Treballar amb dades: un repte complex i apassionant

Manuel Terrádez Gurrea

PID_00235997

Temps mínim previst de lectura i comprensió: **4 hores**

UOC
Universitat Oberta de Catalunya

Índex

Introducció

Vivim en una època en què la gestió de la informació i les dades és cada vegada més important. Els clàssics projectes d'anàlisi estadística de dades s'han sofisticat fins al punt que cal una revisió de totes les fases del projecte, des de la recopilació i l'emmagatzematge de la informació fins a la visualització dels resultats, passant per les tècniques de tractament i anàlisi. I és que el volum, la tipologia i la rapidesa amb la qual es generen les dades requereixen un nou paradigma.

Per això en aquest mòdul tractarem alguns dels conceptes que provoquen els principals debats en les publicacions especialitzades de l'àmbit, així com també en els cursos i congressos relacionats i en els articles divulgatius dels quals es fan ressò la majoria de mitjans de comunicació.

El contingut d'aquest mòdul s'estructura en tres grans blocs. En primer lloc, parlarem sobre el repte que representa, per a les persones i les organitzacions que treballen amb dades, haver d'afrontar el tractament i l'explotació de dades massives. Per descomptat, parlarem amb profunditat del concepte de *big data* (dades massives) i les aplicacions associades, si bé també dedicarem sengles apartats al cicle de vida de les dades, i a com les organitzacions tracten el problema del govern de les dades.

Posteriorment, dedicarem un ampli espai a les noves tendències en l'àmbit de la visualització de dades i les gairebé omnipresents infografies: parlarem dels principals tipus i característiques, sobretot mitjançant l'exposició de diversos exemples.

Finalment, desenvoluparem un apartat sobre les publicacions científiques com a mitjà de divulgació i difusió dels resultats de recerca.

Objectius

Després de la lectura d'aquest mòdul l'estudiant podrà:

1. Entendre el concepte de *big data* i les seves principals aplicacions, així com també altres conceptes associats.

2. Comprendre les principals fases del cicle de vida de les dades i com es relacionen.

3. Conèixer com les organitzacions que gestionen informació estan tractant el govern de les dades.

4. Iniciar-se en les principals tècniques i eines de visualització de dades.

5. Aprendre a fer difusió dels resultats de les recerques científiques de diverses formes.

1. Plantejament inicial

Per a començar, recordarem que un **estudi quantitatiu** és el que utilitza la recollida de dades, el mesurament numèric i l'anàlisi estadística per a demostrar hipòtesis o establir-ne patrons de comportament.

Referència bibliogràfica

Es pot trobar una definició més completa d'estudi quantitatiu a:

Roberto Hernández; Carlos Fernández; Pilar Baptista (2010). *Metodología de la investigación*. Mèxic: McGraw-Hill.

Les diverses teories sobre la planificació d'un projecte acaben amb una fase d'explotació del model que inclou la publicació i difusió dels resultats. Per tant, sembla obvi que un estudi quantitatiu no pot finalitzar amb l'obtenció d'un model «de laboratori», el coneixement del qual quedi en l'àmbit de l'investigador que l'ha dut a terme o només del seu entorn, sinó que ha d'haver-hi una última fase dins del projecte d'investigació que consisteixi en la difusió i explotació dels resultats de l'estudi, així com també en la seva representació i visualització mitjançant gràfics clàssics o tècniques modernes com, per exemple, les infografies.

Així mateix, els nous temps obliguen a replantejar-se els clàssics estudis basats en dades, ja que el ritme de generació de nova informació és tal que sovint l'explotació de grans volums de dades (l'anomenat *big data*) suposa un important repte. A més, en les organitzacions que treballen amb dades sorgeixen nous conceptes, noves tècniques i fins i tot nous rols, amb la qual cosa és important no perdre de vista el cicle de vida de les dades.

Per això aquest mòdul té tres blocs diferenciats:

- Explotació de grans volums de dades i els seus reptes.
  - *Big data* i les seves implicacions.
  - Cicle de vida de les dades.
  - Govern de les dades.

- Representació i visualització de dades.
- Difusió i publicació dels resultats d'un estudi quantitatiu.

2. Explotació de grans volums de dades i els seus reptes

2.1. *Big data* i les seves implicacions

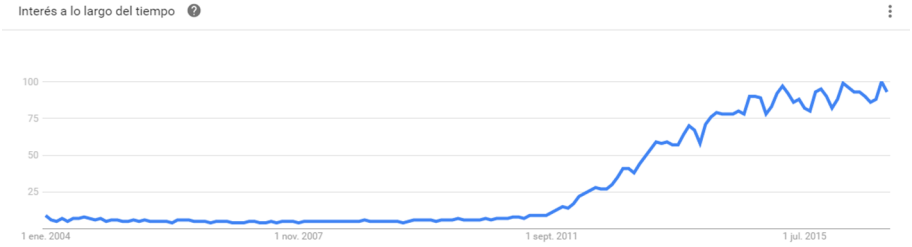
Big data (dades massives) és sens dubte l'expressió de moda en els últims anys en l'àmbit de la informació i les dades, i fins i tot podríem dir que en el món tecnològic en general. Sembla que gairebé qualsevol curs, llibre o article que contingui la paraula *data* ha de tenir necessàriament l'adjectiu *big* al davant per a no semblar antic o passat de moda.

La majoria d'universitats ofereixen titulacions relacionades d'una manera o una altra amb *big data*, i no són poques les empreses que diuen estar immerses en projectes de *big data*. Ara bé, què hi ha de cert i de bombolla en tot això?

El terme es va popularitzar quan van proliferar notícies que es feien ressò que en els últims anys es generava més informació que la generada en tota la història anterior.

El següent gràfic mostra l'evolució en els últims anys de les cerques a Google relacionades amb *big data*. Es pot comprovar que és un terme recent, pràcticament sorgit l'any 2011 (si bé hi ha múltiples referències i antecedents anteriors, com es pot comprovar en l'interessant article «Historia cronológica del Big Data»), però que en poc temps ha tingut un increment considerable, i s'ha situat en aquests moments com una de les «estrelles» en l'àmbit de les dades i la informació per davant d'altres que van tenir un gran èxit anteriorment com, per exemple, *data mining* o *data science*.

Gràfic 1: Evolució de les cerques a Google relacionades amb *big data* en el període 2004-2016



Font: dades de *Google Trends*.

Un exemple seria que fins i tot qui fou president dels EUA, Barack Obama, va parlar (febrer de 2015) sobre aquests termes amb motiu de la presentació del Dr. D. J. Patil com el primer *Chief Data Scientist and Deputy Chief Technology Officer for Data Policy* del govern, o que dues de les principals i més prestigioses revistes de divulgació científica (*Nature* i *Science*) hi han dedicat sengles

monogràfics: monogràfic sobre *big data* a *Nature* i monogràfic sobre *big data* a *Science*; també *The Economist* va dedicar un informe especial, «A special report on managing information», a tractar aquests temes.

2.1.1. Definició

En realitat, abans de res, hauríem de començar pel principi: què és *big data*? Tothom en parla, però pocs defineixen el concepte amb claredat per a saber de què estem parlant perquè no hi ha una definició comunament acceptada del terme.

Disposem d'una bona introducció al concepte i les definicions principals a *Introducción al big data* (Gómez i Conesa) i en el següent article: «Big Data: la próxima "gran cosa" en la gestión de la información».

Lectures recomanades

José Luis Gómez; Jordi Conesa (2015). *Introducción al big data*. Barcelona: UOC.

Julio Alonso; Marta Vázquez (2016, juny). «Big Data: la próxima "gran cosa" en la gestión de la información». *BiD: textos universitaris de biblioteconomia i documentació* (núm. 36).

D'acord amb un estudi dut a terme l'any 2012 per l'IBM Institute for Business Value i la SAID Business School de la Universitat d'Oxford, el terme genera confusió perquè «s'ha utilitzat per a traslladar al públic tot tipus de conceptes entre els quals s'inclouen grans quantitats de dades, analítica de xarxes socials, eines d'última generació per a gestionar les dades, dades en temps real i molt més» (Schroeck i altres, 2012).

No obstant això, per a la definició es recorre a les tres *V* que tradicionalment s'han associat al *big data*:

- **Volum**: la gran quantitat de dades és una de les característiques principals del terme, si bé aquesta quantitat pot variar en funció del sector.

- **Varietat**: la diversitat de tipus i formats de les dades, estructurades o no, suposa un gran repte per la complexitat d'integrar-les i gestionar-les.

- **Velocitat**: la immediatesa de les dades, moltes vegades generades en temps real, fa que es redueixi el temps entre la seva creació i l'accessibilitat.

A les quals s'afegeix una quarta:

- **Veracitat**: moltes dades estan subjectes a un alt grau d'incertesa, que cal tenir en compte, ja que de vegades és difícil o directament impossible evitar-la malgrat les pràctiques de neteja i preprocessament de la informació.

No falten, òbviament, els qui afegeixen noves *V* a la llista com, per exemple, la *visualització* (en dediquem un apartat més endavant) o el *valor*.

Un altre estudi del McKinsey Global Institute simplifica la definició de la següent forma:

> «*Big data* fa referència a conjunts de dades la grandària de les quals està per sobre de la capacitat de les eines i bases de dades tradicionals per a capturar-les, emmagatzemar-les, gestionar-les i analitzar-les.»
>
> J. Manyika; M. Chui; B. Brown; J. Bughin; R. Dobbs; C. Roxburgh; A. H. Byers (2011). *Big data: The next frontier for innovation, competition, and productivity*. Nova York: McKinsey Global Institute.

