Conjunts de dades, gestió documental i arxius: proposta de tractament

María del Pilar Campos Martínez,
Arxivera de l'Institut Municipal d'Hisenda de Barcelona, Ajuntament de Barcelona

Resum

L'article és una introducció conceptual a la gestió de dades i la seva relació amb la gestió documental, identificant reptes i propostes de solucions per a l'aplicació de les funcions arxivístiques sobre *datasets*. Es presenta la publicació *Directrius per al tractament de conjunts de dades als arxius*, en què es fa una proposta de la revisió de la qualitat de les dades prèvia a la transferència.

PARAULES CLAU: *datasets*, conjunts de dades, qualitat de les dades, gestió documental, gestió de dades, arxiu de dades

Résumé

L'article est une introduction conceptuelle à la gestion des données et à sa relation avec la gestion documentaire, identifiant les défis et les propositions de solutions pour l'application des fonctions d'archivage sur les jeux de données. La publication *Directives pour le traitement des jeux de données dans les archives* est présentée, où une proposition de révision de la qualité des données avant leur transfert est faite.

MOTS CLÉS: *datasets*, jeux de données, qualité des données, gestion documentaire, gestion des données, archivage des données

Resumen

El artículo es una introducción conceptual a la gestión de datos y su relación con la gestión documental, identificando retos y propuestas de soluciones para la aplicación de las funciones archivísticas sobre *datasets*. Se presenta la publicación *Directrices para el tratamiento de conjuntos de datos en los archivos*, donde se hace una propuesta de la revisión de la calidad de los datos previa a la transferencia.

PALABRAS CLAVE: *datasets*, conjuntos de datos, calidad de los datos, gestión documental, gestión de datos, archivo de datos

Abstract

This article is a conceptual introduction to data management and its relationship to records management. It seeks to identify challenges and proposals for solutions regarding the application of archival functions to datasets. The publication *Guidelines for the treatment of datasets in archives* is presented, which proposes reviewing the quality of data prior to its transfer to an archive.

KEYWORDS: dataset, data quality, records management, data management, data archiving

Introducció

La transformació digital va deixar de ser innovació per ser una obligació legal a les administracions de l'estat amb la Llei 11/2007 d'accés electrònic dels ciutadans als Serveis Públics (derogada) i, posteriorment i de forma definitiva, amb les Lleis 39/2015 del procés administratiu comú. A l'entorn empresarial, de ser una oportunitat de negoci a tenir un cost d'oportunitat tant alt que no assumir-ho comporta la desaparició. Per tant, hem de trencar el discurs de "noves tecnologies" quan es parla de documents digitals i d'innovació el fet de treballar amb sistemes de bases de dades, al cap i a la fi, estem parlant de sistemes popularitzats els anys 80 (normalització com estàndard ISO del llenguatge SQL és de 1987). El treball amb dades porta temps a les organitzacions, però és en els darrers anys que pren una nova dimensió al tenir la capacitat de visualitzar-les i explotar-les d'una forma més àgil i situant-les en el centre de la estratègia empresarial i social.

En el nou mercat de les dades la mateixa Unió Europa ha establert que aquesta sigui la «Dècada Digital» d'Europa, amb una inversió en sobirania digital, establiment normatiu i tecnològic centrat en les dades (Reglament 2021/694). Al gener del 2022 la Comissió proposava la *Declaració Europea de Drets Digitals i Principis de la Dècada Digital* que marca els principis que regularà la política digital, i al juny de 2022 va entrar en vigor la Llei de governança de dades (Reglament 2022/868) que serà plenament aplicable a partir de setembre de 2023. Per tant, estem en un moment on a nivell legislatiu i d'inversions hi ha un impuls a passar a entorns governats per dades.

La transformació, per tant, ja no és del món analògic paper al món digital, sinó evolucionar a un entorn on els documents no es trobin estructurats i en llenguatge natural, passant a documents estructurats i dades enllaçades. Per la gestió documental representa canvis més profunds que la simple migració de format i ens hem de preguntar si sabrem traduir les nostres formes de treball.

Conceptes previs

En aquest article s'utilitzen els termes següents amb el significat que s'indica. En alguns casos, per tal d'evitar errors, s'han separat termes que són pràcticament sinònims de *dada* però que tenen un matís diferenciat.

Base de dades (angl. *database*). 1. Col·lecció de dades amb una estructura per ingerir, emmagatzemar, proveir o demanar dades per múltiples usuaris. 2. Col·lecció de dades interrelacionades organitzades segons un esquema de bases de dades per donar servei a una o més aplicacions. (Butterfield *et al*., 2016). Per definir, manipular i accedir a les dades de la bases de dades, així com mantenir la seva integritat o seguretat es fa mitjançant un **Sistema de Gestió de Bases de Dades (SGBD)** que actua com 'interfície' entre usuari i base de dades. Per a la simplificació, a aquest article ho tractarem com un conjunt sota el concepte 'base de dades'.

Conjunt de dades o *dataset*: Conjunt estructurat de dades creades per a un propòsit específic. Els conjunts de dades es poden emmagatzemar, administrar i publicar en una varietat de formats i tecnologies. (The National Archives, 2011)

La diferència amb la base de dades és que la primera inclou lògiques internes d'execució, explotació, cerca i interrelació de dades que no hi ha en el *dataset*; sovint, els *datasets* són taules exportades de la base de dades.

