Col·legi Oficial
de Bibliotecaris-
Documentalistes
de Catalunya

La descripció de *datasets*

Pont entre les disciplines de la informació i la ciència de dades

Eva Torrent Castellano

Per què parlem de descripció avui?

La descripció de *datasets*

- Les disciplines de la informació —la biblioteconomia, la documentació i l'arxivística—descriuen recursos des de sempre: llibres, documents històrics o administratius, fotografies, registres sonors...

- La descripció és una tasca essencial que permet trobar, avaluar i interpretar els recursos

Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya

Per què parlem de *datasets* avui?

La descripció de *datasets*

- Definició de conjunt de dades (*dataset*):

  col·lecció estructurada d'informació organitzada de manera que pugui ser processada, analitzada o reutilitzada de forma sistemàtica. Habitualment es presenta en formats llegibles per màquina —com **CSV, JSON o XML**— i pot contenir dades de naturalesa diversa: textos, imatges, xifres, metadades...

- Avui, els *datasets* són un nou tipus de recurs que també necessita ser descrit
- Dos contextos on això és urgent:
  - la ciència de dades
  - el patrimoni cultural digitalitzat

Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya

Índex

Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya

El biaix algorítmic

La descripció de *datasets* dins la ciència de dades

La IA generativa perpetua els biaixos socials existents, hi ha casos detectats en: ocupació, sanitat, educació, participació política

Algorismes de selecció de personal que discriminen per nom o accent (Bertrand et al, 2004)

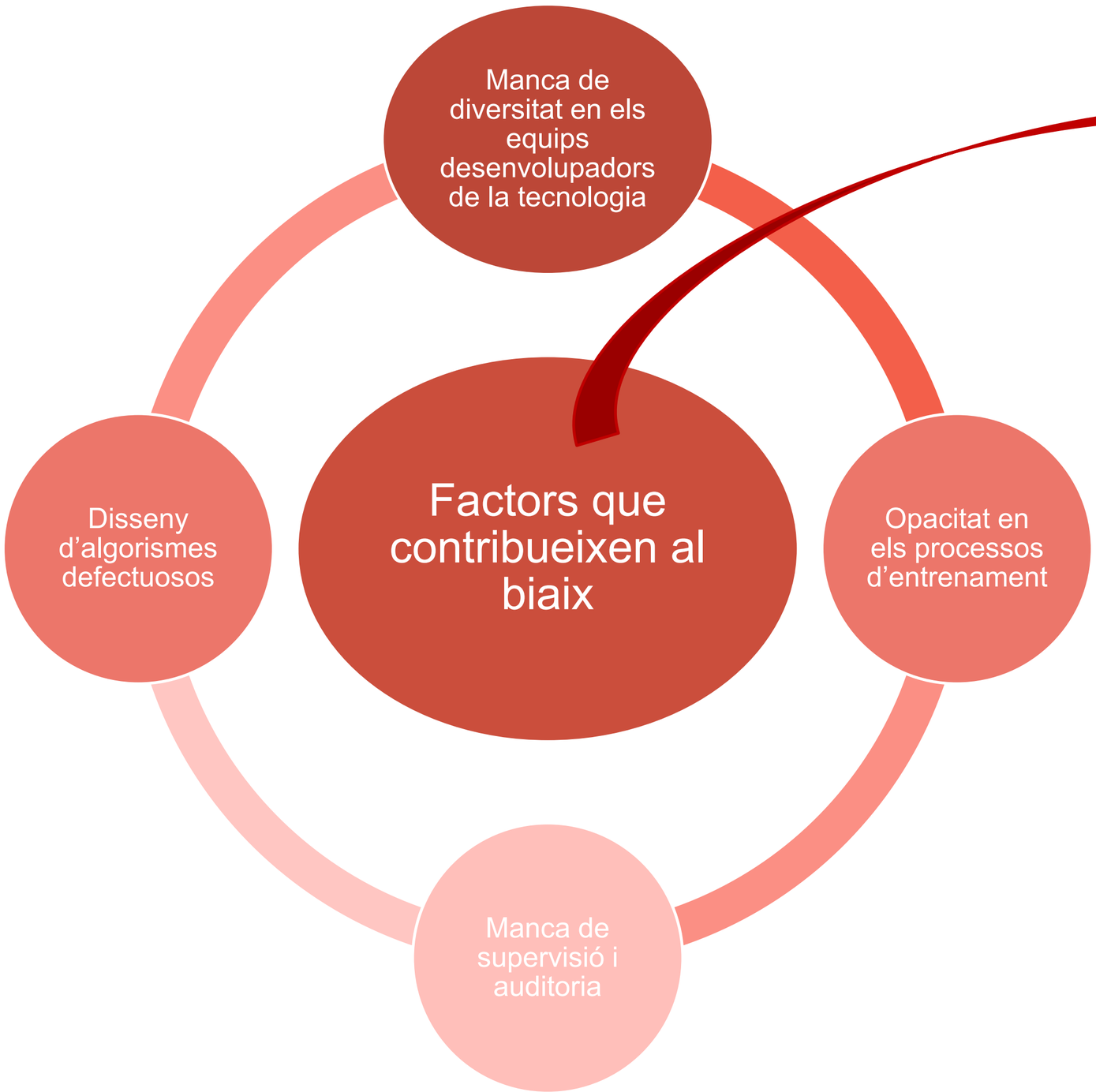
Sistemes de triatge sanitari que infravaloren pacients afroamericans (Obermeyer et al, 2019)

