GOIB CONSELLERIA PRESIDÈNCIA I ADMINISTRACIONS PÚBLIQUES
ESCOLA BALEAR ADMINISTRACIÓ PÚBLICA

UNITAT 9

EINES BÀSIQUES PER AL TRACTAMENT DE DADES

CONTINGUTS

1.1 OBJECTIUS

1. **Conèixer les fases per al tractament de dades**
2. **Descobrir les eines existents per al tractament de dades**

1 Introducció

El fet que tant la ciutadania com les empreses puguin accedir a grans volums de dades públiques d'una manera cada vegada més àmplia i variada, i que les puguin emprar, ha facilitat el naixement d'un nou mercat: **el de la reutilització de la informació pública**.

Apareixen els agents infomediaris, que prenen les dades públiques i les reutilitzen (recopilen i tracten aquestes dades) per a finalitats diferents, i que generen productes, serveis de valor afegit i aplicacions.

Els darrers anys, han sorgit moltes eines que permeten el tractament de les dades i faciliten la creació d'aquests nous productes i serveis. Hi ha eines per a l'extracció, el tractament, l'anàlisi i la visualització de dades, com veurem en aquesta unitat.

2 Transformació de dades públiques

Els agents infomediaris transformen les dades públiques i creen productes, serveis i aplicacions. A continuació, es descriu cada un d'aquests elements:

- **Productes.** Els productes basats en dades obertes són la reutilització d'una o de diverses fonts d'informació pública per generar utilitats de valor. Alguns exemples de productes són:
  - **Dades de valor tractades.** S'encreuen dades de diferents fonts per aconseguir dades més completes i de més qualitat. Aquestes dades, tractades i millorades, es venen.
  - **Mapes.** Les dades geogràfiques es representen de manera gràfica damunt mapes i, d'aquesta manera, es poden vendre.
  - **Publicacions.** A partir de les dades públiques, s'extreuen resums i anàlisis útils que es poden vendre a altres agents infomediaris.
- **Serveis.** Els serveis sobre dades obertes es basen principalment en serveis de consultoria sobre matèries específiques. Alguns exemples de serveis són:
  - **Informes personalitzats:** reutilització d'informació sobre la base de censos i directoris públics, informes o butlletins oficials per oferir informació personalitzada (per exemple, informes de solvència empresarial).
  - **Assessorament:** serveis d'assessoria, investigació i estudis de mercat que es desenvolupen per als clients, que es basen en el tractament de la informació i que estan orientats a donar suport als processos de presa de decisions.
  - **Comparacions:** serveis de comparació de preus a internet de diferents productes o serveis, com ara hotels, assegurances, cotxes, etc.

- ***Clipping* (retalls):** el *clipping* sobre dades públiques es duu a terme recopilant «retalls» d'informació o dades públiques de manera sistemàtica per posar-los a la venda, o bé a petició d'un client.

- **Aplicacions:** eines informàtiques (programari) desenvolupades per a una utilitat específica. Poden estar dissenyades per a ús general o bé ser desenvolupades a mida per resoldre un problema específic d'un client. Com a exemples d'aplicacions destaquen:
  - **Programari del client:** aplicacions que consumeixen dades o informació pública mitjançant una connexió directa al servidor de dades o informació de l'Administració pública.
  - **Aplicacions mòbils:** aplicacions que consumeixen dades públiques i que s'han creat per ser utilitzades des dels mòbils.
  - **Altres aplicacions:** aplicacions de GPS, aplicacions de subscripcions, etc.

3 Eines per al tractament de dades

Les dades públiques que són tractades poden passar per quatre fases diferents: extracció, tractament, anàlisi i visualització.

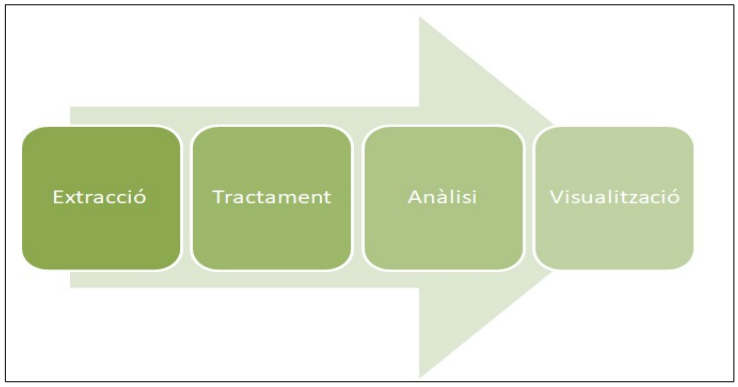

Figura 1. Fases per al tractament de dades.