Dades (angl. *data*): Representació interpretable d'informació de manera adequada per a la comunicació, la interpretació o el processament. (ISO/IEC, 1993)

Dada (angl. *data item*): Ítem que es considera indivisible en un context concret. És una unitat de dades per a la qual la definició, la identificació, la representació i els valors permesos s'especifiquen mitjançant un conjunt d'atributs (metadades). (ISO/IEC, 1994)

Dades estructurades (angl. *estructured data*): Dades organitzades amb diferents elements establerts mitjançant un estàndard o model de dades. Les dades estructurades generalment es refereixen a dades atòmiques i que es troben en diferents camps segons el seu caràcter o ús. Alguns exemples són les bases de dades relacionals, els fulls de càlcul. En comparativa, XML és un contenidor que pot donar estructura a conjunts de dades utilitzant etiquetes (Interpares Trust AI, 2022).

Datalake: Una col·lecció de grans conjunts de dades en el seu format original, sense refinar, normalitzar o tractar, que permet als usuaris la cerca, selecció i anàlisis de les dades originals en seu context de creació (Interpares Trust AI, 2022).

Data wharehouse: repositori que guarda informació de diverses fonts, que ha sigut extreta, normalitzada i transformada segons un esquema comú i que dona suport a anàlisis predefinits i informes. (Interpares Trust AI, 2022).

La diferència entre aquest i l'anterior és que el *datalake* té les dades originals tal i com es van generar o tramitar, en aquest segon hi ha una transformació que permet la explotació més fàcilment essent un element clau per a les aplicacions de *business intelligence*.

Diccionari de dades (angl. *data dictionary*): Repositori centralitzat d'informació sobre dades com ara el significat, les relacions amb altres dades, l'origen, l'ús i el format. Dona suport a la gestió, als administradors de bases de dades, als analistes de sistemes, als programadors d'aplicació i a la planificació, el control, l'avaluació, l'emmagatzemament i l'ús de les dades. (IBM, 2001)

Document (angl. *Record*): s'entén per document tota expressió en llenguatge oral, escrit, d'imatges o de sons, natural o codificat, recollida en qualsevol mena de suport material, i qualsevol altra expressió gràfica que constitueixi un testimoni de les funcions i les activitats socials de l'home i dels grups humans, amb exclusió de les obres d'investigació o de creació. (Llei 9/1993 del Patrimoni cultural català). Per ampliar la definició, s'inclou l'apartat 4. d'informàtica per on s'entén document el conjunt de dades relacionades tractades com una unitat, com els camps d'una taula a una base de dades (SAA, 2005-2022)

La gestió de les dades i la gestió documental

Una de les primeres accions per aplicar metodologia a la gestió de conjunt de dades és conceptual: entendre la diferència entre gestió de les dades i la gestió documental. Quan s'entra a parlar de la gestió de les dades és fàcil caure en la confusió d'unificar les accions que en el document analògic estaven més clarament diferenciades: creació, tractament, accés i explotació. Al produir-se gairebé simultàniament, la separació per fases deixa de tenir sentit, i en molts casos és difícil diferenciar també si es treballa sobre diferents entorns tecnològics ja que les integracions faciliten no haver de canviar de plataforma. Per poder veure quin rol tindrà la gestió documental cal poder definir les accions i identificar a quina part de la gestió de les dades s'ha de poder incidir des de la gestió documental:

- **Creació:** en un entorn administratiu és la tramitació de l'expedient, el moment en el que la organització ingesta la dada al sistema i la treballa. Aquesta pot venir de múltiples fonts, com veurem més endavant. El propietari de les dades és el negoci o organització que les està consumint i necessitant.

- **Tractament:** és la definició de les dades, el seu context, relació entre d'altres, establiment del cicle de vida, accessos, etc. En entorns analògics el tractament es fa després de la ingesta dels documents a l'arxiu i amb l'administració electrònica s'evoluciona per a que s'apliqui de forma automatitzada i des del disseny. En entorns datificats, el tractament es troba amb la definició del catàleg de dades i els flux de treball en primera instància, i posteriorment amb les normes d'avaluació, preservació, migracions, anonimització, avaluació, etc. El tractament es troba en el disseny del sistema i l'establiment de regles.

- **Accés i explotació:** la consulta de les dades, ja sigui pels mateixos usuaris creadors com per altres sistemes o la analítica i la explotació d'aquestes dades. Són el producte derivat de les dades que s'han creat i es poden plasmar a quadres de comandament, visualitzacions de dades o informes per a la presa de decisions. Aquí entren els processos d'extracció de dades, transformació i càrrega a altres sistemes (ETL), l'accés per interoperar dades per un tercer, etc. sempre i quan no hi hagi una gravació o transformació al nostre sistema origen.

La gestió documental aplica a la part de tractament, i aquesta s'influirà tant pels requeriments dels usuaris de negoci com dels usuaris que tenen accés a les dades. No és objecte de la nostra disciplina la creació de les dades de negoci o la tramitació dels expedients, però tampoc l'analítica o visualització que queden pels àmbits dels científics de dades, comandaments, periodistes de dades o sistemes de tercers. Però tal i com es tramita o es consumeixen les dades, influirà en el disseny i, per tant, en el tractament. És encara més imprescindible estar en els equips multidisciplinaris per al disseny de sistema, sense perdre de vista l'abast i responsabilitat.

