Reutilización de datos abiertos y documentos

Tutoriales interactivos en R (webR)

Pablo Sánchez-Núñez
Universidad Carlos III de Madrid

26 de noviembre de 2026

Índice

Presentación 1
Primeros pasos con R en el navegador 5
Tema 0. Introducción a los datos abiertos y la reutilización de la información 9
Tema 1. Reutilización de la información del sector público (RISP) 22
Tema 2. Proyectos relevantes de reutilización de datos abiertos 35
Tema 3. Gestión documental y *open data* 47
Tema 4. Normalización e interoperabilidad 62
Tema 5. Empresas infomediarias y oportunidades de negocio 77
Caso práctico: las bibliotecas públicas municipales de Madrid 89
Laboratorio de limpieza de datos 95
Chuleta de R 97
Guía para el profesorado 100
Glosario 105
Referencias 110
Bibliografía 110

Presentación

Ficha del libro

- **Autor:** Pablo Sánchez-Núñez
- **ORCID:** https://orcid.org/0000-0001-7845-9506
- **Web del autor:** https://pablosancheznunez.com
- **Afiliación:** Universidad Carlos III de Madrid
- **DOI:** https://doi.org/10.5281/zenodo.24500934
- **Licencia:** CC BY 4.0
- **Versión web interactiva:** https://pablosancheznunez.com/reutilizacion-datos/
- **Cómo citarlo:** Sánchez-Núñez, P. (2026). *Reutilización de datos abiertos y documentos: tutoriales interactivos en R (webR)*. Zenodo. https://doi.org/10.5281/zenodo.24500934

Sobre esta versión en PDF

Esta es la versión para leer e imprimir. El código de R aparece sin ejecutar y las preguntas de autoevaluación se muestran sin corrección automática. Para ejecutar el código, hacer los ejercicios con corrección y responder a las preguntas, usa la versión web: https://pablosancheznunez.com/reutilizacion-datos/, o descarga el script `.R` de cada capítulo.

Este libro acompaña a la asignatura **Reutilización de datos abiertos y documentos** (código 17300) del Máster Universitario en Bibliotecas, Archivos y Continuidad Digital *Online* de la Universidad Carlos III de Madrid: optativa de 3 ECTS, primer curso, segundo cuatrimestre.

Cada capítulo corresponde a un tema de la asignatura y combina tres cosas:

1. **Una síntesis de los contenidos**, con la lectura obligatoria del tema y los recursos de referencia.
2. **Tutoriales interactivos en R** que se ejecutan en el propio navegador gracias a webR. No hay que instalar nada.
3. **Ejercicios autocorregidos**, con pistas y solución, para practicar lo que después se evalúa en los cuestionarios de Aula Global, y **preguntas de autoevaluación** (verdadero/falso, opción múltiple y completar) al final de cada tema.

Cada capítulo enlaza además a su parte del glosario y ofrece **descargas**: el script `.R` del capítulo, una plantilla de informe `.qmd` y su tema de estilo `.scss`.

Objetivos de la asignatura

Según la ficha de la asignatura, al terminar el curso deberías:

- Conocer los principios jurídicos y éticos, y la legislación, que sustentan la reutilización de la información.
- Conocer las principales iniciativas y proyectos de reutilización de información del sector público en distintos ámbitos administrativos y geográficos.
- Ser capaz de plantear un proyecto de reutilización de información y documentos.
- Aplicar normas y herramientas que favorecen la normalización y la interoperabilidad de datos y documentos.
- Conocer los principios de los datos abiertos.

Mapa del libro

Cada tema sigue la estructura del temario oficial: empieza por sus **objetivos**, desarrolla los **contenidos** con la misma numeración del temario y termina con una sección de **práctica en R** cuyos pasos indican a qué apartado corresponden.

Capítulo	Contenidos del temario	Práctica en R
Primeros pasos	—	Ejecutar código, vectores, tablas, leer un CSV
Tema 0. Introducción a los datos abiertos y la reutilización de la información	Sociedad de la información y del conocimiento · Internet y la WWW · Evolución de la Web · Gobierno abierto y datos abiertos · Cinco estrellas · FAIR · Datos abiertos y públicos · Formatos abiertos y propietarios · Licencias abiertas · Recursos y portales	Auditar el CSV: estrellas, licencias, formatos, público frente a abierto
Tema 1. Reutilización de la información del sector público	Conceptos básicos · Evolución de las políticas · Marco europeo · Marco español · Transparencia y acceso · Derechos y obligaciones · Licencias de reutilización	Cronología normativa, plazos en meses y en días hábiles, régimen sancionador
Tema 2. Proyectos relevantes de reutilización	Open Data en España · En la UE · Portales internacionales · Casos de éxito · Iniciativas públicas y privadas · Datos para la investigación y la innovación	Clasificar casos con Safarov et al., autoevaluación FAIR

Capítulo	Contenidos del temario	Práctica en R
Tema 3. Gestión documental y *open data*	Gestión documental aplicada · Ciclo de vida · Metadatos · Calidad · Publicación · Preservación y actualización	Laboratorio de calidad, metadatos DCAT, sumas de control y versiones
Tema 4. Normalización e interoperabilidad	Estándares abiertos · Interoperabilidad técnica, semántica y organizativa · ENI · Formatos CSV, XML, JSON y RDF · Linked Open Data · Vocabularios y ontologías	XML, JSON, JSON-LD y Turtle; consultas sobre un grafo; tesauro en SKOS
Tema 5. Empresas infomediarias y oportunidades de negocio	Sector infomediario · Modelos de negocio · Servicios de valor añadido · Emprendimiento e innovación · IA y reutilización · Tendencias de futuro	Enriquecer datos, punto de equilibrio *freemium*, licencias combinadas
Caso práctico: bibliotecas de Madrid	Recorrido por los seis temas con datos reales del Ayuntamiento de Madrid	Codificación, completitud, JSON-LD con coordenadas, límites del texto libre

El conjunto de datos de práctica

Todo el libro usa el mismo fichero, `prestamos_bibliotecas_practica.csv`: préstamos y usuarios activos mensuales (enero a junio de 2025) de seis bibliotecas municipales de cinco municipios del sur de Madrid.

> ! Datos sintéticos
>
> Las bibliotecas y las cifras son **inventadas** para la docencia. El fichero contiene, a propósito, errores típicos de los datos reales: variantes del mismo municipio, fechas en dos formatos, valores ausentes codificados de dos maneras, un valor negativo y filas duplicadas. Los descubrirás y corregirás en el Tema 3.

Cómo funcionan las celdas interactivas

Las cajas de código de este libro son editables. Al abrir una página, el navegador descarga un entorno de R (tarda unos segundos la primera vez). Después:

- **Run Code** ejecuta el código de la caja.
- **Start Over** devuelve la caja a su estado inicial.
- En los ejercicios, los huecos `_____` son lo que tienes que completar. **Show Hint** muestra pistas progresivas; **Show Solution**, la solución. Al ejecutar, el sistema comprueba tu respuesta y te dice si es correcta.

Prueba ahora: cambia el número y pulsa **Run Code**.

```r
anio <- 2026
paste("Bienvenida al curso", anio)
```

> Si algo no funciona
>
> Usa un navegador de escritorio actualizado (Chrome, Firefox, Edge o Safari). Si una caja no responde, recarga la página. Las celdas de un mismo capítulo comparten memoria: si una celda usa un objeto creado antes, ejecuta primero las celdas anteriores.

Relación con la evaluación

Los ejercicios del libro **no cuentan para la nota**: son práctica. La evaluación continua (50 %) se hace con los cuestionarios de Aula Global, los foros y los trabajos; el examen final (50 %) es presencial en la UC3M y hay que aprobarlo para superar la asignatura. Los cuestionarios de Aula Global preguntan sobre los mismos conceptos y, en los Temas 0, 3 y 4, sobre el mismo CSV de práctica.

Autoría y cómo citar

Autor: Pablo Sánchez-Núñez, Universidad Carlos III de Madrid
ORCID: 0000-0001-7845-9506
Web: pablosancheznunez.com

DOI: 10.5281/zenodo.24500934.

> Sánchez-Núñez, P. (2026). *Reutilización de datos abiertos y documentos: tutoriales interactivos en R (webR).* Universidad Carlos III de Madrid. Zenodo. https://doi.org/10.5281/zenodo.24500934

DOI de todas las versiones: 10.5281/zenodo.24500934; apunta siempre a la última edición. Cada edición tiene además su propio DOI, que encontrarás en la ficha de Zenodo.

Normativa básica de la asignatura

- Directiva (UE) 2019/1024, relativa a los datos abiertos y la reutilización de la información del sector público.
- Ley 37/2007, sobre reutilización de la información del sector público (texto consolidado).
- Ley 19/2013, de transparencia, acceso a la información pública y buen gobierno.

Primeros pasos con R en el navegador

No hace falta saber programar para seguir este libro. Este capítulo enseña lo mínimo para entender y modificar el código de los tutoriales: unos 30 minutos.

> **i** Glosario del tema
>
> R · webR · Data frame · Valor ausente

> **Descargas**
>
> Script R del capítulo Plantilla de informe (.qmd) Tema de estilo (.scss) CSV de práctica
>
> El script reúne todo el código del capítulo para trabajar en RStudio o Positron. La plantilla `.qmd` y su tema `.scss` sirven para redactar las prácticas opcionales: guárdalos en la misma carpeta.

R como calculadora

Escribe una operación y pulsa **Run Code**. R devuelve el resultado precedido de `[1]` (el número indica la posición del primer valor de la línea).

```
2637 + 2626 + 2374
```

Objetos: guardar un valor con un nombre

La flecha `<-` guarda un valor en un objeto. Después puedes usar el nombre en lugar del valor.

```
prestamos_enero <- 2637
prestamos_febrero <- 2626
prestamos_enero + prestamos_febrero
```

Vectores: varios valores del mismo tipo

`c()` («combinar») crea un vector. Muchas funciones trabajan sobre vectores enteros: `sum()`, `mean()`, `length()`, `max()`…

```
prestamos <- c(2637, 2626, 2374, 2227, 2656, 2319)
sum(prestamos)
mean(prestamos)
length(prestamos)
```

Los textos van entre comillas:

```
municipios <- c("Getafe", "Leganés", "Alcorcón")
toupper(municipios)
```

Ejercicio 1

Calcula el **máximo** de préstamos del vector `prestamos`.

> **i** Ejercicio: completa los huecos `_____`
>
> ```
> _____(prestamos)
> ```

Pista

La función se llama igual que en una hoja de cálculo, en minúsculas.

Solución

```r
max(prestamos)
```

Comparaciones y valores lógicos

Una comparación devuelve `TRUE` o `FALSE` para cada elemento. `sum()` sobre un vector lógico cuenta los `TRUE`.

```r
prestamos > 2500
sum(prestamos > 2500)
```

El operador `%in%` pregunta si cada elemento está en una lista:

```r
licencias <- c("CC BY 4.0", "CC BY-NC 4.0", "CC0 1.0")
licencias %in% c("CC BY 4.0", "CC0 1.0")
```

Ejercicio 2

¿Cuántos meses tuvieron **menos de 2400** préstamos?

Ejercicio: completa los huecos _____

```r
sum(prestamos ______ 2400)
```

Solución

```r
sum(prestamos < 2400)
```

Tablas de datos (*data frames*)

Un *data frame* es una tabla: cada columna es un vector. Con `$` se accede a una columna.

```r
bibliotecas <- data.frame(
  nombre = c("Centro", "Sur", "Norte"),
  municipio = c("Getafe", "Getafe", "Leganés"),
  prestamos = c(14839, 15319, 7547)
)
bibliotecas
bibliotecas$prestamos
nrow(bibliotecas)
```

Para quedarse con algunas filas se usa `subset()`:

```r
subset(bibliotecas, municipio == "Getafe")
```

Ejercicio 4

¿Cuántos préstamos suman las dos bibliotecas de **Getafe**? Combina `subset()`, `$` y `sum()`.

ℹ Ejercicio: completa los huecos `_____`

```
sum(subset(bibliotecas, municipio == _____)$prestamos)
```

ℹ Pista

El nombre del municipio es un texto: va entre comillas.

💡 Solución

```
sum(subset(bibliotecas, municipio == "Getafe")$prestamos)
```

Leer un fichero CSV

El CSV de práctica está disponible en la carpeta `datos`. `read.csv()` lo convierte en un *data frame*; `head()` muestra las primeras filas y `str()` describe las columnas.

```
prestamos_df <- read.csv("datos/prestamos_bibliotecas_practica.csv")
head(prestamos_df)
str(prestamos_df)
```

Fíjate en `str()`: la columna `prestamos` aparece como `chr` (texto), no como número. Algo en el fichero impide que R la reconozca como numérica. Lo resolverás en el Tema 3.

Ejercicio 3

¿Cuántas filas tiene el CSV de práctica?

ℹ Ejercicio: completa los huecos `_____`

```
_____(prestamos_df)
```

ℹ Pista

Lo usaste unas líneas más arriba con la tabla `bibliotecas`.

💡 Solución

```
nrow(prestamos_df)
```

Ejercicio 5

¿Cuántas **bibliotecas distintas** aparecen en el CSV de práctica? `unique()` elimina los valores repetidos de un vector; después, cuenta sus elementos.

Ejercicio: completa los huecos `_____`

```
______(unique(prestamos_df$biblioteca))
```

Pista

Para contar los elementos de un vector usaste `length()`.

Solución

```
length(unique(prestamos_df$biblioteca))
```

Seis bibliotecas, cada una con varios meses de datos.

Autoevaluación

Responde en el papel y comprueba después con las explicaciones del recuadro. En la versión web estas preguntas se corrigen automáticamente. No cuentan para la nota.

1. En R, `x <- 5` guarda el valor 5 en un objeto llamado `x`. *[V / F]*
2. ¿Qué devuelve `length(c(3, 8, 1))`? _____
3. Para leer un fichero CSV se usa la función *[write.csv() / read.csv() / csv() / open()]*
4. Completa el código para contar los meses con más de 2500 préstamos: `sum(prestamos _____ 2500)`
5. ¿Qué operador selecciona una columna de un *data frame*?

(a) @ (b) $ (c) # (d) %in%

Explicaciones

- `<-` asigna; el valor queda guardado con ese nombre.
- `length()` cuenta los elementos del vector: 3.
- `read.csv()` lee; `write.csv()` escribe.
- La comparación `>` devuelve TRUE/FALSE y `sum()` cuenta los TRUE.
- `df$columna` devuelve la columna como vector.

Para seguir

El Apéndice A reúne en una página todas las funciones que usa el libro.

Tema 0. Introducción a los datos abiertos y la reutilización de la información

Nivelación

Objetivos

- Comprender el contexto en el que surge el movimiento *open data*.
- Diferenciar los conceptos fundamentales relacionados con los datos abiertos.
- Conocer los principios básicos que favorecen la reutilización de la información.

Glosario del tema

Datos abiertos · Open Definition · Datos públicos · Gobierno abierto · Cinco estrellas · FAIR · Formato abierto · Legible por máquina · Licencia abierta · Web Semántica

Descargas

Script R del capítulo Plantilla de informe (.qmd) Tema de estilo (.scss) CSV de práctica

El script reúne todo el código del capítulo para trabajar en RStudio o Positron. La plantilla `.qmd` y su tema `.scss` sirven para redactar las prácticas opcionales: guárdalos en la misma carpeta.

Lectura obligatoria

Janssen, M.; Charalabidis, Y.; Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29(4), 258-268. Acceso por suscripción a través de la Biblioteca UC3M.

Complementaria: Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018 (acceso abierto).

Lecturas complementarias

- Attard et al. (2015) — Revisión de iniciativas de datos abiertos de gobierno y de su ciclo de vida.
- Zuiderwijk y Janssen (2014) — Marco para comparar políticas de datos abiertos.

1. Sociedad de la información y sociedad del conocimiento

La expresión **sociedad de la información** se generaliza en las décadas de 1970 a 1990 para describir sociedades en las que producir, tratar y transmitir información se convierte en la actividad económica y social central, apoyada en las tecnologías digitales. El énfasis está en la **infraestructura**: redes, ordenadores, volumen de datos que circula.

La **sociedad del conocimiento**, término que la UNESCO popularizó en su informe *Hacia las sociedades del conocimiento* (2005), desplaza el foco hacia la capacidad de las personas y las instituciones para **convertir la información en conocimiento** y usarlo: educación, alfabetización informacional, diversidad cultural, acceso equitativo. Tener datos no basta; hace falta saber encontrarlos, evaluarlos y combinarlos.

Las bibliotecas están en el centro de ese paso: median entre la información disponible y las competencias para usarla. La reutilización de datos abiertos es una prolongación natural de esa misión.

2. Internet y la World Wide Web

Conviene no confundirlos:

	Internet	World Wide Web
Qué es	La red de redes: infraestructura que conecta ordenadores	Un servicio que funciona sobre Internet: documentos enlazados
Origen	ARPANET (1969); protocolos TCP/IP (1983)	Tim Berners-Lee en el CERN (1989-1991)
Piezas básicas	TCP/IP, DNS, encaminamiento	URL/URI, HTTP, HTML
Otros servicios	Correo, transferencia de ficheros, mensajería…	—

Las tres piezas de la Web son las mismas que, años después, sostendrán los datos enlazados: **identificadores** (URI), un **protocolo** para pedirlos (HTTP) y un **formato** para representarlos (HTML para documentos; RDF para datos).

3. Evolución de la Web: de la Web 1.0 a la Web Semántica

Etapa	Rasgo principal	Ejemplo en bibliotecas
Web 1.0 (años 90)	Páginas estáticas de lectura; pocos publican y muchos leen	Web institucional con horarios en HTML
Web 2.0 (desde 2004 aprox.)	Participación, contenido generado por usuarios, servicios y API	OPAC social con reseñas; catálogo en redes sociales
Web Semántica / Web de datos	Datos con significado explícito y enlazados entre sí, procesables por máquinas	datos.bne.es; registros enlazados con Wikidata y VIAF

La Web Semántica, propuesta por Berners-Lee, Hendler y Lassila en 2001, pretende que las máquinas no solo muestren documentos sino que **entiendan** los datos: que «Getafe» sea una entidad identificada, no una cadena de texto. Sus tecnologías (RDF, vocabularios, SPARQL) se ven en el Tema 4. Algunos autores hablan después de una «Web 3.0» o de una web descentralizada; el término no tiene un significado único y no lo usaremos.

4. Gobierno abierto y datos abiertos

El **gobierno abierto** (*open government*) es una forma de gobernar basada en tres pilares: **transparencia** (rendir cuentas y publicar información), **participación** (implicar a la ciudadanía en las decisiones) y **colaboración** (trabajar con ciudadanía, empresas y otras administraciones). Se impulsa a partir del memorando de la administración Obama de 2009 y de la **Alianza para el Gobierno Abierto** (*Open Government Partnership*, 2011), de la que España forma parte y en la que aprueba planes de acción periódicos.

Los **datos abiertos** son una herramienta del gobierno abierto, no su sinónimo. La Open Definition 2.1 de la Open Knowledge Foundation dice que el conocimiento es abierto si cualquiera puede acceder a él, usarlo, modificarlo y compartirlo libremente, sujeto como mucho a requisitos que preserven su procedencia y su apertura. En la práctica, un conjunto de datos es abierto cuando cumple a la vez:

- **Licencia abierta**: permite uso, redistribución y modificación para cualquier fin, también comercial.
- **Acceso**: se puede descargar entero por Internet, sin coste o a un coste razonable de reproducción.
- **Legible por máquina** y en **formato abierto**.

Beneficios, barreras y mitos

Janssen et al. contrastan los beneficios que se atribuyen a los datos abiertos (políticos y sociales, económicos, operativos y técnicos) con las barreras reales de adopción y con cinco creencias extendidas que matizan. En síntesis, los mitos son que publicar datos produce beneficios de forma automática, que toda la información debe publicarse sin restricciones, que basta con poner los datos en la red, que cualquier ciudadano puede aprovecharlos y que los datos abiertos traen por sí solos un gobierno abierto.

5. El modelo de las cinco estrellas de Tim Berners-Lee

Berners-Lee propuso en 2010 una escala acumulativa para valorar cómo se publican los datos (5 ★ Open Data):

Estrellas	Requisito (acumulativo)	Ejemplo
★	En la web, con licencia abierta, en cualquier formato	PDF escaneado de la memoria anual
★★	Datos estructurados	Hoja de Excel (.xlsx)
★★★	Formato no propietario	CSV
★★★★	URI para identificar las cosas	RDF con URI para cada biblioteca
★★★★★	Enlazados con otros datos	RDF enlazado a Wikidata o a datos.bne.es

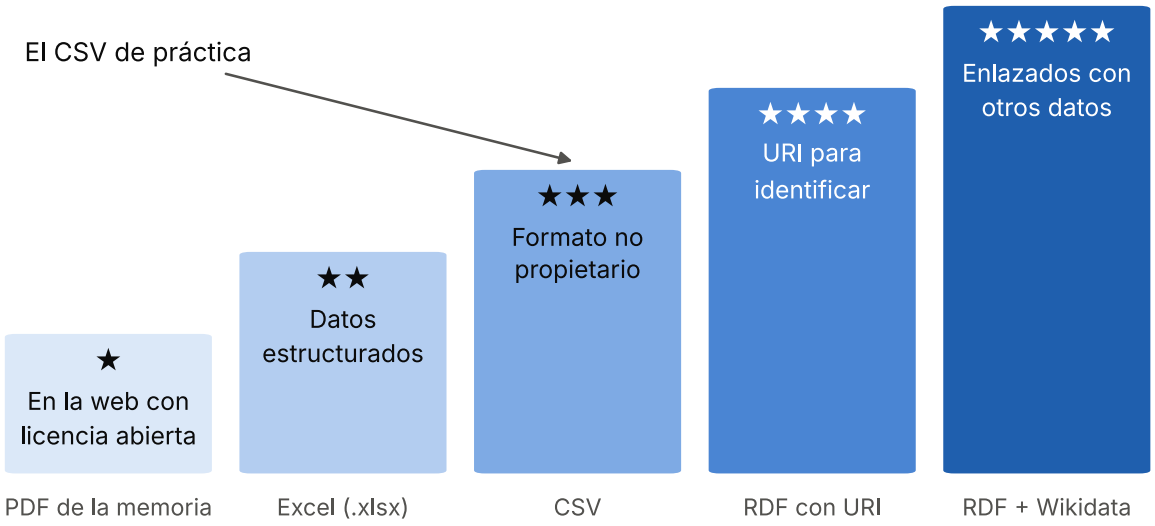

Figura 1: El modelo de cinco estrellas es acumulativo: cada nivel exige los anteriores. El CSV de práctica se queda en tres estrellas; para subir a cuatro habría que identificar cada biblioteca con una URI.

6. Principios FAIR

FAIR no es sinónimo de abierto: describe datos **localizables** (*Findable*), **accesibles** (*Accessible*), **interoperables** (*Interoperable*) y **reutilizables** (*Reusable*), aunque el acceso esté restringido. Los cuatro principios se desarrollan en 15 subprincipios (F1-F4, A1-A2, I1-I3, R1), recogidos por GO FAIR. Uno de ellos (A2) pide que los metadatos sigan accesibles aunque los datos ya no lo estén. Nacieron en el ámbito de la investigación, pero hoy se aplican a cualquier dato.

Principio	Qué exige, en síntesis
Localizable	Identificador persistente y metadatos ricos, registrados en un buscador o catálogo
Accesible	Recuperable por su identificador con un protocolo abierto, con autenticación si hace falta
Interoperable	Lenguaje formal y vocabularios compartidos; referencias a otros datos

Principio	Qué exige, en síntesis
Reutilizable	Licencia clara, procedencia documentada y normas de la comunidad

Laboratorio: autoevaluación FAIR

> ℹ Laboratorio interactivo
>
> Piensa en un conjunto de datos que conozcas (de un portal, de tu trabajo o de un TFM) y responde. El perfil muestra qué principio está más flojo y qué hacer para mejorarlo. Es una orientación inspirada en los subprincipios de GO FAIR, no una certificación.
>
> Este laboratorio es interactivo en la versión web: pablosancheznunez.com/reutilizacion-datos

7. Datos abiertos y datos públicos

Público no significa **abierto**. Un dato es *público* cuando lo produce o custodia una administración (o es accesible al público); es *abierto* cuando, además, cumple la Open Definition. Las dos categorías se cruzan:

	Abierto	No abierto
Público	Préstamos por biblioteca en CSV con CC BY 4.0	Memoria en PDF sin licencia; datos personales de los usuarios
No público (privado)	OpenStreetMap; datos que una empresa publica con ODbL	Registros internos de una editorial

Hay además información pública que **no debe** abrirse: datos personales, información clasificada, documentos sujetos a derechos de propiedad intelectual de terceros. La reutilización de información del sector público (Tema 1) se ocupa precisamente de qué información pública puede abrirse y en qué condiciones.

8. Formatos abiertos y formatos propietarios

Un **formato abierto** tiene una especificación pública y documentada que cualquiera puede implementar sin pagar ni pedir permiso; un **formato propietario** depende de un fabricante, que controla su especificación. Para reutilizar, interesa además que el formato sea **legible por máquina**, es decir, que un programa pueda extraer los datos sin intervención humana.

Formato	¿Abierto?	¿Legible por máquina?	Comentario
CSV	Sí	Sí	Tabla en texto plano (RFC 4180)
JSON	Sí	Sí	Estándar ECMA-404 / RFC 8259
XML	Sí	Sí	Recomendación del W3C
RDF (Turtle, JSON-LD…)	Sí	Sí	Recomendaciones del W3C
ODS	Sí	Sí	OpenDocument (ISO/IEC 26300)
XLSX	Sí (ISO/IEC 29500)	Sí	Estandarizado, pero ligado en la práctica a un fabricante

Formato	¿Abierto?	¿Legible por máquina?	Comentario
XLS (antiguo)	No	Sí	Binario propietario
PDF	Sí (ISO 32000)	Poco	Pensado para leer, no para extraer datos
DOCX, imagen escaneada	—	No	No sirven para publicar datos

El modelo de cinco estrellas trata Excel como «estructurado pero propietario» (dos estrellas). XLSX es hoy una norma ISO, pero sigue siendo preferible publicar CSV u ODS.

9. Licencias abiertas

Una licencia es abierta si permite usar, redistribuir y modificar para cualquier fin. Puede exigir **atribución** o **compartir igual**, pero no prohibir el uso comercial ni las obras derivadas.

Licencia	¿Conforme con la Open Definition?	Motivo
CC0 1.0	Sí	Dedicación al dominio público
CC BY 4.0	Sí	Solo exige atribución
CC BY-SA 4.0	Sí	Atribución y compartir igual
ODbL 1.0	Sí	Licencia de bases de datos con atribución y compartir igual
CC BY-NC 4.0	**No**	Prohíbe el uso comercial
CC BY-ND 4.0	**No**	Prohíbe obras derivadas
Sin licencia	**No**	Sin autorización expresa, la reutilización no está clara

Práctica en R: audita el CSV de práctica

Paso 1. Cargar y describir el fichero

```
prestamos_df <- read.csv("datos/prestamos_bibliotecas_practica.csv")
dim(prestamos_df)        # filas y columnas
names(prestamos_df)      # nombres de las columnas
head(prestamos_df, 3)
```

Paso 2. Comprobar la licencia

Supongamos que el portal municipal que publica el CSV declara la licencia **CC BY 4.0**. Guardamos en un vector las licencias conformes con la Open Definition y comprobamos:

```
licencias_abiertas <- c("CC0 1.0", "CC BY 4.0", "CC BY-SA 4.0", "ODbL 1.0", "ODC-BY
1.0", "PDDL 1.0")
licencia_csv <- "CC BY 4.0"
licencia_csv %in% licencias_abiertas
```

Paso 3. Calcular las estrellas

La función `estrellas()` aplica la escala del apartado 5 de forma acumulativa: si falla un nivel, no se sigue contando.

```r
estrellas <- function(licencia_abierta, estructurado, no_propietario, usa_uri, enlazado)
{
  criterios <- c(licencia_abierta, estructurado, no_propietario, usa_uri, enlazado)
  if (!criterios[1]) return(0)
  # cuenta los TRUE consecutivos desde el principio
  sum(cumprod(criterios))
}

# Un Excel con licencia abierta:
estrellas(TRUE, TRUE, FALSE, FALSE, FALSE)
```

Ejercicio 0.1

Completa los argumentos para el CSV de práctica (licencia CC BY 4.0, formato CSV, sin URI ni enlaces).

ℹ Ejercicio: completa los huecos _____

```r
estrellas(
  licencia_abierta = ______,
  estructurado     = ______,
  no_propietario   = ______,
  usa_uri          = ______,
  enlazado         = ______
)
```

ℹ Pista

Cada argumento es `TRUE` o `FALSE`. Un CSV es estructurado y no propietario, pero no identifica las bibliotecas con URI.

Solución

```r
estrellas(TRUE, TRUE, TRUE, FALSE, FALSE)
```

El CSV tiene **tres estrellas**. Para llegar a cuatro habría que identificar cada biblioteca (y cada municipio) con una URI y publicarlo en RDF, algo que harás en el Tema 4.

Paso 4. Auditar un catálogo

Un portal ficticio publica estos ocho conjuntos de datos de bibliotecas:

```r
catalogo <- data.frame(
  conjunto = c("Préstamos mensuales", "Horarios", "Memoria anual", "Catálogo de fondos",
               "Visitas por sede", "Encuestas de satisfacción", "Agenda cultural",
"Sedes"),
  formato  = c("CSV", "XLSX", "PDF", "RDF", "CSV", "XLSX", "JSON", "RDF"),
  licencia = c("CC BY 4.0", "CC BY-NC 4.0", "CC BY 4.0", "CC0 1.0",
               "Sin licencia", "CC BY-ND 4.0", "CC BY 4.0", "ODbL 1.0")
)
catalogo
```

Ejercicio 0.2

¿Cuántos conjuntos del catálogo tienen una licencia conforme con la Open Definition? Usa `licencias_abiertas` y `%in%`.