Predictors d'abandonament escolar amb taxes d'error superiors per a estudiants d'origen afroamericà i hispà (Gándara et al, 2023)

Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya

L’arrel del problema: les dades

La descripció de *datasets* dins la ciència de dades



Les dades d’entrenament

Un algorisme entrenat amb dades desiguals aprèn i automatitza la desigualtat

Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya

La solució: documentar les dades de partida

La descripció de *datasets* dins la ciència de dades

- De forma prèvia a la tria i selecció de les dades per entrenar un algorisme, convé conèixer els seus continguts i limitacions
- Eines concretes: **models de descripció**

- Models de descripció d'algorismes:
  - Cartes de model (*model card*)
  - Fitxa informativa (*factsheets*)
- Models de descripció de conjunts de dades:
  - Biografia de dades (*data biography*); cartes de dades (*data card*); declaració de dades (*data statement*); fitxa de dades (*datasheet*).

Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya

Conjunts de dades de patrimoni digitalitzat

La descripció de *datasets* patrimonials

- Nova metodologia en recerca històrica: mètodes computacionals sobre fons digitalitzats
- Institucions que publiquen *datasets*:
  - Library of Congress
  - Bibliothèque nationale du Luxembourg
  - Koninklijke Bibliotheek (Països Baixos)
  - British Library Labs

- Problema: cada institució ho fa d'una manera diferent: heterogeneïtat de formats, accessos, llicències

Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya

La necessitat de descriure els datasets

La descripció de *datasets* patrimonials

- Un investigador que vol usar un *dataset* d'una biblioteca no sap:
  - Quin percentatge del fons original ha estat digitalitzat
  - Amb quins criteris es van prioritzar uns materials per sobre d'altres
  - Quin processament han rebut les dades
  - Quines llicències s'hi apliquen
  - Quins col·lectius estan infrarepresentats o absents

- Sense aquesta informació, el risc és el mateix que a la ciència de dades: conclusions esbiaixades amb aparença de rigor

Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya

Model de descripció “Datasheets for Digital Cultural Heritage”

La descripció de *datasets* patrimonials

- Un conjunt d’autors d’Europeana i biblioteques universitàries europees han adaptat el model del *datasheet* per a institucions patrimonials (Alkemade et al, 2023)

- Compatible amb DCAT
- Camps: dades bàsiques · distribució · composició · procés de recollida · vocabularis controlats · manteniment · consideracions ètiques i biaixos · exemples d'ús

Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya

Model de descripció “Datasheets for Digital Cultural Heritage”

La descripció de *datasets* patrimonials

Títol

[Proporcioneu un títol del conjunt de dades.]

Descripció

[Proporcioneu una breu descripció del conjunt de dades. Resumiu de manera concisa el conjunt de dades, el seu ús previst i les tasques que permet o a les que dona suport. Ofereix una visió general dels continguts del conjunt de dades (dades i metadades, només dades o metadades; originades a partir de fons físics, digitalitzats o originalment electrònics, etc.), com i per a què ha estat creat el conjunt de dades. El resum hauria de descriure el domini, tema, àmbit, paraules clau i altres metadades rellevants. Articula de forma clara las raons de creació del conjunt de dades i promou la transparència entorn els interessos de finançament. Per a quin propòsit es va crear el conjunt de dades? Hi havia alguna tasca específica en ment? Hi havia algun buit o necessitat concreta que calgués cobrir? Qui ha creat el conjunt de dades (per ex. quin equip o grup de recerca, etc.) i de part de quina entitat (per ex. empresa, institució, organització)? Qui ha subvencionat la creació del conjunt de dades?]

Pàgina web

[Incloeu la URL de la pàgina web si està disponible.]

Editors

[S’ha d’indicar l’editor individual o l’organització responsable —preferiblement aquesta darrera. Tingueu en compte que probablement serà diferent dels “Curadors del conjunt de dades” i dels “Altres col·laboradors”, que s’han d’emplenar en altres seccions.]