Abast: dades i bases de dades

Les dades no es troben de forma abstracta disperses a les organitzacions, estan estructurades en bases de dades, taules sobre les que les aplicacions interactuen i els usuaris hi poden treballar. La complexitat d'aquesta estructura estarà marcada per l'arquitectura que s'hagi dissenyat, els volums de tramitació o l'abast del sistema. Sobre aquestes bases de dades no només hi ha les taules, també hi ha la definició de les pròpies dades (encara que sigui en el nom de capçalera de les taules o el seu format (per exemple, si són camps numèrics, format de data, longitud), si ha de ser unívoc (per exemple, si és el codi identificador), etc. però a més de la definició de les dades hi pot haver moltes més lògiques de negoci: camps calculats (per exemple, que calculi la diferència entre el sou d'un contribuent i el salari mínim interprofessional), aspectes relacionats amb l'accés (per exemple, qui pot fer els canvis sobre certa taula), la seguretat (com es registren els canvis) i un llarg etcètera. Les funcionalitats i lògiques de les bases de dades, per tant, aporten el valor al funcionament de les taules, i no entrem a la creació de formularis, consultes, informes, interfícies de treball i altres experiències d'usuari.

La dada (en singular) i les dades (en plural) necessiten de la resta de dades per tenir context i significat. La seva desvinculació total del sistema que les ha creat ha d'anar acompanyada d'una metodologia que garanteixi la seva integritat i autenticitat, puguin ser interpretables i mantinguin el sentit.

A nivell de gestió documental s'ha de tornar a reflexionar sobre quin és l'objecte de la nostra disciplina: el sistema que suporta les dades, o les dades en sí mateixes, tenint en compte la dificultat d'aquesta desvinculació i si la pèrdua d'informació que pot haver és assumible. Però aquest dilema no ha d'impedir la ingesta de dades als arxius mitjançant extraccions, ja que la pèrdua possible per la desvinculació, és menor que la pèrdua per no preveure la seva transferència. Quan fem la extracció de la base de dades, aquestes taules passen a ser conjunts de dades o *datasets* que podem tractar des de l'arxiu.

L'entorn tecnològic ja està trobant sortides per a l'emmagatzemament de dades i que es troben mig camí entre l'arxiu (tot i que de vegades, en el seu concepte "dipòsit") i l'anàlisi de dades. Els *datalake* i *data wharehouse* són dos entorns on transferir els conjunts de dades externs sistema de gestió que els ha creat i la seva funció originàriament no és pròpiament la de la gestió documental i arxiu però conèixer el seu funcionament ens pot aproximar a la solució que necessitem.

Conjunts de dades als arxius

Origen de les dades

A l'arxiu poden ingestar-se típicament des de tres orígens diferents:

- **Dades de la organització:** En els sistemes de gestió documental la procedència habitual és l'activitat de la mateixa organització que genera les dades com a fruit de les seves funcions, per tant la captació d'aquests conjunts per s'hauria de fer de manera natural (The National Archives, 2011). Els fons de l'arxiu han de ser més amplis i no només s'ha de restringit a l'àmbit del procediment administratiu i procediments reglats ja que quedarien fora les dades de sistema, sensors, fotografies, analítiques web, sistemes BIM, comunicacions, etc. valuoses per la institució productora i que formen part del seu patrimoni documental.

A l'arxiu poden arribar dins la pròpia política de transferència establerta pel sistema de gestió documental al ser el repositori de tots els documents-dades que es generin, per extraccions periòdiques de dades o per transferències extraordinàries fruit, per exemple, d'una migració de sistemes o discontinuïtat d'un programari.

- **Dades-producte:** són les dades creades a partir d'altres dades de l'arxiu. Són el producte del tractament de les mateixes i poden ser tant explotacions directes de les dades de negoci o històriques. A l'arxiu trobaríem, per exemple, el dataset producte de la extracció d'inventaris, guies o els catàleg de l'arxiu que ja es troben en format dades estructurades i que es poden oferir com ja es fa amb les taules dels quadres de classificació o els registres d'autoritat. Però a banda de les dades d'eines arxivístiques, l'arxiu pot tenir la extracció en format dades de part del seu arxiu analògic. Un cas exemplar és el llibre del Sindicat de Remences de 1448 passat a format dades obertes per l'Arxiu Municipal de Girona Aquestes dades si bé no són fruit directe de funcions originals, sí són dades a tenir en compte ingestar a l'arxiu i tenir-les accessibles.

- **Dades-donació:** els arxius poden custodiar no només els fons producte de les funcions i activitats de la seva organització sinó també fons de tercers, especialment els arxius històrics. Dins les donacions de fons documentals també podran haver extraccions de bases de dades privades o particulars, per exemple, el fons històric de publicacions d'una xarxa social o la comptabilitat d'una empresa sobre un determinat programa. Aquestes dades seran reflex de les funcions i activitats d'una organització o poden ser

col·leccions, el tractament serà similar al primer conjunt però al tenir un origen no directament vinculat a la organització titular de l'arxiu, possiblement no es poden aplicar els mateixos requisits de transferència o controls previs.

L'origen i la forma d'extracció dels conjunts de dades caldrà que es trobi documentat abans de fer la ingesta a l'arxiu, a més d'una sèrie de recomanacions de qualitat de dades que tractarem més endavant.

Les funcions arxivístiques i els conjunts de dades

La UNE-ISO 15489 ens marca les funcions arxivístiques que ha de tenir un sistema de gestió documental: incorporació i registre, classificació i descripció, assignació de categories d'accés i seguretat, avaluació i disposició, emmagatzemament i accés. En els casos de les extraccions de conjunts de dades, les funcions poden mantenir-se però hi ha alguns aspectes on s'acaba entrant en incongruències si seguim la metodologia actual de gestió documental basada en entorns docucèntrics paral·lels als expedients en suport físic.