Ejercicio: completa los huecos _____

```
sum(catalogo$licencia ______ licencias_abiertas)
```

Pista

`%in%` devuelve `TRUE`/`FALSE` para cada fila; `sum()` cuenta los `TRUE`.

Solución

```
sum(catalogo$licencia %in% licencias_abiertas)
```

Cinco conjuntos. «Sin licencia» no es abierto: sin licencia expresa, la reutilización no está autorizada de forma clara.

Ejercicio 0.3

Ahora filtra el catálogo para quedarte con los conjuntos que tienen **licencia abierta y formato CSV, JSON o RDF** (tres estrellas o más). Devuelve solo la columna `conjunto`.

Ejercicio: completa los huecos _____

```
seleccion <- subset(catalogo,
                    licencia %in% licencias_abiertas & formato %in% ______)
seleccion$conjunto
```

Pista

El hueco es un vector con los tres formatos: `c("CSV", ...)`.

Solución

```
seleccion <- subset(catalogo,
                    licencia %in% licencias_abiertas & formato %in% c("CSV", "JSON",
"RDF"))
seleccion$conjunto
```

Paso 5. Formatos abiertos y propietarios

Una tabla de referencia con los formatos del apartado 8:

```
formatos <- data.frame(
  formato = c("CSV", "JSON", "XML", "RDF", "ODS", "XLSX", "XLS", "PDF", "DOCX"),
```

```
  abierto = c(TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, NA),
  legible_maquina = c(TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, FALSE)
)
formatos
```

Ejercicio 0.4

Con la tabla `formatos`, obtén los nombres de los formatos que son **a la vez abiertos y legibles por máquina**.

ℹ Ejercicio: completa los huecos _____

```
subset(formatos, abierto ______ legible_maquina)$formato
```

ℹ Pista

Se tienen que cumplir las dos condiciones: el operador «y» es &.

Solución

```
subset(formatos, abierto & legible_maquina)$formato
```

Paso 6. ¿Público, abierto o ambas cosas?

Ejercicio 0.5

Clasifica cada caso en una de las cuatro casillas del apartado 7. Escribe para cada uno `"publico_abierto"`, `"publico_no_abierto"`, `"privado_abierto"` o `"privado_no_abierto"`.

ℹ Ejercicio: completa los huecos _____

```
casos <- c(
  "Horarios de la red municipal de bibliotecas en CSV con CC BY 4.0",
  "Listado de socios con nombre y domicilio que custodia la biblioteca",
  "Mapa colaborativo OpenStreetMap (ODbL)",
  "Estadísticas internas de ventas de una distribuidora de libros"
)
respuesta <- c(______, ______, ______, ______)
respuesta
```

ℹ Pista

Los datos personales que custodia una administración son información pública en sentido amplio, pero no pueden abrirse. OpenStreetMap lo mantiene una comunidad y una fundación, no una administración.

Solución

```
respuesta <- c("publico_abierto", "publico_no_abierto", "privado_abierto",
"privado_no_abierto")
```

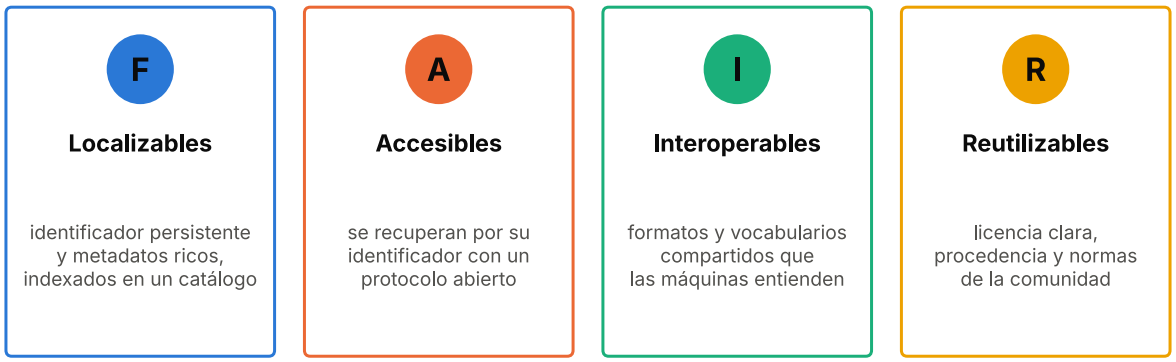

Figura 2: Los principios FAIR (Wilkinson et al., 2016).

Caso práctico resuelto

Valorar un conjunto con las cinco estrellas de Berners-Lee

Un ayuntamiento publica el calendario de actividades culturales de sus bibliotecas. Revisión paso a paso:

1. **¿Licencia abierta?** La página dice «Reutilización permitida citando la fuente» y remite al aviso legal con CC BY 4.0: sí, ★.
2. **¿Estructurado?** Se publica en PDF: legible por personas, no por máquinas. Se queda en ★.
3. **Mejora propuesta:** publicarlo también en XLSX (★★) y en CSV o JSON (★★★, formato no propietario).
4. **Siguiente paso:** dar URI a cada biblioteca y actividad (★★★★) y enlazarlas con otros conjuntos, por ejemplo con el directorio de equipamientos municipales (★★★★★).

Cada estrella facilita la reutilización, pero ya con tres (CSV con licencia abierta) se pueden construir servicios útiles.

Ejercicios graduados

Básico. ¿Cuántas estrellas tiene una hoja de cálculo XLSX con licencia CC BY?

Explicaciones

Dos: datos estructurados con licencia abierta, pero en un formato propietario. En CSV tendría tres.

Avanzado. Elige un conjunto de datos de un portal autonómico y valóralo con las cinco estrellas, proponiendo mejoras concretas para subir dos niveles.

Explicaciones

La valoración debe justificar cada estrella con evidencias (licencia, formato, uso de URI, enlaces) y las mejoras deben ser realistas para el organismo (por ejemplo, añadir CSV y un identificador estable).

Actividades

Actividad: ¿Cuántas estrellas tiene este conjunto de datos?

Categorías: ★ · ★★ · ★★★ · ★★★★ · ★★★★★

- Un PDF escaneado con una tabla de préstamos y licencia abierta
- Una hoja de Excel (.xlsx) con licencia abierta
- El mismo fichero en CSV
- Un RDF en el que cada biblioteca tiene su propio URI
- Ese RDF, además, enlazado con Wikidata y con el callejero oficial

Solución

- Un PDF escaneado con una tabla de préstamos y licencia abierta → ★. Disponible en la web con licencia abierta, pero no estructurado.
- Una hoja de Excel (.xlsx) con licencia abierta → ★★. Estructurado, pero en formato propietario.
- El mismo fichero en CSV → ★★★. Formato abierto y no propietario.
- Un RDF en el que cada biblioteca tiene su propio URI → ★★★★
- Ese RDF, además, enlazado con Wikidata y con el callejero oficial → ★★★★★

Actividad: ¿Qué principio FAIR mejora?

Categorías: Localizable · Accesible · Interoperable · Reutilizable

- Asignar un DOI al conjunto de datos
- Publicarlo en datos.gob.es con metadatos completos
- Ofrecer la descarga por HTTPS sin registro
- Usar el vocabulario DCAT-AP y códigos INE de municipio
- Indicar la licencia CC BY 4.0 y quién generó los datos

Solución

- Asignar un DOI al conjunto de datos → **Localizable**
- Publicarlo en datos.gob.es con metadatos completos → **Localizable**
- Ofrecer la descarga por HTTPS sin registro → **Accesible**
- Usar el vocabulario DCAT-AP y códigos INE de municipio → **Interoperable**
- Indicar la licencia CC BY 4.0 y quién generó los datos → **Reutilizable**

Actividad: ¿Formato abierto o propietario?

Categorías: Abierto · Propietario

- CSV
- JSON
- SAV (SPSS)
- GeoJSON
- DWG (AutoCAD)

Solución

- CSV → **Abierto**
- JSON → **Abierto**
- SAV (SPSS) → **Propietario**
- GeoJSON → **Abierto**
- DWG (AutoCAD) → **Propietario**

Actividad: La evolución de la Web

- Web 1.0: páginas estáticas para leer
- Web 2.0: participación y contenidos de las personas usuarias
- Web semántica: datos enlazados que entienden las máquinas

Solución

1. Web 1.0: páginas estáticas para leer
2. Web 2.0: participación y contenidos de las personas usuarias
3. Web semántica: datos enlazados que entienden las máquinas

Reto evaluable

Reto del Tema 0 · Auditoría de apertura

Elige en datos.gob.es **tres conjuntos de datos de tres publicadores distintos**: uno en formato propietario (XLS o PDF), uno en CSV y uno que ofrezca RDF, una API o un SPARQL. Para cada uno:

1. Identifica la licencia y explica qué permite y qué exige.
2. Calcula sus estrellas con la función `estrellas()` y justifica cada criterio.
3. Propón un cambio concreto y viable que le haría subir una estrella.
4. Cierra con un párrafo comparando los tres: ¿qué publicador facilita más la reutilización y por qué?

Entrega: un informe hecho con la plantilla de informe, renderizado a HTML o PDF, con el código visible y el conjunto de datos citado. Se evalúa con la rúbrica que incluye la plantilla. Este reto **no tiene solución en el libro**: cada estudiante trabaja con datos distintos y la corrección la hace el profesorado.

Espacio de trabajo (no se corrige automáticamente):

```r
source("datos/datareuseR.R")
auditoria <- data.frame(
  conjunto         = c("", "", ""),
  url              = c("", "", ""),
  licencia_abierta = c(NA, NA, NA),
  estructurado     = c(NA, NA, NA),
  no_propietario   = c(NA, NA, NA),
  usa_uri          = c(NA, NA, NA),
  enlazado         = c(NA, NA, NA)
)
# Cuando completes la tabla con TRUE y FALSE, quita los # de las líneas siguientes:
# auditoria$estrellas <- mapply(estrellas, auditoria$licencia_abierta,
auditoria$estructurado,
#                               auditoria$no_propietario, auditoria$usa_uri,
auditoria$enlazado)
# auditoria[, c("conjunto", "estrellas")]
```

Autoevaluación

Responde en el papel y comprueba después con las explicaciones del recuadro. En la versión web estas preguntas se corrigen automáticamente. No cuentan para la nota.

1. Un conjunto de datos con licencia CC BY-NC 4.0 es abierto según la Open Definition. *[V / F]*
2. Los datos FAIR son siempre datos abiertos. *[V / F]*
3. Un CSV con licencia abierta, sin URI ni enlaces, tiene _____ estrellas.
4. Internet es la red de redes; la World Wide Web es *[otro nombre de Internet / un servicio de documentos enlazados que funciona sobre Internet / un protocolo de correo]*
5. ¿Cuál de estos formatos es abierto **y** legible por máquina?

(a) PDF escaneado (b) XLS (Excel antiguo) (c) JSON (d) Imagen JPG de una tabla

6. Completa el código para contar las licencias abiertas de un catálogo: `sum(catalogo$licencia _____ licencias_abiertas)`

> **Explicaciones**
>
> - NC prohíbe el uso comercial, así que no cumple la Open Definition.
> - FAIR se refiere a cómo se describen y gestionan los datos; pueden ser FAIR con acceso restringido.
> - Licencia abierta (★), estructurado (★★) y formato no propietario (★★★); la cuarta exige URI.
> - La Web (URI, HTTP, HTML) es uno de los servicios de Internet, junto al correo o la transferencia de ficheros.
> - JSON tiene especificación pública y se procesa automáticamente; el PDF escaneado y la imagen no son legibles por máquina, y XLS es propietario.
> - `%in%` comprueba si cada licencia está en la lista de licencias abiertas.

Para practicar fuera del libro

> **Tip**
>
> En tu ordenador tienes las funciones de este tema en el paquete `datareuseR`: `leer_csv_abierto()` y `estrellas()`.

Repite la auditoría con un conjunto de datos de bibliotecas de datos.gob.es o de un portal autonómico: anota su licencia y su formato, y aplica `estrellas()`.

10. Recursos y portales de referencia

Portales de datos. datos.gob.es (catálogo nacional, que federa portales autonómicos y locales), data.europa.eu (portal europeo), y portales autonómicos y municipales como datos.madrid.es o Aragón Open Data. En el ámbito bibliotecario, datos.bne.es.

Referencias y formación:

Recurso	Para qué sirve
5 ★ Open Data	Explicación y ejemplos de cada nivel
GO FAIR – FAIR Principles	Texto de los 15 subprincipios FAIR
Open Definition 2.1	Definición de «abierto»
Open Data Handbook	Guía introductoria, disponible en castellano
Creative Commons	Licencias abiertas y selector de licencias
datos.gob.es	Catálogo nacional y sección «Conocimiento»
data.europa.eu academy	Cursos y seminarios web gratuitos

Tema 1. Reutilización de la información del sector público (RISP)

Objetivos

- Comprender el marco jurídico que regula la reutilización de la información pública.
- Identificar los derechos y obligaciones asociados a la reutilización.

Glosario del tema

Reutilización · RISP · Derecho de acceso · Publicidad activa · Silencio negativo · Licencia tipo · Alto valor · Datos dinámicos · Transposición · Infomediario

Descargas

Script R del capítulo Plantilla de informe (.qmd) Tema de estilo (.scss) CSV de práctica

El script reúne todo el código del capítulo para trabajar en RStudio o Positron. La plantilla `.qmd` y su tema `.scss` sirven para redactar las prácticas opcionales: guárdalos en la misma carpeta.

Lectura obligatoria

Clabo, N.; Ramos-Vielba, I. (2015). Reutilización de datos abiertos en la administración pública en España y uso de licencias-tipo. *Revista Española de Documentación Científica*, 38(3), e097. Versión en acceso abierto (Digital.CSIC).

El artículo es anterior a la Directiva (UE) 2019/1024. Léelo junto con la normativa vigente: lo que ha cambiado desde entonces (datos de alto valor, datos de investigación, tarifas, licencias) es justo lo que trabaja este capítulo.

Lecturas complementarias

- Janssen (2011) — Cómo la primera directiva de reutilización se relaciona con el movimiento de datos abiertos.

1. Conceptos básicos de reutilización de la información

La Ley 37/2007 entiende por **reutilización** el uso, por personas físicas o jurídicas, de documentos en poder del sector público con fines comerciales o no comerciales **distintos del propósito inicial** de servicio público para el que se produjeron. El intercambio de documentos entre administraciones en el ejercicio de sus funciones no es reutilización.

Concepto	Significado
Documento	Cualquier contenido, sea cual sea su soporte (papel, electrónico, sonoro, audiovisual), o una parte de él. Incluye los datos
Reutilizador	Persona o entidad que reutiliza; puede ser una empresa, una ONG, un periodista, un investigador o un ciudadano
Agente infomediario	Reutilizador que crea productos o servicios de valor añadido a partir de la información (Tema 5)
Datos abiertos	Documentos en formato abierto que pueden usarse, reutilizarse y compartirse libremente para cualquier fin

Concepto	Significado
Datos dinámicos	Datos que se actualizan con frecuencia o en tiempo real (tráfico, sensores)
Conjuntos de datos de alto valor	Datos cuya reutilización tiene grandes beneficios para la sociedad y la economía; tienen un régimen reforzado

Conviene distinguir dos derechos que a menudo se confunden:

- **Derecho de acceso** (Ley 19/2013): cualquier persona puede solicitar y conocer información pública.
- **Reutilización** (Ley 37/2007): usar esa información para otro fin. La regla general es que la información del sector público es reutilizable salvo excepciones.

El acceso es previo a la reutilización, pero no la garantiza: puedo tener derecho a ver un documento y no poder reutilizarlo, por ejemplo si contiene propiedad intelectual de terceros.

2. Evolución de las políticas de reutilización

Etapa	Rasgos
Información del sector público (PSI), 2003-2012	La Directiva 2003/98/CE armoniza las condiciones de reutilización, pero no obliga a autorizarla. España la transpone con la Ley 37/2007 y la desarrolla para la Administración del Estado con el Real Decreto 1495/2011. Nace el Proyecto Aporta (2009) y el portal datos.gob.es (2011)
Obligación de reutilizar, 2013-2018	La Directiva 2013/37/UE convierte la reutilización en obligación y extiende la norma a bibliotecas (también universitarias), museos y archivos. España aprueba la Ley 19/2013 de transparencia, la NTI-RISP (2013) y la Ley 18/2015
Datos abiertos, 2019-2021	La Directiva (UE) 2019/1024 pasa de «información del sector público» a «datos abiertos»: datos dinámicos y API, datos de investigación, datos de alto valor, empresas públicas. España la transpone con el Real Decreto-ley 24/2021
Economía del dato, desde 2020	La Estrategia Europea de Datos (2020) amplía el foco a datos protegidos y privados: Reglamento de Gobernanza de Datos (2022), datos de alto valor (2023), Ley de Datos (2023), espacios comunes de datos y Ley de IA (2024)

La tendencia es clara: de **permitir** reutilizar a **obligar** a publicar, de documentos a **datos** y API, y del sector público a todo el **ecosistema de datos**.

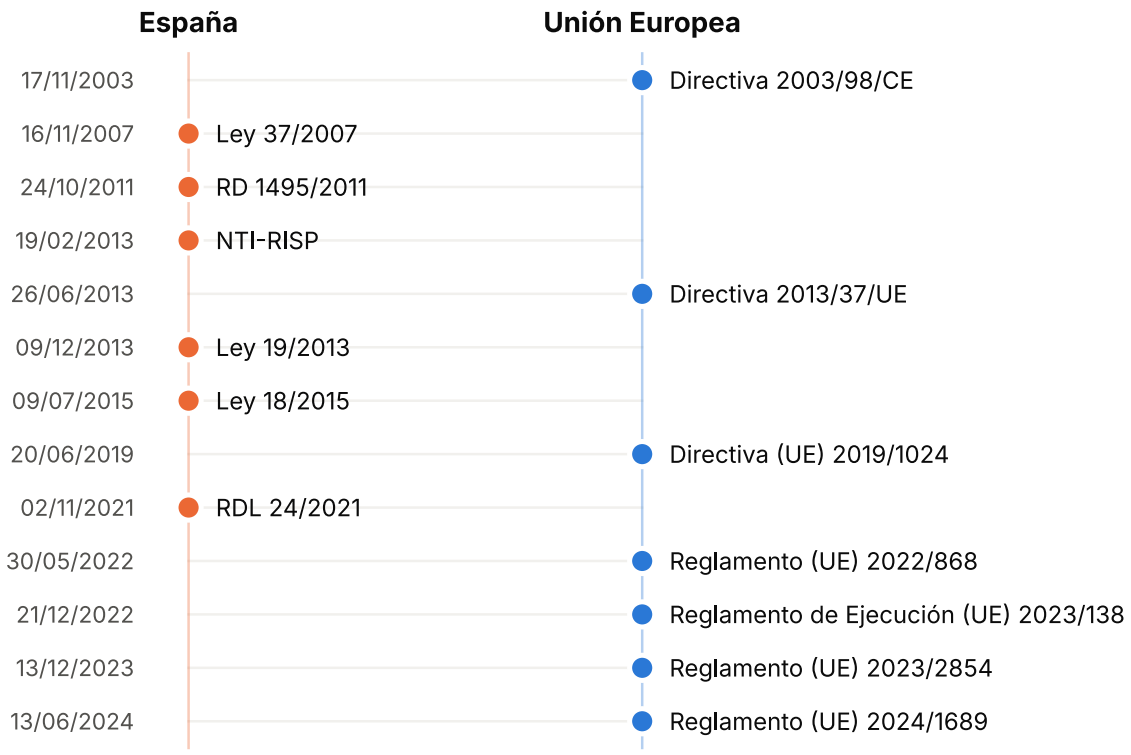

Figura 3: Cronología de las normas europeas y españolas sobre reutilización. Cada directiva europea va seguida de su transposición española; desde 2022 dominan los reglamentos europeos, que se aplican sin transposición.

3. Marco normativo europeo

Norma	Fecha	Qué aporta
Directiva 2003/98/CE	17/11/2003	Primera directiva de reutilización (PSI)
Directiva 2013/37/UE	26/06/2013	Obligación de autorizar la reutilización; bibliotecas, museos y archivos
Directiva (UE) 2019/1024	20/06/2019	Datos abiertos: alto valor, API, datos de investigación (refunde las anteriores)
Reglamento (UE) 2022/868	30/05/2022	Gobernanza de Datos: reutilización de datos protegidos, intermediarios de datos, altruismo de datos
Reglamento de Ejecución (UE) 2023/138	21/12/2022	Lista de conjuntos de datos de alto valor (aplicable desde el 9/6/2024)
Reglamento (UE) 2023/2854	13/12/2023	Ley de Datos: datos generados por productos conectados
Reglamento (UE) 2024/1689	13/06/2024	Ley de Inteligencia Artificial

Las **directivas** obligan a los Estados a transponerlas a su derecho interno; los **reglamentos** se aplican directamente.

Datos de alto valor

El Reglamento de Ejecución (UE) 2023/138 concreta seis categorías temáticas: **geoespaciales**, **observación de la Tierra y medio ambiente**, **meteorológicos**, **estadísticos**, **sociedades y propiedad de sociedades** y **movilidad**. Estos conjuntos deben ofrecerse gratis, en formato legible por máquina, mediante API y, cuando proceda, en descarga masiva, con una licencia CC BY 4.0, CC0 o equivalente.

4. Marco normativo español

Norma	Fecha	Qué aporta
Ley 37/2007	16/11/2007	Norma básica de reutilización (texto consolidado vigente)
Real Decreto 1495/2011	24/10/2011	Desarrolla la Ley 37/2007 para el sector público estatal
NTI de reutilización de recursos de información (NTI-RISP)	19/02/2013	Pautas técnicas comunes: identificación, metadatos, formatos, condiciones de uso (BOE de 4/3/2013)
Ley 19/2013	09/12/2013	Transparencia, acceso a la información pública y buen gobierno
Ley 18/2015	09/07/2015	Modifica la Ley 37/2007 para transponer la Directiva 2013/37/UE
Real Decreto-ley 24/2021	02/11/2021	Modifica la Ley 37/2007 para transponer la Directiva (UE) 2019/1024

Las comunidades autónomas y muchos ayuntamientos tienen además sus propias leyes y ordenanzas de transparencia y datos abiertos, y sus portales se federan en datos.gob.es.

5. Transparencia y acceso a la información pública

La Ley 19/2013 tiene tres partes: **publicidad activa** (información que las administraciones deben publicar sin que nadie la pida: organización, contratos, subvenciones, presupuestos…), **derecho de acceso** a la información pública y **buen gobierno**. Su órgano de garantía estatal es el Consejo de Transparencia y Buen Gobierno (CTBG).

Plazos del derecho de acceso:

- **Resolver** (art. 20.1): un mes desde que la solicitud llega al órgano competente, ampliable **otro mes** si el volumen o la complejidad lo justifican.
- **Silencio** (art. 20.4): si no hay resolución expresa en plazo, la solicitud se entiende **desestimada**.
- **Reclamar** ante el CTBG (art. 24): un mes desde la notificación o desde que se producen los efectos del silencio. El Consejo tiene tres meses para resolver; su silencio también es desestimatorio.

6. Derechos y obligaciones de reutilizadores y administraciones

La Ley 37/2007 (texto consolidado) distribuye derechos y obligaciones así:

Modalidades de reutilización. Hay cuatro modalidades: documentos publicados sin condiciones; documentos publicados con licencias tipo; documentos reutilizados previa solicitud (art. 10); y acuerdos exclusivos (art. 6). Las condiciones que se impongan deben ser objetivas, proporcionadas, no discriminatorias y justificadas por un objetivo de interés público.

Obligaciones de las administraciones:

- **Formatos.** Los documentos se ofrecen en el formato existente, preferiblemente abierto y legible por máquina, con sus metadatos; no hay obligación de crear o adaptar documentos. Los datos dinámicos deben estar disponibles justo después de su recogida mediante API, y los conjuntos de datos de alto valor, en formato legible por máquina y mediante API.
- **Catálogos.** Deben ofrecer catálogos en línea y herramientas de búsqueda de los documentos reutilizables (art. 4.5).

- **No exclusividad.** Los derechos exclusivos están en general prohibidos; solo se admiten si son necesarios para un servicio de interés público, revisándolos al menos cada tres años. Para la digitalización de recursos culturales, la exclusividad tiene un límite general de 10 años.
- **Tarifas.** La reutilización es gratuita, aunque pueden cobrarse tarifas por el suministro limitadas a los costes marginales; ese límite no se aplica a bibliotecas, museos y archivos. Las tarifas deben publicarse en línea.
- **Resolver las solicitudes.** La resolución debe dictarse en veinte días desde la entrada en el registro del órgano competente; si la solicitud es voluminosa o compleja, el plazo puede ampliarse otros veinte días, comunicándolo con su motivación en un plazo de diez días. Si no se resuelve en plazo, el solicitante puede entender desestimada la solicitud.

Obligaciones de los reutilizadores. Las condiciones generales que pueden imponerse son no alterar el contenido ni los metadatos, no desnaturalizar su sentido, citar la fuente, indicar la fecha de última actualización, concretar los fines si hay datos personales y no revertir la anonimización.

Régimen sancionador (art. 11, aplicable a la Administración General del Estado):

Tipo	Ejemplos	Multa
Muy grave	Desnaturalizar el sentido de la información; alteración muy grave	50.001 - 100.000 €
Grave	Reutilizar sin licencia cuando es obligatoria o para otra finalidad; alteración grave; incumplimiento grave de la licencia	10.001 - 50.000 €
Leve	No indicar la fecha de actualización; no citar la fuente; alteración o incumplimiento leve	1.000 - 10.000 €

Las infracciones muy graves y graves pueden conllevar además la prohibición de reutilizar documentos sometidos a licencia durante uno a cinco años y la revocación de la licencia.

Datos de investigación. Los organismos de investigación financiados con fondos públicos deben adoptar medidas para que los datos de investigación sean reutilizables, interoperables y abiertos, con los límites de la propiedad intelectual, la protección de datos y la confidencialidad; si ya se han publicado en un repositorio institucional o temático, pueden reutilizarse con fines comerciales o no comerciales, y su reutilización es gratuita.

Lo que afecta a las bibliotecas

Las bibliotecas (también las universitarias), los museos y los archivos están incluidos en la ley, pero con un régimen propio: conservan sus derechos de propiedad intelectual, aunque deben ejercerlos facilitando la reutilización; pueden cobrar por encima del coste marginal, con un límite de costes más un margen razonable; y la información sujeta a derechos de terceros queda fuera.

7. Licencias de reutilización

Cuando se reutiliza con condiciones, estas se recogen en una **licencia**. La ley anima a usar licencias abiertas con las mínimas restricciones posibles, y toda licencia debe indicar, entre otras cosas, la finalidad, si se admite el uso comercial, la duración, las obligaciones, la responsabilidad y las tarifas (art. 9).

En la práctica se usan **licencias tipo**: textos estándar, iguales para todos los reutilizadores, publicados en el aviso legal del portal. Clabo y Ramos-Vielba analizan cómo las administraciones españolas fueron pasando de condiciones propias a licencias Creative Commons. Hoy lo más frecuente es **CC BY 4.0**, que exige citar la fuente, y es una de las licencias admitidas para los datos de alto valor junto con CC0.

Opción	Cuándo usarla
Sin condiciones / CC0	Datos sin riesgo de desnaturalización; máxima reutilización
CC BY 4.0 (licencia tipo)	Opción por defecto: solo exige atribución
Licencia específica previa solicitud	Casos excepcionales: datos sensibles, condiciones particulares

Práctica en R

Paso 1. La normativa como datos

```
normas <- data.frame(
  norma = c("Directiva 2003/98/CE", "Ley 37/2007", "Directiva 2013/37/UE", "Ley
19/2013",
            "Ley 18/2015", "Directiva (UE) 2019/1024", "RDL 24/2021",
            "RD 1495/2011", "NTI-RISP (Resolución 19/02/2013)",
            "Reglamento (UE) 2022/868", "Reglamento de Ejecución (UE) 2023/138",
            "Reglamento (UE) 2023/2854", "Reglamento (UE) 2024/1689"),
  fecha = as.Date(c("2003-11-17", "2007-11-16", "2013-06-26", "2013-12-09",
                    "2015-07-09", "2019-06-20", "2021-11-02",
                    "2011-10-24", "2013-02-19",
                    "2022-05-30", "2022-12-21", "2023-12-13", "2024-06-13")),
  ambito = c("UE", "ES", "UE", "ES", "ES", "UE", "ES", "ES", "ES", "UE", "UE", "UE",
"UE")
)
normas
```

`as.Date()` convierte un texto con formato `AAAA-MM-DD` en una fecha con la que se puede calcular.

Paso 2. Ordenar y medir intervalos (apartado 2)

```
normas <- normas[order(normas$fecha), ]
normas[, c("fecha", "ambito", "norma")]
# Años transcurridos entre la Directiva 2003/98/CE y su transposición
as.numeric(normas$fecha[2] - normas$fecha[1]) / 365.25
```

Ejercicio 1.1

¿Cuántas de las normas de la tabla son de ámbito europeo (`"UE"`)?

Ejercicio: completa los huecos _____

```
sum(normas$ambito == _____)
```

Solución

```
sum(normas$ambito == "UE")
```

Ejercicio 1.2

Calcula cuántos **días** pasaron entre la Directiva (UE) 2019/1024 y el Real Decreto-ley 24/2021 que la transpuso. (El plazo de transposición terminaba el 17 de julio de 2021: compara.)

Ejercicio: completa los huecos _____

```
rdl <- as.Date("_____")
as.numeric(rdl - directiva)
```

Pista

La fecha del RDL 24/2021 está en la tabla de normas, en formato `AAAA-MM-DD`.

Solución

```
rdl <- as.Date("2021-11-02")
as.numeric(rdl - directiva)
```

Fueron 866 días: España transpuso la directiva unos tres meses y medio después de que venciera el plazo.