Hi ha tot un ecosistema d'eines que els agents infomediaris empren per a l'elaboració de productes, serveis o aplicacions basats en dades obertes.

En aquest apartat presentarem una selecció d'eines que es poden fer servir en les diferents fases de l'elaboració de productes, serveis o aplicacions que reutilitzen dades públiques.

- **Eines d'extracció:** extracció de dades des de fonts d'informació diferents. A vegades, aquestes dades estan en formats no apropiats per a la reutilització i necessiten ser tractades.
- **Eines de tractament:** conversió de l'estructura d'un conjunt de dades a una altra de diferent que ens resulti més apta per a la reutilització i per poder crear productes, serveis o aplicacions.

- **Eines d'anàlisi:** procés de transformació de les dades en coneixement mitjançant una anàlisi descriptiva, inferencial o predictiva, entre altres.
- **Eines de visualització:** presentació dels productes, serveis o aplicacions en formes aprofitables per a l'usuari final, com ara mapes, gràfics o línies del temps.

Seguidament, entram en detall en cada una d'aquestes categories d'eines i n'exposam un conjunt d'exemples.

3.1 Eines d'extracció

Hi ha dos mecanismes principals per extreure dades: les API REST (*application programming RESTful*) i el que anomenam *scraping* (raspat de dades). A continuació, explicarem aquests dos mecanismes i les eines en què es basen:

- **API REST.** És un conjunt de funcions que permeten accedir i consultar les dades emmagatzemades en una base de dades. Aquestes funcions s'executen mitjançant enllaços web (URL). Gràcies a les API REST, les dades es poden consultar de manera automatitzada. Quan una aplicació informàtica necessita dades d'una altra aplicació, és molt probable que empri una API REST per consultar-les.
  - **Avantatges.** Permet consultar dades de manera automatitzada i freqüent. En cas de necessitat, permet aplicar-hi filtres i afinar la consulta. El format de les dades és estructurat (normalment, en format JSON) i preparat perquè la reutilització sigui fàcil i còmoda. Les consultes mitjançant API REST asseguren que sempre s'accedeix a les dades actualitzades de la base de dades.
  - **Inconvenients.** Una API REST necessita una implementació. Si l'API no està ben definida —no té filtres per afinar la consulta o l'estructura de les dades que retorna és molt complexa— i no està ben documentada, pot ser difícil d'emprar.
- ***Scraping* (raspat de dades).** L'*scraping* és un conjunt de tècniques de programació que permeten extreure dades d'un document (*PDF scraping*) o d'un lloc web (*web scraping*) mitjançant enginyeria inversa i que les presenten en un format reutilitzable. Se sol aplicar en cas que no hi hagi API REST de consulta, ja que presenta un seguit d'inconvenients, com ara els següents:
  - Les dades que s'extreuen poden estar poc estructurades i això en pot complicar el tractament.
  - Si les dades estan repartides en diferents llocs web, l'*scraping* pot arribar a ser complex.
  - El joc de caràcters en què estan representades les dades pot dificultar l'extracció (accents, eles geminades, ces trencades, caràcters especials, etc.).

A continuació, es presenten diferents eines per fer *scraping*. Aquestes eines estan classificades en dues categories: 1) *scraping* de documents (*PDF scraping*) i 2) *scraping* de llocs web (*web scraping*).