Per tal de fer una primera aproximació de com tractar els conjunts de dades i acotar els reptes per proposar solucions, ens aproximarem des del model 'reactiu' d'ingesta dels conjunts de dades una vegada finalitzada la seva tramitació i com productes extrets del sistema de gestió. Aquest model és oposat al que s'implanta als sistemes de gestió documental en entorns digitals on la relació entre creació del document i el gestor documental és en temps real: la creació ja implica que aquest es desi al gestor. Replicar aquest model en l'entorn de les dades seria complex i poc sostenible ja que al final estaríem duplicant l'aplicació de gestió amb les dades tant al sistema de gestió com a "l'arxiu únic". Un altre escenari és entendre que el lloc on es troben les dades de gestió (ja sigui durant la tramitació com una vegada finalitzades, com els *datalake,* són també responsabilitat de l'arxiu però cal madurar la seva gestió).

La ingesta de dades a l'arxiu vindrà, per tant, donada pels diferents orígens que s'han tractat anteriorment, en transferències periòdiques (com extraccions) o extraordinàries (com donacions, creacions de dades ex-professo o accions tecnològiques concretes). A la ingesta és on s'aplicaran el conjunt de polítiques, metadades i requeriments d'accés i ús que identifica la UNE-ISO 15489 i la política de gestió de documents de cada organització. Aquest és el moment de incidir en la qualitat de les dades ja que és quan es té relació directa amb el productor.

Pel que fa a la classificació a priori no hauria de ser diferent a la que s'aplica a altres documents i suports: els *datasets* es generen en un context determinat i per una funció. Per exemple: el *dataset* del padró d'habitants és la evolució del cens d'habitants que abans es trobava en llistats i prèviament en llibres de registre, o els exàmens que es fan via l'aula virtual i allà es posen les notes, la extracció d'aquestes dades estarà vinculada a les sèries d'exàmens i l'avaluació acadèmica.

En aquests casos la classificació funcional i descripció d'aquests conjunts de dades és senzilla i es pot fer assignant les metadades descriptives al fitxer, amb descripcions amb estructures multinivell (sèrie – unitat documental) tenint en compte que les jerarquies poden ser més flexibles ja que un dataset pot ser un conjunt d'expedients. Revisant diferents arxius nacionals, es detecta que no utilitzen sistemes de descripció diferents per a les dades que per a la resta de documents dels seus fons (Niu, 2016). De fet, la norma de descripció ISAD(G) (ICA, 2001) permet la descripció dels conjunts de dades i és la que utilitza la Secció de Datasets Governamentals dels Arxius Nacionals del Regne Unit. Des de l'any 2000 amb una proposta d'adaptació on s'amplia el quadre de classificació dos nivells per sota de la sèrie: el nivell «expedient» que seria la captura concreta del conjunt de dades, i un nivell múltiple anomenat «ítem», en el qual es descriurien les diferents taules que componen el conjunt de dades si estan desvinculades entre elles partint d'un model relacional (Shepherd, Smith, 2000).

Les dificultats respecte a la classificació i descripció és en les extraccions de dades que responen a diverses funcions o sèries documentals i que s'han gestionat amb el mateix sistema o no té sentit, en un format de dades, la seva separació. Un exemple podria ser la extracció de dades d'un sistema que gestiona la contractació administrativa d'una institució: el dataset contindrà tant la contractació menor, les obres, els subministraments, els serveis, etc. i és possible que a un camp indiqui quin tipus d'expedient era. Un altre cas pot ser les llicències d'obres majors i menors d'un ajuntament o la base de dades de diferents ajudes: en un format lògic sí es veu clarament un flux de treball i una funció associada, en entorns datificats, poden acabar en un mateix sistema compartint la mateixa estructura de dades i diferenciant-se únicament per alguns valors en camps. En aquests casos es pot valorar fer la extracció separant les dades segons funció i no mantenir-les agrupades o, per contra, considerar-los documents recapitulatius com seria un llibre de registre d'entrada i sortida d'un organisme. S'ha d'analitzar cada cas però estàndards que treballen la descripció arxivística amb eines ontològiques que permeten descriure no només el document sinó també el seu context mitjançant relacions, com s'espera que sigui el model conceptual *Records in Context*, per adaptar-se a les necessitats de flexibilitat de la descripció d'aquestes noves agrupacions documentals (Pastor, 2017). La descripció passada a un model conceptual més similar a les dades enllaçades que a jerarquies pot ser d'utilitat.

Sobre la descripció assenyalar que els documents digitals no només són el fitxer sinó també el conjunt de metadades que el descriuen. El mateix passa amb els datasets i pot ser rellevant incorporar no només les metadades descriptives que es trobin recollides als esquemes de metadades i normes de descripció de la organització sinó també aquelles metadades que descriuen el contingut i context de creació, la explicació del camps i altres dades rellevants per a la seva comprensió.

Més difícil és adaptar la metodologia actual a l'avaluació de les dades. Si centrem la problemàtica en els conjunts de dades que ingestin a l'arxiu (i no l'avaluació a sistema, que seria un altre escenari) poden ser dades de conservació permanent i, per tant, les accions que s'han de fer són de preservació. Per contra, si són dades que en algun moment es podrien eliminar tenim la següent situació:

- **Dataset d'una única sèrie documental i del mateix any**: s'aplica la disposició quan marqui la norma.

Exemple: una extracció anual de totes les devolucions d'ingressos indeguts d'un municipi. S'eliminaria completament el dataset passat els 6 anys des de la data de remissió a la Sindicatura de Comptes o òrgans de control extern.