Paso 3. Plazos del derecho de acceso (apartado 5)

Los plazos de la Ley 19/2013 se expresan en meses, que se cuentan de fecha a fecha. En R, `seq()` con `by = "1 month"` avanza un mes natural:

```
recepcion <- as.Date("2026-02-16")
fin_plazo <- seq(recepcion, by = "1 month", length.out = 2)[2]
fin_plazo
weekdays(fin_plazo)   # día de la semana (en inglés en webR)
```

Nota

Es una simplificación: si el último día es inhábil, el plazo pasa al primer día hábil siguiente (art. 30.5 de la Ley 39/2015). El calendario de días inhábiles no se modela aquí.

Ejercicio 1.3

Una biblioteca pública recibe el **10 de marzo de 2026** una solicitud de acceso al número de préstamos por sede. Por su complejidad, el órgano **amplía el plazo**. ¿Cuál es la fecha límite para resolver?

Ejercicio: completa los huecos _____

```
recepcion <- as.Date("2026-03-10")
meses <- _____
seq(recepcion, by = "1 month", length.out = meses + 1)[meses + 1]
```

Pista

Un mes de plazo ordinario más un mes de ampliación.

Solución

```
recepcion <- as.Date("2026-03-10")
meses <- 2
seq(recepcion, by = "1 month", length.out = meses + 1)[meses + 1]
```

El 10 de mayo de 2026. Como es domingo, en la práctica el plazo terminaría el lunes 11.

Ejercicio 1.4

De esta lista, selecciona **solo** las categorías de datos de alto valor del Reglamento de Ejecución (UE) 2023/138. Escribe sus posiciones en el vector.

Ejercicio: completa los huecos _____

```
candidatas
mis_hvd <- candidatas[c(_____)]
mis_hvd
```

Pista

Son seis categorías. Escribe los números separados por comas, por ejemplo `c(1, 3, 4)`.

Solución

```
mis_hvd <- candidatas[c(1, 3, 4, 6, 8, 9)]
```

Paso 4. Plazos de una solicitud de reutilización (apartado 6)

Los plazos de la Ley 37/2007 se expresan en **días**, que según la Ley 39/2015 son **hábiles** salvo que se diga lo contrario: no cuentan los sábados, los domingos ni los festivos. Esta función suma días hábiles saltando los fines de semana (no conoce el calendario de festivos):

```
sumar_habiles <- function(inicio, n) {
  dia <- inicio
  while (n > 0) {
    dia <- dia + 1
    # weekdays() devuelve el día en inglés en webR
    if (!weekdays(dia) %in% c("Saturday", "Sunday")) n <- n - 1
  }
  dia
}
sumar_habiles(as.Date("2026-01-09"), 3)   # viernes + 3 días hábiles = miércoles
```

Ejercicio 1.5

Un investigador presenta el lunes **2 de marzo de 2026** una solicitud de reutilización de las estadísticas de préstamo de una biblioteca pública estatal. El plazo no se amplía. ¿Hasta qué fecha tiene la administración para resolver? (Supón que no hay festivos en esas semanas.)

Ejercicio: completa los huecos _____

```
sumar_habiles(as.Date("2026-03-02"), ______)
```

Pista

El plazo ordinario de resolución del art. 10 de la Ley 37/2007 es de veinte días.

Solución

```
sumar_habiles(as.Date("2026-03-02"), 20)
```

El 30 de marzo de 2026. Si se ampliara, serían otros veinte días hábiles, y entonces sí habría que descontar los festivos de Semana Santa.

Ejercicio 1.6

Un portal de noticias reutiliza el catálogo de una biblioteca estatal con licencia tipo, pero **no cita la fuente**. Según el art. 11 de la Ley 37/2007, ¿qué tipo de infracción es y cuál es la multa máxima? Completa con la tabla del apartado 6.

Ejercicio: completa los huecos _____

```
infraccion <- c(tipo = "______", multa_maxima = "______")
infraccion
```

Pista

El tipo se escribe en minúsculas (`"leve"`, `"grave"` o `"muy grave"`) y la multa como número, sin puntos ni símbolo de euro.

Solución

```
infraccion <- c(tipo = "leve", multa_maxima = "10000")
```

Laboratorio: ¿puedo combinar estos datos?

Laboratorio interactivo

Quieres crear un conjunto de datos nuevo combinando dos fuentes. Elige sus licencias y el uso que harás: el laboratorio te dice si, en principio, puedes combinarlas y con qué licencia tendrías que publicar el resultado.

Este laboratorio es interactivo en la versión web: pablosancheznunez.com/reutilizacion-datos

Caso práctico resuelto

¿Se puede reutilizar este conjunto para un servicio comercial?

Una empresa quiere crear una app de pago con los datos de aparcamientos de un ayuntamiento. Comprobaciones según la normativa de reutilización (Ley 37/2007 y la Directiva (UE) 2019/1024):

1. **Condiciones de uso:** el portal aplica las condiciones generales de reutilización con licencia CC BY 4.0. Permiten el uso comercial con atribución.
2. **Datos personales:** los datos de ocupación son agregados por aparcamiento; no hay datos personales.
3. **Tarifas:** la regla general es la reutilización gratuita o limitada a los costes marginales.
4. **Obligaciones:** citar la fuente, no desnaturalizar los datos, no sugerir que el ayuntamiento patrocina la app e indicar la fecha de la última actualización.

Conclusión: puede reutilizarlos comercialmente si cumple las condiciones de atribución.

Ejercicios graduados

Básico. ¿Qué tipo de datos promueve especialmente la Directiva (UE) 2019/1024 por su alto potencial económico y social?

Explicaciones

Los conjuntos de datos de alto valor (*high-value datasets*), definidos por un reglamento de ejecución de 2023 en seis categorías: geoespaciales, observación de la Tierra y medio ambiente, meteorológicos, estadísticos, sociedades y propiedad de sociedades, y movilidad.

Avanzado. Analiza las condiciones de reutilización de dos portales de datos (uno estatal y uno municipal) y compáralas en una tabla.

Explicaciones

Compara licencia, uso comercial, atribución exigida, existencia de tarifas, exclusiones (por ejemplo, datos con derechos de terceros) y cómo se comunican las actualizaciones.

Actividades

Actividad: Ordena cronológicamente las normas

- Directiva (UE) 2019/1024, de datos abiertos
- Directiva 2003/98/CE sobre reutilización de la información del sector público
- Ley 19/2013, de transparencia, acceso a la información pública y buen gobierno
- Ley 37/2007, sobre reutilización de la información del sector público
- Reglamento de Ejecución (UE) 2023/138, de conjuntos de datos de alto valor

Solución

1. Directiva 2003/98/CE sobre reutilización de la información del sector público
2. Ley 37/2007, sobre reutilización de la información del sector público
3. Ley 19/2013, de transparencia, acceso a la información pública y buen gobierno
4. Directiva (UE) 2019/1024, de datos abiertos
5. Reglamento de Ejecución (UE) 2023/138, de conjuntos de datos de alto valor

Actividad: ¿Obligación de la Administración o de quien reutiliza?

Categorías: Administración · Reutilizador

- Citar la fuente de los datos
- Publicar los datos en formatos abiertos cuando sea posible
- No desnaturalizar el sentido de la información
- Indicar la fecha de la última actualización
- Responder a las solicitudes de reutilización en plazo
- No sugerir que la Administración patrocina el uso

Solución

- Citar la fuente de los datos → **Reutilizador**
- Publicar los datos en formatos abiertos cuando sea posible → **Administración**
- No desnaturalizar el sentido de la información → **Reutilizador**
- Indicar la fecha de la última actualización → **Reutilizador**
- Responder a las solicitudes de reutilización en plazo → **Administración**
- No sugerir que la Administración patrocina el uso → **Reutilizador**

Reto evaluable

! Reto del Tema 1 · Una solicitud de reutilización

Busca una información del sector público que te interese profesionalmente y que **no esté publicada** como datos abiertos (por ejemplo, las estadísticas de préstamo por sucursal de la red de bibliotecas de tu municipio). Prepara, sin enviarla, la solicitud de reutilización:

1. Redacta la solicitud con los elementos que exige el art. 10 de la Ley 37/2007: identificación, información pedida, finalidad y formato preferido.
2. Identifica el órgano competente y justifica que la información es reutilizable (y no está excluida por el art. 3).
3. Calcula con `sumar_habiles()` la fecha límite de resolución, **descontando los festivos nacionales, autonómicos y locales reales** de tu municipio, y la fecha si el plazo se ampliara.
4. Indica qué condiciones de reutilización esperarías (licencia, tarifas) y por qué.

Entrega: un informe hecho con la plantilla de informe, renderizado a HTML o PDF, con el código visible y el conjunto de datos citado. Se evalúa con la rúbrica que incluye la plantilla. Este reto **no tiene solución en el libro**: cada estudiante trabaja con datos distintos y la corrección la hace el profesorado.

Espacio de trabajo (no se corrige automáticamente):

```
source("datos/datareuseR.R")
presentacion <- as.Date("2026-02-02")   # cámbiala por la fecha en que presentarías la solicitud
festivos <- as.Date(c())                # añade los festivos del periodo, p. ej. "2026-04-02"
ordinario <- sumar_habiles(presentacion, 20, festivos)
ampliado  <- sumar_habiles(ordinario, 20, festivos)
c(ordinario = ordinario, ampliado = ampliado)
```

Autoevaluación

Responde en el papel y comprueba después con las explicaciones del recuadro. En la versión web estas preguntas se corrigen automáticamente. No cuentan para la nota.

1. Tener derecho de acceso a un documento garantiza poder reutilizarlo. *[V / F]*
2. Plazo ordinario para resolver una solicitud de reutilización según la Ley 37/2007 (en días): _____
3. Si la administración no resuelve en plazo una solicitud de acceso (Ley 19/2013), la solicitud se entiende *[estimada / desestimada / archivada]*
4. La Directiva (UE) 2019/1024 se transpuso en España mediante el *[Real Decreto 1495/2011 / Real Decreto-ley 24/2021 / Ley 18/2015]*
5. ¿Cuál de estas **no** es una categoría de datos de alto valor del Reglamento de Ejecución (UE) 2023/138?

(a) Meteorológicos (b) Movilidad (c) Préstamos bibliotecarios (d) Sociedades y propiedad de sociedades

6. No citar la fuente al reutilizar con licencia tipo es una infracción *[leve / grave / muy grave]*
7. Completa el código para sumar un mes a una fecha: `seq(recepcion, by = "1 _____ ", length.out = 2)[2]`

Explicaciones

- El acceso permite conocer el documento; la reutilización puede estar limitada, por ejemplo, por derechos de terceros.
- Art. 10 de la Ley 37/2007: veinte días, ampliables otros veinte.
- Art. 20.4 de la Ley 19/2013: silencio negativo.
- El RDL 24/2021 modificó la Ley 37/2007; la Ley 18/2015 transpuso la Directiva 2013/37/UE.
- Las seis categorías son geoespaciales, observación de la Tierra y medio ambiente, meteorológicos, estadísticos, sociedades y movilidad.
- Art. 11.3 de la Ley 37/2007: multa de 1.000 a 10.000 €.
- `by = "1 month"` avanza un mes natural.

Para practicar fuera del libro

Tip

En tu ordenador tienes las funciones de este tema en el paquete `datareuseR`: `sumar_habiles()`, que además descuenta festivos.

Redacta una solicitud de acceso a un dato bibliotecario que no esté publicado y busca en el Consejo de Transparencia y Buen Gobierno una resolución sobre un caso parecido. Calcula con el código de este capítulo hasta cuándo tendrías para reclamar si la solicitud se desestima por silencio.

Recursos complementarios

Recurso	Para qué sirve
Ley 37/2007 (texto consolidado, BOE)	Norma básica española de reutilización
Directiva (UE) 2019/1024	Directiva europea de datos abiertos
Reglamento de Ejecución (UE) 2023/138	Conjuntos de datos de alto valor
Ley 19/2013 (BOE)	Publicidad activa y derecho de acceso
Consejo de Transparencia y Buen Gobierno	Resoluciones y criterios interpretativos
Portal de la Transparencia	Derecho de acceso ante la AGE
datos.gob.es: convalidación del RDL 24/2021	Novedades de la transposición
Reglamento (UE) 2022/868 y Reglamento (UE) 2023/2854	Gobernanza de Datos y Ley de Datos

Tema 2. Proyectos relevantes de reutilización de datos abiertos

Objetivos

- Analizar iniciativas nacionales e internacionales de reutilización de datos.
- Valorar el impacto de los datos abiertos en diferentes ámbitos.

Glosario del tema

Portal de datos · Federación · Open Data Maturity · Caso de uso · B2G · Altruismo de datos · Ciencia abierta · Repositorio · Data paper · API · Tipo MIME

Descargas

Script R del capítulo Plantilla de informe (.qmd) Tema de estilo (.scss) CSV de práctica

El script reúne todo el código del capítulo para trabajar en RStudio o Positron. La plantilla `.qmd` y su tema `.scss` sirven para redactar las prácticas opcionales: guárdalos en la misma carpeta.

Lectura obligatoria

Safarov, I.; Meijer, A.; Grimmelikhuijsen, S. (2017). Utilization of open government data: A systematic literature review of types, conditions, effects and users. *Information Polity*, 22(1), 1-24. Versión en acceso abierto.

Lecturas complementarias

- Ruijer et al. (2017) — Marco teórico sobre el uso democrático de los datos abiertos.
- Zuiderwijk et al. (2015) — Factores que explican la adopción de tecnologías de datos abiertos.

Un marco para analizar los proyectos

Safarov et al. revisan de forma sistemática 101 estudios sobre el uso de datos abiertos de gobierno y los ordenan en cuatro dimensiones. Sirven como guion para analizar cualquier proyecto de este tema sin quedarse en la descripción:

Dimensión	Pregunta que responde	Ejemplos
Tipo de utilización	¿Qué se hace con los datos?	Crear un servicio o aplicación, analizar, visualizar, investigar, fiscalizar
Condiciones	¿Qué facilita o frena el uso?	Calidad y actualidad, licencia, formatos, competencias de los usuarios, apoyo institucional
Efectos	¿Qué cambia gracias a la reutilización?	Valor económico, valor social, transparencia y buen gobierno
Usuarios	¿Quién reutiliza?	Ciudadanía, empresas, periodistas, investigadores, administraciones, desarrolladores

1. Open Data en España

- **Iniciativa Aporta** (desde 2009): programa estatal para impulsar la apertura y la reutilización, hoy gestionado desde el Ministerio con competencias en transformación digital. Es el origen del portal nacional.

- **datos.gob.es** (desde 2011): catálogo nacional que **federa** los catálogos de la Administración General del Estado, las comunidades autónomas, las entidades locales y las universidades. Publica además guías, informes, un directorio de aplicaciones y de empresas reutilizadoras y métricas de impacto.
- **Portales autonómicos y locales**: Aragón Open Data, Open Data Euskadi, datos.madrid.es, entre muchos otros, con conjuntos sobre bibliotecas.
- **Desafíos Aporta** (ediciones de 2017 a 2021): concurso nacional de ideas y prototipos basados en datos abiertos. No se ha vuelto a convocar desde 2021.
- **Posición internacional**: en el informe anual Open Data Maturity, España figura de forma sostenida en el grupo de cabeza; datos.gob.es resumió la edición de 2025 como una confirmación de su papel de referencia europea.

2. Open Data en la Unión Europea

- **data.europa.eu** (desde 2021): portal europeo que reúne los datos de las instituciones de la UE y recolecta los catálogos nacionales. Incluye formación (data.europa.eu academy) y un observatorio de casos de uso.
- **Open Data Maturity** (desde 2015): informe anual que compara la madurez de 36 países europeos en cuatro dimensiones: **políticas**, **impacto**, **portal** y **calidad**.
- **Estrategia Europea de Datos (2020)** y **espacios comunes europeos de datos**: sanidad, movilidad, energía, patrimonio cultural, investigación (EOSC)… En el ámbito cultural, Europeana es el antecedente y la base del espacio de datos del patrimonio.

3. Portales internacionales de datos abiertos

Portal o iniciativa	Ámbito	Interés
data.gov (2009)	Estados Unidos	Primer gran portal nacional
data.gov.uk (2010)	Reino Unido	Pionero en licencias y en el Open Data Institute
Open Data Charter (2015)	Internacional	Seis principios para gobiernos (abiertos por defecto, oportunos, accesibles, comparables, para mejorar la gobernanza, para el desarrollo inclusivo)
Global Data Barometer	Internacional	Indicadores de gobernanza, disponibilidad y uso de datos
Banco Mundial, OCDE, Naciones Unidas	Internacional	Portales estadísticos de referencia para comparar países

4. Casos de éxito

Un «caso de éxito» debe poder explicarse con las cuatro dimensiones de Safarov et al.: quién usa qué datos, en qué condiciones y con qué efecto. Las fuentes para encontrarlos son el directorio de aplicaciones de datos.gob.es, el observatorio de casos de uso de data.europa.eu y las historias de datos (*datastories*) del portal europeo. Recuerda los «mitos» del Tema 0: que un portal publique datos no garantiza que se usen, y un caso aislado no prueba un impacto general.

5. Iniciativas públicas y privadas

La reutilización no es solo cosa de administraciones:

Tipo	Ejemplos de iniciativa
Públicas	Portales de datos, concursos y retos, laboratorios de datos municipales, programas de formación

Tipo	Ejemplos de iniciativa
Privadas que publican datos	Comunidades como OpenStreetMap; empresas que publican conjuntos con licencia abierta
Colaboración público-privada	Intercambio de datos de empresas con administraciones (*business-to-government*, B2G) para fines de interés general, por ejemplo movilidad o emergencias
Altruismo de datos	Ciudadanos y empresas que ceden datos para fines de interés general a través de organizaciones reconocidas (Reglamento de Gobernanza de Datos)
Tercer sector y periodismo	Fundaciones de transparencia, periodismo de datos, comunidades de desarrolladores

6. Datos abiertos para la investigación y la innovación

En investigación, los datos abiertos forman parte de la **ciencia abierta**: publicar los datos que sustentan los resultados permite comprobarlos, reproducirlos y reutilizarlos para nuevas preguntas.

- **Principio**: «tan abiertos como sea posible, tan cerrados como sea necesario», y siempre FAIR (Tema 0).
- **Marco normativo**: la Directiva (UE) 2019/1024 y la Ley 37/2007 hacen reutilizables los datos de investigación financiados con fondos públicos que ya están en un repositorio (Tema 1). En España, la Ley 17/2022, que modifica la Ley de la Ciencia, refuerza la difusión en acceso abierto, y el Consejo de Ministros aprobó el 3 de mayo de 2023 la primera Estrategia Nacional de Ciencia Abierta (ENCA), para el periodo 2023-2027.
- **Infraestructuras**: repositorios generalistas como Zenodo, repositorios institucionales (en la UC3M, e-Archivo), el registro de repositorios re3data, el grafo de OpenAIRE y la Nube Europea de la Ciencia Abierta (EOSC).
- **Publicación**: los *data papers* y los planes de gestión de datos son ya habituales en revistas y convocatorias.
- **Innovación**: los datos de investigación abiertos alimentan nuevas líneas de investigación, empresas de base tecnológica y, cada vez más, el entrenamiento de modelos de IA (Tema 5).

Las bibliotecas universitarias son un actor clave: gestionan repositorios, asesoran en planes de gestión de datos y asignan identificadores persistentes.

Práctica en R

Paso 1. Clasificar casos de reutilización (apartado 4)

La tabla siguiente recoge **doce casos ilustrativos** (inventados para el ejercicio, aunque inspirados en tipos de proyecto frecuentes) ya clasificados con las dimensiones de Safarov et al.

```
casos <- data.frame(
  caso = c("App de horarios de bibliotecas", "Mapa de wifi público", "Visor de
presupuestos municipales",
           "Recomendador de lecturas", "Estudio de brecha digital", "Alertas de calidad
del aire",
           "Buscador de ayudas públicas", "Reportaje sobre contratos", "Panel de
préstamos regionales",
           "Planificador de rutas accesibles", "Análisis de citas a datasets", "Catálogo
enriquecido con Wikidata"),
  usuario = c("Desarrolladores", "Empresas", "Ciudadanía", "Empresas", "Investigadores",
"Desarrolladores",
              "Empresas", "Periodistas", "Administraciones", "Ciudadanía",
"Investigadores", "Administraciones"),
  utilizacion = c("Servicio", "Servicio", "Visualización", "Servicio", "Análisis",
```

```
"Servicio",
                  "Servicio", "Fiscalización", "Visualización", "Servicio", "Análisis",
"Servicio"),
  efecto = c("Social", "Económico", "Buen gobierno", "Económico", "Social", "Social",
             "Económico", "Buen gobierno", "Buen gobierno", "Social", "Social",
"Social"),
  ambito = c("Local", "Local", "Local", "Nacional", "Nacional", "Local",
             "Nacional", "Nacional", "Autonómico", "Local", "Europeo", "Autonómico")
)
casos[, c("caso", "usuario", "efecto")]
```

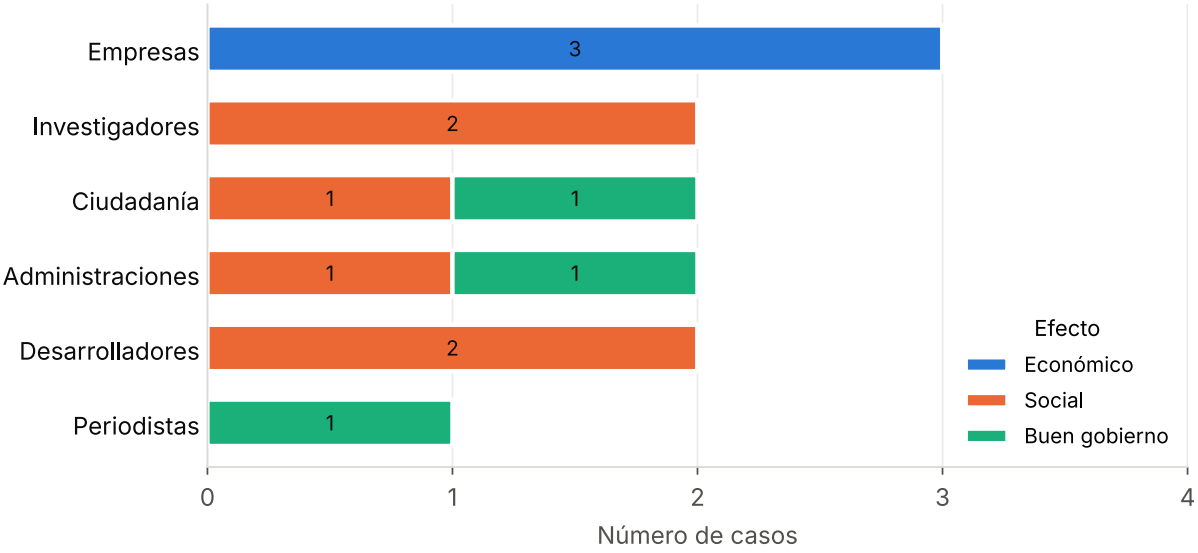

Figura 4: Los doce casos ilustrativos por tipo de usuario y efecto. En esta muestra inventada, el valor económico procede solo de empresas, y el buen gobierno, de periodistas, administraciones y ciudadanía.

La figura se ha hecho con estos mismos datos; en el paso siguiente aprenderás a obtener los recuentos con R.

Paso 2. Contar con `table()`

`table()` cuenta cuántas veces aparece cada valor; `prop.table()` lo convierte en proporciones.

```
table(casos$efecto)
round(100 * prop.table(table(casos$efecto)), 1)
```

Con dos variables, `table()` construye una tabla de contingencia:

```
table(casos$usuario, casos$efecto)
```

Y `barplot()` la dibuja:

```
barplot(sort(table(casos$usuario), decreasing = TRUE),
        las = 2, col = "#000e78", ylab = "Casos", cex.names = 0.8,
        main = "Casos por tipo de usuario")
```

Ejercicio 2.1

¿Qué **porcentaje** de los casos tiene un efecto de tipo **económico**? Redondea a un decimal.

Ejercicio: completa los huecos ______

```
porcentajes <- round(100 * prop.table(table(casos$efecto)), 1)
porcentajes[______]
```

Pista

Puedes seleccionar un elemento de una tabla por su nombre, entre comillas.

Solución

```
porcentajes <- round(100 * prop.table(table(casos$efecto)), 1)
porcentajes[["Económico"]]
```

Tres de doce casos: un 25 %.

Ejercicio 2.2

¿Cuántos casos tienen como usuarios a **empresas** y como efecto el **valor económico**? Usa `subset()` y `nrow()`.

Ejercicio: completa los huecos ______

```
nrow(subset(casos, usuario == "Empresas" ______ efecto == "Económico"))
```

Pista

Las dos condiciones deben cumplirse a la vez: el operador «y» es &.

Solución

```
nrow(subset(casos, usuario == "Empresas" & efecto == "Económico"))
```

Ejercicio 2.3 (sin corrección automática, apartado 2)

Abre la ficha de España del Open Data Maturity 2025, anota las puntuaciones de España y de la media europea en las cuatro dimensiones y calcula la diferencia. ¿En qué dimensión está España más lejos de la media? Este resultado lo necesitarás en el cuestionario del Tema 2 de Aula Global.

```
dimensiones <- c("Políticas", "Impacto", "Portal", "Calidad")
espana <- c(NA, NA, NA, NA)   # escribe aquí las puntuaciones de España (en %)
media  <- c(NA, NA, NA, NA)   # y aquí las de la media de los 36 países
diferencia <- espana - media
data.frame(dimensiones, espana, media, diferencia)
```

Paso 3. Autoevaluación FAIR de un registro de datos (apartado 6)

Una lista de comprobación muy simplificada para un conjunto de datos de investigación depositado en un repositorio:

```
evaluar_fair <- function(registro) {
  c(F = registro$identificador_persistente && registro$metadatos_descriptivos,
    A = registro$acceso_por_protocolo_abierto,
    I = registro$formato_abierto && registro$vocabulario_comun,
    R = registro$licencia_declarada && registro$procedencia_documentada)
}

registro_zenodo <- list(
  identificador_persistente = TRUE,   # DOI
  metadatos_descriptivos = TRUE,
  acceso_por_protocolo_abierto = TRUE, # HTTPS
  formato_abierto = TRUE,              # CSV
  vocabulario_comun = FALSE,           # columnas sin definir con un vocabulario
  licencia_declarada = TRUE,           # CC BY 4.0
  procedencia_documentada = TRUE       # README con método de obtención
)
evaluar_fair(registro_zenodo)
```

Ejercicio 2.4

Un grupo de investigación publica sus datos en su web en un fichero CSV, **sin DOI**, sin licencia y sin descripción de cómo se obtuvieron (sí tiene metadatos descriptivos y usa un vocabulario común). Completa el registro y ejecuta la evaluación. ¿Cuántos de los cuatro principios cumple?

ℹ Ejercicio: completa los huecos _____

```
registro_web <- list(
  identificador_persistente = ______,
  metadatos_descriptivos = TRUE,
  acceso_por_protocolo_abierto = TRUE,
  formato_abierto = TRUE,
  vocabulario_comun = TRUE,
  licencia_declarada = ______,
  procedencia_documentada = ______
)
sum(evaluar_fair(registro_web))
```

ℹ Pista

Los tres huecos describen lo que **no** tiene el conjunto de datos.

Solución

```
registro_web <- list(
  identificador_persistente = FALSE,
  metadatos_descriptivos = TRUE,
  acceso_por_protocolo_abierto = TRUE,
  formato_abierto = TRUE,
  vocabulario_comun = TRUE,
  licencia_declarada = FALSE,
  procedencia_documentada = FALSE
)
sum(evaluar_fair(registro_web))
```

Cumple dos principios (A e I). Falla F, por no tener identificador persistente, y R, por no tener licencia ni procedencia. Depositarlo en un repositorio con DOI y licencia CC BY 4.0 lo arreglaría casi todo.

Ejercicio 2.5

¿Qué **ámbito territorial** es el más frecuente entre los doce casos? `which.max()` devuelve la posición del valor más alto de una tabla y `names()`, su etiqueta.

Ejercicio: completa los huecos _____

```
names(which.max(table(casos$______)))
```

Pista

La columna del ámbito territorial se llama `ambito`, sin tilde.

Solución

```
names(which.max(table(casos$ambito)))
```

El ámbito local: cinco de los doce casos. Los servicios más cercanos a la ciudadanía (horarios, wifi, rutas) se apoyan sobre todo en datos municipales.

Paso 4. Consultar el catálogo nacional en directo (apartado 1)

datos.gob.es tiene una **API**: una dirección web que, en lugar de una página para personas, devuelve datos para programas. Esta consulta pide los 50 primeros conjuntos de datos con «bibliotecas» en el título:

`https://datos.gob.es/apidata/catalog/dataset/title/bibliotecas.json?_pageSize=50`

La respuesta llega en JSON. R la lee con `url()` y `readLines()`, y `jsonlite::fromJSON()` la convierte en listas. La consulta es **en directo**: el resultado puede cambiar de un día a otro. Si el portal no responde, el código usa una copia guardada el 24 de noviembre de 2026.

```
url_api <- "https://datos.gob.es/apidata/catalog/dataset/title/bibliotecas.json?_
pageSize=50"
respuesta <- tryCatch({
```

```
  con <- url(url_api, encoding = "UTF-8")
  texto <- paste(readLines(con, warn = FALSE), collapse = "\n")
  close(con)
  jsonlite::fromJSON(texto, simplifyVector = FALSE)
}, error = function(e) NULL)
if (is.null(respuesta)) {
  message("Sin conexión con datos.gob.es: uso la copia del 24/11/2026.")
  respuesta <- jsonlite::fromJSON("datos/datosgob_bibliotecas.json", simplifyVector =
FALSE)
}
items <- respuesta$result$items
length(items)
str(items[[1]][c("title", "publisher")], max.level = 2)
```

Cada elemento describe un conjunto con el vocabulario **DCAT** que verás en el Tema 4: título (en varios idiomas), publicador y distribuciones, cada una con su formato. Para no repetir este trabajo en cada consulta, el paquete `datareuseR` reúne unas funciones auxiliares. Aquí las cargamos con `source()`:

```
source("datos/datareuseR.R")
tabla <- tabla_datosgob(items)
head(tabla[, c("titulo", "formatos")], 8)
sort(table(unlist(lapply(items, formatos_de))), decreasing = TRUE)
```

Los formatos no vienen escritos igual: unos publicadores usan el vocabulario de tipos de fichero de la UE (`.../file-type/CSV`) y otros, tipos MIME (`text/csv`). `formatos_de()` los unifica con `normalizar_formato()`. Es un problema de **interoperabilidad semántica** que retomarás en el Tema 4.

