de Investigación

Marisa Pérez Aliende
Universidad Autónoma de Madrid
mp.aliende@uam.es
Sevilla, 13 de Junio 2017

UNIVERSIDAD D SEVILLA
Biblioteca