OPEN TUSK

La comunità per i dati aperti pugliesi

DOCUMENTO 2 DI 3

Come costruire un buon CSV

Preparare i dati, dalla tabella disordinata al dato aperto

OPENTUSK

Come costruire un buon CSV

Documento 2 di 3: preparare i dati, dalla tabella disordinata al dato aperto

A chi serve questo documento

A chiunque debba preparare un dataset da pubblicare, partendo da quello che ha già: un foglio Excel, una tabella su Word, un elenco, dei dati estratti da un gestionale. Non servono competenze tecniche.

Come gli altri materiali della raccolta, si rivolge in particolare agli enti che pubblicano open data seguendo il perimetro nazionale (le Linee Guida AgID) ed europeo, come comuni, unioni di comuni, università, centri di ricerca e imprese.

A cosa serve

A trasformare dati "grezzi" in un **CSV ben fatto**, che è la base di ogni buon dato aperto. Qui si parla soltanto di preparare il file, mentre il caricamento sul portale è trattato nel Documento 3.

Cosa leggere prima

Il Documento 1 (open data e principi FAIR) dà il perché di molte regole che troverai qui. Non è obbligatorio, ma aiuta. Prima di mettere mano al file, il documento "Prima di iniziare: quale dataset pubblicare" aiuta a scegliere il dataset giusto e a trovare un template di riferimento da cui partire.

1. Cos'è un CSV (e perché proprio il CSV)

CSV sta per *Comma-Separated Values*, cioè "valori separati da virgola". Si tratta semplicemente di una tabella salvata come testo, in cui ogni riga è una riga della tabella e le colonne sono separate da virgole.

Una tabella come questa:

comune	abitanti	provincia
Lecce	95000	LE
Bari	315000	BA

salvata in CSV diventa questo testo:

```
comune, abitanti, provincia
Lecce,95000,LE
Bari,315000,BA
```

Tutto qui. Niente colori, niente formule, niente formattazione: solo i dati, puliti.

Proprio questa semplicità è la sua forza. Un CSV si apre con qualsiasi programma (Excel, Fogli Google, un editor di testo, un software statistico) senza bisogno di nulla di proprietario. È il formato che soddisfa il principio di **interoperabilità** (la “I” di FAIR vista nel documento 1) e che fa scattare la **terza stella** della qualità dei dati aperti, mentre un Excel o un PDF, da soli, si fermano prima.

> **In sintesi:** il CSV è il minimo indispensabile di un dato aperto fatto bene, e tutto questo documento serve a costruirne uno corretto.

2. Come si fa una buona tabella: cinque regole d’oro

Prima ancora del formato del file, conta come è organizzata la tabella, perché un CSV salvato male parte quasi sempre da una tabella pensata male. Ecco le cinque regole che fanno la differenza.

Regola 1: una riga corrisponde a un solo record

Ogni riga deve descrivere **una sola cosa**, che sia una farmacia, un evento o un impianto. Non vanno messi due eventi sulla stessa riga, e un evento non va spezzato su due righe.

✓ Una riga per ogni farmacia.

✗ Una riga che contiene “Farmacia Rossi e Farmacia Bianchi”.

Regola 2: una colonna corrisponde a un solo tipo di informazione

Ogni colonna contiene un’unica informazione, sempre dello stesso tipo. Non bisogna mischiare indirizzo e telefono nella stessa colonna, né mettere il comune in mezzo al nome.

✓ Colonne separate: `indirizzo`, `telefono`, `comune`.

✗ Una colonna `contatti` con dentro “Via Roma 12 - tel. 0832...”

Un errore tipico è usare una **riga “titolo”** per raggruppare (per esempio “Provincia di Lecce” con sotto le sue righe). In una tabella ben fatta quel raggruppamento diventa una **colonna** `provincia` ripetuta su ogni riga. È più noioso da scrivere, ma è l’unico modo perché un computer capisca a quale provincia appartiene ogni record.

Regola 3: un solo valore per cella

Niente elenchi dentro una cella, niente “e/o”, niente note tra parentesi mischiate al dato.

✓ Cella `posti`: `1200`.

✗ Cella `posti`: `1200 (circa, da verificare)`.

Se un’informazione aggiuntiva ti serve davvero, mettila in una **colonna a parte** (per esempio una colonna `note`), così il dato principale resta pulito e utilizzabile.

Regola 4: intestazioni chiare e semplici

La prima riga deve contenere i **nomi delle colonne**, che devono essere chiari e “macchina-compatibili”: meglio lettere minuscole, senza spazi (al loro posto si usa l’underscore _), senza simboli o unità di misura nel nome.

✓ `data_rilevazione,numero_iscritti,importo_euro`

✗ `Data di rilevazione (gg/mm/aaaa),N° iscritti,Importo €`

Le unità di misura e le spiegazioni non vanno nel nome della colonna, ma nella **descrizione del dataset** (i metadati), di cui parla il Documento 3.