Ejercicio 2.6

¿Cuántos de los conjuntos devueltos ofrecen **al menos una distribución en CSV**? Completa la etiqueta de formato.

Ejercicio: completa los huecos _____

```
tiene_csv <- sapply(items, function(it) "______" %in% formatos_de(it))
sum(tiene_csv)
```

Pista

`formatos_de()` devuelve etiquetas en mayúsculas, como `"JSON"` o `"XLS"`.

Solución

```
tiene_csv <- sapply(items, function(it) "CSV" %in% formatos_de(it))
sum(tiene_csv)
```

Con la copia del 24 de noviembre de 2026, 46 de 50. La cifra puede variar si la consulta es en directo. Que casi todos ofrezcan CSV no quiere decir que todos tengan la misma calidad: muchos son estadísticas en formato PC-Axis convertidas.

Caso práctico resuelto

Analizar un proyecto de reutilización como un caso de negocio

Proyecto de ejemplo: un servicio que combina datos abiertos de calidad del aire y de tráfico para recomendar rutas en bicicleta. Análisis:

Elemento	Respuesta
Datos de origen	Calidad del aire (ayuntamiento, por horas), intensidad de tráfico, carril bici (geoespacial)
Valor añadido	Integración, cálculo de rutas, alertas personalizadas
Usuarios	Ciclistas urbanos; también ayuntamientos interesados en la información agregada
Modelo de ingresos	App gratuita con informes de pago para administraciones
Riesgos	Cambios en el formato de los datos de origen, interrupciones del servicio, calidad desigual entre ciudades

La dependencia de datos externos se mitiga con contratos de servicio, si existen, supervisión de las fuentes y validación automática al cargar.

Ejercicios graduados

Básico. Identifica los datos de origen y el valor añadido de un servicio que muestra en tiempo real la llegada de autobuses.

Explicaciones

Datos: posiciones de los vehículos y horarios publicados por el operador (por ejemplo, en GTFS y GTFS-Realtime). Valor añadido: cálculo de tiempos de llegada, interfaz cómoda, avisos y combinación con otras redes.

Avanzado. Elige un proyecto real de reutilización del catálogo de datos.gob.es y analízalo con la tabla del caso.

Explicaciones

Documenta las fuentes reales que usa, cómo genera valor, quiénes son sus usuarios, cómo se financia y qué riesgos tiene por depender de datos públicos.

Actividades

Actividad: ¿De qué ámbito es cada portal?

Categorías: España · Unión Europea · Internacional o de otro país

- datos.gob.es
- data.europa.eu
- data.gov
- datos.madrid.es
- Zenodo

Solución

- datos.gob.es → **España**
- data.europa.eu → **Unión Europea**
- data.gov → **Internacional o de otro país**
- datos.madrid.es → **España**
- Zenodo → **Internacional o de otro país**

Actividad: ¿Quién reutiliza y para qué?

Categorías: Empresa · Ciudadanía · Investigación · Periodismo

- Una aplicación de transporte que muestra los horarios en tiempo real
- Un vecino que consulta la calidad del aire de su barrio
- Un estudio sobre la relación entre bibliotecas y nivel educativo
- Un reportaje que cruza contratos públicos y empresas adjudicatarias

Solución

- Una aplicación de transporte que muestra los horarios en tiempo real → **Empresa**
- Un vecino que consulta la calidad del aire de su barrio → **Ciudadanía**
- Un estudio sobre la relación entre bibliotecas y nivel educativo → **Investigación**
- Un reportaje que cruza contratos públicos y empresas adjudicatarias → **Periodismo**

Errores frecuentes

- **Pensar que un portal es una fuente única**: muchos conjuntos se repiten en el portal local, el autonómico y el nacional; elige el original.
- **No mirar la fecha de actualización** antes de basar un proyecto en un conjunto de datos.
- **Confundir «publicado» con «reutilizable»**: comprueba la licencia y el formato.

Reto evaluable

! Reto del Tema 2 · Radiografía de un tema en el catálogo nacional

Elige una palabra relacionada con tu interés profesional que **no sea «bibliotecas»** (por ejemplo, *archivos*, *museos*, *patrimonio* o *lectura*) y consulta en directo la API de datos.gob.es:

1. ¿Cuántos conjuntos devuelve y qué publicadores concentran la oferta? Resúmelo en una tabla.
2. ¿Qué formatos predominan? Haz un gráfico y comenta qué dice del nivel de apertura (Tema 0).
3. Localiza en la sección de aplicaciones y empresas de datos.gob.es tres reutilizaciones reales relacionadas con tu tema y clasifícalas con las dimensiones de Safarov et al. (usuario, utilización, efecto).
4. Concluye: ¿qué conjunto de datos falta para que surja una reutilización de valor en tu tema?

Entrega: un informe hecho con la plantilla de informe, renderizado a HTML o PDF, con el código visible y el conjunto de datos citado. Se evalúa con la rúbrica que incluye la plantilla. Este reto **no tiene solución en el libro**: cada estudiante trabaja con datos distintos y la corrección la hace el profesorado.

Espacio de trabajo (no se corrige automáticamente; necesita conexión):

```
source("datos/datareuseR.R")
palabra <- "archivos"   # cámbiala por tu tema
items <- consultar_datosgob(palabra, n = 50, copia = FALSE)
tabla <- tabla_datosgob(items)
length(items)
head(sort(table(tabla$publicador), decreasing = TRUE))
formatos <- sort(table(unlist(lapply(items, formatos_de))), decreasing = TRUE)
barplot(formatos, las = 2, col = "#2a78d6", border = NA, ylab = "Conjuntos")
```

Autoevaluación

Responde en el papel y comprueba después con las explicaciones del recuadro. En la versión web estas preguntas se corrigen automáticamente. No cuentan para la nota.

1. datos.gob.es solo publica datos de la Administración General del Estado. *[V / F]*
2. El Open Data Maturity compara _____ países europeos.
3. En el marco de Safarov et al., «la licencia y la calidad de los datos» corresponden a la dimensión de *[efectos / condiciones / usuarios / tipo de utilización]*
4. ¿Qué es re3data?

(a) Un portal de datos abiertos de la UE (b) Un registro internacional de repositorios de datos de investigación (c) Un repositorio de software libre (d) Un concurso de reutilización

5. Los Desafíos Aporta se siguen convocando cada año. *[V / F]*
6. Completa el código para calcular proporciones a partir de una tabla de frecuencias: _____ `(table(casos$efecto))`

Explicaciones

- Es el catálogo nacional y federa portales estatales, autonómicos, locales y universitarios.
- El informe de 2025 compara 36 países en cuatro dimensiones.
- Las condiciones son lo que facilita o frena el uso: calidad, licencia, formatos, competencias.
- re3data registra repositorios de datos de investigación de todo el mundo.
- La última edición fue la de 2021.
- `prop.table()` divide cada frecuencia entre el total.

Para practicar fuera del libro

> **Tip**
>
> En tu ordenador tienes las funciones de este tema en el paquete `datareuseR`: `consultar_datosgob()`, `tabla_datosgob()` y `evaluar_fair()`.

Elige un caso español y uno internacional de las fuentes anteriores y descríbelos con las cuatro dimensiones. Añádelos como filas nuevas a `casos` con `rbind()` y repite las tablas.

Recursos complementarios

Recurso	Para qué sirve
datos.gob.es – Aplicaciones y Empresas	Casos de reutilización en España
datos.gob.es – Desafíos Aporta	Ediciones 2017-2021 y ganadores
Open Data Maturity 2025	Informe comparado (resumen en datos.gob.es)
data.europa.eu – casos de uso	Casos europeos con seguimiento de impacto
data.gov y data.gov.uk	Portales internacionales de referencia
Global Data Barometer	Indicadores de gobernanza y uso de datos
datos.madrid.es, Aragón Open Data, Open Data Euskadi	Portales con datos de bibliotecas

Tema 3. Gestión documental y *open data*

Objetivos

- Comprender el papel de la gestión documental en la publicación de datos abiertos.
- Aplicar criterios de calidad y documentación de conjuntos de datos.

Glosario del tema

Gestión documental · Ciclo de vida · Metadatos · DCAT · Distribución · Completitud · Conformidad · Exactitud · Valor ausente · Checksum · OAIS · Identificador persistente

Descargas

Script R del capítulo Plantilla de informe (.qmd) Tema de estilo (.scss) CSV de práctica

El script reúne todo el código del capítulo para trabajar en RStudio o Positron. La plantilla `.qmd` y su tema `.scss` sirven para redactar las prácticas opcionales: guárdalos en la misma carpeta.

Lectura obligatoria

Vetrò, A.; Canova, L.; Torchiano, M.; Orozco Minotas, C.; Iemma, R.; Morando, F. (2016). Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, 33(2), 325-337. Postprint en acceso abierto.

Norma de referencia: ISO/IEC 25012:2008, modelo de calidad de datos (norma de pago; basta con la descripción).

Lecturas complementarias

- Batini et al. (2009) — Revisión clásica de metodologías de evaluación de la calidad de datos.
- Umbrich et al. (2015) — En *2015 3rd International Conference on Future Internet of Things and Cloud*, 404-411. Métricas de calidad aplicadas a portales de datos abiertos.

1. Gestión documental aplicada a los datos abiertos

Un conjunto de datos publicado es, a todos los efectos, un **documento** de la administración: tiene un productor, un contexto de creación, un valor (administrativo, informativo, histórico) y un plazo de conservación. Por eso la publicación de datos abiertos no empieza en el portal, sino en el **sistema de gestión documental** de la organización.

Aportación de la gestión documental	Aplicación a los datos abiertos
Identificación y clasificación (cuadro de clasificación)	Inventario de la información que produce cada unidad: qué datos existen y dónde
Valoración	Decidir qué conjuntos tienen valor para la reutilización y cuáles deben protegerse
Descripción	Metadatos que garantizan autenticidad, contexto y comprensión
Calendario de conservación	Cuánto tiempo se mantiene cada versión publicada y cuándo se transfiere o elimina
Control de acceso	Separar lo publicable de lo que contiene datos personales o información protegida

Normas de referencia:

- **ISO 15489-1:2016**, conceptos y principios de la gestión de documentos: autenticidad, fiabilidad, integridad y disponibilidad.
- **ISO 30301**, sistemas de gestión para los documentos.
- **Esquema Nacional de Interoperabilidad** y sus normas técnicas de **documento electrónico**, **expediente electrónico** y **política de gestión de documentos electrónicos**, que fijan cómo se crean, describen y conservan los documentos en la administración española (Tema 4).

En la práctica, muchas administraciones elaboran un **plan de apertura**: inventario de la información, priorización de los conjuntos con más demanda o valor, calendario de publicación y responsables de mantenerlos.

2. Ciclo de vida de la información

La gestión documental clásica describe el **ciclo vital del documento** por las fases de archivo que recorre: oficina o gestión, central, intermedio e histórico, según disminuye su uso administrativo y aumenta su valor informativo o histórico.

Para los datos se usan modelos más específicos. El modelo de ciclo de vida del DCC ordena las acciones secuenciales de curación: **crear o recibir → valorar y seleccionar → ingresar → acciones de preservación → almacenar → acceso, uso y reutilización → transformar**. Alrededor hay acciones continuas (descripción, planificación de la preservación, seguimiento de la comunidad) y ocasionales (eliminar, revalorar, migrar).

En el laboratorio de este capítulo, la limpieza es una acción de preservación sobre datos ya ingresados; publicar el resultado con sus metadatos facilita el acceso y la reutilización.

3. Metadatos

Un conjunto de datos sin metadatos no se encuentra ni se entiende. Hay metadatos **descriptivos** (qué es), **administrativos** (quién lo publica, con qué licencia), **estructurales** (qué columnas tiene y qué significan) y **de preservación** (versiones, sumas de control, formatos).

El vocabulario del W3C DCAT describe catálogos, conjuntos de datos (`dcat:Dataset`) y distribuciones (`dcat:Distribution`, cada fichero descargable), y reutiliza términos de **Dublin Core** (`dct:`). En Europa se usa el perfil DCAT-AP y en España DCAT-AP-ES, basado en DCAT-AP 2.1.1 y en la extensión de datos de alto valor. Según datos.gob.es, será el modelo de metadatos de la nueva versión de la NTI-RISP; comprueba si la norma modificada ya se ha publicado en el BOE.

Propiedad	Significado
`dct:title`	Título del conjunto
`dct:description`	Descripción
`dct:publisher`	Organismo que publica
`dct:license`	Licencia (URI)
`dct:accrualPeriodicity`	Periodicidad de actualización
`dct:temporal`	Cobertura temporal
`dct:spatial`	Cobertura geográfica
`dcat:downloadURL`	URL de descarga directa del fichero
`dcat:mediaType`	Tipo MIME del fichero (por ejemplo, `text/csv`)

Los metadatos **estructurales** de un CSV (nombre, tipo y significado de cada columna) se describen con un diccionario de datos o con especificaciones como CSV on the Web o Frictionless Data.

4. Calidad de los datos

Vetrò et al. proponen medir la calidad de cada conjunto de datos, no solo la del portal, con siete características y 17 métricas:

Característica	Qué mide
Trazabilidad	Si constan la creación y las modificaciones del conjunto
Actualidad	Si los datos se publican a tiempo respecto al periodo que describen
Expiración	Si el conjunto sigue vigente o ya ha sido sustituido
Completitud	Proporción de celdas y filas sin valores ausentes
Conformidad	Si los valores siguen el formato esperado (fechas, códigos, unidades)
Exactitud	Si los valores son correctos y plausibles
Comprensibilidad	Si hay metadatos y descripciones que permitan entenderlos

En el laboratorio calcularás dos métricas de completitud y corregirás problemas de conformidad y exactitud en el CSV de práctica.

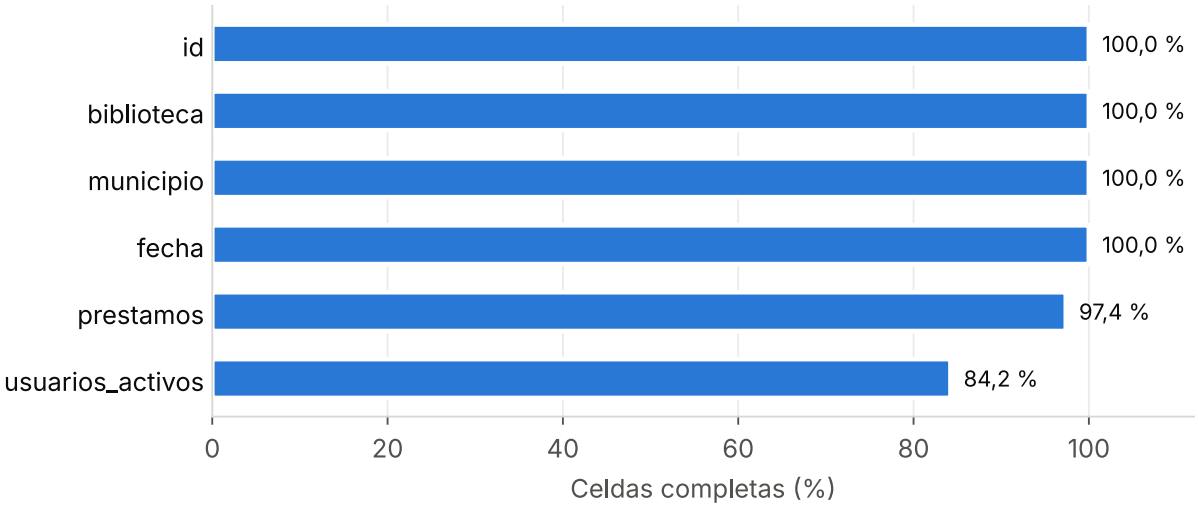

Figura 5: Completitud por columna del CSV de práctica: las columnas descriptivas están completas, pero faltan datos en préstamos y, sobre todo, en usuarios activos.

5. Publicación de conjuntos de datos

Las Buenas prácticas para datos en la Web del W3C y la experiencia de los portales permiten resumir la publicación en una secuencia:

1. **Seleccionar** el conjunto a partir del inventario y del plan de apertura.
2. **Revisar** la protección de datos: anonimizar o agregar si hay datos personales.
3. **Depurar** y comprobar la calidad (laboratorio de este capítulo).
4. **Preparar las distribuciones** en formatos abiertos (CSV, JSON, RDF) y, si procede, una API.
5. **Describir** con metadatos DCAT-AP-ES y un diccionario de datos.
6. **Asignar una licencia** (normalmente CC BY 4.0).

7. **Publicar** en el portal de la organización, a menudo construido con software como CKAN, y **federarlo** con datos.gob.es.
8. **Difundir** y **recoger la opinión** de los reutilizadores para mejorar.

6. Preservación y actualización

Publicar no es el final: un conjunto de datos debe seguir siendo **íntegro**, **accesible** y **actual**.

- **Modelo OAIS** (ISO 14721): modelo de referencia para archivos digitales que deben preservar información a largo plazo; distingue los paquetes de información que se reciben, se archivan y se difunden.
- **Integridad (fijeza)**: sumas de control (*checksums*) como MD5 o SHA-256 permiten comprobar que un fichero no ha cambiado ni se ha corrompido.
- **Identificadores persistentes**: un DOI o una URI estable evitan que los enlaces se rompan cuando el portal cambia.
- **Versionado**: cada actualización debe quedar registrada (qué cambió y cuándo); las versiones anteriores se conservan según el calendario de conservación.
- **Actualización periódica**: la periodicidad declarada en `dct:accrualPeriodicity` es un compromiso. Un conjunto que deja de actualizarse pierde **actualidad** y acaba **expirado**, dos características de calidad de Vetrò et al.
- **Formatos de preservación**: los formatos abiertos y documentados (CSV, XML) son también los más fáciles de conservar.

Práctica en R: laboratorio de calidad (apartados 3 a 6)

El fichero tiene errores introducidos a propósito. El objetivo es detectarlos, medirlos y corregirlos. Cada paso indica la operación equivalente en OpenRefine, la herramienta que usarás en el cuestionario de Aula Global.

Paso 1. Leer todo como texto

Leemos todas las columnas como texto (`colClasses = "character"`) para que R no transforme nada antes de que lo veamos.

```
bruto <- read.csv("datos/prestamos_bibliotecas_practica.csv", colClasses = "character")
dim(bruto)
head(bruto, 8)
```

Paso 2. Valores ausentes y completitud

Los ausentes aparecen de dos formas: celdas vacías y el texto `"n/d"`. Una matriz lógica marca cuáles lo son:

```
ausente <- bruto == "" | bruto == "n/d"
colSums(ausente)          # ausentes por columna
sum(ausente)              # total de celdas ausentes
```

Las dos métricas de completitud de Vetrò et al. son:

- **Porcentaje de celdas completas** = celdas no vacías / celdas totales × 100.
- **Porcentaje de filas completas** = filas sin ningún ausente / filas totales × 100.

Ejercicio 3.1

Calcula el porcentaje de celdas completas del fichero original, redondeado a dos decimales.

Ejercicio: completa los huecos _____

```
celdas_totales <- nrow(bruto) * ncol(bruto)
celdas_completas <- celdas_totales - _______
round(100 * celdas_completas / celdas_totales, 2)
```

Pista

Resta el número total de celdas ausentes: `sum(ausente)`.

Solución

```
celdas_totales <- nrow(bruto) * ncol(bruto)
celdas_completas <- celdas_totales - sum(ausente)
round(100 * celdas_completas / celdas_totales, 2)
```

Ejercicio 3.2

Calcula ahora el porcentaje de **filas** completas (dos decimales). Una fila es completa si no tiene ningún ausente.

Ejercicio: completa los huecos _____

```
fila_con_ausente <- rowSums(ausente) > 0
filas_completas <- sum(_______)
round(100 * filas_completas / nrow(bruto), 2)
```

Pista

`fila_con_ausente` vale `TRUE` en las filas incompletas. Necesitas lo contrario: el operador de negación es `!`.

Solución

```
fila_con_ausente <- rowSums(ausente) > 0
filas_completas <- sum(!fila_con_ausente)
round(100 * filas_completas / nrow(bruto), 2)
```

Paso 3. Filas duplicadas

En OpenRefine: ordenar, «Blank down» y eliminar las filas en blanco, o facetar por duplicados.

`duplicated()` marca las filas idénticas a otra anterior.

```
duplicated(bruto)
```

Ejercicio 3.3

¿Cuántas filas duplicadas hay? Después, crea `sin_dup` sin ellas y comprueba cuántas filas quedan.

Ejercicio: completa los huecos _____

```
n_dup <- sum(duplicated(bruto))
sin_dup <- bruto[______, ]
c(duplicadas = n_dup, filas_restantes = nrow(sin_dup))
```

Pista

Quieres las filas que **no** están duplicadas: `!duplicated(bruto)`.

Solución

```
n_dup <- sum(duplicated(bruto))
sin_dup <- bruto[!duplicated(bruto), ]
c(duplicadas = n_dup, filas_restantes = nrow(sin_dup))
```

Paso 4. Normalizar los municipios

En OpenRefine: «Edit cells → Common transforms → Trim whitespace» y «Cluster and edit» con el método key collision / fingerprint.

```
sin_dup <- bruto[!duplicated(bruto), ]
table(sin_dup$municipio)
```

El mismo municipio aparece escrito de varias formas. Construimos una **clave** como hace el método *fingerprint* de OpenRefine: quitar espacios sobrantes, pasar a minúsculas y eliminar tildes.

```
clave <- function(x) {
  x <- trimws(x)
  x <- tolower(x)
  chartr("áéíóúüñ", "aeiouun", x)
}
table(clave(sin_dup$municipio))
```

Después asignamos a cada clave su forma correcta con un vector con nombres (una tabla de equivalencias):

```
forma_correcta <- c(getafe = "Getafe", leganes = "Leganés", alcorcon = "Alcorcón",
                    mostoles = "Móstoles", fuenlabrada = "Fuenlabrada")
sin_dup$municipio <- unname(forma_correcta[clave(sin_dup$municipio)])
table(sin_dup$municipio)
```

Ejercicio 3.4

¿Cuántas formas distintas de escribir el municipio había en el fichero original y cuántos municipios hay realmente? Devuelve un vector con los dos números.

Ejercicio: completa los huecos _____

```
c(variantes  = length(unique(bruto$municipio)),
  municipios = length(unique(______)))
```

Pista

Aplica la función `clave()` a la columna antes de contar los valores únicos.

Solución

```
c(variantes  = length(unique(bruto$municipio)),
  municipios = length(unique(clave(bruto$municipio))))
```

Paso 5. Unificar las fechas

En OpenRefine: «Edit cells → Transform» con una expresión GREL que detecte el formato, o «To date» con varios formatos.

```
table(grepl("/", sin_dup$fecha))   # TRUE = formato dd/mm/aaaa
```

Convertimos cada formato con su patrón y combinamos el resultado:

```
es_dmy <- grepl("/", sin_dup$fecha)
fecha <- as.Date(sin_dup$fecha, format = "%Y-%m-%d")
fecha[es_dmy] <- as.Date(sin_dup$fecha[es_dmy], format = "%d/%m/%Y")
sin_dup$fecha <- fecha
sum(is.na(sin_dup$fecha))   # ninguna fecha sin convertir
range(sin_dup$fecha)
```

Paso 6. Valores numéricos

En OpenRefine: «To number», facetas numéricas para localizar valores no numéricos y negativos.

```
num <- function(x) as.numeric(ifelse(x %in% c("", "n/d"), NA, x))
sin_dup$prestamos <- num(sin_dup$prestamos)
sin_dup$usuarios_activos <- num(sin_dup$usuarios_activos)
summary(sin_dup$prestamos)
subset(sin_dup, prestamos < 0)
```

Un número de préstamos negativo es imposible: es un error de **exactitud**. Sin acceso a la fuente no podemos saber el valor correcto (¿un signo menos erróneo?), así que lo marcamos como ausente y lo documentamos.

```
sin_dup$prestamos[which(sin_dup$prestamos < 0)] <- NA
limpio <- sin_dup
str(limpio)
```

Ejercicio 3.5

Con el fichero limpio, calcula el **total de préstamos de Getafe** entre enero y junio. Ten en cuenta que puede haber `NA`.

Ejercicio: completa los huecos _____

```
getafe <- subset(limpio, municipio == _____)
sum(getafe$prestamos, na.rm = _____)
```

Pista

El primer hueco es el nombre del municipio entre comillas; el segundo, el valor lógico que indica a `sum()` que ignore los `NA`.

Solución

```
getafe <- subset(limpio, municipio == "Getafe")
sum(getafe$prestamos, na.rm = TRUE)
```

Ejercicio 3.6

¿Cuántos **préstamos por usuario activo** tuvo cada municipio? Calcula la ratio con `aggregate()` (que suma por grupos) y ordénala de mayor a menor. Este ejercicio no tiene una única respuesta: interpreta el resultado.

Ejercicio: completa los huecos _____

```
# Solo filas con ambos valores, para que numerador y denominador sean comparables
completas <- subset(limpio, !is.na(prestamos) & !is.na(usuarios_activos))
por_muni <- aggregate(cbind(prestamos, usuarios_activos) ~ municipio, data =
completas, FUN = sum)
por_muni$ratio <- round(por_muni$prestamos / por_muni$usuarios_activos, 2)
por_muni[order(por_muni$ratio, decreasing = TRUE), ]
```

Nota

¿Por qué se usan solo las filas con ambos valores? Si sumaras todos los préstamos pero solo los usuarios conocidos, la ratio de los municipios con más ausentes saldría inflada. Es un ejemplo de cómo la completitud afecta a los indicadores que se derivan de los datos.

Resultado de la limpieza

Con los datos ya depurados se puede, por fin, comparar municipios. Los huecos de las líneas son los meses en que el dato era ausente o negativo: la limpieza no los inventa, los deja a la vista.

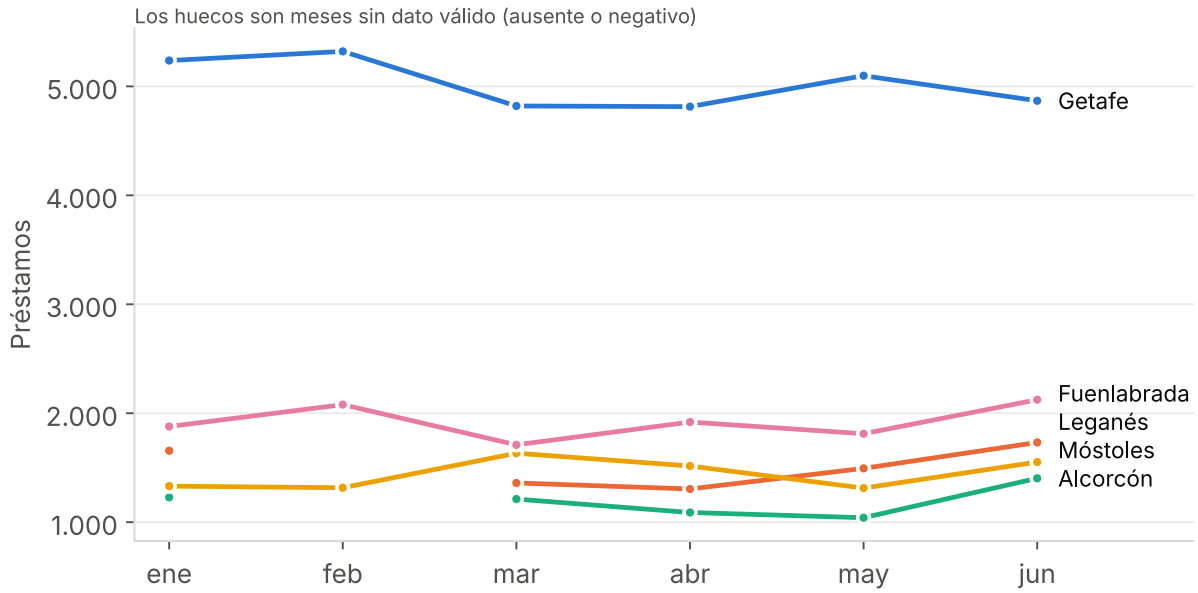

Figura 6: Préstamos mensuales por municipio tras la limpieza, enero a junio de 2025 (datos sintéticos). Getafe suma dos bibliotecas; el resto, una.

Explora los datos limpios

Marca y desmarca municipios: la gráfica se recalcula en tu navegador con R.

> **i** Gráfica interactiva
>
> Esta gráfica es interactiva en la versión web: pablosancheznunez.com/reutilizacion-datos

Activa la **media por biblioteca**: Getafe deja de destacar, porque tiene dos bibliotecas en la muestra. Elegir bien el indicador es parte de la calidad de un servicio de datos.