- **ParseHub.** És una aplicació web amb propietari (de pagament) que permet fer *web scraping*. Disposa d'un assistent visual per obtenir les dades presents en web de tercers en forma de dades estructurades (*https://www.parsehub.com/*).
- **Import.io.** És una aplicació web d'ús lliure (gratuïta) per realitzar *web scraping*. És un convertidor automàtic de pàgines web en dades estructurades. Té una versió web i una altra d'escriptori multiplataforma. L'aplicació d'escriptori permet fer *scrapings* de forma il·limitada (*https://www.import.io/*).
- **PDF Tables.** És una aplicació web amb propietari (de pagament) per fer *PDF scraping*. És un convertidor automàtic de fitxers PDF en dades estructurades. El pla gratuït només suporta *scraping* per a 25 pàgines per document (*https://pdftables.com/*).
- **Tabula.** És una aplicació d'escriptori multiplataforma d'ús lliure (gratuïta) per fer *PDF scraping*. És un assistent per convertir taules que es troben dins de fitxers PDF en dades estructurades (*https://tabula.technology/*).

3.2 Eines de tractament

A vegades, els conjunts de dades representats en taules no són perfectes, estan expressats de diferents formes, empren abreviatures, tenen errors de codificació, etc., i corregir-los de forma manual no és viable.

En aquests casos, cal emprar eines de transformació per millorar la qualitat de les dades i fer-les completament reutilitzables, perquè siguin bones de filtrar, millorar i classificar. Per a aquest tractament de les dades, hi ha un conjunt de tècniques que es presenten a continuació:

- ***Data cleansing* (neteja de dades).** Corregeix errors de les dades que afecten la qualitat. Per exemple, elimina caràcters estranys que poden dificultar la recerca o normalitzen noms de ciutats o codis postals a fi que tots segueixin la mateixa nomenclatura.
- ***Data wrangling* (transformació de dades).** Converteixen l'estructura de les dades en una de diferent, més apta per a la reutilització en forma de visualització o anàlisi estadística.
- ***Record linkage* (enllaç de dades).** Vincula registres entre conjunts de dades diferents per relacionar-los. Per exemple, relaciona les adjudicacions (conjunt de dades de contractació pública) amb les partides pressupostàries corresponents (conjunt de dades de pressuposts).

A continuació, s'han seleccionat dues eines que permeten fer el tractament de dades. S'han triat aquestes eines perquè són d'ús senzill sense necessitat de programar, són

aplicacions web o multiplataforma, i són de llicència lliure o ofereixen una versió gratuïta.

- **Open Refine.** És una aplicació d'escriptori multiplataforma d'ús lliure (gratuïta) per al tractament de dades. Permet netejar i transformar dades desordenades o completar-les. A més a més, permet fer operacions de *data cleansing* (neteja), *data wrangling* (transformació) i *record linkage* (enllaçament de dades). Open Refine és la versió d'ús lliure de Google Refine (*https://code.google.com/archive/p/google-refine/*). Google va abandonar el desenvolupament de Google Refine i el va alliberar com a programari lliure. Des de llavors, el projecte Open Refine manté una versió evolucionada de Google Refine (*https://openrefine.org/*).
- **Data Wrangler.** És una aplicació web amb propietari (de pagament) per al tractament de dades. És un assistent visual per a la neteja (*data cleansing*) i la transformació de dades (*data wrangling*) capaç d'exportar-les a taules d'anàlisi, a fi que puguin ser tractades amb aplicacions com Excel, R o Tableau (*http://vis.stanford.edu/wrangler/*).

3.3 Eines d'anàlisi

Les eines d'anàlisi estadística permeten fer exploracions avançades de conjunts de dades grans i formular models que permetin correlacionar variables o fer prediccions. Per tant, seran de gran valor a l'hora de treure el màxim partit de les dades, perquè permeten dur a terme anàlisis complexes.

Les activitats que típicament es duen a terme amb les eines d'anàlisi estadística són les següents:

- ***Clustering* (anàlisi d'agrupament).** Consisteix a agrupar un conjunt d'objectes, de manera que els objectes del mateix grup (anomenat *clúster*) siguin més similars entre si (en un sentit o un altre) que amb els dels altres grups. Per exemple, es poden agrupar les parades d'autobusos en funció de si es troben a 0,5 km, 1 km o més d'1 km de distància per analitzar quins barris tenen més bones comunicacions.
- **Anàlisi de regressió.** És una eina que s'empra freqüentment en estadística i que permet investigar les relacions entre una variable dependent i una o diverses variables independents. És a dir, l'objectiu és poder estimar el valor futur de la variable dependent tenint en compte els canvis de les variables independents.
- **Anàlisi predictiva.** S'empra per analitzar dades actuals i històriques i poder fer prediccions sobre el futur o sobre esdeveniments no coneguts.