- **Dataset d'una única sèrie documental i diversos anys o diferents terminis de conservació :** la disposició és més difícil ja que la integritat del dataset quedaria en dubte si s'eliminen dades parcialment. Aquest cas es pot donar si s'ha fet una extracció massiva per una migració de dades o no és una extracció periòdica. En aquest cas, segurament caldria esperar el termini màxim per a fer la eliminació o documentar la modificació. Igual passaria si conté dades d'un expedient afectat per un recurs o inspecció i que aturen els terminis d'aplicació de la norma.

Exemple: una extracció de totes les devolucions d'ingressos indeguts de diversos anys o amb les dades d'una devolució que s'ha recorregut. No es podria fer la eliminació completa del dataset fins passats els anys màxims que apliqui la norma.

Aquest escenari entra en conflicte amb la obligatorietat d'eliminació de les dades personals una vegada finalitzats els terminis i, per tant, cal que el Delegat de Protecció de Dades de la organització estigui informat.

Si la norma marca un mostreig entre els expedients (per exemple, guardar una mostra dels expedients acabats en un any o expedients que compleixin certes condicions) caldria fer-ne la extracció del data set original. En aquest cas, seria més convenient que fossin dues extraccions: una amb els expedients de conservació i una altra amb els que seran d'eliminació per facilitar-ne la gestió.

- **Dataset de diverses sèries documentals, mateix any i mateix termini de conservació**: Si l'abast és anual, únicament s'ha de tenir en compte per poder informar correctament la eliminació per sèrie documental al registre d'eliminacions.

- **Dataset de diverses sèries documentals i diferents terminis de conservació**: és en aquest cas on fer l'avaluació esdevé impossible sense trencar la integritat de la extracció de dades. És per això que cal seguir els passos de fer macroavaluació i anar a les grans funcions de les organitzacions ja que segurament les extraccions de dades dels sistemes estaran relacionades entre sí per funcions.

Contingut del dataset	Terminis avaluació	Abast temporal	Problema	Solució provisional
Una sèrie	El mateix termini	Anual	No hi ha problema si totes les dades tenen igual termini.	
Una sèrie	Diversos terminis	Diversos anys o expedients afectats per recursos	No es pot aplicar parcialment la eliminació	Aplicar terminis màxims o demanar extraccions "a mida".
Diverses sèries	El mateix termini	Anual	No hi ha problema si totes les dades tenen igual termini.	**Únicament poder diferenciar-les per informar el registre d'eliminacions**
Diverses sèries	Diversos terminis	Diversos anys o expedients afectats per recursos	No es pot aplicar parcialment la eliminació	Aplicar terminis màxims

Taula *1*: Avaluació dels conjunts de dades segons contingut (sèries), terminis d'avaluació i abast temporal. Identificació possible problemàtica i proposta d'actuació o solució provisional.

Sol·licitar les extraccions de dades de sistema vinculant-les a l'avaluació documental i terminis de conservació pot ser una solució temporal que faciliti la funció a l'arxiu que ha de gestionar aquests conjunts de dades. Cal tenir en compte que a la extracció que en fem ja s'haurà d'estudiar si correspon a sèries documental, una tipologia que apunta a diverses sèries, és un registre, etcètera. Si es pot delimitar per sèries, sol·licitarem la extracció de forma que sigui més senzill l'accés i disposició.

En els casos de conservació permanent, la gestió de la preservació d'aquestes dades és més senzilla que altres documents al tenir formats tabulars molt estàndards com són el csv, xml o el json. La conversió de formats ofimàtics és una pràctica senzilla i no ha d'implicar pèrdues de dades, sempre revisant que no s'hagin produït errors ni en el moment de la extracció ni en la migració.

Directrius per al tractament de conjunts de dades als arxius

El textos *Directrius per al tractament de conjunts de dades als arxius* publicat per l'Associació de Professionals de l'Arxivística i Gestió de Documents de Catalunya l'any 2022 (Campos, 2022) tenia com objectiu exposar el repte d'adaptació de les metodologies actuals a les funcions de la gestió documental i arxivística que continuen plenament vigents tal i com hem recollit a aquest article. A més, la proposta d'unes directrius de tractament centrades sobretot en la qualitat de les dades al moment de la transferència, ja que és el moment previ a la ingesta on es poden incidir en aspectes que posteriorment seria problemàtic, quan no impossible, recuperar aquesta informació.

La qualitat de les dades és un aspecte molt més ampli dins el cicle de vida de les dades i en el moment de la transferència alguns aspectes ja no es podran modificar. En aquests casos, com ja passava en entorns analògic, des de l'arxiu s'ha de poder documentar la situació per recollir-ho en cas de consultes posteriors i evitar males interpretacions. Els arxius no poden ser els auditors del contingut dels expedients, però sí els afecten algunes de les característiques de la qualitat recollides, per exemple, a la ISO/IEC 25012 com són la exactitud, completesa, consistència, precisió o traçabilitat.