Regola 5: niente “abbellimenti”

Celle unite, colori, righe vuote di separazione, righe di “totale”, titoli e loghi in cima al foglio: tutto questo è pensato per l’occhio umano, ma **rompe** un CSV. Un computer che legge il file non capisce una cella unita, e una riga “TOTALE” viene scambiata per un dato.

> Una tabella pronta per diventare CSV è “nuda”: solo l’intestazione e le righe di dati, niente altro.

3. Date, numeri e testo: le insidie classiche

Anche con una tabella ben strutturata, ci sono alcuni punti dove ci si fa male più spesso.

Le date

Le date vanno scritte sempre nello **stesso formato** in tutta la colonna. Lo standard internazionale, raccomandato anche per i dati aperti, è **AAAA-MM-GG** (anno-mese-giorno):

✓ `2026-03-15`

✗ Mischiare `15/03/2026`, `15-3-26`, `15 marzo 2026` nella stessa colonna.

Il formato anno-mese-giorno ha anche un vantaggio pratico: ordina correttamente da solo, perché segue lo stesso ordine delle cifre.

I numeri

Tre attenzioni:

- **Niente unità di misura dentro la cella.** Si scrive `1500` e non `1500 €` o `1500 mq`, perché l’unità si specifica nella descrizione o nel nome della colonna.
- **Attenzione al separatore delle migliaia.** In Italia siamo abituati a scrivere `33.876`, ma in un CSV quel punto può essere scambiato per un decimale. Meglio scrivere il numero senza separatore, cioè `33876`, e soprattutto essere coerenti in tutta la colonna.
- **Il separatore decimale.** Va scelta una convenzione (il punto è lo standard internazionale: `3.5`) e usata ovunque. L’importante è non avere mezza colonna con la virgola e mezza con il punto.

Il testo e gli accenti (la codifica UTF-8)

Qui serve una parola su un concetto invisibile ma importante, cioè la **codifica**: il modo in cui il computer salva le lettere. Se salvi un file con la codifica sbagliata, le lettere accentate si rovinano e “Nardò” diventa “NardÃ²” o “Nardo?”.

La codifica giusta, universale, si chiama **UTF-8** ed è quella che preserva à, è, ì, ò, ù e qualsiasi altro carattere. Quando salvi un CSV, scegli sempre l’opzione **UTF-8**.

C’è poi una piccola regola di forma: se un valore di testo **contiene una virgola** (per esempio “Ingresso libero, su prenotazione”), va racchiuso tra **virgolette doppie**, altrimenti la virgola interna verrebbe scambiata per un separatore di colonna. La maggior parte dei programmi lo

fa in automatico quando si salva in CSV, ed è utile saperlo per capire perché ogni tanto si vedono le virgolette nel file.

La coerenza dei valori in una colonna

Dentro una stessa colonna, lo stesso concetto va scritto sempre allo stesso modo. Se la colonna `comune` a volte dice “Lecce”, a volte “LECCE”, a volte “Lecce (LE)”, per una persona è chiaro che si tratta sempre dello stesso comune, ma per un computer sono **tre valori diversi**, e il dato diventa impossibile da contare o da incrociare correttamente.

✓ Sempre `Lecce`, su tutte le righe.

✗ Mescolare `Lecce`, `LECCE`, `Lecce (LE)`, `lecce`.

Vale per tutte le colonne “categoriche” (comune, provincia, tipologia, categoria): si sceglie una forma e la si mantiene identica ovunque. È uno degli errori che più spesso rende “sporco” un dataset, ed è anche uno dei più facili da evitare.

4. Il catalogo degli errori più comuni

Per comodità, ecco raccolti in un’unica tabella gli errori che si incontrano più spesso nella costruzione dei CSV, con un esempio di come si presentano e la relativa correzione. È una sintesi delle regole viste fin qui, utile da tenere sotto mano durante la pulizia di un file.

Errore	Come si presenta	Come si corregge
Intestazioni non standard	`N° iscritti (2024)`, `Data (gg/mm)`	`numero_iscritti`, `data_rilevazione`
Campi misti	`Via Roma 12 - tel. 0832 000000` in un’unica colonna	Due colonne separate: `indirizzo` e `telefono`
Valori non normalizzati	`Lecce`, `LECCE`, `Lecce (LE)` nella stessa colonna	Una sola forma, identica ovunque: `Lecce`
Formati data incoerenti	`3/2/2025`, `15-03-2025`, `2025/04/01` mescolati	Un unico formato: `AAAA-MM-GG`
Celle unite	Una cella “Provincia di Lecce” che copre più righe	Una colonna `provincia` ripetuta su ogni riga
Righe di totale	`TOTALE: 35.076` in mezzo ai dati	Si elimina: è un calcolo, non un dato
Segnaposto nei vuoti	`N/D`, `n.d.`, `-`, `0` al posto del dato mancante	Cella davvero vuota
Unità dentro la cella	`1500 €`, `1500 mq`	`1500`, con l’unità nella descrizione

Se riconosci uno o più di questi errori nel tuo file, conviene correggerli prima dell’esportazione in CSV, direttamente nel foglio di lavoro: è molto più semplice che intervenire dopo.