Curadors del conjunt de dades

[Enumereu les persones clau implicades en la recopilació del conjunt de dades i la seva afiliació professional. Es recomana que els curadors del conjunt de dades incloguin informació sobre el seu càrrec, per exemple en forma d’una breu biografia amb dades sobre la seva formació, rol i funció dins de la institució de patrimoni cultural[59] o del projecte, així com les tasques assumides en la creació del conjunt de dades.]

Altres contribuents

Model de descripció “Datasheets for Digital Cultural Heritage”

La descripció de *datasets* patrimonials

Procés de recopilació de dades

[Obtingueu informació que pugui ajudar investigadors i professionals a reutilitzar aquest *dataset* o crear-ne de similars.]

Raó de la curació

[Quina necessitat va motivar la creació del *dataset*? Quines van ser algunes de les raons que van influir en les decisions preses en la seva creació?]

Font de les dades

[Aquesta secció descriu les dades d’origen (per ex., textos de notícies i titulars, frases traduïdes, etc.).]

Recopilació inicial de dades

[Descriviu el procés de recopilació de dades. Expliqueu criteris per la selecció, filtratge o extracció (en particular, llistant paraules clau o termes de cerca utilitzats). Si és possible i rellevant, incloeu informació de temps d’execució del procés de recopilació. Si les dades es van obtenir d’altres *datasets* preexistents, enllaceu la font aquí.]

Agents productors de la font de les dades

[Indiqueu si les dades o objectes d’origen van ser produïts per humans (com ara autors de llibres) o generats per màquina i proporcioneu-ne una descripció. Si està disponible, incloeu informació demogràfica o d’identitat dels creadors de les dades reportada per ells mateixos, però eviteu inferir-la. En lloc d’això, indiqueu que aquesta informació és desconeguda. Vegeu aquest article[68] per a l’ús de categories d’identitat com a variables, especialment de gènere.]

Cadena de digitalització

[Indiqueu informació (*paradata*) sobre el procés de digitalització, rellevant per a la reutilització; per exemple, motivació de la digitalització (conservació, projecte de recerca, etc.) i metadades tècniques com equipament i programari utilitzat per capturar imatges i/o el processament OCR/HTR[69], i/o el nivell de qualitat dels resultats. Si escau, descriviu com la digitalització constitueix una capa addicional de selecció respecte el fons complet disponible en la institució patrimonial; indiqueu criteris i mètriques de selecció que mostrin com això ha afectat a la transformació del fons en un conjunt de

Beneficis

La descripció de *datasets*

Per què s'han de descriure els conjunts de dades?

Descoberta i selecció: saber si un dataset és adequat sense haver-lo d'examinar sencer

Transparència i reproductibilitat: auditar algorismes, reproduir estudis, detectar errors

Responsabilitat ètica: documentar les limitacions d'un *dataset* és un requisit bàsic per la transparència

Gestió del coneixement a llarg termini: les col·leccions d'avui tindran usos imprevistos demà

Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya

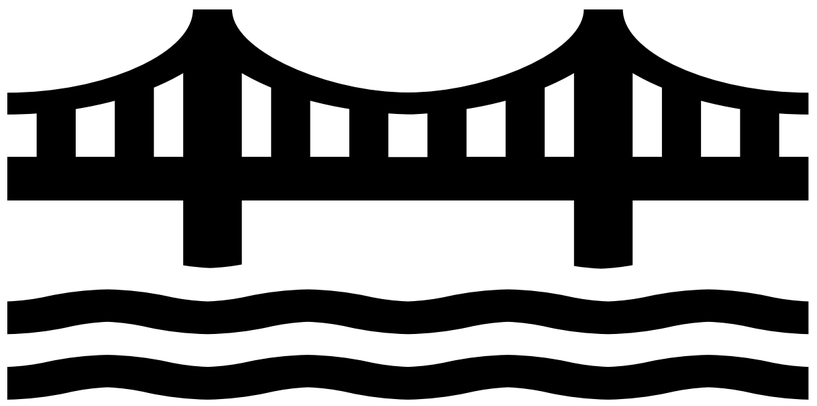
Conclusions

La descripció de *datasets: pont entre les disciplines de la informació i la ciència de dades*

- La descripció és una tradició consolidada de les disciplines de la informació
- Avui, aquest saber professional és necessari ja que documentar *datasets* és condició per dos camps:

Ciència de dades

Sistemes d'IA més justos i auditables



Disciplines de la informació

Recerca computacional més rigorosa

Col·legi Oficial de Bibliotecaris-Documentalistes de Catalunya