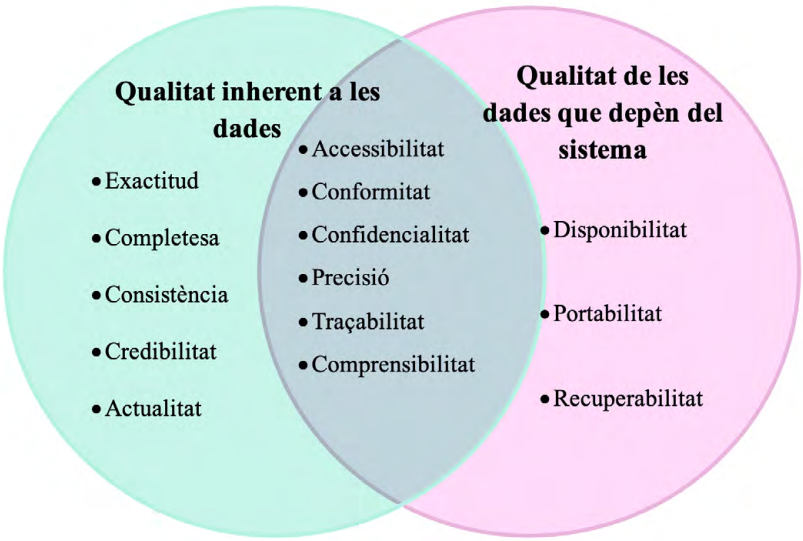

Figura 2: El model de qualitat de producte de dades definit per l'estàndard ISO/IEC 25012 amb els valor de les dades

Les directrius proposen un seguit de comprovacions prèvies a la ingesta agrupades segons la propietat a la que afecten per a que els arxius que comencen a acceptar conjunts de dades les puguin adaptar als seus procediments interns de transferència. Igual que algunes administracions han establert controls per a la publicació de les dades als seus portals de dades obertes, com la *Guia de publicació de dades obertes de la Generalitat Valenciana* o la *Guía práctica para la mejora de la calidad de datos abiertos*, els arxius també han de poder elaborar adaptar la seva norma al respecte.

La comprovació de la qualitat de les dades seria similar als controls que es fan als arxius en paper de les transferències. Igual que s'exigeix que els departaments conformin expedients complets i tancats, nets de còpies, esborranys, es trobin identificats, etc. i als arxius digitals es determinen les metadades i es fan les comprovacions de tipologies documentals, documents obligatoris o formats, es pot demanar uns mínims en els casos dels conjunts de dades. Un dels aspectes, és que no s'entreguin aquests documents amb «dades errònies».

El concepte «dada errònia» en aquest text s'entén com a les dades inexactes o inapropiades (per exemple, dades en columnes incorrectes o que no respecten les llistes de valors), files duplicades inexplicablement, dades incorrectes per mala ortografia, errors tipogràfics, etc. Errors de forma i en cap cas errades en el contingut o sentit de la dada. Si una multa es va posar per error, aquella fila no conté dades errònies, ja que sí recull un tràmit administratiu correctament, malgrat que la sanció no ho fos.

Per a la elaboració de la llista es van recollir exemples de bones pràctiques i errors habituals que es donen quan es treballa amb conjunts de dades i revisió de la bibliografia, especialment basada en l'article *The Quartz guide to bad data* (Graskopf, 2015) en el qual hi ha una primera llista i el debat posterior que va haver-hi a Github al voltant del mateix.

Els errors s'agrupen segons els valor de les dades que entraria en risc (validesa, exactitud, integritat, consistència o uniformitat) i en taules amb els camps següents:

- Identificació del problema
- Definició del problema
- Exemples
- Acció mitigadora o de correcció.

Integritat

PROBLEMA	DEFINICIÓ	EXEMPLES	ACCIÓ
Valor en blanc o nul	Dades no introduïdes.	Es donen dades anuals i hi ha un any està en blanc: cal saber si és perquè no es van recollir dades aquell any o per algun altre motiu. En el cas de les enquestes, saber si el valor blanc és perquè l'entrevistat no sap la resposta o ha refusat contestar-la, o si és possible que no trobin la resposta a la llista controlada (per exemple, per la qüestió del gènere home/dona en les persones que es defineixen com a no binàries).	Detecció: cercar els valors en blanc i veure si tenen una explicació. Solució: preguntar a la font el significat i descriure-ho en el diccionari de dades.
Les dades en blanc es reemplacen per zeros	Ús de valors arbitraris per substituir els valors en blanc o nuls.	Un valor fals per expressar la data s'acostuma a trobar com a «1970-01-01-01T00:00:00Z» o «1969-12-31T24:59:59ZZ», que és on comença el registre temporal Unix o Unix epoch for timestamps. Un fals 0 per a la localització es pot representar com a 0º 00´ 00˝N+ 0º 00´ 00˝E o simplement com a 0ºN 0ºE, que és el punt de l'oceà Atlàntic just al sud de Ghana conegut com a Null Island.	Detecció: cercar els valors 0, –1, «false» i «null» i comprovar si són coherents. Cercar si apareix la data del registre temporal Unix o la geolocalització de Null Island. Solució: preguntar a la font el significat.
Manquen dades que hi haurien de ser	Falta de dades que per la temàtica el o concepte hi haurien de ser.	Dades geogràfiques: tots els municipis, comunitats autònomes, països o estats han d'estar representats, o ha d'haver-hi alguna raó per tal que no hi siguin. Si es treballa amb un *dataset* sobre les empreses de l'IBEX35, s'ha de comprovar que no en manca cap.	Detecció: tenir el coneixement del conjunt de dades. Solució: si hi ha mancances, demanar a la font el *dataset* complet o l'explicació dels buits per poder-ho documentar.

Taula 2: Exemple d'una de les taules de la publicació. Es veu el valor, la identificació dels riscos, la definició, exemples i accions accions mitigadores.