Paso 7. Publicar el resultado con metadatos (apartados 3 y 5)

Guardamos el fichero limpio y lo describimos con propiedades DCAT. `jsonlite` convierte una lista de R en JSON.

```
write.csv(limpio, "prestamos_limpio.csv", row.names = FALSE)

metadatos <- list(
  "@type" = "dcat:Dataset",
  "dct:title" = "Préstamos mensuales de bibliotecas municipales (enero-junio 2025)",
  "dct:description" = "Datos sintéticos para docencia. Versión depurada: municipios
normalizados, fechas ISO 8601, duplicados eliminados.",
  "dct:license" = "https://creativecommons.org/licenses/by/4.0/",
  "dct:accrualPeriodicity" = "http://publications.europa.eu/resource/authority/
frequency/MONTHLY",
  "dct:temporal" = list(startDate = "2025-01-01", endDate = "2025-06-30"),
  "dct:spatial" = "Comunidad de Madrid",
  "dcat:distribution" = list(
    "@type" = "dcat:Distribution",
    "dcat:downloadURL" = "https://ejemplo.org/datos/prestamos_limpio.csv",
    "dcat:mediaType" = "text/csv",
    "dct:format" = "CSV"
  )
```

```
)
cat(jsonlite::toJSON(metadatos, auto_unbox = TRUE, pretty = TRUE))
```

Ejercicio 3.7

Completa la distribución en **JSON** del mismo conjunto: su tipo MIME es `application/json`. Y su licencia es CC0 1.0, cuya URI es `https://creativecommons.org/publicdomain/zero/1.0/`.

Ejercicio: completa los huecos _____

```
distribucion <- list(
  "@type" = "dcat:Distribution",
  "dcat:downloadURL" = "https://ejemplo.org/datos/prestamos_limpio.json",
  "dcat:mediaType" = "_____",
  "dct:license" = "_____"
)
distribucion
```

Solución

```
distribucion <- list(
  "@type" = "dcat:Distribution",
  "dcat:downloadURL" = "https://ejemplo.org/datos/prestamos_limpio.json",
  "dcat:mediaType" = "application/json",
  "dct:license" = "https://creativecommons.org/publicdomain/zero/1.0/"
)
```

Paso 8. Comprobar la integridad de los ficheros (apartado 6)

`tools::md5sum()` calcula la suma de control MD5 de un fichero. Si dos ficheros tienen la misma suma, su contenido es idéntico; si cambia un solo carácter, la suma cambia por completo.

```
write.csv(limpio, "prestamos_limpio.csv", row.names = FALSE)
sumas <- tools::md5sum(c("datos/prestamos_bibliotecas_practica.csv",
"prestamos_limpio.csv"))
sumas
```

Un registro de versiones mínimo combina la fecha, la descripción del cambio y la suma de control:

```
versiones <- data.frame(
  version = c("1.0", "1.1"),
  fecha = as.Date(c("2025-07-01", Sys.Date())),
  cambio = c("Publicación inicial (fichero original)",
             "Depuración: duplicados, municipios, fechas y valores negativos"),
  md5 = unname(sumas)
)
versiones
```

Ejercicio 3.8

Un reutilizador descarga el fichero original y quiere comprobar que no se ha alterado. Compara la suma de control del fichero con la que publica el portal (`suma_publicada`). El resultado debe ser `TRUE` o `FALSE`.

Ejercicio: completa los huecos ______

```
suma_descarga <- unname(tools::md5sum("______"))
suma_descarga ______ suma_publicada
```

Pista

El primer hueco es la ruta del CSV original (`datos/...`). El segundo, el operador que compara si dos valores son iguales.

Solución

```
suma_descarga <- unname(tools::md5sum("datos/prestamos_bibliotecas_practica.csv"))
suma_descarga == suma_publicada
```

`TRUE`: el fichero está íntegro. Si comparas con `prestamos_limpio.csv` obtendrás `FALSE`, porque el contenido es distinto.



Figura 7: El ciclo de vida de los datos: cada fase tiene responsables y documentos propios.

Laboratorio: comprueba la calidad de un CSV

Laboratorio interactivo

Pega un CSV (o abre uno de tu ordenador) y pulsa **Revisar**. El laboratorio detecta el separador y revisa cinco dimensiones de calidad: **completitud** (celdas vacías), **unicidad** (filas duplicadas), **validez** (números y fechas mal escritos), **consistencia** (variantes de un mismo valor) y **formato** (espacios sobrantes).

Este laboratorio es interactivo en la versión web: pablosancheznunez.com/reutilizacion-datos

Caso práctico resuelto

Una ficha de metadatos que permite reutilizar

Conjunto: «Préstamos en bibliotecas municipales, 2024». Metadatos mínimos para que alguien pueda reutilizarlo sin preguntar:

- **Título y descripción** clara: qué se cuenta (préstamos, no usuarios), unidad y periodo.
- **Organismo publicador** y contacto.
- **Licencia** (CC BY 4.0).
- **Frecuencia de actualización** (mensual) y **fecha de la última actualización**.
- **Cobertura temporal y geográfica.**
- **Formato y distribución** (CSV, UTF-8, separador «;»).
- **Diccionario de datos**: cada columna con su significado, tipo y valores permitidos.
- **Calidad**: limitaciones conocidas (por ejemplo, la biblioteca X cerró por obras en marzo).

Sin diccionario de datos, una columna llamada «TIPO» con valores 1, 2 y 3 es inutilizable.

Ejercicios graduados

Básico. ¿Qué información falta si un CSV no indica su codificación ni su separador?

Explicaciones

Cómo leerlo correctamente: sin la codificación (UTF-8, Latin-1) pueden corromperse las tildes y la «ñ»; sin el separador (coma, punto y coma) las columnas se leen mal, sobre todo cuando se usa la coma decimal.

Avanzado. Redacta el diccionario de datos completo de un conjunto que hayas usado en el curso.

Explicaciones

Cada columna con nombre, descripción, tipo, unidad, valores permitidos o rango, tratamiento de los valores perdidos y ejemplo.

Actividades

Actividad: ¿Qué dimensión de la calidad falla?

Categorías: Completitud · Exactitud · Consistencia · Actualidad · Unicidad · Validez

- La columna de préstamos tiene celdas vacías
- «Madrid», «madrid » y «MADRID» aparecen como municipios distintos
- El conjunto se actualizó por última vez en 2019
- La misma biblioteca aparece dos veces en el mismo mes
- Una fecha dice 31/02/2024
- Los préstamos de una biblioteca están multiplicados por diez por un error de carga

Solución

- La columna de préstamos tiene celdas vacías → **Completitud**
- «Madrid», «madrid » y «MADRID» aparecen como municipios distintos → **Consistencia**
- El conjunto se actualizó por última vez en 2019 → **Actualidad**
- La misma biblioteca aparece dos veces en el mismo mes → **Unicidad**
- Una fecha dice 31/02/2024 → **Validez**
- Los préstamos de una biblioteca están multiplicados por diez por un error de carga → **Exactitud**

Actividad: Ordena la publicación de un conjunto de datos

- Describir el conjunto con metadatos DCAT-AP
- Elegir el formato abierto y la licencia
- Mantener la periodicidad de actualización anunciada
- Publicarlo en el portal y federarlo en datos.gob.es
- Revisar la calidad y la protección de datos personales

Solución

1. Revisar la calidad y la protección de datos personales
2. Elegir el formato abierto y la licencia
3. Describir el conjunto con metadatos DCAT-AP
4. Publicarlo en el portal y federarlo en datos.gob.es
5. Mantener la periodicidad de actualización anunciada

Reto evaluable

! Reto del Tema 3 · Limpieza documentada de un conjunto real

Elige en un portal de datos abiertos un CSV **de al menos 100 filas y 6 columnas** que no se use en este libro. Trabaja en tu ordenador con el paquete `datareuseR`:

1. Léelo con `leer_csv_abierto()` y documenta la codificación y el separador detectados.
2. Mide su calidad con `informe_calidad()` antes de tocarlo.
3. Aplica al menos **tres tipos de corrección** distintos (duplicados, variantes de texto con `clave_agrupacion()`, fechas, números, valores imposibles…) y justifica cada decisión en una tabla: problema, dimensión de calidad, corrección y número de filas afectadas.
4. Vuelve a medir la calidad y compara el antes y el después.
5. Describe el conjunto depurado con metadatos DCAT, incluida su procedencia (`dct:source`).

Entrega: un informe hecho con la plantilla de informe, renderizado a HTML o PDF, con el código visible y el conjunto de datos citado. Se evalúa con la rúbrica que incluye la plantilla. Este reto **no tiene solución en el libro**: cada estudiante trabaja con datos distintos y la corrección la hace el profesorado.

Espacio de trabajo, con el CSV de práctica para probar el código antes de usarlo con tus datos (no se corrige automáticamente):

```
source("datos/datareuseR.R")
datos <- leer_csv_abierto("datos/prestamos_bibliotecas_practica.csv", colClasses =
"character")
attr(datos, "lectura")
informe_calidad(datos)
# 1. Duplicados:        datos <- datos[!duplicated(datos), ]
# 2. Variantes de texto: table(clave_agrupacion(datos$municipio))
# 3. Fechas y números:  ...
# informe_calidad(datos)
```

Autoevaluación

Responde en el papel y comprueba después con las explicaciones del recuadro. En la versión web estas preguntas se corrigen automáticamente. No cuentan para la nota.

1. Un número de préstamos negativo es un problema de *[completitud / exactitud / trazabilidad / comprensibilidad]*
2. Que las fechas aparezcan en dos formatos distintos afecta a la *[conformidad / actualidad / expiración]*
3. En DCAT, cada fichero descargable de un conjunto de datos es una _____
4. El tipo MIME de un fichero CSV es _____
5. Si la suma de control MD5 de un fichero descargado coincide con la publicada, el fichero no ha cambiado. *[V / F]*
6. ¿Qué norma ISO describe el modelo de referencia OAIS para archivos digitales?

(a) ISO 15489 (b) ISO/IEC 25012 (c) ISO 14721 (d) ISO 30301

7. Completa el código para quedarte solo con las filas no duplicadas: `bruto[ _____ , ]`

Explicaciones

- Un valor imposible es un error de exactitud.
- La conformidad mide si los valores siguen el formato esperado.
- `dcat:Distribution` describe cada forma de obtener el conjunto: un CSV, un JSON, una API.
- `text/csv` es el tipo MIME registrado para CSV.
- La suma cambia si cambia un solo carácter del fichero.
- ISO 14721 es OAIS; ISO 15489 es gestión de documentos; ISO/IEC 25012 es calidad de datos; ISO 30301 es sistemas de gestión para los documentos.
- `duplicated()` marca las repeticiones y `!` las excluye.

Resumen de problemas encontrados

Problema	Característica de Vetrò et al.	Corrección
Ausentes como vacío y como `"n/d"`	Completitud y conformidad	Codificar todos como `NA`
Variantes del mismo municipio	Conformidad	Clave normalizada y tabla de equivalencias
Fechas en dos formatos	Conformidad	Convertir a ISO 8601 (`AAAA-MM-DD`)
Préstamos negativos	Exactitud	Marcar como ausente y documentarlo
Filas duplicadas	Exactitud	Eliminar duplicados

Para practicar fuera del libro

Tip

En tu ordenador tienes las funciones de este tema en el paquete `datareuseR`: `informe_calidad()`, `completitud()` y `clave_agrupacion()`.

Recursos complementarios

Recurso	Para qué sirve
Guía de implementación de DCAT-AP-ES y validación	Modelo de metadatos para catálogos españoles
datos.gob.es – DCAT-AP-ES, un paso adelante	Novedades respecto a la NTI-RISP de 2013
ISO/IEC 25012:2008	Modelo de calidad de datos
OpenRefine	Limpieza y transformación de datos
Frictionless Data	Describir y validar ficheros tabulares
CKAN	Software de catálogo de muchos portales
DCC Curation Lifecycle Model	Ciclo de vida de los datos
W3C – Data on the Web Best Practices	Buenas prácticas de publicación

Tema 4. Normalización e interoperabilidad

Objetivos

- Conocer los principales estándares utilizados en la publicación de datos abiertos.
- Comprender los fundamentos de la interoperabilidad técnica y semántica.

Glosario del tema

Estándar abierto · Interoperabilidad · ENI · URI · RDF · Tripleta · Turtle · JSON-LD · Linked Open Data · SPARQL · Vocabulario · Ontología · SKOS

Descargas

Script R del capítulo Plantilla de informe (.qmd) Tema de estilo (.scss) CSV de práctica

El script reúne todo el código del capítulo para trabajar en RStudio o Positron. La plantilla `.qmd` y su tema `.scss` sirven para redactar las prácticas opcionales: guárdalos en la misma carpeta.

Lectura obligatoria

Bizer, C.; Heath, T.; Berners-Lee, T. (2009). Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 1-22. Acceso por suscripción a través de la Biblioteca UC3M.

Lecturas complementarias

- Hallo et al. (2016) — Estado de los datos enlazados en bibliotecas digitales.

1. Estándares abiertos

Un **estándar** (o norma) es una especificación acordada que permite que productos y sistemas distintos funcionen juntos. Es **abierto** cuando su especificación es pública y gratuita, se ha desarrollado en un proceso abierto a todas las partes interesadas y puede implementarse sin pagar derechos ni pedir permiso. No hay que confundirlo con el software libre: un estándar abierto puede implementarse en software privativo.

Organismo	Estándares relevantes para datos abiertos
W3C (World Wide Web Consortium)	HTML, XML, RDF, SPARQL, OWL, SKOS, DCAT, JSON-LD, CSV on the Web
IETF (Internet Engineering Task Force)	HTTP, URI, CSV (RFC 4180), JSON (RFC 8259)
ISO	Dublin Core (ISO 15836), OAIS (ISO 14721), PDF (ISO 32000), códigos de país y lengua
OGC (Open Geospatial Consortium)	Servicios de mapas (WMS, WFS), GeoPackage
Ámbito bibliotecario (Library of Congress, IFLA, OCLC)	MARC 21, BIBFRAME, IFLA LRM, Z39.50, OAI-PMH

Los estándares de facto (los que se imponen por uso, como un formato de un fabricante dominante) pueden ser útiles, pero no garantizan acceso ni continuidad.

2. Interoperabilidad técnica, semántica y organizativa

La **interoperabilidad** es la capacidad de los sistemas y las organizaciones de intercambiar información y usarla. El Marco Europeo de Interoperabilidad (EIF, revisado en 2017) distingue cuatro niveles; publicar en un formato abierto resuelve solo uno de ellos:

Nivel	Pregunta	Ejemplo en bibliotecas
Jurídico	¿Las normas y licencias permiten combinar los datos?	Licencias compatibles entre dos portales
Organizativo	¿Los procesos y responsabilidades están alineados?	Acuerdo entre la red de bibliotecas y el portal autonómico para actualizar cada mes
Semántico	¿Los datos significan lo mismo para todos?	Usar `schema:Library` y la misma definición de «préstamo»
Técnico	¿Los sistemas pueden intercambiar datos?	API, formatos, protocolos

El Reglamento (UE) 2024/903, de 13 de marzo de 2024, conocido como **Reglamento sobre la Europa Interoperable**, se aplica en su mayor parte desde el 12 de julio de 2024. Convierte la interoperabilidad de los servicios públicos transfronterizos en una obligación con estructuras de gobernanza propias.

3. Esquema Nacional de Interoperabilidad

En España, el **Esquema Nacional de Interoperabilidad (ENI)**, previsto en el artículo 156 de la Ley 40/2015 y regulado por el Real Decreto 4/2010, fija los criterios que deben seguir las administraciones para que sus sistemas interoperen. Se desarrolla en **Normas Técnicas de Interoperabilidad (NTI)**, entre ellas:

NTI	Qué regula
Documento electrónico	Componentes, formato y metadatos mínimos del documento
Expediente electrónico	Estructura e índice del expediente
Digitalización y copiado auténtico	Requisitos de imágenes y copias
Política de gestión de documentos electrónicos	Procesos de gestión documental (Tema 3)
Catálogo de estándares	Formatos y estándares admitidos en la administración
Reutilización de recursos de información (NTI-RISP)	Identificación con URI, metadatos, formatos y condiciones de uso para publicar datos (Temas 1 y 3)

El texto del ENI y de sus normas está en el Portal de Administración Electrónica. El ENI tiene su paralelo en el **Esquema Nacional de Seguridad**, que regula la seguridad de los mismos sistemas.

4. Formatos de intercambio: CSV, XML, JSON y RDF

El mismo registro, en los cuatro formatos:

CSV

```
biblioteca,municipio,prestamos
Biblioteca Municipal Centro,Getafe,2637
```

XML

```
<biblioteca>
  <nombre>Biblioteca Municipal Centro</nombre>
  <municipio>Getafe</municipio>
  <prestamos>2637</prestamos>
</biblioteca>
```

JSON

```
{"biblioteca": "Biblioteca Municipal Centro", "municipio": "Getafe", "prestamos": 2637}
```

RDF (Turtle)

```
<https://ejemplo.org/biblioteca/centro> a schema:Library ;
    schema:name "Biblioteca Municipal Centro" ;
    schema:containedInPlace wd:Q8802 .
```

Formato	Característica principal	Puntos fuertes	Limitaciones
CSV	Tabla en texto plano	Sencillo; lo abre cualquier programa	Sin tipos ni significado; problemas de separadores y codificación
XML	Elementos jerárquicos con etiquetas	Validación con esquemas (XSD); estándar en metadatos bibliográficos (MARCXML, Dublin Core, METS)	Verboso
JSON	Objetos de clave y valor anidados	Formato habitual de las API; ligero	Sin esquema obligatorio ni semántica
RDF (Turtle, RDF/XML, JSON-LD)	Tripletas sujeto–predicado–objeto con URI	Significado explícito; enlazable	Curva de aprendizaje mayor

JSON-LD es JSON con un `@context` que asigna a cada clave una URI de un vocabulario: es RDF con aspecto de JSON.

5. Linked Open Data

Bizer, Heath y Berners-Lee recogen los cuatro principios de *Linked Data*:

1. Usar URI para nombrar las cosas.
2. Usar URI HTTP, para que cualquiera pueda consultarlas.
3. Al consultar una URI, ofrecer información útil con estándares (RDF, SPARQL).
4. Incluir enlaces a otras URI, para descubrir más cosas.

Linked Open Data (LOD) son datos enlazados publicados con licencia abierta: las cinco estrellas del Tema 0. La nube LOD muestra los conjuntos que se enlazan entre sí, con Wikidata y DBpedia en el centro. En bibliotecas, datos.bne.es, VIAF o id.loc.gov publican autoridades y registros como datos enlazados. Los datos RDF se consultan con **SPARQL**, el lenguaje de consulta del W3C.

6. Vocabularios y ontologías

Para que los datos signifiquen lo mismo para todos (interoperabilidad semántica) no se inventan propiedades: se **reutilizan vocabularios** existentes.

- Un **vocabulario** define un conjunto de clases y propiedades con URI (por ejemplo, `schema:Library` o `dct:title`).
- Una **ontología** es un vocabulario con una semántica formal más rica: jerarquías de clases, restricciones y reglas que permiten inferir información nueva. Se expresa en **RDFS** u **OWL**.
- Un **sistema de organización del conocimiento** (tesauro, clasificación, lista de encabezamientos) se publica en la Web con **SKOS**, que modela conceptos y sus relaciones (`skos:broader`, `skos:narrower`, `skos:related`, `skos:prefLabel`).

Vocabulario u ontología	Uso
Dublin Core (`dct:`)	Metadatos descriptivos básicos
DCAT	Catálogos y conjuntos de datos
schema.org	Descripción general de entidades en la Web (lo usan los buscadores)
FOAF	Personas y relaciones
SKOS	Tesauros y clasificaciones (EuroVoc, LCSH, tesauros de la BNE)
BIBFRAME, IFLA LRM / RDA	Descripción bibliográfica en datos enlazados
Vocabularios básicos de SEMIC	Personas, organizaciones, direcciones y servicios públicos en la UE

Para encontrar el vocabulario adecuado se usa Linked Open Vocabularies (LOV).

Práctica en R

Paso 1. Generar XML (apartado 4)
Partimos del CSV limpio y escribimos cada fila como un elemento `<registro>` con `sprintf()`:

```
cargar_limpio <- function() {
  b <- read.csv("datos/prestamos_bibliotecas_practica.csv", colClasses = "character")
  b <- b[!duplicated(b), ]
  clave <- chartr("áéíóúüñ", "aeiouun", tolower(trimws(b$municipio)))
  b$municipio <- unname(c(getafe = "Getafe", leganes = "Leganés", alcorcon = "Alcorcón",
                          mostoles = "Móstoles", fuenlabrada = "Fuenlabrada")[clave])
  es_dmy <- grepl("/", b$fecha)
  f <- as.Date(b$fecha, "%Y-%m-%d"); f[es_dmy] <- as.Date(b$fecha[es_dmy], "%d/%m/%Y")
  b$fecha <- format(f)
  for (v in c("prestamos", "usuarios_activos"))
    b[[v]] <- as.numeric(ifelse(b[[v]] %in% c("", "n/d"), NA, b[[v]]))
  b$prestamos[which(b$prestamos < 0)] <- NA
  b
}
limpio <- cargar_limpio()

a_xml <- function(d) {
  filas <- sprintf(
    '  <registro>\n    <biblioteca>%s</biblioteca>\n    <municipio>%s</municipio>\n
<fecha>%s</fecha>\n    <prestamos>%s</prestamos>\n  </registro>',
    d$biblioteca, d$municipio, d$fecha, ifelse(is.na(d$prestamos), "", d$prestamos))
  paste0('<?xml version="1.0" encoding="UTF-8"?>\n<prestamos_bibliotecas>\n',
         paste(filas, collapse = "\n"), '\n</prestamos_bibliotecas>')
}
xml <- a_xml(limpio)
cat(substr(xml, 1, 520))
```

Un documento XML está **bien formado** si tiene un único elemento raíz y cada etiqueta que se abre se cierra. Contar etiquetas de apertura y cierre es una comprobación rápida:

```
contar <- function(texto, patron) lengths(regmatches(texto, gregexpr(patron, texto,
fixed = TRUE)))
c(abren = contar(xml, "<registro>"), cierran = contar(xml, "</registro>"))
```

Ejercicio 4.1

¿Cuántos elementos `<prestamos>` **vacíos** (sin valor, porque el dato era ausente o erróneo) hay en el XML? Busca la cadena exacta de un elemento vacío.

Ejercicio: completa los huecos _____

```
contar(xml, "_____")
```

Pista

Un elemento vacío es la etiqueta de apertura seguida inmediatamente de la de cierre: `<prestamos></prestamos>`.

Solución

```
contar(xml, "<prestamos></prestamos>")
```

Dos: el «n/d» y el valor negativo que marcamos como ausente en el Tema 3. En XML también podría omitirse el elemento o marcarse con el atributo `xsi:nil="true"`.

Paso 2. De CSV a JSON (apartado 4)

`jsonlite::toJSON()` convierte cada fila en un objeto JSON:

```
cat(jsonlite::toJSON(head(limpio, 2), pretty = TRUE))
```

El JSON es más fácil de consumir desde una API, pero `"municipio": "Getafe"` sigue siendo solo una cadena de texto: ¿qué Getafe? (en Filipinas hay otro). Ahí entran las URI.

Paso 3. JSON-LD con schema.org (apartados 4 y 6)

Asignamos a cada municipio su URI de Wikidata y describimos las bibliotecas como `schema:Library`.

```
wikidata <- c(Getafe = "Q8802", "Leganés" = "Q77433", "Alcorcón" = "Q10372",
              "Móstoles" = "Q187826", Fuenlabrada = "Q54902")

bibliotecas <- unique(limpio[, c("biblioteca", "municipio")])
bibliotecas$id <- paste0("https://ejemplo.org/biblioteca/",
                         gsub(" ", "-", tolower(sub("Biblioteca Municipal ", "",
bibliotecas$biblioteca))))
bibliotecas
```

```
a_jsonld <- function(fila) {
  list(
    "@context" = "https://schema.org/",
    "@type" = "Library",
    "@id" = fila$id,
    "name" = fila$biblioteca,
    "containedInPlace" = list(
      "@id" = paste0("http://www.wikidata.org/entity/", wikidata[[fila$municipio]]),
      "name" = fila$municipio
```

```
    )
  )
}
cat(jsonlite::toJSON(a_jsonld(bibliotecas[1, ]), auto_unbox = TRUE, pretty = TRUE))
```

El `@context` dice que `name` significa `https://schema.org/name` y `Library` significa `https://schema.org/Library`. Copia el resultado en el JSON-LD Playground y mira la pestaña *N-Quads*: son las tripletas RDF.

Paso 4. Turtle (apartado 5)
La misma información en Turtle. `sprintf()` rellena una plantilla de texto con valores:

```
plantilla <- '@prefix schema: <https://schema.org/> .
@prefix wd: <http://www.wikidata.org/entity/> .

<%s> a schema:Library ;
    schema:name "%s"@es ;
    schema:containedInPlace wd:%s .'

b <- bibliotecas[1, ]
cat(sprintf(plantilla, b$id, b$biblioteca, wikidata[[b$municipio]]))
```

Ejercicio 4.2

Completa el Turtle de la Biblioteca Municipal Norte, de Leganés: falta la clase de schema.org y la entidad de Wikidata del municipio (está en el vector `wikidata`).

ℹ Ejercicio: completa los huecos _____

```
ttl <- '@prefix schema: <https://schema.org/> .
@prefix wd: <http://www.wikidata.org/entity/> .

<https://ejemplo.org/biblioteca/norte> a schema:______ ;
    schema:name "Biblioteca Municipal Norte"@es ;
    schema:containedInPlace wd:______ .'
cat(ttl)
ttl
```

ℹ Pista

La clase es la misma que en el ejemplo. Para la entidad, escribe `wikidata[["Leganés"]]` en otra línea para consultarla.

Solución

```
ttl <- '@prefix schema: <https://schema.org/> .
@prefix wd: <http://www.wikidata.org/entity/> .

<https://ejemplo.org/biblioteca/norte> a schema:Library ;
    schema:name "Biblioteca Municipal Norte"@es ;
    schema:containedInPlace wd:Q77433 .'
```

Paso 5. Consultas sobre un grafo (apartado 5)

Un grafo RDF es, en el fondo, una tabla de tres columnas: sujeto, predicado y objeto. Este pequeño grafo describe las bibliotecas y sus actividades culturales:

```
grafo <- read.table(header = TRUE, stringsAsFactors = FALSE, text = '
s               p                       o
ex:centro       rdf:type                schema:Library
ex:centro       schema:containedInPlace wd:Q8802
ex:sur          rdf:type                schema:Library
ex:sur          schema:containedInPlace wd:Q8802
ex:norte        rdf:type                schema:Library
ex:norte        schema:containedInPlace wd:Q77433
ex:parque       rdf:type                schema:Library
ex:parque       schema:containedInPlace wd:Q10372
ex:act1         rdf:type                schema:Event
ex:act1         schema:location         ex:centro
ex:act1         schema:about            "Club de lectura"
ex:act2         rdf:type                schema:Event
ex:act2         schema:location         ex:centro
ex:act2         schema:about            "Taller de datos abiertos"
ex:act3         rdf:type                schema:Event
ex:act3         schema:location         ex:sur
ex:act3         schema:about            "Cuentacuentos"
ex:act4         rdf:type                schema:Event
ex:act4         schema:location         ex:norte
ex:act4         schema:about            "Club de lectura"
ex:act5         rdf:type                schema:Event
ex:act5         schema:location         ex:centro
ex:act5         schema:about            "Alfabetización digital"
ex:act6         rdf:type                schema:Event
ex:act6         schema:location         ex:parque
ex:act6         schema:about            "Taller de datos abiertos"
')
grafo
```

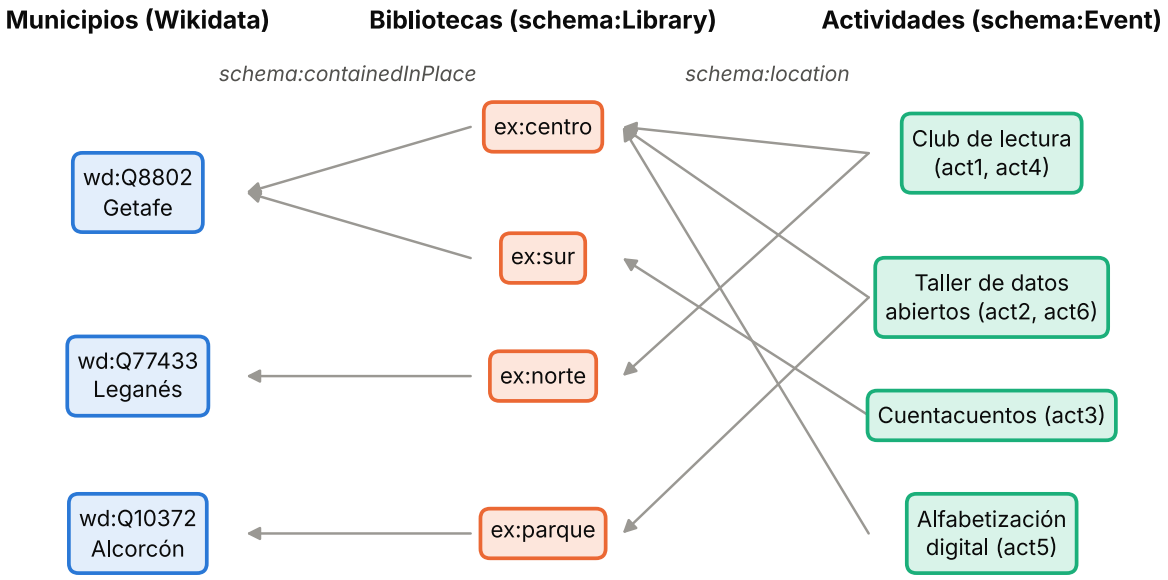

Figura 8: El grafo de la práctica: las actividades apuntan a la biblioteca donde se celebran (`schema:location`) y las bibliotecas, al municipio de Wikidata donde están (`schema:containedInPlace`). Seguir las flechas es lo que hace una consulta SPARQL.