L'ús d'aquestes eines està recomanat quan es vol dissenyar un model matemàtic o estadístic per elaborar simulacions, prediccions o sistemes de recomanació. El veritable aprofitament d'aquestes eines implica coneixements mitjans o avançats en estadística inferencial i probabilitat.

A continuació, mostrarem dues eines d'anàlisi de dades que han estat seleccionades per la seva àmplia acceptació per part de professionals de l'estadística en el món acadèmic i empresarial, i també perquè són aplicacions d'escriptori fàcils d'instal·lar.

- **RStudio.** És una eina d'ús lliure (gratuïta) i multiplataforma que permet emprar el llenguatge de programació R. Disposa de totes les funcionalitats necessàries per fer estadística descriptiva (també gràfics), inferencial i probabilística. Posseeix funcionalitats específiques per fer anàlisi d'agrupament, regressió i predicció (*https://posit.co/products/open-source/rstudio/*).
- **MatLab.** És un entorn multiplataforma amb propietari (de pagament) que funciona amb el llenguatge de programació MATLAB. Permet fer càlcul numèric, anàlisi i visualització de dades, programació i desenvolupament d'algorismes. La seva funcionalitat és més àmplia que R, tot i que ofereix característiques similars. Posseeix funcionalitats específiques per fer anàlisis d'agrupament, regressió i predicció (*https://www.mathworks.com/products/matlab.html*).

3.4 Eines de visualització

La visualització engloba les tècniques que s'empren per crear imatges, diagrames o animacions. El seu objectiu és crear una representació visual que ajuda a transmetre un missatge.

En el llibre *The Visual Display of Quantitative Information* (1983, Edward Tufte), es defineix l'efectivitat de les visualitzacions com a idees complexes que són comunicades amb claredat, precisió i eficiència, i que permeten l'anàlisi visual, la comparativa i la causalitat.

Segons el missatge que es vol comunicar, es poden definir unes visualitzacions o unes altres. A continuació, repassam algunes d'aquestes visualitzacions:

- **Visualitzacions exploratòries.** Es construeixen sobre una anàlisi descriptiva de dades i tenen l'objectiu de comunicar propietats (rellevància, relacions, etc.), patrons (tendències, correlacions, etc.) o estratègies de modelatge de les dades visualitzades.
- **Visualitzacions expositives.** Volen transmetre com a missatge els resultats d'una anàlisi o investigació. L'objectiu és comunicar de forma visual la informació descoberta com a resultat de la investigació.
- **Aplicacions genèriques.** La representació visual de dades es fa amb gràfics típicament emprats en l'estadística descriptiva, com ara gràfics de barres, de columnes, de dispersió, d'àrea o de línies.
- **Visualització temporal.** Són visualitzacions especialment dissenyades per a la representació temporal.
  - **Línia del temps.** Són representacions visuals de successos ocorreguts al llarg del temps, normalment presentats de forma cronològica.

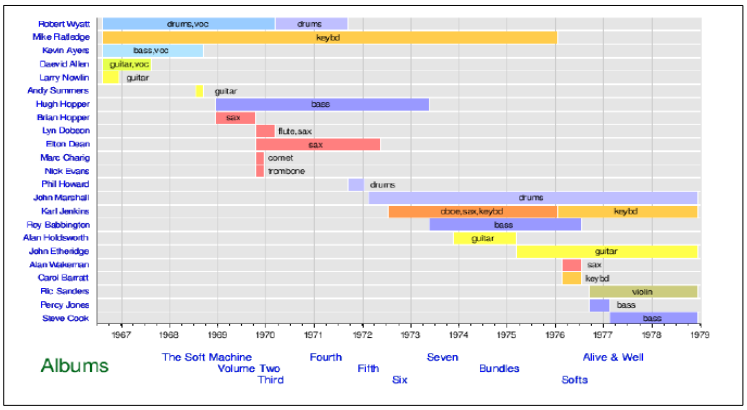

Figura 2. Línia del temps.