Per tant, és una definició «dinàmica» i que pot variar segons el sector.

2.1.2. Aplicacions

D'acord amb les conclusions de l'estudi d'IBM citat, el *big data* ha deixat d'estar limitat al món de la tecnologia per a ser una prioritat empresarial en la majoria de sectors, si bé també es reconeix que l'àmplia cobertura mediàtica que rep no permet distingir amb claredat el mite de la realitat, i que resulta difícil trobar informació exhaustiva sobre el que les empreses fan realment en aquest àmbit.

De l'estudi es dedueixen cinc recomanacions clau a l'hora d'obtenir el màxim valor de les iniciatives de *big data*:

- **Dedicar els esforços inicials a obtenir resultats del client**. *Big data* proporciona la capacitat per a comprendre i predir millor els comportaments dels clients i, en conseqüència, millorar-ne l'experiència.

- **Desenvolupar un pla de *big data* per a tota l'organització**. La integració i l'emmagatzematge de les dades són reptes clau en la part tecnològica.

- **Començar amb dades ja existents per a aconseguir resultats a curt termini**. En aquest sentit es destaca com a conclusió interessant i una mica sorprenent l'impacte relativament petit de les dades procedents de les xarxes socials, ja que les empreses utilitzen en el seu lloc fonts internes de dades (transaccions, registres, etc.).

- **Desenvolupar capacitats analítiques sobre la base de prioritats de negoci**. Consulta i generació d'informes, visualització de dades, models predictius, etc.

- **Crear un cas de negoci sobre la base de resultats quantificables**. S'identifiquen quatre fases: educar (crear una base de coneixement), explorar (definir el cas de negoci i el full de ruta), interactuar (fase de proves) i executar (fase d'implementació).

Quant a l'estudi de McKinsey, es destaca que malgrat que per a molts ciutadans l'explosió de les dades es percep gairebé exclusivament com una intrusió en la privadesa, les dades poden generar un considerable valor per a l'economia mundial, incloent-hi el sector públic i, per tant, la ciutadania. Ara bé, cal que les organitzacions afrontin importants reptes: entre ells, el principal és l'escassetat de talent, almenys a curt termini; si bé no és l'únic, ja que l'ús de les infraestructures adequades o la salvaguarda de la seguretat i la privadesa són també molt rellevants.

De l'estudi es dedueixen set principals aspectes a tenir en compte a l'hora d'obtenir valor dels projectes de *big data*:

- Abast global, per a tots els sectors industrials.

- Varietat de possibilitats d'obtenció de valor, entre elles:
  - Major transparència.
  - Millora dels resultats en obtenir dades més precises.
  - Millora de la segmentació de col·lectius.
  - Millora de la presa de decisions i minimització de riscos en reemplaçar el factor humà per algorismes, en certs àmbits.
  - Creació de nous productes i serveis, transformació de sectors o fins i tot creació de noves formes de negoci.

- Factor clau per a superar la competència.

- Noves possibilitats de millora de la productivitat i satisfacció del client, en dissenyar productes que encaixen millor amb les necessitats de l'usuari.

- Identificació de sectors amb major potencial de millora.
  - D'acord amb un indicador format per cinc dimensions (quantitat de dades disponibles, variabilitat de rendiment, nombre de parts interessades, intensitat de transaccions i turbulència del sector) s'identifiquen com a sectors amb major potencial els de la informació i la tecnologia, seguits per les finances i el govern.

- Escassetat de talent necessari, que a més és difícil de produir.

- Necessitat de superar determinats obstacles, especialment els relatius a privadesa, seguretat i propietat intel·lectual.

Com podem observar, hi ha diversos aspectes recurrents i coincidents en els estudis citats i en altres de similars que s'han realitzat recentment.

D'exemples d'aplicacions concretes, n'hi ha moltíssims, però en destaquem els següents pel seu interès i actualitat:

- «Humanizar los datos». Una aplicació en els sondejos d'opinió i les enquestes electorals.

- «Un tuit...,¿un voto?». Similar a l'anterior, però afegint la relació amb les xarxes socials.

- «Sabemos (antes que tú) qué harás el próximo verano». Predicció amb *big data*. Sobre la intel·ligència artificial i l'aplicació en models predictius.

- «Big data y smartcities: así se obtienen los datos de las ciudades». Les ciutats intel·ligents i internet de les coses, dues de les principals aplicacions del *big data*.

O també aquestes recopilacions força completes:

- «How companies use big data». Interessant anàlisi amb casos pràctics sobre com empreses capdavanteres a nivell mundial estan transformant els seus sectors mitjançant l'ús de la gestió de les dades.

- «Eres un dato y las empresas te quieren». Valor de les dades personals. Aplicacions actuals i futures en diversos sectors, i com tenen en compte les dades personals.

I sobre altres implicacions i connexions amb conceptes relacionats:

L'aspecte humà

Una gran part de les dades són generades per les persones, i arran d'això sorgeixen debats sobre com utilitzen les empreses les nostres dades, i fins i tot sobre com nosaltres podem utilitzar-les:

- «The data-driven life». Excel·lent article del *New York Times* que analitza l'auge del mesurament de les dades personals, associant-lo a quatre principals causes: la millora en grandària i qualitat dels sensors, la proliferació dels dispositius mòbils, la tendència a compartir-ho tot en les xarxes socials, i l'aparició del «núvol» com a forma d'emmagatzematge.

- «My data (human side)». Sobre l'ús i aprofitament de les dades personals.

- «Big data. Nuestros datos en la red». Les nostres dades a la xarxa. Com s'utilitzen les dades que generem en les xarxes socials o aplicacions mòbils.

- «Cuantificarse para vivir a través de los datos: los datos masivos (big data) aplicados al ámbito personal». Anàlisi de les perspectives social i tecnològica de fenòmens com el *lifelogging* i el *quantified self*.

Ètica i privadesa

La seguretat de les dades, i especialment la seva privadesa, constitueix una part fonamental de la gestió de la informació. A més, les normatives sobre privadesa estan en constant evolució i poden variar molt segons el país:

- «Ética y big data». Reflexió sobre els principals conflictes ètics associats a l'ús de les dades massives.

- «Facebook y Apple podrán tener el control que la KGB nunca tuvo sobre los ciudadanos». Control sobre els ciutadans. Opinions sobre el possible mal ús de les dades per part d'algunes empreses.

Open data

Un altre àmbit relacionat amb el *big data* és el de les «dades obertes», és a dir, com les entitats públiques, generalment (i excepcionalment algunes de privades), ofereixen accés a les seves dades de forma universal i gratuïta:

- *Big data*, *open data*...: «*Open government*, *open data*, *big data* i transparència: la informació com a nexe d'unió». Conceptes bàsics de la relació entre ambdós termes.

- *Open data* per a combatre el frau: «Hervé Falciani: "El fraude fiscal puede ser detectado desde la matemática sencilla"». Una aplicació directa de l'*open data*.

- El descodificador de Nova York: «Amen ra Mashariki, el descodificador de Nueva York». Com Nova York gestiona les seves dades obertes.

Finalment, recollim altres articles interessants relacionats amb el tema:

- «El big data, ¿nos viene aún grande?». Reflexió sobre la importància que les dades siguin correctes més que sobre el seu volum.

- «You need the right data». Similar a l'anterior.

- «El Dark Data, el lado oscuro del Big Data». El costat fosc. Sobre el desaprofitament d'algunes de les dades emmagatzemades.

- «La nueva y valiosa vida de los "datos basura"». Dades brossa. Similar a l'anterior.

- «Los conocimientos de Big Data, Internet of Things y Predictive Analytics, los más demandados en los portales de empleo». *Big data* i mercat de treball. Nous perfils professionals associats al món de la gestió de les dades.

2.2. El cicle de vida de les dades

Les dades, des del seu origen fins a l'ús final, passen per diferents fases en el cicle de vida, i s'identifiquen habitualment les set següents:

- Fase 1. Creació/identificació.
- Fase 2. Recollida/captura.
- Fase 3. Emmagatzematge/conservació.
- Fase 4. Tractament/preprocessament.
- Fase 5. Anàlisi.
- Fase 6. Visualització.
- Fase 7. Publicació/difusió.

Tot i que acostumen a presentar-se en un ordre, cadascuna de les fases pot aparèixer en diverses seqüències i en diversos moments del projecte, i algunes poden repetir-se més que d'altres.

Entendre el cicle de vida de les dades és fonamental tant per als usuaris com per a la resta de col·laboradors amb els quals s'interactuï en qualsevol de les fases: informàtics, investigadors, enquestadors, enquestats, mitjans de comunicació, etc.

Una adequada gestió de les dades i el seu cicle de vida redunda en una major facilitat per a localitzar, utilitzar, analitzar, compartir i reutilitzar les dades, i, en definitiva, en uns resultats de major qualitat. A continuació descriurem cadascuna de les fases, estenent-nos més en les tres primeres.

No l'hem citada entre les fases identificades, tot i que alguns autors sí que la inclouen, però un pas previ és el de la planificació, que és diferent de la planificació del projecte i que consisteix a dissenyar el procés i els recursos implicats en tot el cicle. Es pot començar pels objectius del projecte i a partir d'ells definir el **pla de gestió de dades**. Per motius obvis és important que aquesta fase s'encari a l'inici del projecte, si bé es pot revisitar i retocar posteriorment; consisteix, entre altres coses, a definir rols i assignar responsabilitats, tenir en compte els costos esperats i el pressupost disponible, identificar les persones i grups que interactuaran amb les dades, definir els resultats esperats i el model de dades, etc. Totes aquestes decisions requereixen un coneixement dels objectius del projecte i una familiaritat amb les dades i com es generen. És a dir, no hi ha «receptes màgiques» que funcionin per a tot tipus de dades, i és important comptar amb persones expertes en l'àmbit del projecte.