Una consulta SPARQL busca tripletas que encajen con un **patrón**. Por ejemplo, «¿qué bibliotecas están en Getafe?»:

```
PREFIX schema: <https://schema.org/>
PREFIX wd: <http://www.wikidata.org/entity/>
SELECT ?biblioteca WHERE {
  ?biblioteca a schema:Library ;
              schema:containedInPlace wd:Q8802 .
}
```

En R, el mismo patrón es un filtro sobre la tabla de tripletas:

```
subset(grafo, p == "schema:containedInPlace" & o == "wd:Q8802")$s
```

Cuando el patrón tiene dos tripletas encadenadas por una variable, la operación equivalente es una **unión** (`merge()`). «¿Qué actividades hay en bibliotecas de Getafe?»:

```
SELECT ?actividad ?tema WHERE {
  ?actividad schema:location ?biblioteca ;
             schema:about ?tema .
  ?biblioteca schema:containedInPlace wd:Q8802 .
}
```

```
lugar <- subset(grafo, p == "schema:location")[, c("s", "o")]
names(lugar) <- c("actividad", "biblioteca")
tema <- subset(grafo, p == "schema:about")[, c("s", "o")]
names(tema) <- c("actividad", "tema")
en_getafe <- data.frame(biblioteca = subset(grafo, p == "schema:containedInPlace" & o ==
"wd:Q8802")$s)
```

```
resultado <- merge(merge(lugar, tema), en_getafe)
resultado
```

Ejercicio 4.3

¿Cuántas filas devuelve esta consulta? Resuélvela en R.

```
SELECT ?actividad WHERE {
  ?actividad schema:location ex:centro .
}
```

ℹ Ejercicio: completa los huecos `_____`

```
nrow(subset(grafo, p == "_____" & o == "_____"))
```

Solución

```
nrow(subset(grafo, p == "schema:location" & o == "ex:centro"))
```

Ejercicio 4.4

¿En cuántas bibliotecas **distintas** se celebra un «Taller de datos abiertos»? Equivale a `SELECT (COUNT(DISTINCT ?biblioteca) AS ?n)`.

ℹ Ejercicio: completa los huecos `_____`

```
actividades <- merge(lugar, tema)
talleres <- subset(actividades, tema == "_____")
length(unique(talleres$_____))
```

ℹ Pista

Primero filtra por el tema exacto; después cuenta los valores únicos de la columna `biblioteca`.

Solución

```
actividades <- merge(lugar, tema)
talleres <- subset(actividades, tema == "Taller de datos abiertos")
length(unique(talleres$biblioteca))
```

Paso 6. Un vocabulario controlado en SKOS (apartado 6)

Un pequeño esquema de conceptos para clasificar las actividades de las bibliotecas. `skos:broader` apunta al concepto más general:

```
skos <- read.table(header = TRUE, stringsAsFactors = FALSE, text = '
concepto                propiedad        valor
ex:actividades          skos:prefLabel   "Actividades culturales"
ex:fomento_lectura      skos:prefLabel   "Fomento de la lectura"
ex:fomento_lectura      skos:broader     ex:actividades
ex:club_lectura         skos:prefLabel   "Club de lectura"
ex:club_lectura         skos:broader     ex:fomento_lectura
ex:cuentacuentos        skos:prefLabel   "Cuentacuentos"
ex:cuentacuentos        skos:broader     ex:fomento_lectura
ex:formacion            skos:prefLabel   "Formación"
ex:formacion            skos:broader     ex:actividades
ex:alfabetizacion       skos:prefLabel   "Alfabetización digital"
ex:alfabetizacion       skos:broader     ex:formacion
ex:taller_datos         skos:prefLabel   "Taller de datos abiertos"
ex:taller_datos         skos:broader     ex:formacion
ex:taller_datos         skos:related     ex:alfabetizacion
')
subset(skos, propiedad == "skos:broader")
```

En SPARQL, los conceptos específicos de «Formación» serían:

```
SELECT ?concepto WHERE { ?concepto skos:broader ex:formacion . }
```

Ejercicio 4.5

Obtén las **etiquetas** (`skos:prefLabel`) de los conceptos cuyo concepto general es `ex:fomento_lectura`. Necesitas dos pasos: localizar los conceptos y buscar sus etiquetas.

ℹ Ejercicio: completa los huecos `_____`

```
especificos <- subset(skos, propiedad == "skos:broader" & valor == "_____")$concepto
subset(skos, propiedad == "skos:prefLabel" & concepto %in% _____)$valor
```

ℹ Pista

El primer hueco es la URI abreviada del concepto general; el segundo, el vector que acabas de crear.

Solución

```
especificos <- subset(skos, propiedad == "skos:broader" & valor ==
"ex:fomento_lectura")$concepto
subset(skos, propiedad == "skos:prefLabel" & concepto %in% especificos)$valor
```



Figura 9: El modelo DCAT: un catálogo contiene conjuntos de datos, y cada conjunto, una o varias distribuciones.

Laboratorio: del CSV al JSON y al RDF

i Laboratorio interactivo

Los mismos datos en tres formatos. Edita el CSV y mira cómo cambian el **JSON** (para las API) y el **RDF en Turtle** (para los datos enlazados, con un URI para cada biblioteca y el municipio enlazado con Wikidata).

En el RDF, `wd:Q2807` no es un texto: es el identificador de Madrid en Wikidata. Esa es la diferencia entre la tercera estrella (un CSV) y la quinta (datos enlazados). Los URI `example.org` son de ejemplo.

Este laboratorio es interactivo en la versión web: pablosancheznunez.com/reutilizacion-datos

Caso práctico resuelto

i Describir un catálogo con DCAT-AP

Para que un portal intercambie metadatos con datos.gob.es y con el portal europeo, cada conjunto se describe con el vocabulario **DCAT-AP** (en España, el perfil DCAT-AP-ES):

Concepto	Clase DCAT	Ejemplo
Catálogo	`dcat:Catalog`	Portal de datos del ayuntamiento
Conjunto de datos	`dcat:Dataset`	Préstamos en bibliotecas, 2024
Distribución	`dcat:Distribution`	El CSV y el JSON de ese conjunto
Servicio	`dcat:DataService`	La API que da acceso a los datos

Propiedades habituales del conjunto: `dct:title`, `dct:description`, `dct:publisher`, `dcat:keyword`, `dcat:theme`, `dct:temporal`, `dct:spatial`, `dct:accrualPeriodicity`. De la distribución: `dcat:accessURL` o `dcat:downloadURL`, `dct:format`, `dct:license`.

Al usar vocabularios controlados (por ejemplo, los temas del sector público), los buscadores de los portales agregadores clasifican bien los datos.

Ejercicios graduados

Básico. ¿Es una distribución o un conjunto de datos el archivo CSV con los préstamos de 2024?

Explicaciones

Una distribución: es una forma concreta (formato, URL) de acceder al conjunto de datos «Préstamos en bibliotecas, 2024», que puede tener otras distribuciones (JSON, API).

Avanzado. Escribe en Turtle la descripción DCAT de un conjunto con dos distribuciones.

Explicaciones

Debe declarar un `dcat:Dataset` con título, descripción, publicador, licencia y periodicidad, y dos `dcat:Distribution` enlazadas con `dcat:distribution`, cada una con formato y URL de descarga.

Actividades

Actividad: ¿Qué formato es?

Categorías: CSV · JSON · XML · RDF (Turtle)

- `municipio;prestamos` seguido de `Madrid;12450`
- `{"municipio": "Madrid", "prestamos": 12450}`
- `<biblioteca><municipio>Madrid</municipio></biblioteca>`
- `ex:eugenio-trias schema:location wd:Q2807 .`

Solución

- `municipio;prestamos` seguido de `Madrid;12450` → **CSV**
- `{"municipio": "Madrid", "prestamos": 12450}` → **JSON**
- `<biblioteca><municipio>Madrid</municipio></biblioteca>` → **XML**
- `ex:eugenio-trias schema:location wd:Q2807 .` → **RDF (Turtle)**

Actividad: ¿A qué propiedad DCAT-AP corresponde?

Categorías: dct:title · dct:publisher · dct:license · dct:modified · dcat:downloadURL

- «Préstamos en las bibliotecas públicas municipales»
- «Ayuntamiento de Madrid»
- «https://creativecommons.org/licenses/by/4.0/»
- «2024-02-15» como fecha de la última actualización
- Enlace directo al fichero CSV

Solución

- «Préstamos en las bibliotecas públicas municipales» → **dct:title**
- «Ayuntamiento de Madrid» → **dct:publisher**
- «https://creativecommons.org/licenses/by/4.0/» → **dct:license**
- «2024-02-15» como fecha de la última actualización → **dct:modified**
- Enlace directo al fichero CSV → **dcat:downloadURL**

Actividad: ¿Qué nivel de interoperabilidad?

Categorías: Técnica · Semántica · Organizativa

- Usar HTTPS y codificación UTF-8
- Usar los códigos de municipio del INE en todos los conjuntos
- Un convenio entre ayuntamientos para publicar con la misma periodicidad
- Describir los datos con el vocabulario DCAT-AP

Solución

- Usar HTTPS y codificación UTF-8 → **Técnica**
- Usar los códigos de municipio del INE en todos los conjuntos → **Semántica**
- Un convenio entre ayuntamientos para publicar con la misma periodicidad → **Organizativa**
- Describir los datos con el vocabulario DCAT-AP → **Semántica**

Reto evaluable

Reto del Tema 4 · Publicar como datos enlazados

Parte del conjunto que depuraste en el reto del Tema 3 (o de otro conjunto abierto que describa entidades: bibliotecas, archivos, museos, centros culturales…):

1. Convierte **al menos diez registros** a JSON-LD con el tipo de schema.org más adecuado y justifica la elección de cada propiedad.
2. Enlaza al menos una propiedad con un identificador externo (por ejemplo, el municipio con su URI de Wikidata).
3. Comprueba el resultado en el validador de schema.org y corrige los avisos.
4. Describe el conjunto con DCAT-AP: título, descripción, publicador, licencia, cobertura temporal y una distribución con su formato del vocabulario de tipos de fichero de la UE.
5. Explica qué gana un reutilizador con esta versión frente al CSV original.

Entrega: un informe hecho con la plantilla de informe, renderizado a HTML o PDF, con el código visible y el conjunto de datos citado. Se evalúa con la rúbrica que incluye la plantilla. Este reto **no tiene solución en el libro**: cada estudiante trabaja con datos distintos y la corrección la hace el profesorado.

Espacio de trabajo (no se corrige automáticamente):

```r
a_jsonld <- function(fila) {
  list("@context" = "https://schema.org/",
       "@type" = "Library",                # cámbialo según tus datos
       "name" = fila$nombre,
       "address" = list("@type" = "PostalAddress", "addressLocality" = fila$municipio),
       "containedInPlace" = list("@id" = "https://www.wikidata.org/entity/Q8802"))  #
Q8802 = Getafe
}
ejemplo <- data.frame(nombre = "Biblioteca Municipal Centro", municipio = "Getafe")
cat(jsonlite::toJSON(a_jsonld(ejemplo[1, ]), auto_unbox = TRUE, pretty = TRUE))
```

Autoevaluación

Responde en el papel y comprueba después con las explicaciones del recuadro. En la versión web estas preguntas se corrigen automáticamente. No cuentan para la nota.

1. Usar el mismo vocabulario (por ejemplo, `schema:Library`) resuelve la interoperabilidad *[técnica / semántica / jurídica / organizativa]*
2. El ENI está previsto en el artículo 156 de la Ley 40/2015. *[V / F]*
3. En JSON-LD, la clave que asigna a cada término una URI de un vocabulario es _____
4. ¿Cuál de estos es el cuarto principio de Linked Data?

(a) Usar formatos propietarios (b) Publicar solo en PDF (c) Incluir enlaces a otras URI (d) Evitar los identificadores

5. La propiedad SKOS que apunta a un concepto más general es `skos:` _____
6. Una ontología es un vocabulario con semántica formal que permite inferencias. *[V / F]*
7. Completa la tripleta Turtle: `<https://ejemplo.org/biblioteca/centro>` _____ `schema:Library .`

> **Explicaciones**
>
> - La interoperabilidad semántica asegura que los datos signifiquen lo mismo para todos.
> - Así es: el ENI se regula en el Real Decreto 4/2010.
> - `@context` convierte las claves JSON en URI.
> - Los cuatro principios: URI, URI HTTP, información útil con estándares y enlaces a otras URI.
> - `skos:broader` apunta al concepto general; `skos:narrower`, al específico.
> - RDFS y OWL permiten expresar jerarquías y restricciones de las que se deduce información nueva.
> - En Turtle, `a` abrevia `rdf:type`.

Para practicar fuera del libro

> **Tip**
>
> En tu ordenador tienes las funciones de este tema en el paquete `datareuseR`: `normalizar_formato()` y `tabla_datosgob()`.

- Copia en el JSON-LD Playground el JSON-LD de todas las bibliotecas (`lapply(split(bibliotecas, seq_len(nrow(bibliotecas))), a_jsonld)`) y comprueba que es válido.
- En el Wikidata Query Service lanza una consulta real de bibliotecas de una provincia. Los resultados cambian cada día porque Wikidata se edita continuamente: por eso los ejercicios de este capítulo usan un grafo fijo.

Recursos complementarios

Recurso	Para qué sirve
Esquema Nacional de Interoperabilidad (PAe)	ENI y sus normas técnicas
Interoperable Europe (SEMIC)	EIF, DCAT-AP y vocabularios básicos
W3C DCAT 3	Vocabulario de catálogos de datos
W3C CSV on the Web	Metadatos para publicar CSV interoperables
Linked Open Vocabularies	Buscador de vocabularios RDF
schema.org	Vocabulario usado en el capítulo
Wikidata Query Service	Consultas SPARQL sobre Wikidata
The Linked Open Data Cloud	Mapa de conjuntos de datos enlazados
datos.bne.es	Datos enlazados de la BNE
JSON-LD Playground	Validar y convertir JSON-LD

Tema 5. Empresas infomediarias y oportunidades de negocio

Objetivos

- Analizar el ecosistema económico de la reutilización de datos.
- Identificar oportunidades de innovación y emprendimiento basadas en datos abiertos.

Glosario del tema

Sector infomediario · Marco 6-V · Freemium · Punto de equilibrio · Valor añadido · Compartir igual · Minería de textos y datos · Espacio de datos · Datos sintéticos

Descargas

Script R del capítulo Plantilla de informe (.qmd) Tema de estilo (.scss) CSV de práctica

El script reúne todo el código del capítulo para trabajar en RStudio o Positron. La plantilla `.qmd` y su tema `.scss` sirven para redactar las prácticas opcionales: guárdalos en la misma carpeta.

Lectura obligatoria

Ahmadi Zeleti, F.; Ojo, A.; Curry, E. (2016). Exploring the economic value of open government data. *Government Information Quarterly*, 33(3), 535-551. Versión en acceso abierto (MURAL).

Lecturas complementarias

- Magalhaes y Roseira (2020) — Modelos de negocio de empresas que usan datos abiertos de gobierno.
- Abella et al. (2017) — Innovación y generación de valor con datos en ecosistemas urbanos.
- Jetzek et al. (2014) — Mecanismos por los que los datos abiertos generan valor.

1. El sector infomediario

Las empresas **infomediarias** crean productos y servicios de valor añadido a partir de información, en buena parte del sector público, para venderlos a terceros. Se agrupan por el tipo de información que reutilizan: geográfica y cartográfica, económica, financiera y mercantil, jurídica, de mercado y consumo, meteorológica, de movilidad y transporte, sociodemográfica, cultural…

En España, el estudio oficial más reciente de caracterización del sector es el del ONTSI de 2016. Para cifras actuales se usa el informe anual de ASEDIE, la asociación del sector, cuya 14.ª edición se presentó en abril de 2026 con datos a 31 de diciembre de 2024. Consulta las cifras en el resumen de datos.gob.es: te harán falta en el cuestionario de Aula Global. Ten en cuenta que los dos estudios no usan la misma metodología, así que sus cifras no son directamente comparables.

El ecosistema no lo forman solo empresas: también participan administraciones que publican, intermediarios de datos, comunidades de desarrolladores, medios de comunicación, centros de investigación y, como mediadoras de la información, las bibliotecas.

2. Modelos de negocio basados en datos abiertos

Ahmadi Zeleti et al. proponen el marco **6-V** para describir cómo un negocio crea y captura valor con datos abiertos:

Componente	Pregunta
Propuesta de valor	¿Qué se ofrece y a quién?
Proceso de creación de valor	¿Qué se hace con los datos para generar ese valor?

Componente	Pregunta
Valor obtenido (*value in return*)	¿Qué recibe la organización a cambio?
Captura de valor	¿Cómo se convierte ese valor en ingresos o beneficios?
Gestión del valor	¿Cómo se mantiene y se protege el valor en el tiempo?
Red de valor	¿Con qué socios, proveedores y clientes se crea?

A partir de ellos agrupan los modelos de negocio en **patrones**, entre ellos *freemium* (servicio básico gratis y funciones avanzadas de pago), *premium* (todo de pago, con valor añadido alto), **ahorro de costes** (los datos abiertos reducen gastos propios), **beneficio indirecto** (los datos se usan para apoyar otro negocio o la reputación) y **«maquinilla y cuchillas»** (algo barato o gratuito que genera ventas recurrentes de otra cosa).

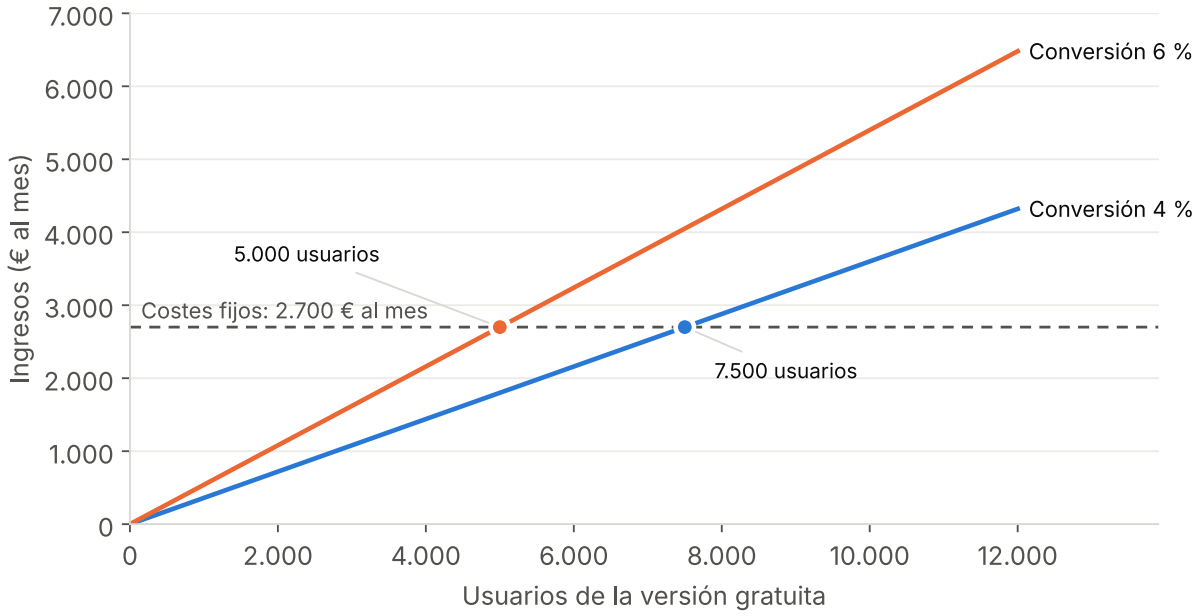

Figura 10: Punto de equilibrio del servicio *freemium* de la práctica, con costes fijos de 2.700 € al mes y un precio de 9 €. Subir la conversión del 4 % al 6 % reduce el número de usuarios gratuitos necesarios de 7.500 a 5.000.

3. Servicios de valor añadido

Los datos abiertos son materia prima: el valor está en lo que se hace con ellos. Los servicios de valor añadido más habituales son:

Tipo de servicio	Qué añade	Ejemplo con datos de bibliotecas
Agregación	Reunir fuentes dispersas en un solo lugar	Panel con los préstamos de todos los municipios de una comunidad
Normalización y limpieza	Datos homogéneos y fiables	Series depuradas como las del Tema 3
Enriquecimiento	Cruzar con otras fuentes	Préstamos por cada 1.000 habitantes con datos de población
Análisis y predicción	Indicadores, tendencias, modelos	Previsión de demanda para planificar compras

Tipo de servicio	Qué añade	Ejemplo con datos de bibliotecas
Visualización	Mapas, gráficos, cuadros de mando	Mapa de cobertura bibliotecaria
Acceso programático	API estables y documentadas	API que sirve los préstamos en JSON
Alertas y personalización	Información a medida	Aviso de novedades o actividades por barrio
Consultoría	Interpretación experta	Informe para la renovación de un servicio municipal

La cadena de valor va de los **datos** (publicados por la administración) a la **información** (datos depurados y combinados) y al **conocimiento** (indicadores y decisiones), una vuelta a la sociedad del conocimiento del Tema 0.

4. Emprendimiento e innovación

Emprender con datos abiertos exige lo mismo que cualquier proyecto (un problema real, unos clientes, un modelo sostenible) y algo más: confianza en que los datos seguirán publicándose y se podrán usar legalmente.

- **Herramientas de diseño**: el lienzo de modelo de negocio (*Business Model Canvas*) o su variante *Lean Canvas* ayudan a ordenar la propuesta; el marco 6-V la adapta a los datos.
- **Vías de impulso**: retos y concursos (como los Desafíos Aporta, 2017-2021), *hackatones*, incubadoras y aceleradoras, compra pública de innovación y programas europeos.
- **Innovación abierta**: administraciones y empresas colaboran con comunidades externas para resolver problemas con datos.
- **Riesgos que hay que gestionar**: discontinuidad o cambios en las fuentes, calidad desigual, dependencia de un único publicador y, sobre todo, licencias.

Licencias al combinar fuentes

Un servicio infomediario suele combinar varias fuentes, y sus obligaciones se acumulan:

Licencia	Atribución	Compartir igual	Uso comercial
CC0 1.0	No	No	Sí
CC BY 4.0	Sí	No	Sí
CC BY-SA 4.0	Sí	Sí	Sí
ODbL 1.0	Sí	Sí (base de datos derivada que se distribuya)	Sí
CC BY-NC 4.0	Sí	No	**No**

Dos matices: la cláusula «compartir igual» de ODbL afecta a la base de datos derivada, no necesariamente a un producto creado a partir de ella (por ejemplo, un mapa impreso); y dos licencias con «compartir igual» distintas pueden ser incompatibles entre sí, por lo que no siempre es posible combinarlas.

5. Inteligencia artificial y reutilización de datos

La relación funciona en los dos sentidos.

Los datos abiertos alimentan la IA. Son una fuente de datos de entrenamiento y evaluación de modelos. Antes de usarlos hay que revisar:

- La **licencia** de cada fuente y sus obligaciones (atribución, compartir igual, no comercial).
- Las excepciones de **minería de textos y datos** de la Directiva (UE) 2019/790 sobre derechos de autor, y la posibilidad de que los titulares se hayan reservado ese uso.

- La presencia de **datos personales** (RGPD), incluso en conjuntos aparentemente anónimos.
- Las obligaciones del Reglamento (UE) 2024/1689 de IA: transparencia frente a los usuarios en ciertos sistemas y, para los modelos de uso general, documentación técnica y un resumen público del contenido usado para el entrenamiento.
- La **calidad** y los **sesgos** de los datos: un recomendador entrenado con préstamos solo de algunas sedes reproducirá sus patrones.

La IA facilita la reutilización. Ayuda a generar y mejorar metadatos, detectar errores, clasificar, traducir y consultar datos en lenguaje natural. A cambio, exige verificar los resultados: un modelo puede inventar cifras o referencias.

6. Tendencias de futuro

Algunas líneas que ya están en marcha y marcarán los próximos años:

- **Datos de alto valor y API**: obligación de publicar datos clave gratis, por API y en tiempo casi real.
- **Espacios comunes europeos de datos**: entornos de confianza para compartir datos, también privados, por sectores (patrimonio cultural, sanidad, movilidad…). El Data Spaces Support Centre los acompaña.
- **Ley de Datos**: acceso de usuarios y terceros a los datos generados por productos conectados, aplicable desde septiembre de 2025.
- **Altruismo de datos e intermediarios de datos**, regulados por el Reglamento de Gobernanza de Datos.
- **IA generativa como interfaz** para encontrar y consultar datos, y como nuevo gran consumidor de datos abiertos.
- **Datos sintéticos** para compartir información sensible protegiendo la privacidad (como el CSV de este libro, a pequeña escala).
- **Medición del impacto**: cada vez se exige más demostrar el valor económico y social de la apertura (dimensión «impacto» del Open Data Maturity).

Práctica en R

Paso 1. Un servicio de valor añadido: enriquecer los préstamos (apartado 3)

Cruzamos el CSV limpio con una tabla de población para calcular **préstamos por cada 1.000 habitantes**, un indicador que ninguna de las dos fuentes ofrece por separado.

> ⚠ Advertencia
>
> Las cifras de población son **ficticias y redondeadas**, inventadas para el ejercicio. En un servicio real se usarían las del Padrón continuo del INE.

```
b <- read.csv("datos/prestamos_bibliotecas_practica.csv", colClasses = "character")
b <- b[!duplicated(b), ]
clave <- chartr("áéíóúüñ", "aeiouun", tolower(trimws(b$municipio)))
b$municipio <- unname(c(getafe = "Getafe", leganes = "Leganés", alcorcon = "Alcorcón",
                        mostoles = "Móstoles", fuenlabrada = "Fuenlabrada")[clave])
b$prestamos <- as.numeric(ifelse(b$prestamos %in% c("", "n/d"), NA, b$prestamos))
b$prestamos[which(b$prestamos < 0)] <- NA
limpio <- b

poblacion <- data.frame(
  municipio = c("Getafe", "Leganés", "Alcorcón", "Móstoles", "Fuenlabrada"),
  habitantes = c(185000, 190000, 170000, 210000, 190000)   # ficticias
)

por_muni <- aggregate(prestamos ~ municipio, data = limpio, FUN = sum)
```

```r
enriquecido <- merge(por_muni, poblacion)
enriquecido
```

Ejercicio 5.1

Añade a `enriquecido` la columna `por_mil` (préstamos por cada 1.000 habitantes, redondeada a un decimal) y devuelve la tabla ordenada de mayor a menor.

ℹ Ejercicio: completa los huecos `_____`

```r
enriquecido$por_mil <- round(enriquecido$prestamos / enriquecido$habitantes * _____,
1)
enriquecido[order(enriquecido$por_mil, decreasing = TRUE), ]
```

ℹ Pista

«Por cada 1.000 habitantes»: multiplica el cociente por 1000.

Solución

```r
enriquecido$por_mil <- round(enriquecido$prestamos / enriquecido$habitantes * 1000,
1)
enriquecido[order(enriquecido$por_mil, decreasing = TRUE), ]
```

Getafe queda en cabeza, pero tiene dos bibliotecas en la muestra y los demás municipios una. Un buen servicio de valor añadido debe explicar estas limitaciones a sus clientes.

Paso 2. ¿Es viable un servicio *freemium*? (apartados 2 y 4)

Imagina una empresa que ofrece a las bibliotecas un panel con sus préstamos comparados con los de otros municipios, a partir de datos abiertos. La versión básica es gratuita; la avanzada cuesta una cuota mensual.

```r
ingresos_mensuales <- function(usuarios, conversion, precio) {
  usuarios * conversion * precio
}
usuarios   <- 5000    # bibliotecas y profesionales que usan la versión gratuita
conversion <- 0.04    # 4 % pasa a la versión de pago
precio     <- 9       # euros al mes
costes_fijos <- 2700  # euros al mes (servidores, mantenimiento, actualización de datos)

ingresos <- ingresos_mensuales(usuarios, conversion, precio)
c(ingresos = ingresos, costes = costes_fijos, resultado = ingresos - costes_fijos)
```

¿Cómo cambia el resultado con el número de usuarios? Prueba otros valores de conversión y precio:

```r
u <- seq(0, 12000, by = 500)
plot(u, ingresos_mensuales(u, conversion, precio), type = "l", lwd = 2, col = "#000e78",
     xlab = "Usuarios de la versión gratuita", ylab = "Euros al mes",
     main = "Ingresos frente a costes fijos")
abline(h = costes_fijos, lty = 2, col = "firebrick")
```

```
legend("topleft", c("Ingresos", "Costes fijos"), lty = c(1, 2), col = c("#000e78",
"firebrick"), bty = "n")
```

Simulador del punto de equilibrio

Mueve los controles y observa cómo cambia el número de usuarios gratuitos necesarios para cubrir los costes.

> **i** Gráfica interactiva
>
> Esta gráfica es interactiva en la versión web: pablosancheznunez.com/reutilizacion-datos

Ejercicio 5.2

¿Cuántos usuarios de la versión gratuita necesita la empresa para cubrir los costes fijos (punto de equilibrio)? Despeja de `usuarios × conversión × precio = costes fijos`.

> **i** Ejercicio: completa los huecos _____
>
> ```
> costes_fijos / (______ * ______)
> ```

> **i** Pista
>
> Lo que aporta cada usuario gratuito, de media, es la conversión por el precio.

> **Solución**
>
> ```
> costes_fijos / (conversion * precio)
> ```
>
> 7.500 usuarios. Cada usuario gratuito aporta de media 0,36 € al mes.

Ejercicio 5.3

La empresa mejora la versión de pago y consigue una **conversión del 6 %** manteniendo el precio. ¿Cuál es ahora el resultado mensual (ingresos menos costes) con 5.000 usuarios?

> **i** Ejercicio: completa los huecos _____
>
> ```
> ingresos_mensuales(usuarios = 5000, conversion = ______, precio = 9) - costes_fijos
> ```

> **Solución**
>
> ```
> ingresos_mensuales(usuarios = 5000, conversion = 0.06, precio = 9) - costes_fijos
> ```

Paso 3. Obligaciones al combinar licencias (apartado 4)

Codificamos la tabla de licencias como un *data frame* y escribimos una función que, dadas varias fuentes, acumula sus obligaciones.

```r
licencias <- data.frame(
  licencia        = c("CC0 1.0", "CC BY 4.0", "CC BY-SA 4.0", "ODbL 1.0", "CC BY-NC
4.0"),
  atribucion      = c(FALSE, TRUE, TRUE, TRUE, TRUE),
  compartir_igual = c(FALSE, FALSE, TRUE, TRUE, FALSE),
  uso_comercial   = c(TRUE, TRUE, TRUE, TRUE, FALSE)
)

obligaciones <- function(fuentes) {
  l <- licencias[licencias$licencia %in% fuentes, ]
  c(atribucion      = any(l$atribucion),
    compartir_igual = any(l$compartir_igual),
    uso_comercial   = all(l$uso_comercial))
}

obligaciones(c("CC0 1.0", "CC BY 4.0"))
```

`any()` es `TRUE` si al menos una fuente lo exige; `all()` es `TRUE` solo si todas lo permiten.