5. Da Excel a CSV, senza rovinare i dati

Quasi sempre il punto di partenza è un foglio Excel. Convertirlo è semplice, ma va fatto con qualche accortezza.

Prima di esportare, applica le regole della Sezione 2: una sola tabella nel foglio, intestazioni pulite nella prima riga, niente celle unite, niente righe di totale, niente righe vuote. In pratica, "spoglia" il foglio fino a lasciare solo i dati.

Per esportare, usa `File → Salva con nome` e scegli il formato **"CSV UTF-8 (delimitato da virgole)"**, che è l'opzione che salva correttamente gli accenti.

Dopo aver esportato, vale la pena aprire il file con un editor di testo semplice (Blocco note o TextEdit) e dare un'occhiata: i valori devono essere separati da virgole, gli accenti devono apparire giusti e la prima riga deve essere quella delle intestazioni.

> **Il tranello più comune:** nelle versioni italiane di Excel a volte il CSV viene salvato con il **punto e virgola** (`;`) al posto della virgola, perché da noi la virgola è il separatore decimale. Se aprendo il file vedi i punti e virgola, scegli esplicitamente l'opzione "CSV UTF-8 (delimitato da virgole)". Lo standard del dato aperto è la virgola.

Anche il nome del file conta. Meglio sceglierne uno chiaro, in minuscolo, senza spazi né accenti, usando l'underscore al posto degli spazi: un nome come `farmacie_muro_leccese_2024.csv` dice già cosa contiene, mentre un nome come `Elenco def (2).csv` non aiuta nessuno.

E se i dati non sono in Excel?

Non sempre si parte da un foglio di calcolo. I casi più comuni sono questi:

- **Dati in un gestionale o in un database.** Quasi tutti i gestionali permettono di "esportare" i dati in CSV o in Excel: cerca questa funzione (spesso si chiama Esporta, Download o Scarica). È la strada migliore, perché i dati restano fedeli all'originale.
- **Dati in un PDF.** Qui serve attenzione. Se il PDF è stato generato da una tabella che esiste ancora altrove (un Excel o un gestionale), conviene risalire a quella fonte e lavorare su quella, perché è molto più sicuro che estrarre dal PDF. Se la fonte originale non esiste più, la tabella va ricostruita a mano, controllando ogni valore.
- **Dati come testo o elenco.** Vanno riorganizzati in tabella applicando le regole della Sezione 2 (una riga corrisponde a un record, una colonna a un'informazione).

Una parola sull'**OCR**, cioè i programmi che "leggono" il testo da un'immagine o da un PDF scansionato: può far risparmiare tempo, ma sbaglia spesso, soprattutto su numeri e lettere accentate. Se lo usi, **verifica ogni dato** rispetto all'originale prima di considerarlo buono, perché su dati ufficiali un OCR non controllato è una fonte di errori.

6. La regola che vale più di tutte: non inventare i dati

Quando si "pulisce" un file, la tentazione è di sistemare anche ciò che non torna. Resisti, perché su dati ufficiali destinati alla pubblicazione vale una regola che viene prima di ogni altra:

> **Non inventare, non dedurre, non "correggere" un valore di tua iniziativa.**

- Se un valore **manca**, lascia la cella **vuota**. Una cella vuota dice “non lo sappiamo”, ed è un’informazione onesta, mentre uno zero inventato è un dato falso. E “vuota” significa **davvero vuota**: niente `0`, niente `N/D`, `n.d.` o trattino, perché questi segnaposto sembrano dati veri.
- Se un valore è **strano o sospetto** (un CAP con quattro cifre, una data impossibile, un anno scritto “anni ’80”), **non aggiustarlo da solo**, ma verificalo alla fonte o segnalalo a chi di dovere.
- Se non sai come interpretare un dato, **chiedi** a chi l’ha prodotto, invece di indovinare.

Tutto questo si collega direttamente alla **provenienza** e all’**accuratezza** del principio di riutilizzo (la “R” di FAIR): un dato di cui ci si può fidare è un dato che riporta fedelmente la realtà, comprese le sue lacune.

7. Attenzione ai dati personali

Un dato aperto è pubblico per definizione: lo vedrà chiunque, per sempre. Per questo **non deve contenere dati personali** che permettano di identificare singole persone, come nomi e cognomi di persone fisiche, codici fiscali, email personali, indirizzi privati, dati su salute o situazioni sensibili.