Fase 1. Creació/identificació

Les formes clàssiques de generar dades susceptibles de ser analitzades amb posterioritat són múltiples, per exemple:

- Es generen dades mitjançant l'elaboració d'enquestes i sondejos predissenyats o el simple emplenament de formularis de recollida d'informació a l'hora de realitzar una gestió.

- Gairebé tots els processos industrials generen multitud de dades que majoritàriament són monitoritzades per màquines que enregistren aquesta informació.

- Els elements naturals (pluja, temperatura, etc.) són una font inesgotable de dades contínues.

- Les transaccions comercials o financeres generen informació de tipologies diverses.

- Els esdeveniments culturals o esportius són una font d'informació d'assistència, venda d'entrades, despeses de tot tipus, etc.

- Els diversos registres públics (educatius, mèdics, de gestió municipal o regional, etc.) generen informació de diversa índole.

A més, en un món digital com el que vivim, cada acció que duu a terme un usuari, o cada servei que és invocat, deixa un rastre en forma de dades que són susceptibles de ser analitzades posteriorment per a ser convertides en coneixement. L'origen de la tipologia de les dades és molt divers, i va des de simples indicadors binaris relatius a una acció concreta, localitzats en el temps i espai (per exemple, succeeix/no succeeix un esdeveniment), fins a escenaris complexos on milers d'usuaris, serveis i recursos interactuen entre ells (per exemple, clients d'un supermercat que observen els productes exposats a la venda).

La recent proliferació de sensors en edificis i altres zones públiques de les ciutats provoca la generació d'una quantitat ingent de dades, l'ús de les quals està relacionat amb àmbits com l'IoT (*Internet of Things* o Internet de les coses) i les *smart cities* (ciutats intel·ligents): «Eres un dato y las empresas te quieren».

Fase 2. Recollida/captura

Per analogia amb l'apartat anterior, podríem parlar de les formes o suports clàssics de recollida d'informació, per exemple:

- La recollida de dades d'enquestes pot ser en format paper o electrònic, i generalment després és capturada i processada en fitxers amb formats predefinits.

- La informació generada per processos industrials habitualment és recollida pels sistemes que creen les dades.

- Les dades dels fenòmens naturals se solen capturar per mitjà de dispositius específics (termòmetres, pluviòmetres, etc.).

- La informació de transaccions comercials o financeres pot ser recollida per dispositius senzills com els TPV (terminal de punt de venda) o per sistemes electrònics molt més sofisticats.

- Les bases de dades públiques recopilen les dades dels diversos registres públics, tant si han estat recollides en paper com en format electrònic.

L'objectiu final d'aquesta fase és obtenir un model de dades que descrigui de quines dades disposem, quins són els formats i com s'organitzaran i processaran. També cal definir quines dades es recolliran (de vegades no cal emmagatzemar totes les dades que es generen), les seves relacions i associacions, i fins i tot tenir en compte com podrien utilitzar-les futurs usuaris.

Tot i que el procés de captura de dades acostuma a estar molt lligat al de creació, i que és necessari capturar-les en el moment de la generació (per exemple, captura de les matrícules dels cotxes que circulen per un carrer), hi ha altres casos en els quals les dades es poden recuperar *a posteriori*, ja que han estat generades en un entorn digital i són persistents (per exemple, extracció d'enllaços d'una pàgina web). Hi ha algunes eines útils per a la captura de dades, com ara:

- Extracció de dades de pàgines web (tècnica coneguda com *scraping*) amb Scrapy.

- Captura de dades de Twitter amb Tweepy.

- Extracció de taules de documents PDF amb Tabula.

La digitalització de documents cada vegada té més importància pels múltiples avantatges que comporta, entre els quals hi ha l'estalvi de costos d'emmagatzematge, una major garantia de conservació de la informació i la possibilitat d'accedir-hi des de diversos llocs i dispositius.

La digitalització suposa un primer pas en la gestió documental, però s'ha de complementar amb una correcta classificació i indexació dels documents, així com també amb eines d'extracció de dades.

Fase 3. Emmagatzematge/conservació

La dada, una vegada capturada, s'ha d'emmagatzemar d'acord amb uns criteris relacionats amb l'explotació posterior, utilitzant alguns dels formats més habituals:

- CSV
- JSON
- XML
- RDF

Amb aquesta finalitat, la majoria d'organitzacions disposen d'un ***data warehouse*** (la traducció literal del qual seria «magatzem de dades», però està més estès l'apel·latiu de «repositori de dades» abreujat amb l'acrònim DWH), que és un conjunt de bases de dades dissenyat per a afavorir l'anàlisi i la divulgació eficient de les dades en les organitzacions.

Les organitzacions distingeixen entre el DWH i els sistemes operacionals o transaccionals, de manera que el DWH no ha de contenir dades d'ús actual i diari.

Les característiques principals d'un DWH són l'orientació en un determinat àmbit o tema (és a dir, dades relatives a un mateix esdeveniment o objectiu unides entre elles), la variabilitat temporal (els canvis en el temps queden enregistrats de manera que els possibles informes que se n'obtinguin reflecteixin aquestes variacions), la volatilitat (és a dir, la informació no es modifica ni s'elimina) i la coherència interna.

També acostumen a integrar dades que provenen de fonts heterogènies i bases de dades amb estructures diferents, amb la qual cosa és una pràctica habitual normalitzar les dades abans de combinar-les en el DWH mitjançant eines ETL (*extraction, transformation and loading*: extracció, transformació i càrrega). És important, per tant, facilitar-ne una descripció global, de manera que els usuaris puguin entendre quines dades estan emmagatzemades i quina és la millor manera d'accedir-hi.

Els components més habituals quan es treballa amb dades són les anomenades **metadades**, és a dir, informació que descriu quina és l'estructura de les dades emmagatzemades i les seves relacions.

Recentment han sorgit altres conceptes com el de *data lake*, associats a la proliferació de les dades desestructurades i massives: «Data Lake vs Data Warehouse: Key Differences».

No obstant això, per a ús personal o domèstic qualsevol base de dades o full de càlcul senzill pot ser un suport admissible i suficient per a l'emmagatzematge de la informació.

Altres aspectes a tenir en compte en aquesta fase:

- Decidir quines dades emmagatzemarem (tal com ja hem comentat, no cal recollir totes les dades generades, ni emmagatzemar totes les dades recollides).

- Mantenir la informació, si és possible, en l'estat original. És a dir, si s'han produït càlculs o transformacions cal emmagatzemar-los, però sense eliminar les dades font (d'altra manera, no seria possible desfer les transformacions o corregir errors).

- Generar i emmagatzemar metadades, és a dir, dades que proporcionen informació sobre el contingut i els formats de les dades disponibles. Utilitzar metadades permet una millor coordinació de les parts implicades, així com també un millor intercanvi i aprofitament de les dades entre diversos usuaris.

- Garantir l'accessibilitat i disponibilitat de les dades en temps raonables.

- Utilitzar versionats, per a poder accedir a situacions passades de la informació.

- Identificar i gestionar dades sensibles, per a preservar-ne la privadesa i la confidencialitat.

Fase 4. Tractament/preprocessat

En general, el coneixement s'extreu combinant diferents fonts de dades, eliminant dades ambigües o parcials, seleccionant els valors desitjats, etc. L'objectiu final d'aquesta fase és disposar d'un conjunt únic de dades per a una posterior anàlisi. Hi ha eines disponibles per a la manipulació de dades com OpenRefine.

També pertanyen a aquesta fase els procediments necessaris de control de qualitat, per a mirar d'identificar potencials errors i com solucionar-los.

És crític documentar convenientment tots els tractaments aplicats a les dades durant aquesta fase.

Fase 5. Anàlisi

Les dades ja preprocessades, lliures d'errors i complint els requisits necessaris, són analitzades utilitzant una bateria de tècniques de l'estadística o la mineria de dades, d'acord amb l'objectiu establert (interpretació, regressió, predicció, classificació, etc.).

Per a l'anàlisi de dades hi ha disponibles multitud d'eines:

- Programes comercials com SPSS, SAS, Minitab, Stata, etc.
- Llenguatges de programació *open-source* com R o Python.
- Eines més específiques, com Gephi, per a l'anàlisi de xarxes i grafs.

També pertanyen a aquesta fase els possibles enriquiments de la informació per mitjà de la combinació amb altres fonts de dades que puguin aportar un valor afegit.

Fase 6. Visualització

Anàlogament, les dades poden visualitzar-se *a priori* (respecte a l'anàlisi) amb l'objectiu de fer-ne una primera inspecció preliminar, detectar-ne patrons, tendències, etc., que permetin afinar els instruments utilitzats per a la seva anàlisi, o *a posteriori*, per a resumir-ne, mostrar-ne o explicar-ne els resultats de manera que sigui més senzill transmetre'n el coneixement extret.

Vegeu també

Per a la visualització de dades hi ha disponibles moltes eines, algunes de les quals es relacionen en l'apartat 3.2.5 «Eines i altres recursos» dins de l'apartat 3 «Representació i visualització de dades».