Ejercicio 5.4

Un servicio combina el padrón municipal (CC BY 4.0) con datos de OpenStreetMap (ODbL 1.0). Calcula sus obligaciones.

Ejercicio: completa los huecos _____

```r
obligaciones(c("______", "______"))
```

Pista

Escribe los nombres de las licencias exactamente como aparecen en la tabla `licencias`.

Solución

```r
obligaciones(c("CC BY 4.0", "ODbL 1.0"))
```

Hay que atribuir ambas fuentes y, si se distribuye la base de datos derivada, hacerlo con ODbL. El uso comercial está permitido.

Ejercicio 5.5

Si se añade una tercera fuente con licencia **CC BY-NC 4.0**, ¿puede la empresa vender el servicio? Devuelve solo el valor de `uso_comercial`.

Ejercicio: completa los huecos _____

```r
o <- obligaciones(c("CC BY 4.0", "ODbL 1.0", ______))
o[["______"]]
```

Solución

```
o <- obligaciones(c("CC BY 4.0", "ODbL 1.0", "CC BY-NC 4.0"))
o[["uso_comercial"]]
```

`FALSE`: basta una fuente no comercial para que el servicio no pueda venderse. Por eso CC BY-NC no es una licencia abierta según la Open Definition (Tema 0).

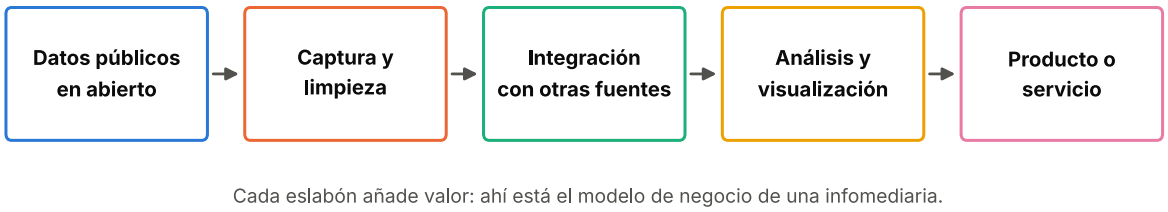

Figura 11: La cadena de valor de los datos abiertos.

Caso práctico resuelto

Diseñar un servicio infomediario viable

Una empresa infomediaria quiere ofrecer a bibliotecas informes comparativos de uso basados en datos abiertos de varias comunidades autónomas.

1. **Necesidad:** las bibliotecas quieren compararse con otras parecidas, pero los datos están dispersos y en formatos distintos.
2. **Datos:** estadísticas de bibliotecas públicas abiertas, población por municipio (INE), equipamientos.
3. **Proceso:** extracción automática, normalización de nombres y unidades, control de calidad, cálculo de indicadores comparables.
4. **Producto:** panel e informe anual con municipios similares (por población y presupuesto).
5. **Modelo:** suscripción para redes de bibliotecas; versión gratuita con indicadores básicos.
6. **Diferenciación:** el valor no está en los datos (abiertos para todos), sino en la integración, la limpieza y el análisis.

Ejercicios graduados

Básico. ¿Por qué una empresa puede cobrar por un servicio basado en datos abiertos gratuitos?

Explicaciones

Porque cobra por el valor añadido: integración de fuentes, limpieza, actualización continua, análisis, visualización y soporte, que tienen un coste y ahorran trabajo al cliente.

Avanzado. Esboza el modelo de negocio (lienzo de propuesta de valor) de un servicio infomediario para el sector de la información.

Explicaciones

Debe describir el segmento de clientes, sus problemas, la propuesta de valor, los datos abiertos que se usan, los recursos y actividades clave, los ingresos y los costes, y los riesgos de dependencia de las fuentes.

Actividades

Actividad: ¿Qué modelo de negocio es?

Categorías: Freemium · Suscripción · Servicios a medida · Publicidad · Venta de informes

- Consulta básica gratuita y funciones avanzadas de pago
- Una cuota mensual por acceder a una API con datos limpios
- Un ayuntamiento encarga un cuadro de mando de sus bibliotecas
- Una aplicación gratuita de horarios que muestra anuncios
- Un estudio sectorial anual elaborado con datos públicos

Solución

- Consulta básica gratuita y funciones avanzadas de pago → **Freemium**
- Una cuota mensual por acceder a una API con datos limpios → **Suscripción**
- Un ayuntamiento encarga un cuadro de mando de sus bibliotecas → **Servicios a medida**
- Una aplicación gratuita de horarios que muestra anuncios → **Publicidad**
- Un estudio sectorial anual elaborado con datos públicos → **Venta de informes**

ℹ Escenario: Una idea de negocio con datos abiertos

Quieres crear una aplicación que recomiende la biblioteca más cercana con salas libres, combinando datos municipales.

1. La idea

¿Qué compruebas primero?

- Si hay personas dispuestas a usarla y qué necesitan → paso 3
- Cómo programar la aplicación → paso 2

2. Empezar por el código

(final mejorable)

Sin saber si alguien la necesita ni si los datos lo permiten, corres el riesgo de construir algo que nadie usará.

3. Las personas usuarias

Las entrevistas confirman el interés. Ahora miras los datos.

- Compruebo la licencia, la frecuencia de actualización y si los datos de ocupación existen → paso 5
- Doy por hecho que, si están en el portal, sirven → paso 4

4. Supuesto arriesgado

(mal final)

Descubres después que la ocupación de las salas no se publica y que el conjunto se actualiza una vez al año: la idea no es viable tal como estaba planteada.

5. Datos revisados

Los horarios y ubicaciones están en abierto (CC BY 4.0), pero la ocupación en tiempo real no se publica.

- Pido al ayuntamiento que publique la ocupación y, mientras, lanzo una versión con horarios y ubicaciones → paso 7
- Copio la ocupación de su web con un programa sin preguntar → paso 6

6. Sin permiso

(mal final)

Extraer datos de una web sin revisar sus condiciones puede incumplirlas y hace tu servicio frágil: un cambio en la web lo rompe.

7. Producto mínimo viable

(buen final)

Empiezas con lo que los datos permiten, citas la fuente como exige la licencia y conviertes la carencia en una petición de datos: así funciona el ecosistema de la reutilización.

Reto evaluable

! Reto del Tema 5 · Propuesta de un servicio infomediario

Diseña un servicio de valor añadido para bibliotecas, archivos o su público que combine **al menos dos fuentes de datos abiertos con licencias distintas**:

1. Describe el servicio, sus usuarios y el valor que añade frente a consultar directamente las fuentes.
2. Identifica las licencias y calcula con `obligaciones()` qué exige y qué permite el servicio resultante.
3. Elige un modelo de negocio y calcula con `punto_equilibrio()` los usuarios necesarios en tres escenarios (pesimista, base y optimista). Justifica los supuestos.
4. Analiza los riesgos: dependencia del publicador, actualización de los datos, calidad y cambios de licencia.

Entrega: un informe hecho con la plantilla de informe, renderizado a HTML o PDF, con el código visible y el conjunto de datos citado. Se evalúa con la rúbrica que incluye la plantilla. Este reto **no tiene solución en el libro**: cada estudiante trabaja con datos distintos y la corrección la hace el profesorado.

Espacio de trabajo (no se corrige automáticamente):

```
source("datos/datareuseR.R")
fuentes <- c("CC BY 4.0", "ODbL 1.0")   # las licencias de tus fuentes
obligaciones(fuentes)
escenarios <- data.frame(
  escenario  = c("pesimista", "base", "optimista"),
  conversion = c(0.02, 0.04, 0.06),
  precio     = c(9, 9, 12),
  costes_mes = c(3000, 2700, 2700)
)
escenarios$usuarios_necesarios <- with(escenarios, punto_equilibrio(costes_mes,
conversion, precio))
escenarios
```

Autoevaluación

Responde en el papel y comprueba después con las explicaciones del recuadro. En la versión web estas preguntas se corrigen automáticamente. No cuentan para la nota.

1. Combinar una fuente CC BY 4.0 con otra CC BY-NC 4.0 permite vender el servicio resultante. *[V / F]*
2. Un servicio con un nivel gratuito y funciones avanzadas de pago sigue el patrón *[premium / freemium / ahorro de costes / maquinilla y cuchillas]*
3. Con costes fijos de 3.000 € al mes, una conversión del 5 % y un precio de 10 €, el punto de equilibrio es de _____ usuarios gratuitos.
4. ¿Qué componente del marco 6-V responde a «¿con qué socios, proveedores y clientes se crea el valor?»?

(a) Propuesta de valor (b) Captura de valor (c) Red de valor (d) Gestión del valor

5. Cruzar los préstamos con la población para obtener préstamos por cada 1.000 habitantes es un servicio de *[agregación / enriquecimiento / alertas]*
6. Completa el código: los préstamos por cada 1.000 habitantes se calculan como `prestamos / habitantes` * _____

Explicaciones

- Una sola fuente no comercial impide el uso comercial del conjunto.
- *Freemium* combina un nivel gratuito y otro de pago.
- 3000 / (0,05 × 10) = 6.000 usuarios.
- La red de valor reúne a los socios, proveedores y clientes.
- Enriquecer es cruzar con otras fuentes para obtener información nueva.
- Se multiplica el cociente por 1000.

Para practicar fuera del libro

Tip

En tu ordenador tienes las funciones de este tema en el paquete `datareuseR`: `punto_equilibrio()` y `obligaciones()`.

En grupo, idead un servicio que combine al menos dos conjuntos de datos abiertos. Describidlo con los seis componentes del marco 6-V, calculad su punto de equilibrio con la función `ingresos_mensuales()` y comprobad las obligaciones de licencia con `obligaciones()`. Presentadlo en 5 minutos.

Recursos complementarios

Recurso	Para qué sirve
ASEDIE – Informes anuales	Informe anual del sector infomediario
datos.gob.es – Así evoluciona el ecosistema infomediario según ASEDIE	Cifras de la 14.ª edición
Estudio de caracterización del sector infomediario 2016 (ONTSI)	Último estudio oficial
datos.gob.es – Empresas reutilizadoras	Directorio de empresas españolas
Desafíos Aporta	Ideas y productos premiados (2017-2021)
Open Data Institute	Modelos de negocio y ecosistemas de datos
data.europa.eu – Open data business models	Informe europeo sobre modelos de negocio
Data Spaces Support Centre	Espacios de datos europeos
Reglamento (UE) 2024/1689 de IA	Obligaciones de transparencia y datos

Caso práctico: las bibliotecas públicas municipales de Madrid

Recorrido completo por los seis temas con un conjunto de datos real

Objetivos

- Aplicar en un caso real lo aprendido en los Temas 0 a 5.
- Reconocer los problemas que presentan los datos abiertos publicados «de verdad».
- Plantear un servicio de valor añadido sobre datos públicos y evaluar sus limitaciones.

Glosario del caso

Licencia abierta · Cinco estrellas · Metadatos · Completitud · Distribución · JSON-LD · Valor añadido

Descargas

Script R del caso CSV original (Latin-1) Plantilla de informe (.qmd) Tema de estilo (.scss)

Los datos

Fuente: Ayuntamiento de Madrid, Portal de Datos Abiertos. *Bibliotecas de Madrid*. Responsable: Dirección General de Bibliotecas, Archivos y Museos. Licencia CC BY 4.0. https://datos.madrid.es/dataset/201747-0--bibliobuses-bibliotecas

Copia descargada el 24 de noviembre de 2026, sin modificar. El portal actualiza el conjunto a diario, así que la versión vigente puede ser distinta.

A diferencia del CSV de práctica, este fichero no se ha preparado para la docencia: es lo que publica una administración. Cada sección del caso corresponde a un tema del libro.

Tema 0. Auditar el conjunto

Leerlo… y fallar

Lo primero que hace cualquier reutilizador es abrir el fichero. Si lo leemos como el CSV de práctica, el resultado no tiene sentido:

```
f <- "datos/bibliotecas_madrid_2026-11-24.csv"
mal <- read.csv(f)
dim(mal)
names(mal)[1:3]
```

R ha leído una sola columna: el separador no es la coma, sino el punto y coma, algo habitual en los portales españoles porque la coma es el separador decimal. Además, el fichero no está en UTF-8, sino en **Latin-1 (ISO-8859-1)**: sin indicarlo, las tildes se estropean.

```
bib <- read.csv(f, sep = ";", fileEncoding = "latin1")
dim(bib)
head(bib$NOMBRE, 4)
```

Ejercicio C.1

¿Cuántas bibliotecas incluye el conjunto? Lee el fichero correctamente y cuenta las filas.

Ejercicio: completa los huecos _____

```
bib <- read.csv("datos/bibliotecas_madrid_2026-11-24.csv", sep = "_____",
fileEncoding = "_____")
nrow(bib)
```

Pista

El separador es el punto y coma. La codificación se llama `"latin1"` en R.

Solución

```
bib <- read.csv("datos/bibliotecas_madrid_2026-11-24.csv", sep = ";", fileEncoding =
"latin1")
nrow(bib)
```

Ejercicio C.4

El paquete `datareuseR` incluye `leer_csv_abierto()`, que detecta la codificación y el separador antes de leer y guarda lo que ha detectado en el atributo `"lectura"`. ¿Qué **codificación** detecta en el fichero de Madrid?

Ejercicio: completa los huecos _____

```
bib2 <- leer_csv_abierto("datos/bibliotecas_madrid_2026-11-24.csv")
attr(bib2, "_____")$encoding
```

Pista

El atributo se llama como el verbo del que viene: *leer*.

Solución

```
bib2 <- leer_csv_abierto("datos/bibliotecas_madrid_2026-11-24.csv")
attr(bib2, "lectura")$encoding
```

`"latin1"`: la función ha comprobado que el texto no es UTF-8 válido. Detectar lo que antes resolviste a mano es útil, pero conviene documentarlo, como hace el atributo.

Valoración

Criterio	Valoración
Licencia	CC BY 4.0: abierta según la Open Definition
Formato	CSV: estructurado y no propietario

Criterio	Valoración
Estrellas	★★★ (las URI de `CONTENT-URL` son páginas web, no identificadores de las bibliotecas como entidades)
FAIR	Localizable y accesible en el portal; la reutilización mejoraría con un diccionario de datos legible por máquina

Tema 1. Condiciones de reutilización

La licencia CC BY 4.0 permite cualquier uso, también comercial, con una obligación: **citar la fuente**. Una atribución correcta indica el autor, el título, la licencia y si se ha modificado el original. Por ejemplo:

> Datos: «Bibliotecas de Madrid», Ayuntamiento de Madrid, Portal de Datos Abiertos, CC BY 4.0. Modificados: selección de columnas y conversión a UTF-8.

Si lo reutilizáramos con licencia tipo y sin citar la fuente, cometeríamos el tipo de infracción leve que estudiaste en el Tema 1.

Tema 2. ¿Quién podría reutilizarlo?

Antes de tocar los datos, conviene pensar en usos posibles con las dimensiones de Safarov et al.:

Uso posible	Usuario	Efecto
Mapa de bibliotecas en una app de ocio	Empresas o desarrolladores	Económico
Estudio de la cobertura bibliotecaria por distrito	Investigadores, administraciones	Social y de buen gobierno
Buscador de bibliotecas accesibles	Ciudadanía, asociaciones	Social

Tema 3. Calidad y metadatos

Columnas vacías y completitud

```
vacia <- function(x) is.na(x) | trimws(as.character(x)) == ""
ausentes <- sapply(bib, function(col) sum(vacia(col)))
sort(ausentes[ausentes > 0], decreasing = TRUE)
```

Tres columnas (`PLANTA`, `PUERTA`, `ESCALERAS`) están **completamente vacías**: forman parte de un modelo de datos genérico de equipamientos que, para las bibliotecas, no se rellena.

Ejercicio C.2

Calcula el porcentaje de celdas completas del conjunto, redondeado a dos decimales, con la misma métrica de Vetrò et al. que en el Tema 3.

ℹ Ejercicio: completa los huecos `_____`

```
celdas_vacias <- sum(sapply(bib, function(col) sum(vacia(col))))
celdas_totales <- _______
round(100 * (celdas_totales - celdas_vacias) / celdas_totales, 2)
```

Pista

Celdas totales = filas × columnas.

Solución

```
celdas_vacias <- sum(sapply(bib, function(col) sum(vacia(col))))
celdas_totales <- nrow(bib) * ncol(bib)
round(100 * (celdas_totales - celdas_vacias) / celdas_totales, 2)
```

Códigos sin explicar

```
table(bib$ACCESIBILIDAD)
```

La columna `ACCESIBILIDAD` contiene códigos («1,6», «1,5,6»…) cuyo significado no está en el propio fichero, sino en un documento PDF del portal («Equipamientos. Contenido y estructura del fichero»). Es un problema de **comprensibilidad**: los datos son abiertos, pero no se entienden sin documentación externa. Un diccionario de datos legible por máquina (por ejemplo, con CSV on the Web o Frictionless) lo resolvería.

Bibliotecas por distrito

```
por_distrito <- sort(table(bib$DISTRITO), decreasing = TRUE)
op <- par(mar = c(4, 11, 1, 1))
barplot(rev(por_distrito), horiz = TRUE, las = 1, col = "#2a78d6", border = NA,
        cex.names = 0.7, xlab = "Bibliotecas públicas municipales")
par(op)
```

Ejercicio C.5

¿Qué **distrito** tiene más bibliotecas públicas municipales?

Ejercicio: completa los huecos _____

```
names(which.max(table(bib$______)))
```

Pista

La columna se llama igual que la variable: `DISTRITO`, en mayúsculas.

Solución

```
names(which.max(table(bib$DISTRITO)))
```

Puente de Vallecas, con cinco. Para valorar la cobertura habría que relacionarlo con la población de cada distrito, que está en otro conjunto del mismo portal.

Metadatos del conjunto derivado

Si publicáramos nuestra versión depurada, deberíamos describirla y declarar su procedencia con `dct:source`:

```
meta <- list(
  "@type" = "dcat:Dataset",
  "dct:title" = "Bibliotecas públicas municipales de Madrid (versión depurada)",
  "dct:source" = "https://datos.madrid.es/dataset/201747-0-bibliobuses-bibliotecas",
  "dct:license" = "https://creativecommons.org/licenses/by/4.0/",
  "dct:modified" = "2026-11-24",
  "dct:description" = "Selección de columnas, conversión a UTF-8 y eliminación de
columnas vacías."
)
cat(jsonlite::toJSON(meta, auto_unbox = TRUE, pretty = TRUE))
```

Tema 4. De CSV a datos enlazados

Cada biblioteca tiene coordenadas. Con schema.org podemos describirla como `Library` con su ubicación geográfica:

```
a_jsonld <- function(b) {
  list("@context" = "https://schema.org/", "@type" = "Library",
       "name" = b$NOMBRE,
       "url" = b$CONTENT.URL,
       "address" = list("@type" = "PostalAddress",
                        "streetAddress" = paste(b$CLASE.VIAL, b$NOMBRE.VIA, b$NUM),
                        "postalCode" = as.character(b$CODIGO.POSTAL),
                        "addressLocality" = "Madrid"),
       "geo" = list("@type" = "GeoCoordinates", "latitude" = b$LATITUD, "longitude" =
b$LONGITUD))
}
cat(jsonlite::toJSON(a_jsonld(bib[1, ]), auto_unbox = TRUE, pretty = TRUE, digits = NA))
```

Fíjate en que R ha cambiado los guiones de los nombres de columna por puntos (`CONTENT-URL` ⟶ `CONTENT.URL`): otro detalle que conviene conocer al reutilizar.

Las coordenadas permiten incluso un «mapa» sin cartografía de fondo:

```
plot(bib$LONGITUD, bib$LATITUD, pch = 19, col = "#2a78d6", asp = 1 / cos(40.4 * pi /
180),
     xlab = "Longitud", ylab = "Latitud", main = "Bibliotecas públicas municipales de
Madrid")
```

Tema 5. Un servicio de valor añadido

Idea: un buscador que diga qué bibliotecas abren los sábados. La columna `HORARIO` es **texto libre**, así que lo más sencillo es buscar la palabra «sábado»:

```
menciona_sabado <- grepl("s[áa]bado", bib$HORARIO, ignore.case = TRUE)
sum(menciona_sabado)
cat(bib$HORARIO[1])
```

Pero mencionar el sábado no significa abrir el sábado: el texto también recoge **días de cierre** («Días de cierre: sábados: …») y periodos de verano. Un servicio fiable necesitaría que el portal publicara los horarios estructurados (por ejemplo, con la propiedad `openingHoursSpecification` de schema.org). Este es el tipo de mejora que un infomediario puede proponer al publicador, o asumir como trabajo propio de valor añadido.

Ejercicio C.3

¿Cuántas bibliotecas mencionan la palabra «domingo» en su horario? Compara el resultado con el de los sábados: ¿te parece verosímil que tantas bibliotecas abran en domingo?

Ejercicio: completa los huecos _____

```
sum(grepl("_____", bib$HORARIO, ignore.case = TRUE))
```

Solución

```
sum(grepl("domingo", bib$HORARIO, ignore.case = TRUE))
```

Salen incluso más menciones del domingo que del sábado. Si lees los horarios, verás que mezclan aperturas («sábados y domingos de 8:30 a 14 horas»), cierres («domingos: todos») y periodos («Semana Santa, de jueves a domingo»). El recuento de palabras no responde a la pregunta: hacen falta horarios estructurados.

Autoevaluación

Responde en el papel y comprueba después con las explicaciones del recuadro. En la versión web estas preguntas se corrigen automáticamente. No cuentan para la nota.

1. El fichero del Ayuntamiento de Madrid está codificado en *[UTF-8 / Latin-1 (ISO-8859-1) / UTF-16]*
2. Con la licencia CC BY 4.0, una empresa puede usar estos datos en una aplicación de pago si cita la fuente. *[V / F]*
3. El conjunto tiene _____ estrellas en el modelo de Berners-Lee.
4. Que la columna `ACCESIBILIDAD` solo se entienda con un PDF aparte es un problema de *[exactitud / actualidad / comprensibilidad]*
5. ¿Por qué no basta con buscar «sábado» en `HORARIO` para saber qué bibliotecas abren ese día?

(a) Porque R no distingue tildes (b) Porque el texto también recoge días y periodos de cierre (c) Porque la columna está vacía

Explicaciones

- El fichero se lee con `fileEncoding = "latin1"`; sin ello, las tildes se estropean.
- CC BY 4.0 permite el uso comercial y solo exige atribución.
- Licencia abierta, estructurado y formato no propietario; no usa URI para identificar las bibliotecas.
- Los datos son correctos, pero no se entienden sin documentación externa.
- El horario es texto libre que mezcla aperturas y cierres; harían falta horarios estructurados.

Para practicar fuera del libro

Tip

En tu ordenador tienes las funciones de este tema en el paquete `datareuseR`: `leer_csv_abierto()`, `informe_calidad()` y el conjunto `bibliotecas_madrid`.

Repite el caso con el conjunto de bibliotecas de otro portal (tu comunidad autónoma o tu ayuntamiento) y redacta el resultado con la plantilla de informe. Compara: ¿qué portal publica mejor sus datos, y por qué?

Laboratorio de limpieza de datos

Antes de reutilizar un conjunto de datos hay que **limpiarlo**: espacios sobrantes, mayúsculas inconsistentes, fechas en cinco formatos, números con separador de miles, valores imposibles y duplicados. Este laboratorio usa un CSV pequeño y deliberadamente sucio de préstamos en bibliotecas municipales (datos ficticios).

Los problemas del fichero

Problema	Ejemplo en el fichero	Por qué importa
Espacios sobrantes	`"  Biblioteca Sur"`	«Biblioteca Sur» y « Biblioteca Sur» cuentan como dos bibliotecas
Mayúsculas inconsistentes	`BIBLIOTECA OESTE`, `centro`	Rompen agrupaciones y uniones
Variantes de un nombre	`Río` y `Rio`	Duplican categorías
Fechas en varios formatos	`01/03/2024`, `3 mar 2024`, `2024/03/02`	No se pueden ordenar ni filtrar por periodo
Separador de miles y texto	`1.250`, `n/d`	R los lee como texto, no como números
Valores imposibles	`-12` préstamos	Sesgan medias y totales
Duplicados	dos filas de la Biblioteca Central del 1 de marzo	Cuentan dos veces el mismo dato

Laboratorio: limpia paso a paso

ℹ Laboratorio interactivo

Activa las operaciones de limpieza. La tabla resalta con ⚠ las celdas que aún tienen problemas y, debajo, aparece el código de R equivalente, listo para copiarlo en la consola del final.

Este laboratorio es interactivo en la versión web: pablosancheznunez.com/reutilizacion-datos

ℹ Actividad: ¿En qué orden conviene limpiar?

Algunas operaciones dependen de otras: por ejemplo, los duplicados solo aparecen cuando las filas ya están normalizadas. Ordena los pasos.

- Convertir fechas a un formato común
- Convertir los números y marcar los valores imposibles
- Documentar en un registro qué se cambió y por qué
- Eliminar los duplicados
- Recortar espacios
- Unificar mayúsculas y variantes de los nombres

Solución

1. Recortar espacios
2. Unificar mayúsculas y variantes de los nombres
3. Convertir fechas a un formato común
4. Convertir los números y marcar los valores imposibles
5. Eliminar los duplicados
6. Documentar en un registro qué se cambió y por qué

Si eliminas duplicados al principio, «Biblioteca Central » y «biblioteca central» siguen pareciendo filas distintas. Y documentar los cambios es lo que hace la limpieza **reproducible**.

Hazlo tú en R

Copia aquí el código que ha generado el laboratorio (o escríbelo tú) y ejecútalo. El fichero está en `datos/bibliotecas_sucio.csv`.

```
df <- read.csv("datos/bibliotecas_sucio.csv", colClasses = "character")
df
```

```
# Escribe aquí tu limpieza y termina con:
str(df)
```

Para ir más allá

Para conjuntos grandes o con muchas variantes de nombres, **OpenRefine** agrupa valores parecidos («Rio», «Río», «RIO») con algoritmos de *clustering* y guarda un historial de operaciones que puedes exportar y reaplicar a otro fichero.

Para ver y seguir aprendiendo

Son enlaces, no vídeos incrustados: el vídeo se abre en la web de quien lo publica, así esta página no carga rastreadores de terceros. La mayoría está en inglés; activa los subtítulos si los necesitas.

- **The next web** — Tim Berners-Lee · TED2009. El creador de la web defiende publicar datos abiertos y enlazados (unos 16 minutos, con transcripción).

Chuleta de R

Todas las funciones que usa el libro, en una página.

Para…	Función	Ejemplo
Guardar un valor	`<-`	`x <- 5`
Crear un vector	`c()`	`c(1, 2, 3)`
Contar elementos	`length()`	`length(x)`
Sumar, media, máximo	`sum()`, `mean()`, `max()`	`sum(x, na.rm = TRUE)`
Redondear	`round()`	`round(3.14159, 2)`
¿Está en la lista?	`%in%`	`"CC BY 4.0" %in% abiertas`
Y / O / NO	`&`, `|`, `!`	`x > 1 & x < 5`
Crear una tabla	`data.frame()`	`data.frame(a = 1:3, b = c("x", "y", "z"))`
Leer un CSV	`read.csv()`	`read.csv("datos/fichero.csv")`
Guardar un CSV	`write.csv()`	`write.csv(df, "salida.csv", row.names = FALSE)`
Dimensiones	`nrow()`, `ncol()`, `dim()`	`nrow(df)`
Ver el principio	`head()`	`head(df, 10)`
Estructura	`str()`	`str(df)`
Una columna	`$`	`df$municipio`
Filtrar filas	`subset()`	`subset(df, municipio == "Getafe")`
Ordenar	`order()`	`df[order(df$fecha), ]`
Valores únicos	`unique()`	`unique(df$municipio)`
Contar valores	`table()`	`table(df$municipio)`
Proporciones	`prop.table()`	`prop.table(table(df$efecto))`
Sumar por grupos	`aggregate()`	`aggregate(prestamos ~ municipio, data = df, FUN = sum)`
Unir tablas	`merge()`	`merge(tabla1, tabla2)`
Duplicados	`duplicated()`	`df[!duplicated(df), ]`
Ausentes	`is.na()`	`sum(is.na(df$prestamos))`
Quitar espacios	`trimws()`	`trimws(" Getafe ")`
Minúsculas	`tolower()`	`tolower("GETAFE")`
Sustituir caracteres	`chartr()`	`chartr("áé", "ae", "Leganés")`
¿Contiene un patrón?	`grepl()`	`grepl("/", "01/03/2025")`
Texto a fecha	`as.Date()`	`as.Date("01/03/2025", format = "%d/%m/%Y")`

Para…	Función	Ejemplo
Secuencia de fechas	`seq()`	`seq(fecha, by = "1 month", length.out = 2)`
Rellenar plantilla	`sprintf()`	`sprintf("Hola, %s", "Getafe")`
Lista a JSON	`jsonlite::toJSON()`	`jsonlite::toJSON(lista, auto_unbox = TRUE, pretty = TRUE)`
Gráfico de barras	`barplot()`	`barplot(table(df$usuario))`
Valores sin repetir	`unique()`	`unique(df$biblioteca)`
Posición del máximo	`which.max()`	`names(which.max(table(df$municipio)))`
Leer de una URL	`url()` + `readLines()`	`readLines(url("https://…"), warn = FALSE)`
JSON a listas	`jsonlite::fromJSON()`	`jsonlite::fromJSON(texto, simplifyVector = FALSE)`

Prueba cualquiera de ellas aquí:

```
df <- read.csv("datos/prestamos_bibliotecas_practica.csv")
head(df)
```

El paquete datareuseR

`datareuseR` reúne los datos y las funciones auxiliares del libro para que repitas las prácticas en tu ordenador, con RStudio o Positron, y las apliques a tus propios datos. Dentro del libro se carga con `source("datos/datareuseR.R")`; en tu ordenador, instálalo una vez y cárgalo con `library()`:

```
install.packages("remotes")
remotes::install_github("pablosancheznunez/reutilizacion-datos", subdir = "datareuseR")
library(datareuseR)
vignette("datareuseR")   # guía rápida
```

Para…	Función	Tema
Leer un CSV con codificación y separador desconocidos	`leer_csv_abierto()`	0, 3
Estrellas del esquema de cinco estrellas	`estrellas()`	0
Plazos en días hábiles, con festivos	`sumar_habiles()`	1
Evaluación FAIR simplificada	`evaluar_fair()`	2
Consultar la API de datos.gob.es (con copia sin conexión)	`consultar_datosgob()`	2
Título, formatos y tabla de los resultados	`titulo_es()`, `formatos_de()`, `tabla_datosgob()`	2, 4
Unificar etiquetas de formato (URI de la UE y tipos MIME)	`normalizar_formato()`	4

Para…	Función	Tema
Celdas vacías y completitud (Vetrò et al.)	`celda_vacia()`, `completitud()`	3
Informe rápido de calidad	`informe_calidad()`	3
Clave de agrupación al estilo de OpenRefine	`clave_agrupacion()`	3
Ingresos y punto de equilibrio *freemium*	`ingresos_mensuales()`, `punto_equilibrio()`	5
Obligaciones al combinar licencias	`obligaciones()`	5

Datos incluidos: `prestamos_practica` (el CSV de práctica, sin limpiar), `bibliotecas_madrid` (copia del 24 de noviembre de 2026, CC BY 4.0, Ayuntamiento de Madrid), `casos_reutilizacion` (los doce casos del Tema 2) y `licencias` (la tabla del Tema 5).