- **Diagrama de Gantt.** És la representació cronològica i ordenada de diferents tasques, en un temps total determinat. Per a cada tasca es defineix una dedicació de temps (hores, dies, setmanes, etc.).

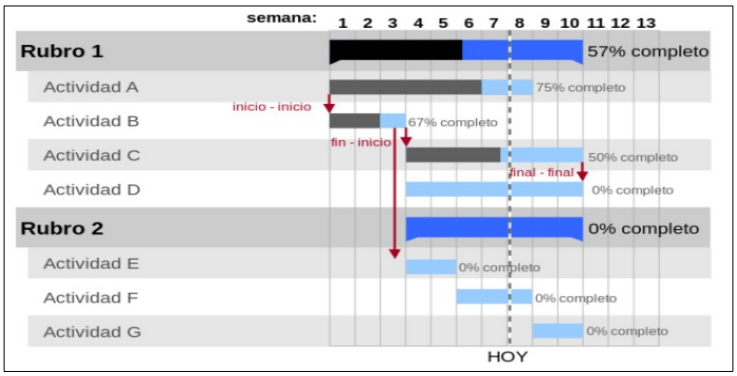

Figura 3. Diagrama de Gantt.

- **Visualització geoespacial.** És una representació de dades sobre mapes. En funció de la geometria emprada per distribuir les dades en el mapa, destacam els tipus de mapes següents:
  - **Mapa de distribució de punts.** Es basen en una dispersió visual de punts per mostrar un patró espacial. A més, els punts poden emprar diferents símbols, grandàries i colors per presentar la distribució d'altres variables.

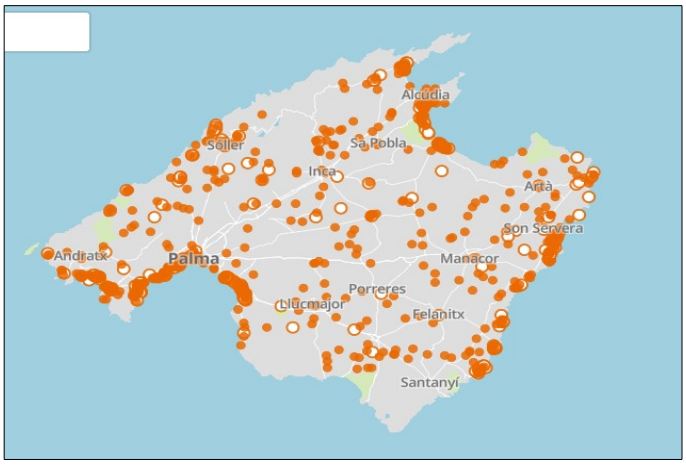

Figura 4. Mapa de distribució de punts.

- **Mapa de calor.** Presenta els valors individuals agrupats en àrees. Segons la seva proximitat s'assigna un color, i es representen les diferents proporcions d'una variable estadística.

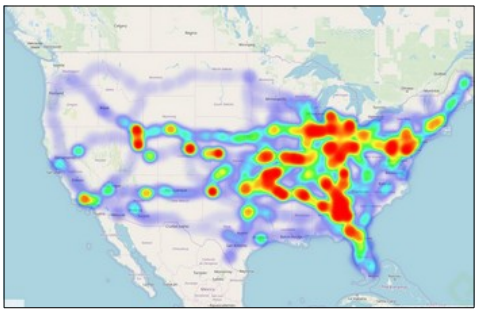

Figura 5. Mapa de calor.

- **Mapa coroplètic.** S'empren àrees ombrejades per presentar les diferents proporcions d'una variable estadística, molt sovint amb divisions administratives.

Figura 6. Mapa coroplètic.

A continuació, presentam quatre eines per fer anàlisis de dades. La selecció es basa en la facilitat d'ús (no cal programar), la disponibilitat de versions en programari de

lliure (gratuït) i la possibilitat d'inserir els resultats a web externs (per exemple, un blog o una pàgina personal).