Fase 7. Publicació/difusió

Finalment, les dades (capturades, preprocessades, analitzades i visualitzades) poden ser publicades, de manera que sigui possible compartir-les amb tercers que puguin reutilitzar-les en un altre context o amb un objectiu diferent.

Vegeu també

Consulteu l'apartat 4 «Difusió i publicació de resultats d'un estudi quantitatiu» d'aquest material per a més detalls.

2.3. El govern de les dades

En el món empresarial, des de fa molts anys, es treballa amb dades i informació, i es mira d'extreure coneixement d'aquestes dades per a fer-ne ús tenint com a objectiu el màrqueting i les projeccions d'informació financera o comercial, entre d'altres.

Això fa que siguin moltes persones i departaments els que utilitzen les dades internes de les empreses, però habitualment aquestes tasques d'accés i explotació de la informació es duen a terme sense gaire control i homogeneïtat, exceptuant alguns casos de «bones pràctiques».

Per això un dels temes dels quals es parla cada vegada més és el del govern de les dades, entès com una manera d'establir una estructura de responsabilitats sobre cadascuna de les fases o moments clau del cicle de vida de les dades, i les accions que es realitzen sobre elles.

Podem definir el **govern de les dades** com un marc d'organització que harmonitza l'estratègia, defineix els objectius i estableix les polítiques per a la informació corporativa.

Per tant, com a disciplina involucra tant les persones com els processos i la tecnologia relacionats amb la gestió de les dades, de manera que tracta la informació corporativa com un recurs de valor empresarial.

Recentment, el govern de dades ha pres una importància sense precedents dins de les organitzacions, atès que cada vegada és més evident que els problemes en el maneig de la informació afecten la presa de decisions, ja que no hi ha processos ni polítiques que permetin garantir la confiança en les dades.

2.3.1. El rol del *Chief Data Officer*

Les empreses han invertit molt de temps, diners i esforç a millorar la part tecnològica, tant el suport de la informació com les aplicacions que la gestionen; no obstant això, en termes generals s'ha invertit molt menys en l'organització de les dades i dels treballadors que les manegen.

Això provoca que, malgrat disposar de moltíssima informació, moltes companyies encara tinguin problemes per a extreure el valor econòmic de les seves dades. És freqüent fins i tot que tinguin dificultats per a trobar i explotar informació respecte a aspectes molt bàsics com l'estat i evolució del negoci.

Hi ha grans oportunitats de negoci que requereixen entendre el poder de les dades, però sembla contradictori navegar en un oceà de dades i informes, i no obstant això no tenir informació, o dubtar-ne. La clau és passar d'entendre les dades com a cost a entendre-les com un actiu important, passar d'un ús operacional de les dades a un ús estratègic. El valor real de les dades és en allò que el negoci pot fer amb elles respecte al que pot fer sense elles.

Evidentment, cada empresa pren les seves decisions segons el tipus de dades de què disposa, les necessitats i la manera com tracta la informació. Tanmateix, comença a estendre's la figura del ***Chief Data Officer*** (CDO), que es podria considerar com **el responsable de les dades**: si tractem les dades com un actiu intangible estratègic, requereixen una gestió professional.

Amb l'auge del *big data*, moltes organitzacions han entès que les dades, si s'utilitzen convenientment, tenen un gran valor per al negoci, fins al punt de poder canviar l'empresa i fins i tot el sector sencer. Per aquest motiu no n'hi ha prou amb les clàssiques figures del *Chief Technology Officer* (CTO) o el *Chief Information Officer* (CIO), amb un perfil marcadament tècnic.

Les exigències regulatives, especialment en certs sectors (com el financer o el sanitari) on s'acostumen a manejar dades sensibles, que han de ser convenientment protegides per a salvaguardar la privadesa i la confidencialitat, han estat un altre impuls per a la creació d'aquesta nova figura.

Tot i que cada companyia concep el rol del CDO de forma diferent (i que fins i tot en determinats casos es crea la figura sense tenir del tot clares les responsabilitats reals), entre les seves funcions més recurrents hi hauria l'inventari dels béns d'informació, el disseny de l'estratègia en aquest àmbit, l'optimització de la gestió de la informació i els recursos humans, tècnics i econòmics implicats.

D'acord amb el que exposa Steele en el llibre *Understanding the Chief Data Officer*, les principals responsabilitats del CDO serien:

- **Amplitud**. Es tracta d'un paper global que implica una àmplia varietat de tasques: des d'associar les necessitats de dades de la companyia a l'objectiu general del negoci de cara a crear valor, fins a treballar juntament amb totes les àrees de la companyia per a assegurar que totes van en la mateixa direcció, passant per habilitar les tecnologies necessàries. Cal, per tant, controlar el cicle de vida de les dades. Òbviament, no és necessari que l'equip del CDO realitzi efectivament totes aquestes tasques, però sí que les coordini.

- **Equilibri**. El CDO ha de buscar i mantenir un cert equilibri entre les estratègies ideals i les implantacions pràctiques, entre el curt i el llarg termini i entre les diverses prioritats de cada àrea.

- **Centralització**. És responsabilitat del CDO garantir la disponibilitat de les dades entre els diversos departaments i la seva centralització, ja que el principal valor rau a combinar dades de múltiples fonts per a aconseguir els millors resultats. Això inclou una visió holística dels clients i de les dades que se'n tenen. També implica ser capaç d'integrar dades externes (de col·laboradors, proveïdors, etc.)

- **Priorització**. El CDO lidera la definició de les prioritats de l'empresa en l'àmbit dels projectes que utilitzen dades. Una adequada estratègia ha de ser prou flexible i realista perquè en posar-la en pràctica no doni problemes.

- **Integració**. L'equip del CDO treballa al costat de les altres parts implicades recopilant informació i generant retorns que permetin cobrir les necessi-

tats, però de manera que això resulti en unes accions i presa de decisions millors per a la companyia. És a dir, es tracta d'utilitzar les dades com una peça integral d'un objectiu comú.

- **Facilitació**. El CDO ha de ser capaç d'eliminar barreres existents per a l'ús de les dades, d'alliberar els recursos que permetin utilitzar-les convenientment, així com també de proveir noves eines.

Però amb una funció tan global i variada, el perfil necessari no és gens fàcil d'aconseguir, perquè ha de combinar coneixements de negoci, de gestió empresarial, tècnics i estadístics, així com també visió estratègica i habilitats comunicatives (ha de ser capaç, per tant, d'ajudar en la sempre difícil comunicació entre les àrees tècniques i de negoci, que habitualment es diu que «parlen llengües diferents»).

L'equip del CDO pot ser complementat amb altres figures que habitualment reben el nom de *Data Stewards*. Es tracta d'experts (en cadascuna de les àrees de la companyia que utilitzen dades) que assumeixen la responsabilitat de les dades de la seva àrea. D'aquesta forma es descentralitza la funció del CDO, la qual cosa pot ser una bona estratègia en determinats sectors en què, per la seva complexitat o pels hàbits adquirits durant molt de temps, resulta complicat, almenys inicialment, establir una nova figura que assumeixi tantes funcions i, en conseqüència, tanta càrrega de treball.

Aquí tenim dos enllaços on es parla d'aquestes figures:

- A «La emergencia del Chief Data Officer», l'autor es fa ressò de recents articles i notícies per a destacar la creixent importància d'aquest nou càrrec: «Aquest rol té la responsabilitat de gestionar totes les iniciatives de dades en l'organització i, com és possible suposar, cada vegada més esdevé un perfil imprescindible per a les organitzacions.»

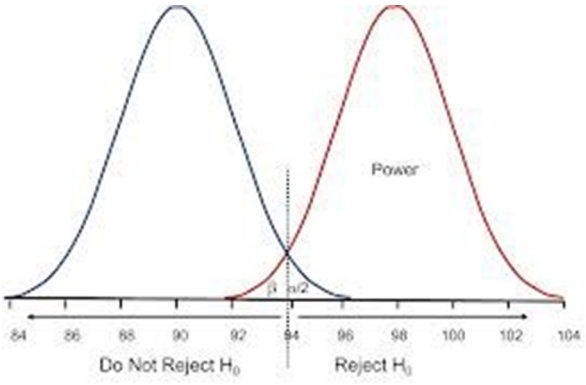
- De forma similar, a «A new corporate position: Chief Data Officer» i basant-se en el cas concret de la ciutat de San Francisco, es parla de la necessitat d'aquest càrrec en general, destacant que és un rol cada vegada més popular en les organitzacions pel fet que l'ús de la informació guanya més pes en la presa de decisions.

3. Representació i visualització de dades

3.1. *Storytelling*, el relat estadístic

Tradicionalment, els científics estem acostumats a presentar les nostres recerques de forma abstracta i, sovint, poc atractiva, amb profusió de fórmules, taules i, en el millor dels casos, alguns gràfics que il·lustren conceptes teòrics:

	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septbre	Octubre	Novbre	Dicbre
1995	9	13	17	15	24	44	10	44	24	24	33	106
1996	90	39	26	32	62	19	11	33	87	16	75	113
1997	123	3	6	70	55	63	24	46	97	20	56	82
1998	59	26	10	32	95	16	1	25	30	13	14	62
1999	20	17	69	32	22	25	31	11	70	70	29	22
2000	31	0	36	57	52	17	3	4	16	175	28	52
2001	45	36	32	38	54	5	9	13	70	52	53	50
2002	33	2	51	94	107	47	16	60	48	39	39	46
2003	35	75	41	67	95	25	8	34	62	93	52	28



Tanmateix, les recents tendències indiquen que, en lloc d'això, hauríem de dedicar-nos a **explicar històries**; i és que l'***storytelling***, present des de sempre en altres disciplines, també s'estén a nous camps en els últims temps, i concretament al que ens ocupa, com es pot comprovar en els següents articles: «Data scientist job: storytelling» i «Statistical storytelling», o també en la creació d'esdeveniments i seminaris específics («Storytelling conference»).