Prueba el paquete aquí:

```
source("datos/datareuseR.R")
datos <- leer_csv_abierto("datos/prestamos_bibliotecas_practica.csv", colClasses =
"character")
informe_calidad(datos)
```

Guía para el profesorado

Cómo encaja el libro con Aula Global

El libro es material de estudio y práctica; la evaluación sigue en los cuestionarios de Aula Global. Cada capítulo prepara la actividad evaluable de su tema:

Capítulo	Actividad de Aula Global	Qué practica el libro
Tema 0	«Clasifica y audita»	Estrellas, Open Definition y auditoría del CSV
Tema 1	«Del derecho de acceso a la reutilización»	Cronología de normas, plazos de la Ley 19/2013, datos de alto valor
Tema 2	«WebQuest de reutilización»	Dimensiones de Safarov et al.; lectura de la ficha ODM 2025
Tema 3	«Laboratorio de calidad con OpenRefine»	Completitud, duplicados, municipios, fechas y total de Getafe en R
Tema 4	«Taller de datos enlazados»	Turtle, JSON-LD y consultas sobre un grafo fijo
Tema 5	«Caso de empresa infomediaria»	Patrones de negocio y obligaciones de licencias

Para enlazarlo desde Aula Global basta un recurso **URL** por tema que apunte al capítulo correspondiente. También puede incrustarse un capítulo en una **Página** con un `<iframe>`, aunque la lectura es más cómoda en pestaña aparte.

El conjunto de datos de práctica

El fichero `datos/prestamos_bibliotecas_practica.csv` de este libro es sintético. Si en Aula Global se usa otra versión del CSV, sustituye el fichero de la carpeta `datos/` por esa versión con el mismo nombre y vuelve a publicar. Las comprobaciones de los ejercicios que dependen del CSV (Primeros pasos, Temas 3, 4 y 5) calculan la respuesta esperada a partir del fichero, así que se adaptan solas, con dos condiciones:

- Las columnas deben llamarse igual (`id`, `biblioteca`, `municipio`, `fecha`, `prestamos`, `usuarios_activos`).
- La tabla de equivalencias de municipios (`forma_correcta`, en los Temas 3, 4 y 5) y la tabla de población ficticia del Tema 5 debe incluir todos los municipios del fichero.

Glosario, descargas y autoevaluación

- **Glosario** (`92-glosario.qmd`): términos agrupados por tema, cada uno con un ancla `#gl-...`. Cada capítulo tiene un recuadro «Glosario del tema» con enlaces a sus términos. Para añadir uno, crea un encabezado `### Término {#gl-termino}` en la sección del tema y enlázalo como `[Término](92-glosario.qmd#gl-termino)`.
- **Descargas** (carpeta `plantillas/`): un script `.R` limpio por capítulo, la plantilla de informe `informe_practica.qmd` y su tema `informe.scss`. Los scripts se generan a partir de los bloques `{webr}` de cada capítulo con `python3 herramientas/extraer_r.py .`; vuelve a ejecutarlo si cambias el código de un capítulo.
- **Autoevaluación**: preguntas con el formato de webexercises, mediante la extensión local `_extensions/webex/`, que genera el mismo marcado sin necesidad de R al renderizar:

```
::: {.webex-check .webex-box}
1. Pregunta de verdadero o falso {{< torf true >}}
2. Desplegable: {{< mcq "*correcta" "incorrecta" "incorrecta" >}}
3. Hueco: {{< fitb "respuesta" "alternativa" ignorecase=true >}}
```

````
4. Opciones en lista:

{{< longmcq "incorrecta" "*correcta" >}}
:::

::: {.webex-solution boton="Ver explicaciones"}
Texto que se despliega con el botón.
:::
````

El asterisco marca la opción correcta. Los huecos no pueden ir dentro de código en línea: escribe el código antes y después, por ejemplo `` `sum(x` {{< fitb ">" >}} `2)` ``.

Figuras y versión en PDF

- **Figuras** (carpeta `figuras/`): una o dos por tema, en SVG, generadas con `python3 herramientas/figuras.py .` (necesita Python con matplotlib). Usan una paleta categórica validada para el daltonismo y llevan texto alternativo en el capítulo. Si cambias el CSV de práctica, vuelve a generarlas.
- **PDF** (`pdf/reutilizacion-datos-abiertos.pdf`): se genera con `python3 herramientas/pdf.py`, que une los capítulos y los compila con Typst, incluido en Quarto, sin LaTeX. El filtro `_extensions/libro-pdf/` adapta el contenido al papel: muestra el código sin ejecutar, convierte pistas y soluciones en recuadros, presenta las preguntas sin corrección automática y transforma los enlaces entre capítulos en enlaces internos. Genera el PDF antes de `quarto render`, para que se copie al sitio.

Orden para publicar una edición nueva:

```
python3 herramientas/extraer_r.py .   # scripts .R de cada capítulo
python3 herramientas/figuras.py .     # figuras
python3 herramientas/moodle.py .      # banco de preguntas para Aula Global
python3 herramientas/pdf.py           # PDF
quarto render                         # sitio web en _book/
```

Banco de preguntas para Aula Global

Las preguntas de autoevaluación de cada capítulo (49 en total) se exportan a **Moodle XML** con `python3 herramientas/moodle.py .`, que deja en la carpeta `moodle/` un fichero por capítulo y otro con todo el banco. Cada pregunta conserva su explicación como retroalimentación general. Para usarlas como cuestionario calificado: *Banco de preguntas* ⟶ *Importar* ⟶ *Moodle XML*.

- Banco completo (Moodle XML)

Tipos: verdadero/falso, opción múltiple con respuesta única y respuesta corta. Las respuestas cortas que aceptan variantes (con y sin tilde, en mayúsculas o minúsculas) las incluyen todas.

Caso práctico y datos reales

El caso práctico usa una copia del conjunto «Bibliotecas de Madrid» del Ayuntamiento de Madrid (CC BY 4.0), descargada el 24 de noviembre de 2026 y sin modificar: `datos/bibliotecas_madrid_2026-11-24.csv`. El portal lo actualiza a diario; para actualizar el caso, descarga una copia nueva, cambia el nombre del fichero en el capítulo y en `herramientas/extraer_r.py`, y vuelve a ejecutar las pruebas: las comprobaciones de los ejercicios calculan la respuesta a partir del fichero.

Accesibilidad, analítica y comprobación automática

- **Accesibilidad.** El libro se ha revisado con axe-core contra WCAG 2.1 AA sin incidencias. El fichero `accesibilidad.html` añade nombres accesibles a los editores de código y a los controles interactivos, y `estilos.scss` ajusta el contraste de pies de figura y citas.

- **Analítica.** El libro carga `/consent.js` de la web: Google Analytics y Microsoft Clarity solo se activan si el visitante acepta las cookies, y la elección es la misma en toda la web. Fuera de pablosancheznunez.com el script no existe y no se carga nada.
- **Comprobación automática.** El flujo `.github/workflows/comprobar.yml` genera el libro en cada cambio y ejecuta en un navegador real los ejercicios de webR, las preguntas de autoevaluación y la auditoría de accesibilidad (`herramientas/pruebas/`). Si algo falla, el cambio aparece con una cruz roja en GitHub.

Retos evaluables

Cada tema termina con un **reto evaluable** pensado para la evaluación continua en Aula Global. A diferencia de los ejercicios, no tiene pista, solución ni corrección automática, y cada estudiante trabaja con datos distintos (un portal, una palabra de búsqueda o un conjunto elegido por él), lo que dificulta copiar. La entrega es un informe hecho con la plantilla `informe_practica.qmd` y se corrige con su rúbrica (pesos 15/25/20/20/10/10).

Tema	Reto	Funciones de `datareuseR`
0	Auditoría de apertura de tres conjuntos de publicadores distintos	`estrellas()`
1	Solicitud de reutilización con cálculo de plazos y festivos reales	`sumar_habiles()`
2	Radiografía de un tema en datos.gob.es con la API en directo	`consultar_datosgob()`, `tabla_datosgob()`
3	Limpieza documentada de un CSV real (≥100 filas)	`leer_csv_abierto()`, `informe_calidad()`, `clave_agrupacion()`
4	Publicación de diez registros como JSON-LD y del conjunto con DCAT-AP	—
5	Propuesta de servicio infomediario con tres escenarios	`obligaciones()`, `punto_equilibrio()`

Las soluciones de referencia y los criterios de corrección de cada reto **no están en el repositorio público**: se entregan aparte al profesorado. Si las guardas en el repositorio, hazlo en una rama o un repositorio privado.

Actividades interactivas y laboratorios

Desde la versión 1.4, cada tema incluye **actividades de corrección inmediata** (ordenar, clasificar y escenarios con decisiones) y **laboratorios** que se usan sin programar: compatibilidad de licencias, comprobador de calidad de un CSV (que también acepta los ficheros del estudiante, sin que salgan de su navegador) y conversión de CSV a JSON y a RDF. En el PDF y el EPUB las actividades aparecen en papel, con su solución. Las actividades están escritas en Markdown con la extensión `interactivos` (`_extensions/interactivos/interactivos.lua`), así que se pueden añadir o cambiar sin programar.

Cada tema tiene además **diapositivas de clase** (objetivos, figuras, tablas principales, lectura, preguntas para debatir y reto) en diapositivas/; se generan con `python3 herramientas/diapositivas.py` después de `quarto render`.

Progreso, uso sin conexión y modo oscuro

- **Progreso del estudiante** (`progreso.html`): cada capítulo muestra cuántos ejercicios ha resuelto y cuántas preguntas ha acertado el estudiante, y la presentación resume todos los capítulos. Se guarda en el almacenamiento local del navegador (`localStorage`), sin cookies ni envío a ningún servidor: el profesorado no lo ve. Sirve de autorregulación, no de evaluación.

- **Uso sin conexión** (`sw.js`): un *service worker* guarda las páginas, webR y sus paquetes la primera vez que se visitan. Las páginas se piden primero a la red y, si no hay conexión, se sirven desde la copia; el resto, al revés. Al publicar una versión nueva, cambia `VERSION` en `sw.js`.
- **Modo claro y oscuro**: el interruptor junto al título, en la barra lateral, cambia entre los dos; por defecto se respeta la preferencia del sistema. El tema oscuro está en `estilos-oscuro.scss` y también se audita con axe-core.
- **PDF y EPUB**: `python3 herramientas/pdf.py` genera los dos formatos; el EPUB lleva portada, metadatos (DOI, licencia) y la misma ficha del libro que el PDF.

El paquete datareuseR

La carpeta `datareuseR/` contiene un paquete de R con los datos y las funciones auxiliares del libro, documentado y con pruebas (`R CMD check --as-cran` sin errores ni avisos). El libro no puede instalarlo dentro de webR, así que `python3 herramientas/datareuser_webr.py` genera `datos/datareuseR.R` con sus funciones, y los capítulos lo cargan con `source()`. Si cambias el paquete, vuelve a ejecutar ese script. Los estudiantes pueden instalarlo con `remotes::install_github("pablosancheznunez/reutilizacion-datos", subdir = "datareuseR")`; para publicarlo en CRAN basta con enviar la carpeta del paquete.

Publicar y actualizar el libro

El libro es un proyecto de Quarto con la extensión quarto-live, que ya va incluida en la carpeta `_extensions/`. Los capítulos usan `engine: markdown`, así que para renderizar no hace falta R instalado: el código se ejecuta en el navegador del estudiante.

```
quarto preview          # vista previa local
quarto render           # genera el sitio en _book/
```

Opciones de publicación:

- **GitHub Pages**: el repositorio incluye `.github/workflows/publicar.yml`, que renderiza y publica el libro en cada *push* a `main`. En *Settings* → *Pages*, elige *GitHub Actions* como origen.
- **Cualquier servidor web**: sube el contenido de `_book/`. Debe servirse por HTTP(S); abrir `index.html` con doble clic no funciona, porque webR necesita un servidor.

Mantenimiento

- **Datos que cambian cada año**: el Tema 2 remite a la ficha del Open Data Maturity 2025 y el Tema 5 al informe de ASEDIE. El libro no incluye esas cifras, solo los enlaces; basta con actualizar los enlaces cuando salga una nueva edición.
- **DCAT-AP-ES y NTI-RISP**: antes de cada curso, comprueba si la NTI modificada se ha publicado en el BOE y ajusta el párrafo del Tema 3.
- **Versión de webR**: el libro sirve webR 0.6.0 desde su propia carpeta `webr/` y los paquetes desde `webr-repo/`, sin depender de servidores externos. Para actualizarlo, sustituye esas carpetas por la nueva distribución de webR y los binarios correspondientes, y cambia `VERSION` en `sw.js` para que los navegadores renueven su copia.
- **Fuentes externas**: el flujo `.github/workflows/semanal.yml` comprueba cada lunes los enlaces del libro, el CSV de bibliotecas de Madrid y la API de datos.gob.es, y comprueba el paquete `datareuseR` con la consulta en directo. Si algo falla, abre una incidencia en GitHub con el informe; también puedes ejecutarlo a mano con `python3 herramientas/comprobar_fuentes.py`.

Añadir un ejercicio

Un ejercicio autocorregido tiene hasta cuatro bloques con la misma etiqueta `exercise`:

````
```{webr}
#| exercise: mi_ejercicio
```
````

````
#| setup: true
datos <- read.csv("datos/prestamos_bibliotecas_practica.csv")
```

```{webr}
#| exercise: mi_ejercicio
nrow(______)
```

::: {.solucion-ejercicio exercise="mi_ejercicio"}
`nrow(datos)`
:::

```{webr}
#| exercise: mi_ejercicio
#| check: true
if (identical(.result, nrow(read.csv("datos/prestamos_bibliotecas_practica.csv")))) {
  list(correct = TRUE, message = "Correcto")
} else {
  list(correct = FALSE, message = "Revisa el objeto que cuentas")
}
```
````

En el bloque `check`, `.result` es el valor de la última línea del código del estudiante y `.user_code` es el texto del código. La documentación de quarto-live explica más opciones.
````

Glosario

Términos técnicos del libro, agrupados por tema. Cada capítulo enlaza aquí desde su recuadro «Glosario del tema», y desde cada término puedes volver al capítulo donde se explica. Usa el buscador del libro (arriba a la izquierda) para localizar cualquier término.

Tema 0. Datos abiertos

Datos abiertos
Datos que cualquiera puede usar, modificar y compartir libremente para cualquier fin, con como mucho la obligación de citar la fuente o de compartir en las mismas condiciones. → Tema 0

Open Definition
Definición de «abierto» de la Open Knowledge Foundation (versión 2.1). Exige licencia abierta, acceso completo y formato legible por máquina y abierto. → Tema 0

Datos públicos
Información producida o custodiada por el sector público. No toda es abierta: puede incluir datos personales o protegidos. → Tema 0

Gobierno abierto
Forma de gobernar basada en la transparencia, la participación y la colaboración. Los datos abiertos son una de sus herramientas. → Tema 0

Modelo de cinco estrellas
Escala acumulativa de Tim Berners-Lee para valorar la publicación de datos: de un documento con licencia abierta (★) a datos enlazados (★★★★★). → Tema 0

Principios FAIR
Datos localizables (*Findable*), accesibles (*Accessible*), interoperables (*Interoperable*) y reutilizables (*Reusable*). No equivale a «abierto». → Tema 0

Formato abierto
Formato con especificación pública que cualquiera puede implementar sin pagar ni pedir permiso (CSV, JSON, XML, ODS). Su contrario es el **formato propietario**. → Tema 0

Legible por máquina
Formato del que un programa puede extraer los datos sin intervención humana. Un PDF escaneado no lo es; un CSV sí. → Tema 0

Licencia abierta
Licencia que permite usar, redistribuir y modificar para cualquier fin. Puede exigir atribución (CC BY) o compartir igual (CC BY-SA, ODbL), pero no prohibir el uso comercial (NC) ni las obras derivadas (ND). → Tema 0

Web Semántica
Extensión de la Web en la que los datos tienen un significado explícito y enlazado que las máquinas pueden procesar. → Tema 0, Tema 4

Tema 1. Reutilización de la información del sector público

Reutilización
Uso de documentos del sector público con fines comerciales o no comerciales distintos del propósito inicial para el que se produjeron (Ley 37/2007). → Tema 1

RISP
Reutilización de la información del sector público. Nombre del marco normativo europeo y español sobre la materia. → Tema 1

Derecho de acceso
Derecho de cualquier persona a solicitar y conocer la información pública (Ley 19/2013). Es previo a la reutilización, pero no la garantiza. → Tema 1

Publicidad activa
Información que las administraciones deben publicar sin que nadie la pida: organización, contratos, subvenciones, presupuestos. → Tema 1

Silencio administrativo negativo
Si la administración no resuelve en plazo, la solicitud se entiende desestimada. Rige tanto en el derecho de acceso como en las solicitudes de reutilización. → Tema 1

Licencia tipo
Texto de licencia estándar, igual para todos los reutilizadores, publicado en el aviso legal de un portal. → Tema 1

Conjuntos de datos de alto valor
Datos con gran beneficio social y económico, que deben ofrecerse gratis, por API y con licencia abierta. Seis categorías: geoespaciales, observación de la Tierra y medio ambiente, meteorológicos, estadísticos, sociedades y movilidad. → Tema 1

Datos dinámicos
Datos que se actualizan con frecuencia o en tiempo real, como los de tráfico o sensores. Deben ofrecerse mediante API. → Tema 1

Transposición
Incorporación de una directiva europea al derecho de un Estado miembro mediante una norma propia. Los reglamentos, en cambio, se aplican directamente. → Tema 1

Agente infomediario
Reutilizador que crea productos o servicios de valor añadido a partir de información, sobre todo pública. → Tema 1, Tema 5

Tema 2. Proyectos de reutilización

Portal de datos abiertos
Sitio web que cataloga y da acceso a conjuntos de datos abiertos, como datos.gob.es o data.europa.eu. → Tema 2

Federación de catálogos
Recolección automática de los metadatos de varios portales en un catálogo común. datos.gob.es federa portales estatales, autonómicos, locales y universitarios. → Tema 2

Open Data Maturity
Informe anual que compara la madurez en datos abiertos de 36 países europeos en cuatro dimensiones: políticas, impacto, portal y calidad. → Tema 2

Caso de uso
Proyecto concreto que reutiliza datos. Conviene describirlo con las dimensiones de Safarov et al.: tipo de utilización, condiciones, efectos y usuarios. → Tema 2

B2G
Business-to-government: intercambio de datos de empresas con administraciones para fines de interés general. → Tema 2

Altruismo de datos
Cesión voluntaria de datos por ciudadanos o empresas para fines de interés general, regulada por el Reglamento de Gobernanza de Datos. → Tema 2

Ciencia abierta
Práctica de la investigación que hace accesibles publicaciones, datos, código y métodos para que puedan comprobarse y reutilizarse. → Tema 2

Repositorio de datos
Servicio que deposita, describe, preserva y da acceso a conjuntos de datos con un identificador persistente (Zenodo, e-Archivo). re3data es el registro mundial de repositorios. → Tema 2

Data paper
Artículo científico que describe un conjunto de datos —origen, método y estructura— para facilitar su reutilización. → Tema 2

API
Interfaz de programación de aplicaciones: una dirección web que devuelve datos estructurados (normalmente JSON o XML) para que los use un programa, no una persona. datos.gob.es ofrece una para consultar su catálogo. → Tema 2

Tema 3. Gestión documental y *open data*

Gestión documental
Conjunto de procesos para controlar la creación, clasificación, valoración, conservación y eliminación de documentos (ISO 15489). → Tema 3

Ciclo de vida de los datos
Secuencia de fases por las que pasan los datos, desde su creación hasta su reutilización, preservación o eliminación (modelo del DCC). → Tema 3

Metadatos
Datos que describen otros datos: qué son, quién los publica, con qué licencia, cómo están estructurados y cómo se conservan. → Tema 3

DCAT
Vocabulario del W3C para describir catálogos, conjuntos de datos (`dcat:Dataset`) y distribuciones (`dcat:Distribution`). DCAT-AP y DCAT-AP-ES son sus perfiles europeo y español. → Tema 3

Distribución
Cada forma concreta de obtener un conjunto de datos: un fichero CSV, un JSON, una API. → Tema 3

Completitud
Característica de calidad que mide la proporción de celdas o filas sin valores ausentes. → Tema 3

Conformidad
Característica de calidad que mide si los valores siguen el formato esperado (fechas, códigos, unidades). → Tema 3

Exactitud
Característica de calidad que mide si los valores son correctos y plausibles. Un número de préstamos negativo es un error de exactitud. → Tema 3

Valor ausente
Celda sin dato. En R se representa con `NA`; en los ficheros aparece de muchas formas («», «n/d», «-»), lo que conviene unificar. → Tema 3

Suma de control (*checksum*)
Huella numérica de un fichero (MD5, SHA-256) que permite comprobar que no ha cambiado ni se ha corrompido. → Tema 3

OAIS
Modelo de referencia para archivos digitales que deben preservar información a largo plazo (ISO 14721). → Tema 3

Identificador persistente
Identificador que no cambia aunque cambie la ubicación del recurso, como un DOI, un ORCID o un Handle. → Tema 3

Tema 4. Normalización e interoperabilidad

Estándar abierto
Especificación pública, desarrollada en un proceso abierto e implementable sin pagar derechos. → Tema 4

Interoperabilidad
Capacidad de sistemas y organizaciones para intercambiar información y usarla. El EIF distingue cuatro niveles: jurídico, organizativo, semántico y técnico. → Tema 4

ENI
Esquema Nacional de Interoperabilidad: criterios que deben seguir las administraciones españolas para que sus sistemas interoperen. Se desarrolla en Normas Técnicas de Interoperabilidad (NTI). → Tema 4

URI
Identificador uniforme de recursos: cadena que identifica algo de forma única en la Web, como `http://www.wikidata.org/entity/Q8802`. → Tema 4

RDF
Modelo del W3C que representa la información como tripletas sujeto–predicado–objeto. Se escribe en sintaxis como Turtle, RDF/XML o JSON-LD. → Tema 4

Tripleta
Unidad mínima de RDF: un sujeto, un predicado y un objeto, por ejemplo «biblioteca — está en — Getafe». → Tema 4

Turtle
Sintaxis compacta y legible de RDF, con prefijos para abreviar las URI. → Tema 4

JSON-LD
JSON con un `@context` que asigna a cada clave una URI de un vocabulario: RDF con aspecto de JSON. → Tema 4

Linked Open Data
Datos enlazados publicados con licencia abierta: las cinco estrellas del modelo de Berners-Lee. → Tema 4

SPARQL
Lenguaje de consulta del W3C para datos RDF, basado en patrones de tripletas. → Tema 4

Vocabulario
Conjunto de clases y propiedades con URI para describir datos (schema.org, Dublin Core, DCAT). → Tema 4

Ontología
Vocabulario con una semántica formal más rica (jerarquías, restricciones, reglas de inferencia), expresado en RDFS u OWL. → Tema 4

SKOS
Vocabulario del W3C para publicar tesauros y clasificaciones en la Web, con relaciones como `skos:broader` y `skos:narrower`. → Tema 4

Tipo MIME
Etiqueta normalizada que indica el formato de un fichero en internet, como `text/csv` o `application/json`. Convive en los catálogos con el vocabulario de tipos de fichero de la UE, lo que obliga a normalizar. → Tema 2, Tema 4

Tema 5. Infomediarios y negocio

Sector infomediario
Conjunto de empresas que crean productos y servicios de valor añadido a partir de información, en buena parte pública. → Tema 5

Marco 6-V
Seis componentes para describir cómo un negocio crea y captura valor con datos abiertos: propuesta de valor, proceso de creación, valor obtenido, captura, gestión y red de valor (Ahmadi Zeleti et al.). → Tema 5

Freemium
Modelo de negocio con un servicio básico gratuito y funciones avanzadas de pago. → Tema 5

Punto de equilibrio
Volumen de actividad a partir del cual los ingresos cubren los costes. → Tema 5

Servicio de valor añadido
Producto que transforma los datos en algo más útil: agregación, limpieza, enriquecimiento, análisis, visualización o API. → Tema 5

Compartir igual (*share-alike*)
Cláusula de licencia que obliga a distribuir las obras o bases de datos derivadas con la misma licencia (CC BY-SA, ODbL). → Tema 5

Minería de textos y datos
Análisis automatizado de textos y datos para extraer patrones. La Directiva (UE) 2019/790 regula sus excepciones de derechos de autor. → Tema 5

Espacio de datos
Entorno de confianza para compartir datos, también privados, entre organizaciones de un sector con reglas comunes. → Tema 5

Datos sintéticos
Datos generados artificialmente que imitan la estructura de los reales sin corresponder a casos reales. El CSV de este libro es un ejemplo. → Tema 5

Herramientas

R
Lenguaje y entorno libre para análisis estadístico y de datos. → Primeros pasos

webR
Versión de R compilada a WebAssembly que se ejecuta en el navegador, sin instalación. → Presentación

Quarto
Sistema de publicación científica de código abierto que combina texto y código para generar libros, artículos, presentaciones y webs. → Presentación

Data frame
Tabla de datos en R: cada columna es un vector del mismo tamaño. → Primeros pasos

OpenRefine
Herramienta libre para limpiar y transformar datos, con agrupación (*clustering*) de valores similares. → Tema 3

Paquete de R
Conjunto de funciones, datos y documentación de R que se instala y se carga con `library()`. El paquete `datareuseR` reúne las funciones y los datos de este libro. → Apéndice A

Service worker
Programa que el navegador ejecuta en segundo plano para una web. En este libro guarda las páginas y el motor de R para que funcionen sin conexión después de la primera visita. → Presentación

Referencias

Bibliografía

Abella, A., Ortiz-de-Urbina-Criado, M., y De-Pablos-Heredero, C. (2017). A model for the analysis of data--driven innovation and value generation in smart cities' ecosystems. *Cities*, *64*, 47-53. https://doi.org/10.1016/j.cities.2017.01.011

Attard, J., Orlandi, F., Scerri, S., y Auer, S. (2015). A systematic review of open government data initiatives. *Government Information Quarterly*, *32*(4), 399-418. https://doi.org/10.1016/j.giq.2015.07.006

Batini, C., Cappiello, C., Francalanci, C., y Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3). https://doi.org/10.1145/1541880.1541883

Hallo, M., Luján-Mora, S., Maté, A., y Trujillo, J. (2016). Current state of Linked Data in digital libraries. *Journal of Information Science*, *42*(2), 117-127. https://doi.org/10.1177/0165551515594729

Janssen, K. (2011). The influence of the PSI directive on open government data: An overview of recent developments. *Government Information Quarterly*, *28*(4), 446-456. https://doi.org/10.1016/j.giq.2011.01.004

Jetzek, T., Avital, M., y Bjørn-Andersen, N. (2014). Data-driven innovation through open government data. *Journal of Theoretical and Applied Electronic Commerce Research*, *9*(2). https://doi.org/10.4067/S0718--18762014000200008

Magalhaes, G., y Roseira, C. (2020). Open government data and the private sector: An empirical view on business models and value creation. *Government Information Quarterly*, *37*(3), 101248. https://doi.org/10.1016/j.giq.2017.08.004

Ruijer, E., Grimmelikhuijsen, S., y Meijer, A. (2017). Open data for democracy: Developing a theoretical framework for open data use. *Government Information Quarterly*, *34*(1), 45-52. https://doi.org/10.1016/j.giq.2017.01.001

Umbrich, J., Neumaier, S., y Polleres, A. (2015,). *Quality assessment and evolution of open data portals*. https://doi.org/10.1109/FiCloud.2015.82

Zuiderwijk, A., Janssen, M., y Dwivedi, Y. K. (2015). Acceptance and use predictors of open data technologies. *Government Information Quarterly*, *32*(4), 429-440. https://doi.org/10.1016/j.giq.2015.09.005

Zuiderwijk, A., y Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, *31*(1), 17-29. https://doi.org/10.1016/j.giq.2013.04.003