Quan parlem de la transferència de dades als arxius, la manipulació de les mateixes per fer-ne correccions no és admissible. Aquesta pràctica entraria en contradicció directa amb el rol que ha de tenir l'arxiu i el respecte a la integritat dels documents, però el llistat de comprovació té un altre objectiu: la detecció de les dades errònies que són fruit de la extracció de les dades del sistema, de manipulacions prèvies a la transferència i migració, tractaments d'agrupacions de dades i errors en el catàleg de dades de la organització.

Si no es pot fer la correcció prèvia, les correccions posteriors que es puguin fer del dataset poden tractar-se com productes de l'arxiu i es poden oferir per als investigadors amb la neteja de dades ja efectuada, sempre i quan s'hagin documentat els canvis. Es considera que el 80% de l'anàlisi de les dades són feines prèvies molt relacionades amb el procés de neteja i preparació de les dades per al processament (Oni, 2019). Si l'arxiu vol esdevenir referent de consulta de dades per a un nou públic, una inversió en la creació d'aquests datasets pre-processats seria una estratègia de servei. Aquest tractament es pot fer amb eines com l'OpenRefine que permet comprovar la consistència de les dades i serveix per detectar i corregir noms erronis (per exemple toponímia mal escrita com Gerona/Girona) o errors en els camps de coordenades que podem haver detectat en el moment de la transferència.

Conclusions

La vinculació entre els arxius i les dades està fora de debat, tal i com s'exposa als diferents articles d'aquesta revista dedicada a la governança de la informació. La gestió documental ha d'adaptar les seves metodologies per tal de seguir donant resposta a les necessitats de les organitzacions i complir amb les seves funcions, que continuen plenament vigents. La gestió amb bases de dades no és nova però a les organitzacions s'utilitzaven com suport a la tramitació paper, actualment el pes de la gestió sobre aplicacions ha evolucionat esdevenint el paper el suport residual de l'administració electrònica.

La gestió documental ha d'adaptar-se de nou a aquest escenari, amb un canvi que és més profund que la substitució de suport ja que per garantir la integritat dels documents, els seus valors probatoris i documentar el context caldrà ampliar els manuals i formes de treball actual. El valor de les dades transferides als arxius també el marcarà l'ús que en puguin fer els usuaris futurs, per tant cal assegurar la qualitat de les dades en el moment de creació. Si no és possible, des de l'arxiu i en el moment de la transferència és un moment de control sobre les mateixes ja que hi ha una extracció dels conjunts de dades i es poden vigilar aspectes de format, detectar dades errònies o que en futur poden resultar confoses per tal de demanar correcció o explicacions per documentar-les.

L'Associació de Professionals de l'Arxivística i Gestió de Documents de Catalunya ha publicat en obert una proposta de directrius de tractament dels conjunts de dades als arxius per oferir un llistat de possibles errors als conjunts de dades per a que els arxius els adaptin per incorporar a les polítiques de transferència. Així, obrir la porta als arxius a rebre les extraccions de dades dels sistemes de gestió de les organitzacions i anar incorporant als fons documentals un nou format de documents que en els propers anys anirà augmentant.

Es comença amb la política d'adquisició i transferència per a la ingesta d'aquests conjunts de dades però cal continuar treballant sobre les metodologies a aplicar pel que fa a descripció, classificació o avaluació. Hem establert que les dades són producte d'arxiu, tema al que hem dedicat diversos articles i congressos, i els professionals de la gestió documental ens hem posicionat en el grups de treball per al disseny dels sistemes cap a entorns datificats optimitzant els procediments. Però a més de fixar-nos ens les aplicacions de negoci, també és hora d'establir la metodologia arxivística per fer el tractament d'aquests documents per quan arribin als arxius.

Bibliografia

BUTTERFIELD, Andrew; EKEMBE NGONDI, Gerard.; KERR, Anne. (ed.) A Dictionary of Computer Science [en línia]. 6a ed. Oxford: Oxford University Press, 2016 <https://doi.org/10.1093/acref/9780199688975.001.0001> [Consulta: 28/12/2022]

Catalunya. Llei 9/1993, de 30 de setembre, del Patrimoni Cultural Català. BOGC, 11/10/1993, núm. 1807) <https://www.boe.es/eli/es-ct/l/1993/09/30/9/con>.

CAMPOS MARTÍNEZ, M. Pilar. Directrius per al tractament de conjunts de dades als arxius [en línia]. Barcelona: AAC, 2022. (Textos;11) <https://arxivers.com/wp-content/uploads/2022/05/textos-11.pdf> [Consulta: 28/12/2022]

CASELLAS I SERRA, Lluís-Esteve. « L'avaluació arxivística en el nou context de les organitzacions». La destrucció d'informació pública. Una mirada multidisciplinària sobre l'eliminació ordenada de la documentació [en línia]. Oficina Antifrau i Associació d'Arxivers Gestors de Documents de Catalunya, 2019. <https://www.girona.cat/sgdap/docs/Antifrau-Publicacio_CASELLAS.pdf> [Consulta: 28/12/2022]

CALABRESE, Julieta; ESPONDA, Silvia; PASINI, Ariel [et al.] «Guía para evaluar calidad de datos basada en ISO/IEC 25012. » dins XXV Congreso Argentino de Ciencias de la Computación (CACIC) [en línia]. Córdoba: Universidad Nacional de Río Cuarto, 14 al 18 de octubre de 2019.. 2019. <https://core.ac.uk/download/pdf/288490953.pdf> [Consulta: 28/12/2022]

Espanya. Llei 11/2007, de 22 de juny, d'accés electrònic dels ciutadans als Serveis Públics. BOE, 23/06/2007, núm. 150). <https://www.boe.es/eli/es/l/2007/06/22/11/con>.