La pregunta que sorgeix és: com podem aplicar aquest tipus de tècniques a un projecte quantitatiu? Doncs bé, al capdavall es tracta simplement de comunicar millor, amb la qual cosa el primer pas és aplicar les normes bàsiques de la comunicació efectiva que provenen d'altres àmbits, com ara:

- Utilitzar un títol atractiu i que capti l'atenció.
- Tenir sempre en compte l'audiència a qui ens dirigim.
- Adoptar un estil narratiu.
- Utilitzar visualitzacions.

Ara bé, també hem de tenir en compte qüestions específiques de l'àmbit quantitatiu:

- No hi ha relat (estadístic) sense una exploració prèvia de les dades. És a dir, és important una fase inicial de coneixement de les dades abans d'afrontar una anàlisi, sigui numèrica o gràfica.

- La història la fan les dades, no la visualització. Per tant, cal evitar que visualitzacions efectistes puguin enterbolir la informació subjacent en les dades.

- Una història no és la veritat (almenys, no l'única possible). És, doncs, recomanable oferir més d'una línia d'anàlisi; o, si no es poden explotar per falta de temps o d'interès, com a mínim suggerir-les.

En els següents enllaços podem veure un parell d'exemples de visualitzacions de dades realment espectaculars, ambdues sobre el mateix tema, les desigualtats econòmiques als EUA:

- En versió dinàmica tenim aquest excel·lent vídeo on no només la visualització està molt ben dissenyada, sinó que la història que hi ha darrere s'explica de forma molt suggeridora, ja que es basa en la comparació de com creiem que haurien de ser les coses, com creiem que són, i com són realment, i demostra que fins i tot en les nostres opinions moltes vegades ens quedem curts: «Wealth inequality in America ».

- En versió estàtica tenim un bon exemple de combinació de gràfics diversos per a il·lustrar diferents punts de vista sobre un mateix tema i amb un mateix fil conductor: «It's the Inequality, Stupid».

3.1.1. Principals gèneres de narrativa visual

Segel i Heer proposen els següents set tipus d'agrupació dels gèneres de narrativa visual:

- Estil revista.
- Gràfic comentat.
- Pòster en parts.
- Flux.
- Tira de còmic.
- Presentació amb diapositives.

- Vídeo/animació.

A continuació, mostrem diferents exemples de visualitzacions, cadascun dels quals pot associar-se a algun dels gèneres comentats:

- «Desmontando mitos sobre el mundo». Es tracta del vídeo de la visita del famós Hans Rosling (un dels principals gurus actuals de la visualització de dades; recomanables també algunes de les seves xerrades a TED) al programa *Redes*, de TVE. Els seus estudis són excel·lents exemples d'animacions.

- «Index of Globalization». Aquest tipus d'infografia és possiblement la més freqüent. S'hi combinen diversos tipus de gràfics en una seqüència (generalment vertical, de dalt a baix), configurant una mena de tira de còmic.

- «Online in 60 seconds». Es tracta d'un típic gràfic amb comentaris.

- «La hidratación del deportista». Clar exemple de pòster en parts.

3.2. La visualització de dades. Principals característiques

3.2.1. La importància de visualitzar dades

La pregunta no hauria de ser si és important visualitzar les dades o no, sinó quin tipus de visualització pot ser la més útil a cada situació. No oblidem que, segons diversos estudis, aproximadament la meitat del cervell humà es dedica a processar informació visual. A més, també hi ha indicis clars que la capacitat persuasiva d'una història és superior si s'utilitzen gràfics i visualitzacions.

La visualització de dades, a més, té l'avantatge que no és un «efecte visual» o una «il·lusió òptica», sinó que està associada a fets que es poden mesurar i comprovar. De fet, un gràfic realitzat a partir de dades no és una simple il·lustració, per tant, cal saber «llegir-lo» i el dissenyador ha de facilitar-ne la interpretació.

A Espanya la importància que es concedeix a la visualització de dades és encara incipient, però també clarament creixent. A l'article «Periodismo de datos, infografía y visualización de la información: un estudio de *El País*, *El Mundo*, *Marca* y *El Correo*» s'analitzen les infografies utilitzades en diversos mitjans nacionals tant des del punt de vista del seu ús com a recurs periodístic, com dels aspectes tècnics.

3.2.2. Avantatges de la visualització de dades

La visualització de dades (si està ben feta) permet, entre altres, els següents avantatges:

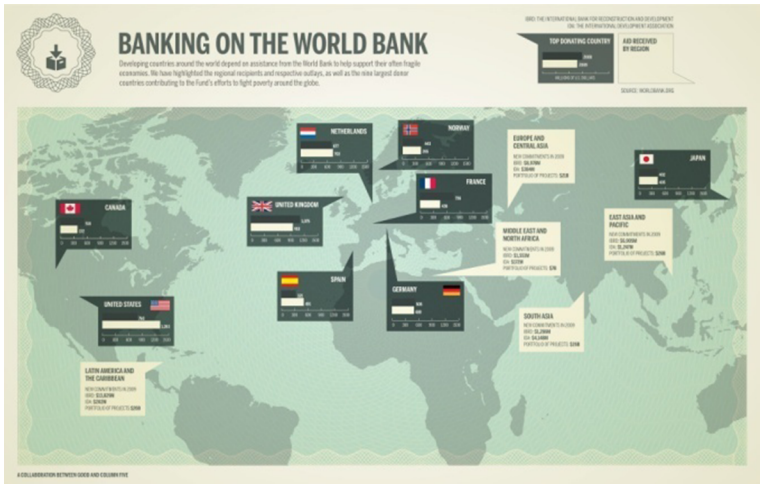
- Veure d'una manera nova allò que resulta familiar.
- Mostrar una dinàmica temporal, els canvis produïts amb el pas del temps.
- Fer comparatives de valors.
- Mostrar connexions, relacions i fluxos.
- Mostrar jerarquies.
- Explorar grans bases de dades.
- Il·lustrar un argument de manera convincent.
- Treure informació tècnica no necessària.
- Oferir transparència sobre el procés de construcció.

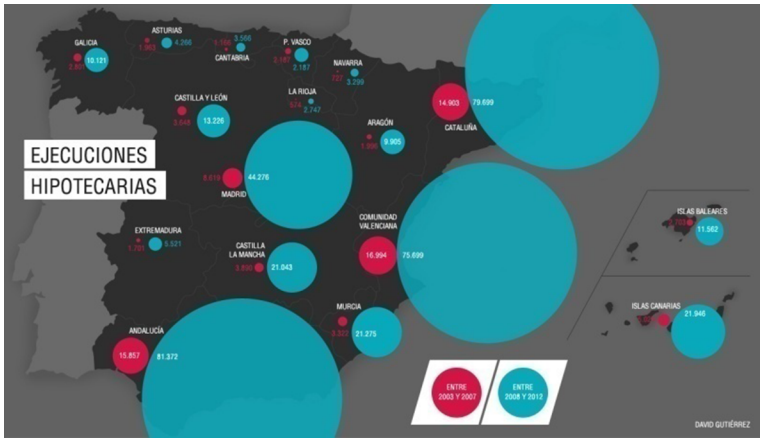
Ara bé, no sempre és adequat fer visualitzacions. Per exemple, no cal fer-les quan:

- Es tenen molt poques dades.
- N'hi ha prou amb una senzilla taula per a mostrar les dades.
- Un mapa no és un mapa (és a dir, no aprofita la dimensió espacial).

Aquí tenim un clar exemple de visualització innecessària:

Aquí, unes visualitzacions no gaire ben plantejades, ja que en cap dels dos casos s'aprofita bé la dimensió espacial i, a més, en el segon es fa ús de mesures absolutes, no relatives:





3.2.3. Tipus de visualitzacions

Segons Alberto Cairo, les visualitzacions es poden agrupar en dos grans tipus:

- **Visualitzacions exploratòries o interactives**. Són les més cridaneres, però normalment només es veuen a les pàgines de diaris especialitzats o empreses que disposen de prou recursos, ja que generalment s'han d'utilitzar llenguatges de programació no gaire senzills per a obtenir-les.

- **Visualitzacions explicatives o infografies**. Malgrat ser estàtiques, si estan ben fetes també són un recurs molt atractiu, i tenen l'avantatge que el disseny és molt més senzill, ja que hi ha diverses eines (un gran nombre d'elles gratuïtes) per a poder fer-les.

D'acord amb el mateix autor, aquests són els quatre principis bàsics de les bones visualitzacions:

- Basar-se en bones dades i informació de qualitat.

- Atreure l'atenció del lector, de l'audiència.

- No decebre/frustrar el lector (o per excessiva complexitat o per massa superficialitat).

- Mostrar la quantitat adequada d'informació.

Aquí mostrem diversos exemples molt interessants de visualitzacions interactives:

- «How different groups spend their day». Tot i que fa uns quants anys de la seva publicació, és un molt bon exemple d'un gràfic d'àrees que incorpora múltiples dimensions.