Tableau Public, DataWrapper, RAW i Power BI ofereixen un ampli ventall de possibilitats per a la visualització de dades. Concretament, Tableau Public i DataWrapper són àmpliament emprats pels professionals del periodisme de dades.

- **Tableau Public.** És una aplicació d'escriptori amb propietari que permet elaborar gràfics de barres, columnes, línies, dispersió, mapes en arbre, bombolles, diagrames de Gantt i altres mapes. Permet, a més, dur a terme operacions d'enllaç de dades (*record linkage*) i anàlisi descriptiva. Els gràfics generats poden ser inserits dins pàgines web externes. Té disponible una versió gratuïta, però només permet treballar amb dades en format full de càlcul (*https://www.tableau.com/*).
- **DataWrapper.** És una aplicació web d'ús lliure (gratuïta) que permet elaborar gràfics de barres, columnes, línies, sectors i mapes. Facilita, a més, certes operacions de neteja de dades. Els gràfics generats poden ser inserits dins pàgines web externes (*https://www.datawrapper.de/*).
- **RAW.** És una aplicació web lliure (gratuïta) que facilita als usuaris sense coneixements de programació la creació de visualitzacions: dendrogrames, bombolles, mapa en arbre, mapa mental, coordenades paral·leles, etc. També té una llibreria de programació per a usuaris avançats que els permet crear tot tipus de visualitzacions interactives. Els gràfics generats poden ser inserits dins pàgines web externes (*https://www.rawgraphs.io/*).
- **Power BI.** És una aplicació amb propietari que permet unir diferents fonts de dades, analitzar-les i presentar-ne una anàlisi a través d'informes i panells. Aquestes anàlisis es poden fer des d'una aplicació d'escriptori gratuïta. Si les volguéssim compartir amb altres usuaris, les hauríem de carregar al repositori que té Power BI al nigul, que és de pagament. Des del nigul, diferents usuaris podrien treballar col·laborativament en la mateixa anàlisi (*https://powerbi.microsoft.com*).

Seguidament presentam una selecció d'eines de visualització geoespacial: CartoDB, Google My Maps i ArcGIS Online. S'han seleccionat en funció de la seva facilitat d'ús (a l'abast d'un usuari amb capacitació tecnològica mitjana) i de la possibilitat d'emmagatzemar els mapes en línia per compartir-los o inserir-los dins webs externs.

Aquestes tres eines tenen una llarga trajectòria i una àmplia acceptació per a projectes comercials i no comercials:

- **CartoDB.** És una aplicació web lliure que permet elaborar mapes de punts, de calor i coroplètics. Facilita, a més, certes operacions de neteja de dades i *record linkage*. Els mapes generats poden ser inserits dins pàgines web externes. La versió en línia de CartoDB permet emmagatzemar mapes de com a màxim 250 MB. Per a més espai, disposa de plans de pagament (*https://carto.com/*).
- **Google My Maps.** És una aplicació web lliure que permet elaborar mapes de punts, de calor i coroplètics. Facilita, a més, certes operacions de *data cleansing*

(neteja de dades) i *record linkage* (enllaç de dades). Els mapes generats poden ser inserits a pàgines web externes.

- **ArcGIS Online.** És una aplicació web amb propietari que permet crear i compartir mapes web interactius. Com que es basa en una arquitectura al nigul, permet que diferents usuaris puguin treballar col·laborativament i de manera simultània sobre el mateix mapa (*https://www.arcgis.com/index.html*).

A continuació, presentam dues eines específiques que permeten crear visualitzacions temporals: TimeFlow i TimelineJS. Els avantatges que presenten aquestes eines són la facilitat d'ús (a l'abast d'un usuari amb capacitació tecnològica mitjana), que són gratuïtes i que permeten inserir els resultats dins webs externs, com un blog o una pàgina personal.

- **TimeFlow.** És una aplicació d'escriptori multiplataforma que permet fer diferents tipus de visualitzacions temporals: línia del temps, calendari, taula i llista a partir de fulls de càlcul. Els gràfics generats poden ser exportats en format HTML (*http://flowingmedia.com/timeflow.html*).
- **TimelineJS.** És una aplicació web que permet fer línies del temps a partir de fulls de càlcul que provenguin de Google Drive. La representació visual final és una col·lecció cronològica de diapositives a les quals es pot afegir imatge o vídeo. Les línies del temps generades poden ser inserides en pàgines web externes (*https://timeline.knightlab.com/*).