Espanya. Llei 39/2015, d'1 d'octubre, del Procediment Administratiu Comú de les Administracions Públiques. BOE, 02/10/2015, núm. 236. <https://www.boe.es/eli/es/l/2015/10/01/39/con>.

DEPARTAMENT DE CULTURA. Quadre de classificació de la documentació. Dades obertes de Catalunya [en línia]. Portal de dades obertes de la Generalitat de Catalunya. <https://analisi.transparenciacatalunya.cat/Sector-P-blic/Quadre-de-Classificaci-de-la-documenta-ci-/5t23-dy8y> [Consulta: 28/12/2022]

Declaració (UE). Declaración Europea sobre los Derechos y Principios Digitales para la Década Digital. Brusel·les, 28/10/2022. <https://ec.europa.eu/newsroom/dae/redirection/document/82888> [Consulta: 28/12/2022]

IBM. (2001). Dictionary of computer science, engineering, and technology. Vol. 38, núm. 10.

GRASKOPF, Christopher. (2015). «The Quartz guide to bad data» dins Quartz [en línia]. (15 de desembre de 2015). <https://qz.com/572338/the-quartz-guide-to-bad-data/>. [Consulta: 28/12/2022]

ICA, International Council on Archives, ISAD(G): Norma Internacional General de Descripció Arxivística = General International Standard Archival Description; adoptada pel Comitè de Normes de Descripció, Estocolm, Suècia, 19-22 de setembre de 1999. Versió catalana a cura de Josep Matas i Jaume Rufí Pagès; amb l'assessorament d'Àngels Bernal et al. 2a ed. Vol. 2a, p. 112. Barcelona: Associació d'Arxivers de Catalunya; Departament de Cultura de la Generalitat de Catalunya. 2001. <http://cataleg.ub.edu/record=b1492091~S1*cat>.

INTERPARES, InterPares Trust AI Terminology Database [en línia]. 2022. <https://interparestrustai.org/terminology>. [Consulta: 28/12/2022]

ISO/IEC 2382-1:1993 – Information Technology – Vocabulary – Part 1: Fundamental Terms. 1993

ISO - ISO/IEC 11179-3:1994 – Information technology – Specification and standardization of data elements. – Part 3: Basic attributes of data elements. 1994 <https://www.iso.org/standard/19184.html>.

ISO/IEC 25012:2008. Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model (2008)

NIU, Jinfang. (2016). «Organisation and description of datasets». Archives and Manuscripts [en línia]. Núm. 44, p. 73-85. <https://doi.org/10.1080/01576895.2016.1179585>. [Consulta: 28/12/2022]

ONI, Samoson.; CHEN, Zhiyuan.; HOBAN, Susan [et al] «A comparative study of data cleaning tools». International Jounral of Data Warehousing and Mining [en línia]. Vol. 15, núm. 4, p. 48-65. <https://doi.org/10.4018/IJDWM.2019100103>. [Consulta: 28/12/2022]

PASTOR-SÁNCHEZ, Juan Antonio; LLANES PADRÓN, Dunia. «Records in Contexts y la publicación de conjuntos de datos archivísticos interoperables». Dins: MELO SIMÕES, Maria da Graça; MANUEL BORGES, María (ed.). Tendências atuais e perspetivas futuras em organização do conhecimento [en línia], 2019 p. 587-599. Centro de Estudos Interdisciplinares do Século XX -

CEIS20. <https://dialnet.unirioja.es/descarga/libro/719647.pdf>. [Consulta: 28/12/2022]

Reglament (UE) 2016/679 del Parlament Europeu i del Consell, de 27 d'abril de 2016, relatiu a la protecció de les persones físiques pel que fa al tractament de dades personals i a la lliure circulació d'aquestes dades i pel qual es deroga la Directiva 95/46/CE (Reglament general de protecció de dades) DOUE L núm.119 de 4/5/2016. <https://www.boe.es/doue/2016/119/L00001-00088.pdf>.

Reglament (UE) 2021/694 del Parlamento Europeo y del Consejo de 29 de abril de 2021 por el que se establece el Programa Europa Digital y por el que se deroga la Decisión (UE) 2015/2240. DOUE, num. 166, de 11/05/2021, pàg. 1 a 34. <https://www.boe.es/doue/2021/166/L00001-00034.pdf>.

Reglament (UE) 2022/868 del Parlamento Europeo y del Consejo de 30 de mayo de 2022 relativo a la gobernanza europea de datos y por el que se modifica el Reglamento (UE) 2018/1724 (Reglamento de Gobernanza de Datos). DOUE L, num. 152 de 3.6.2022, pàg. 1 a 44). <http://data.europa.eu/eli/reg/2022/868/oj>

SAA Society of American Archivists, Dictionary of Archives Terminology [en línia] (2005-2022) <https://dictionary.archivists.org/entry/record.html> [Consulta: 28/12/2022]

SHEPHERD, E.; Smith, C. (2000). «The application of ISAD(G) to the description of archival datasets». Journal of the Society of Archivists [en línia]. 2000. Vol. 21, núm. 1, p. 55-86. <https://doi.org/10.1080/00379810050006911>. [Consulta: 28/12/2022]

THE NATIONAL ARCHIVES. Managing the Continuity of Datasets [en línia]. 2011 .p. 1-35. < https://cdn.nationalarchives.gov.uk/documents/information-management/managing-continuity-of-datasets.pdf>. [Consulta: 28/12/2022]>.

UNE-ISO/TR 15489-1:2006. Gestión de documentos. Parte 1: Generalidades