- «Index of Economic Freedom». Un excel·lent exemple de com haurien de ser els projectes de visualitzacions basats en dades: incorpora gràfics estàtics i interactius, publica les dades i en permet la descàrrega. D'aquesta forma, les necessitats de gairebé qualsevol tipus d'usuari de la informació queden satisfetes.

- «Confira a evolução da população do mundo desde 1950». Població mundial. Visualització (en *flash*) sobre l'evolució de la població mundial, que també permet l'exploració de la informació des de diversos angles.

- «Music Timeline» de Google. Un exemple molt semblat al primer d'aquest llistat, i que després ha estat copiat o reutilitzat amb dades d'altres temàtiques.

- «Globally known people». Un exemple de com es poden incorporar múltiples dimensions en un gràfic, jugant amb la grandària, el color, etc.

- «World migration». Gràfic interactiu molt original sobre els fluxos migratoris entre països del món.

- «World economy». Un altre projecte que permet visualitzar les dades de forma interactiva amb un desplegament tècnic i visual que resulta molt atractiu.

- «Timeless songs». Curiós exemple que combina també diverses formes de dades i com es poden mostrar.

Ara, alguns exemples d'infografies estàtiques:

- «Las 100 mayores empresas del mundo». Mitjançant una combinació de gràfics senzills i que podríem considerar clàssics, s'ofereix una panoràmica força clara del tema tractat.

- «Daily routines». Un senzill gràfic de barres apilades permet comparar les rutines diàries de diversos genis de la història.

També hi ha visualitzacions que podríem denominar mixtes, és a dir, dotades de certa interactivitat sense arribar a ser del tot animades o exploratòries:

- «Globalization Index». Aquest estudi sobre l'índex de globalització mundial és un bon exemple d'informació en diverses capes, on després de clicar cada país s'obté un detall de les dades i gràfics addicionals.

- «Global Traffic Map». Una combinació de diversos gràfics per a oferir diferents punts de vista, en aquest cas sobre els moviments entre països.

- «US Elections». La web del *New York Times* ha convertit en un clàssic les seves visualitzacions sobre els resultats de les eleccions presidencials.

Hi ha multitud de webs i blogs que recopilen visualitzacions de diverses fonts; és recomanable fer-hi un cop d'ull per a captar idees interessants. Per exemple:

- Viziometrics.

- Dades públiques dels EUA.

- Certamen de premis de periodisme de dades.

- Eager Eyes.

3.2.4. Suggeriment per a fer visualitzacions

Crear una visualització és un procés iteratiu de refinament. És habitual començar dissenyant un gràfic senzill per a anar transformant-lo i millorant-lo progressivament.

A continuació, presentem una compilació de recomanacions i errors comuns que cal evitar en la realització de gràfics:

- **Ser precís** (sense obsessionar-se). És molt important perquè, de vegades, un ús incorrecte de les escales o els eixos pot confondre.

- **Utilitzar les mesures adequades** a cada cas. Mesures relatives si cal, ràtios, percentatges, etc.

- **Minimitzar l'ús de gràfics de sectors**. Només s'han d'utilitzar quan volem representar les parts d'un total, però malgrat l'èxit que tenen i l'ús molt freqüent que se'n fa, no gaudeixen de gaire bona fama entre els experts.

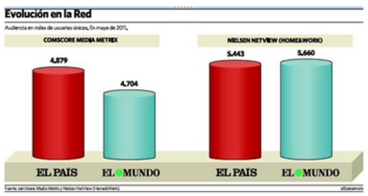
- **Utilitzar ordre**. Jerarquies, cronologies, etc.
- Connectar i **relacionar informació**.
- **Evitar correlacions** (i causalitats) **equivocades**.
- **Utilitzar les escales múltiples amb coherència**, ja que moltes vegades confonen si no s'expliquen bé.

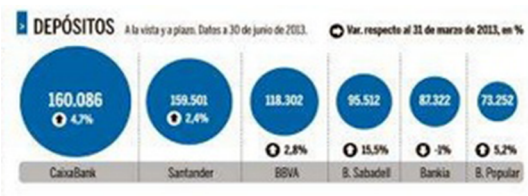
Tot seguit, una sèrie de recomanacions sobre el disseny:

- **Evitar elements i efectes innecessaris**. Moltes vegades els efectes tridimensionals o els colors cridaners desvien l'atenció del que és realment important.
- **Respectar els convencionalismes** (per exemple, codis de colors).
- **Explorar les dades de maneres diferents**.
- **Transmetre una idea** i, després, simplificar.
- Dissenyar pensant en **dos tipus d'audiències** (els experts i els que no ho són tant).
- **Interactuar amb les dades**, no amb les aplicacions. Les eines han d'ajudar l'usuari a aprofundir, no fer-li-ho més difícil.
- **Afegir notes que aportin significat**. De vegades les etiquetes no són gaire informatives.

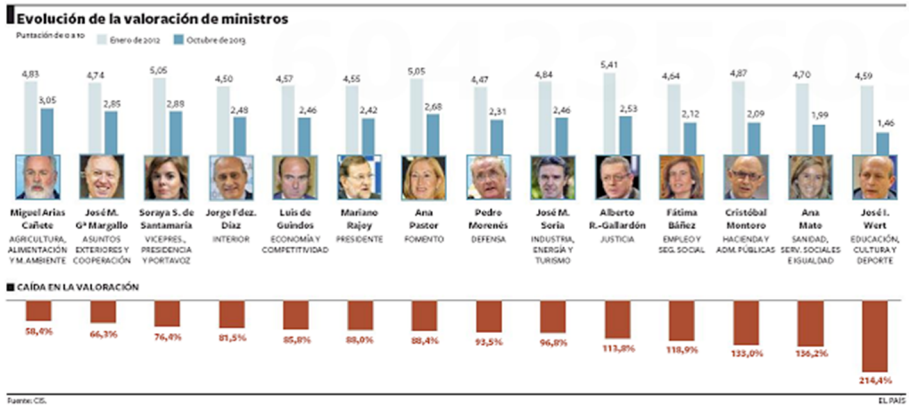
A continuació, mostrem alguns exemples d'errors molt comuns.

En aquestes dues imatges, els gràfics no respecten les escales i, per tant, provoquen confusió:

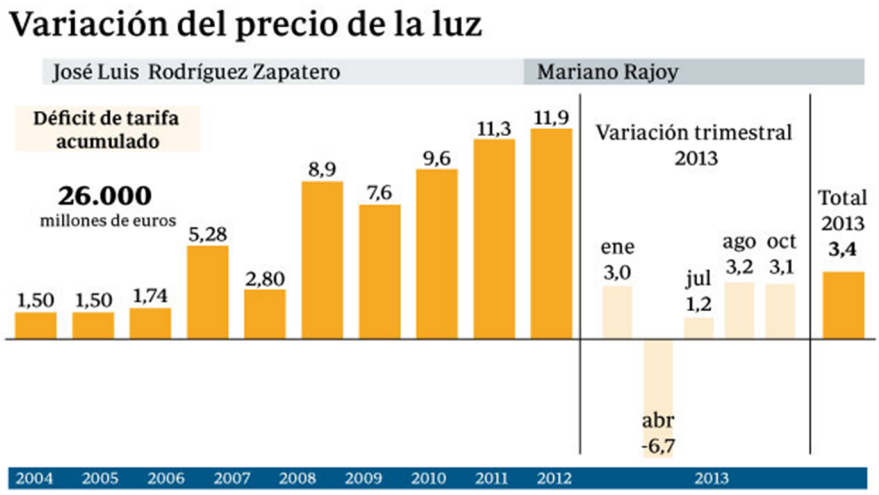




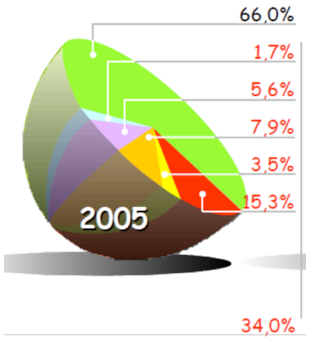
La següent imatge és un exemple d'ús de mesures incorrectes, ja que no es pot perdre més del 100% de la valoració:

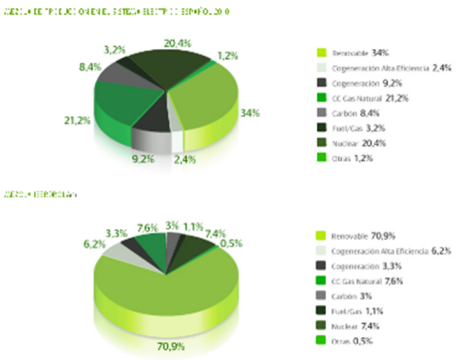


En la següent imatge el canvi d'escala (anual enfront de trimestral) pot crear confusió (l'última columna, la del resum de 2013, no estava en la versió publicada originalment i es va afegir després de les queixes d'alguns usuaris):