Els gràfics de xarxes permeten descobrir xarxes i patrons existents en una sèrie de dades relacionades. Visualment, es componen de nodes i arestes. Una aresta que uneix dos nodes representa una relació existent entre tots dos.

Els gràfics de xarxes (o grafs) són objecte d'estudi de la matemàtica discreta, i s'han emprat per modelar i resoldre problemes, com ara trobar la distància més curta entre dos nodes, el mínim recorregut que passi per tots els nodes o l'agrupament per similitud o proximitat. Es tracta, per tant, d'eines que permeten simultàniament l'anàlisi i la visualització d'informació.

Les xarxes de dades s'empren tant per a l'exploració visual de les relacions entre nodes com per a l'anàlisi matemàtica i estadística d'aquestes relacions. Si l'objecte de la visualització o l'anàlisi no és la relació entre nodes, es recomana l'ús d'altres eines de visualització. Per poder emprar eines de visualització o anàlisi de xarxes de dades, és requisit previ tenir un conjunt de dades que contengui una llista tant de nodes com d'arestes (relacions entre nodes).

A continuació, presentam dues eines per fer la visualització de xarxes de dades: Graphviz i Gephi. La primera és una eina que ja hem presentat, i que permet una visualització molt bàsica de xarxes de dades de forma molt senzilla. La segona és una suite específica de visualització i anàlisi de xarxes de dades. S'han seleccionat en funció de la facilitat d'ús (a l'abast d'un usuari amb capacitació tecnològica mitjana) i pel fet de ser de llicència lliure (gratuïtes).

- **Graphviz.** És un programa d'ús lliure (gratuït) que permet visualitzar informació estructural mitjançant diagrames de gràfics i xarxes abstractes. Aquesta aplicació es pot emprar en diferents àmbits: bioinformàtica, enginyeria del programari, disseny de bases de dades i web, aprenentatge automàtic i altres dominis tècnics. Té moltes funcions útils, com ara opcions de colors, tipus de lletra, dissenys de nodes tabulars, estils de línia, enllaços i formes personalitzades (*https://graphviz.org/*).
- **Gephi.** És una aplicació d'escriptori multiplataforma que serveix per a la creació, la visualització i l'anàlisi de xarxes de dades. Permet importar dades des de fulls de càlcul i connexió directa a bases de dades. És capaç de treballar amb desenes de milers de nodes i fer qualsevol tipus d'anàlisi típica de xarxes: agrupament, camins mínims, etc. És de programari lliure multiplataforma i les visualitzacions que es generen poden ser exportades en format imatge i vectorial per ser incrustades en llocs externs (*https://gephi.org/*).

BIBLIOGRAFIA

ArcGIS Online. Disponible a: *https://www.arcgis.com/index.html*.

CartoDB. Disponible a: *https://carto.com/*.

Data Wrangler. Disponible a: *http://vis.stanford.edu/wrangler/*.

DataWrapper. Disponible a: *https://www.datawrapper.de/*.

Gephi. Disponible a: *https://gephi.org/*.

Graphviz. Disponible a: *https://graphviz.org/*.

Import.io. Disponible a: *https://www.import.io/*.

MatLab. Disponible a: *https://www.mathworks.com/products/matlab.html*.

Open Refine. Disponible a: *https://openrefine.org/*.

ParseHub. Disponible a: *https://www.parsehub.com/*.

PDF Tables. Disponible a: *https://pdftables.com/*.

Power BI. Disponible a: *https://powerbi.microsoft.com*.

Tableau Public. Disponible a: *https://www.tableau.com/*.

Tabula. Disponible a: *https://tabula.technology/*.

TimeFlow. Disponible a: *http://flowingmedia.com/timeflow.html*.

TimelineJS. Disponible a: *https://timeline.knightlab.com/*.

RAW. Disponible a: *https://www.rawgraphs.io/*.

RStudio. Disponible a: *https://posit.co/products/open-source/rstudio/*.