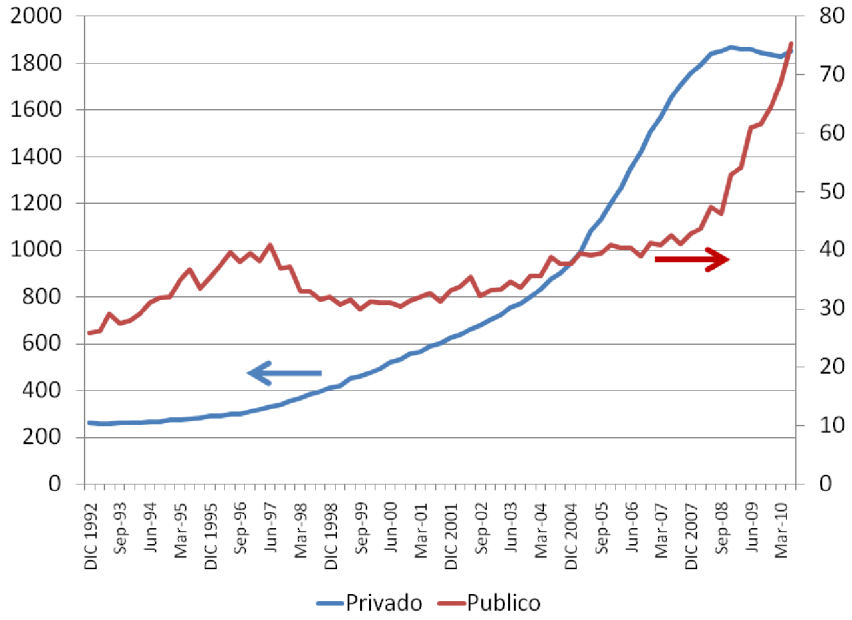


A continuació, dos exemples molt evidents de mal ús dels gràfics de sectors, i també d'efectes absolutament innecessaris i contraproduents:

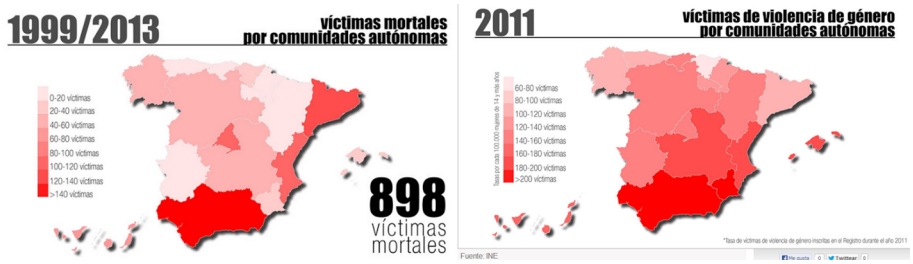




Ara, dos casos de mal ús de les escales: en el primer cas, escales múltiples, i, en el segon, canvi d'escala a l'eix horitzontal:



Finalment, un curiós exemple d'error autocorregit. En la primera versió publicada s'utilitzaven mesures absolutes (i, per tant, les regions amb més població tenien les dades més altes), mentre que en la segona ja es fa ús de mesures relatives, molt més adequades:



3.2.5. Eines i altres recursos

Hi ha moltíssims recursos sobre infografies i visualitzacions, ja que és un *trending topic* dels últims anys.

A continuació, un parell de recursos curiosos que ens ajuden a triar el gràfic més adequat per a cada ocasió:

- A periodic table of visualization methods.
- Chart chooser.

Quant a eines i aplicacions, també n'hi ha moltes i contínuament n'apareixen de noves; per exemple:

- Piktochart.
- Infogr.am.
- Visualize free.
- Stat Silk.
- Raw.
- Icharts.
- JoliCharts.
- Google Data Studio.
- Mapchart.
- Visme.
- Venngage.
- InZight.
- Quadrigram.

Alguns blogs interessants:

- Flowing Data.
- Visualising data.
- The Functional Art.

Compilacions de recursos:

- UK Data Explorer.
- GapMinder.
- Policyviz.
- Keshif.
- Seeing data.

4. Difusió i publicació de resultats d'un estudi quantitatiu

La publicació o comunicació científica és un dels últims passos de qualsevol recerca. En el procés de comunicació científica sovint se sol distingir entre difusió i divulgació, que si bé en ocasions es poden confondre i utilitzar com a sinònims, són conceptes diferents encara que no antagònics, sinó complementaris. La difusió és la manera de fer arribar els treballs científics a la comunitat investigadora que, com a públic expert, és capaç d'entendre l'abast dels mateixos i també provocar revisions o discussions que puguin tant millorar-los com posar-los en dubte, arribat el cas. La difusió generalment parteix de l'investigador o grup de recerca, pel seu interès a donar coneixement dels resultats i conclusions dels seus estudis. La divulgació tracta d'estendre els coneixements científics a un públic més general, que pot comprendre la seva importància i les principals argumentacions, però que no compta amb un domini profund del camp en qüestió. A l'hora de divulgar, per tant, és important fer assequibles i amenes (mitjançant exemples, visualitzacions, etc.) les conclusions dels treballs. La divulgació està més relacionada amb els mitjans de comunicació i el món del periodisme, encara que també la pot fer el mateix investigador o grup de recerca per mitjà de blogs o xarxes socials.

Les revistes científiques (*journals*) són les que major reconeixement tenen a l'hora de difondre el coneixement d'aquest tipus. Hi ha molta literatura sobre com es pot escriure un article de recerca, ja que en els últims anys la publicació en determinades revistes de prestigi s'ha convertit en un dels principals requisits per a avaluar la qualitat dels investigadors, i proliferen els rànquings i classificacions de les revistes d'acord amb el seu impacte. Poden trobar-se articles enfocats en aquest àmbit com, per exemple, els de Derntl (2014), Jalalian i Aslam (2012), o Torres-Salinas i Cabezas-Clavijo (2013).

Entre les principals característiques de les publicacions científiques cal destacar-ne la claredat, la precisió, la verificabilitat, la universalitat i l'objectivitat. No obstant això, hi ha un intens debat no exempt de certa controvèrsia sobre fins a quin punt aquesta obsessió per la publicació en revistes d'alt impacte és o no positiva per a la difusió dels treballs de recerca i del coneixement en general, i si el sistema de *peer-review* (revisió d'experts), que és un estàndard en aquest àmbit, és realment tan objectiu i transparent com es pretén.

En l'article «Publish-or-perish: Peer review and the corruption of science», de David Colquhoun (2011), l'autor es fa ressò de diversos casos de publicacions fraudulentes per a proposar un debat sobre el procés de revisió d'experts, la seva utilitat i les possibles alternatives.

En l'article «¿Qué es la producción científica?», de Juan Carlos Argüelles (2008), es desgranen la majoria dels aspectes rellevants d'aquest àmbit: des de conceptes generals i podríem dir que filosòfics sobre l'acceptació del valor universal de la ciència i si la seva essència és crear o produir («la disjuntiva radica en si s'investiga per a descobrir o per a publicar»), fins a aspectes històrics com la substitució del científic individual pel grup de recerca finançat amb fons públics o privats, passant per qüestions de més actualitat com ara que «l'abundància i nivell de les publicacions les ha convertit en un paràmetre crucial de política científica per a mesurar la qualitat i decidir quines línies i equips de treball mereixen ser finançats», o que els procediments habituals de citacions i revisió per experts no estan exempts d'inconvenients.

L'interessant article «How much better are the most-prestigious journals? The statistics of academic publication» d'Starbuck (2005) conclou, mitjançant una anàlisi estadística, que si bé les revistes científiques de gran prestigi acostumen a publicar articles de major valor que les menys prestigioses, també les primeres sovint publiquen força articles de menor qualitat i les últimes no pocs excel·lents articles.

El molt recomanable blog Nada es Gratis dedica també amb certa periodicitat alguna de les seves entrades a tractar aquests temes, i ho fa a més des d'un punt de vista original i amb alguna dosi d'humor, la qual cosa fa la lectura força amena, com, per exemple a «El (largo, tortuoso y entretenido) making-of de un artículo de investigación», «El impacto del papel couché en la ciencia» i «No, el proceso de revisión por pares tampoco es perfecto».

Malgrat tot, les revistes científiques no són l'únic mitjà per a la comunicació científica. Existeixen altres canals clàssics que faciliten la divulgació dels coneixements, com per exemple presentacions i ponències en congressos, seminaris i cursos; mitjans de comunicació tradicionals (documentals o reportatges en premsa, ràdio i televisió); llibres divulgatius; museus temàtics; etc. Les noves tecnologies estan promovent altres canals, que permeten que la informació flueixi de manera més àgil i ràpida, i que el receptor pugui accedir a la ciència de forma més senzilla mitjançant Internet i els recursos electrònics a la seva disposició (vídeos, visualitzacions, animacions, etc.): un exemple d'això últim són les anomenades revistes científiques electròniques, que es poden considerar una evolució de les tradicionals, però també la Wikipedia («La Wikipedia es la principal fuente de información científica»), els blogs científics (Naukas, Xataka), les xarxes socials («Cómo la ciencia está usando las redes sociales para la investigación»), etc.

En l'àmbit que ens ocupa especialment en aquest document, que és el del treball amb dades, cal destacar especialment la recent irrupció de noves revistes especialitzades en dades, els anomenats *data journals*. En l'article «Data journals: eclosión de nuevas revistas especializadas en datos», de García-García, López-Borrull i Peset (2015), es fa una completa revisió de l'estat actual d'aquest tipus de publicacions, destacant entre les causes de la seva eclosió la necessitat

de les revistes d'adoptar una estratègia per a oferir les dades utilitzades pels autors dels articles perquè qualsevol usuari pugui reutilitzar-les en honor de la transparència, així com també l'increment de la quantitat de dades que els científics produeixen. Com a conclusions, destaquen que amb aquest tipus de revistes especialitzades «es reforça la visibilitat dels aspectes metodològics que van associats a la recerca i es fan visibles dades», així com també que «el nou model de revista fa més confiable la comunicació científica».

4.1. Característiques bàsiques

La Comissió Econòmica de les Nacions Unides per a Europa (UNECE) indica que la difusió d'informació estadística es duu a terme des de les oficines nacionals d'estadística i altres organismes oficial, però també hi ha altres centres difusors de dades i estudis estadístics, que poden ser públics o privats, individuals o grupals.

D'acord amb UNECE, la difusió d'informació estadística per als mitjans de comunicació es basa en els principis següents:

Referència bibliogràfica

Crompton, V.; Ellison, D.; Flanders, J.; Pedersen, H.; van donin Elshout, S.; Weijers, D. (2005). «Making Data Meaningful». UNECE.

- **Rellevància**. La informació ha de ser rellevant per a la vida social, econòmica i per a les condicions generals del país, i satisfer les necessitats dels responsables de presa de decisions públiques i privades. Per als mitjans de comunicació, la rellevància es tradueix en l'interès periodístic o valor informatiu. No obstant això, cal anar amb compte a presentar la informació d'una manera que no trivialitzi les dades o resultats. L'objectiu és informar els ciutadans sobre la disponibilitat de les dades o informació. La cobertura mediàtica és desitjable, ja que augmenta l'audiència del missatge, incrementa el coneixement i estimula el debat entre el públic.

- **Confidencialitat**. És fonamental protegir la confidencialitat dels informants, ja siguin persones o empreses, per a totes les dades recollides. No s'ha de divulgar informació que identifiqui un individu o grup sense consentiment previ. L'organització tampoc ha de revelar informació que infringeixi la confidencialitat dels enquestats. Aquesta restricció s'aplica als mitjans de comunicació igual que a qualsevol altre usuari.

- **Independència i objectivitat**. La informació ha de presentar-se de manera objectiva i imparcial, i ser independent de qualsevol control o influència política.

- **Puntualitat**. La informació hauria de ser actual i estar a la disposició del públic tan aviat com sigui possible. La puntualitat de la informació influirà en la seva rellevància.

- **Accessibilitat i claredat**. En principi, tots els usuaris han de tenir accés a les dades i les metadades. La informació ha de ser pública i estar disponible en formats apropiats per mitjà dels canals de distribució adequats, i ha

d'estar escrita en llenguatge senzill i comprensible que s'adapti al nivell de comprensió dels grups d'usuaris principals. És important assegurar-se que els mitjans, igual que altres usuaris, siguin capaços d'accedir i interpretar correctament la informació sobre els mètodes estadístics, conceptes, variables i classificacions utilitzades en la producció de resultats estadístics.

- **Coherència**. L'ús de conceptes estàndard, classificacions i poblacions objectiu promou la coherència i la credibilitat de la informació estadística, així com també l'ús d'una metodologia comuna en les enquestes. L'adhesió a aquests principis fonamentals de difusió augmentarà la credibilitat de la informació i fomentarà la confiança en la seva fiabilitat.

4.2. Objectius

Cada vegada més, les organitzacions i els individus reconeixen la importància d'utilitzar estadístiques per a prendre decisions basades en l'evidència dels resultats.

Molts ciutadans només accedeixen a estadístiques pels mitjans de comunicació. Per tant, és fonamental que hi hagi una comunicació eficaç amb els mitjans per a aconseguir tres importants objectius de la difusió:

- Informar el públic en general sobre els resultats obtinguts i com els poden afectar en la vida quotidiana.

- Mostrar la rellevància de la informació estadística per a la presa de decisions, tant en l'àmbit públic com en el privat.

- Augmentar en el públic el coneixement i la cultura estadística, de manera que creixi la confiança en els estudis basats en l'anàlisi de dades.

L'assoliment d'aquests objectius es veurà facilitat en la mesura en què hi hagi una comunicació eficaç amb els mitjans de comunicació i a través d'ells. En això radica el gran interès per a construir una sòlida relació de treball amb els mitjans, i que així als periodistes els resulti fàcil donar informació estadística d'una forma exacta, oportuna i informativa, i perquè s'adoptin mesures per a augmentar la cobertura dels mitjans com una forma d'arribar a la societat en general amb informació estadística important.

L'audiència de dades estadístiques s'expandeix. Abans d'internet, els usuaris de dades i estudis estadístics eren una petita elit formada per un grup d'experts amb un alt grau d'interès i coneixement de les dades. Per contra, amb internet, la base de clients de la informació estadística s'ha expandit fins al punt que qualsevol persona amb un ordinador i una connexió a internet pot accedir a la informació.

No obstant això, com a resultat d'aquesta major disponibilitat de les dades, molts dels nous usuaris no tenen coneixements estadístics, ni estan tan familiaritzats amb els continguts o el llenguatge estadístic com hi estaven els usuaris anteriors.

Aquesta àmplia gamma d'usuaris obliga els estadístics a estar més centrats en l'usuari a l'hora d'oferir la difusió dels resultats de les seves anàlisis.

Ara que el públic pot accedir a la informació directament a la web, el paper dels mitjans de comunicació en la difusió de dades ha canviat, encara que continua essent important. Els professionals dels mitjans tenen el paper d'intermediaris en transformar les dades brutes en coneixements.

A més d'interpretar les dades i redactar articles en un llenguatge comprensible per al públic, els periodistes poden esmentar les fonts originals de les dades en cas que calguin més detalls. Abans que el públic tingués accés directe a les dades, els usuaris havien de posar-se en contacte directament amb les organitzacions estadístiques per a obtenir les dades i la seva interpretació. Aquesta modalitat assegurava que les estadístiques anaven acompanyades de metadades adequades (incloent-hi la metodologia apropiada) limitacions, explicacions i definicions.

Amb la web, l'accés de tipus autoservei permet als usuaris fer un cop d'ull pel seu compte i, en aquest context, descarregar fàcilment metadades adequades i comprensibles. És una necessitat de màxima importància. Els periodistes i altres usuaris continuen posant-se en contacte amb els organismes estadístics perquè els ajudin a interpretar les dades.

Resum

En aquest mòdul hem revisat alguns dels conceptes que estan provocant els principals debats en l'àmbit de la gestió de dades en els últims temps.

El primer apartat introdueix el lector en el concepte de *big data* i les seves aplicacions associades, aportant una sèrie d'articles recents en els quals es pot comprovar que és un tema de màxima actualitat i, per tant, en constant evolució.

Posteriorment, es dedica un apartat al cicle de vida de les dades, aspecte important per a entendre totes les fases per les quals pot passar la informació des de la seva creació fins a la seva explotació.

Com a part del repte que suposa per a les organitzacions que treballen amb dades el fet d'haver d'afrontar el govern de les dades, s'introdueix com les organitzacions encaren aquesta situació amb la creació de figures específiques com, per exemple, el *Chief Data Officer*.

El capítol central es dedica a la visualització de dades, en què s'expliquen les tècniques per a la creació d'un relat estadístic i també s'analitzen amb detall diversos exemples d'infografies i visualitzacions interactives, per a poder descriure quines són les principals característiques positives i també els principals errors que s'han d'evitar. Acabem aquesta secció amb una relació de recursos i eines disponibles per a la realització d'aquest tipus de visualitzacions.

Finalment, es tracten aspectes relacionats amb la difusió dels resultats de recerca, tant en publicacions especialitzades com en altres mitjans divulgatius.

Bibliografia

Argüelles, J. C. (2008). «¿Qué es la producción científica?». *El País*.

Cairo, A. (2011). *El arte funcional. Infografía y visualización de información*. Madrid: Alamut.

Colquhoun, D. (2011). «Publish-or-perish: Peer review and the corruption of science». [article en línia]. *The Guardian*. [Data de consulta: 5 de setembre de 2011]. <https://www.theguardian.com/science/2011/sep/05/publish-perish-peer-review-science>

Coyoli, C. G. (2014). «Analytics: el uso de big data en el mundo real». *Boletín Científico de las Ciencias Económico Administrativas del ICEA* (vol. 3, núm. 5).

Crompton, V.; Ellison, D.; Flanders, J.; Pedersen, H.; van donin Elshout, S.; Weijers, D. (2005). *Making Data Meaningful*. UNECE.

Derntl, M. (2014). «Basics of research paper writing and publishing». *International Journal of Technology Enhanced Learning* (vol. 6, núm. 2, pàg. 105-123).

García-García, A.; López-Borrull, A.; Peset, F. (2015). «Data journals: eclosión de nuevas revistas especializadas en datos». *El profesional de la información* (vol. 24, núm. 6, pàg. 845-854) [article en línia]. <http://recyt.fecyt.es/index.php/EPI/article/view/epi.2015.nov.17>

Gómez, J. L.; Conesa, J. (2015). *Introducción al big data*. Barcelona: UOC.

Jalalian, M.; Danial, A. H. (2012). «Writing for academic journals: A general approach». *Electronic physician* (vol. 4, núm. 2, pàg. 474-476).

Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C.; Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. Nova York: McKinsey Global Institute.

Segel, E.; Heer, J. (2010). *Narrative visualization: Telling stories with data*. IEEE transactions on visualization and computer graphics (vol. 16, núm. 6, pàg. 1.139-1.148) [article en línia]. <http://vis.stanford.edu/files/2010-narrative-infovis.pdf>

Starbuck, W. H. (2005). «How much are the most-prestigious journals? The statistics of academic publication». *Organization Science* (vol. 16, núm. 2, pàg. 180-200).

Steele, J. (2015). *Understanding the Chief Data Officer*. Boston: O'Reilly Media, Inc.

Torres-Salinas, D.; Cabezas-Clavijo, Á. (2013). «Cómo publicar en revistas científicas de impacto: consejos y reglas sobre publicación científica». *EC3 Working Papers* (núm. 13).