Se i tuoi dati di partenza contengono informazioni del genere (per esempio l’elenco nominativo dei beneficiari di un contributo), prima della pubblicazione vanno:

- **rimossi**, oppure
- **aggregati** (per esempio, al posto dei singoli beneficiari, “numero di beneficiari e importo totale per categoria”), oppure
- **anonimizzati** secondo una valutazione formale.

La scelta su cosa si può pubblicare **non è tecnica, ma di responsabilità dell’ente**: in caso di dubbio va coinvolto il Responsabile della Protezione dei Dati (RPD/DPO) o il referente privacy. Nel dubbio, non si pubblica.

8. Un esempio completo: prima e dopo

Mettiamo insieme tutto con un esempio (illustrativo, non basato su dati reali). Immagina di ricevere una tabella come questa, verosimile e piena dei problemi tipici:

Prima, la tabella disordinata:

```
Impianti sportivi - Provincia di LECCE
Comune      Impianto               Posti     Anno
Lecce       Stadio Via del Mare    33.876    1966
Nardo'      Palazzetto "G. Pinto"  1.200     anni '80
Galatina    Piscina Comunale                 2003
Totale posti: 35.076
```

I problemi: la provincia è un titolo invece che una colonna, c’è una riga di totale, i numeri hanno il punto delle migliaia, un anno è scritto a parole e manca un valore (i posti di Galatina).

Dopo, lo stesso dato come CSV pulito:

```
provincia,comune,impianto,posti,anno
```

```
Lecce,Lecce,Stadio Via del Mare,33876,1966
Lecce,Nardò,"Palazzetto ""G. Pinto""",1200,anni '80
Lecce,Galatina,Piscina Comunale,,2003
```

Cosa è cambiato, e perché:

- La **provincia**, che era un titolo, è diventata una colonna ripetuta su ogni riga *(Regola 2)*.
- La riga **“Totale posti” è stata eliminata**, perché è un calcolo e non un dato *(Regola 5)*.
- I **numeri** hanno perso il punto delle migliaia e ora sono coerenti: `33876`, `1200` *(Sezione 3)*.
- Il **valore mancante** dei posti di Galatina è rimasto una cella vuota, perché non è stato inventato *(Sezione 6)*.
- L’**anno “anni ’80”** di Nardò è stato lasciato com’era: non conoscendo la data esatta non possiamo inventarla, e andrà semmai verificata alla fonte *(Sezione 6)*.
- Il nome con le **virgolette interne** (“G. Pinto”) è racchiuso tra virgolette doppie raddoppiate, come richiede il formato CSV *(Sezione 3)*.
- L’**accento di “Nardò”** è preservato grazie al salvataggio in UTF-8 *(Sezione 3)*.

Questo è, in piccolo, tutto il lavoro che questo documento descrive: prendere dati pensati per l’occhio umano e renderli puliti, coerenti e fedeli, pronti per una macchina.

9. Esempi di struttura per tipologia di dato

Ogni ente pubblica dati diversi, quindi ecco alcuni esempi di struttura per le tipologie più comuni. Per ciascuna trovi le colonne consigliate e una riga di esempio. **Tutti i valori sono fittizi e dichiaratamente dimostrativi** (nomi come “Comune Demo” o “Via degli Esempi” servono proprio a ricordarlo): mostrano la struttura, non dati reali da copiare. Adatta liberamente le colonne al tuo caso, mantenendo le regole delle sezioni precedenti.

Farmacie

```
nome,indirizzo,comune,provincia,telefono,orario_apertura,turno_festivo
Farmacia Demo Centrale,Via degli Esempi 1,Comune Demo,XX,0000 000000,08:30-
20:00,si
```

Biblioteche

```
nome,indirizzo,comune,provincia,tipologia,posti_lettura,accesso_disabili,si
to_web
Biblioteca Civica Demo,Piazza degli Esempi 2,Comune
Demo,XX,comunale,40,si,https://esempio.example
```

Scuole

```
denominazione,codice_meccanografico,indirizzo,comune,provincia,grado,numero
_alunni
Istituto Demo,XXXX000000,Via degli Esempi 3,Comune Demo,XX,primaria,250
```

Piste ciclabili

```
nome_percorso,comune,provincia,lunghezza_km,tipologia,fondo,anno_realizzazi
one
Ciclovia Demo,Comune Demo,XX,4.2,in sede propria,asfalto,2020
```

Eventi

```
titolo,descrizione,comune,luogo,data_inizio,data_fine,categoria,ingresso_gr
atuito
Festa Demo,Evento dimostrativo,Comune Demo,Piazza degli Esempi,2026-07-
01,2026-07-03,cultura,si
```

Parcheggi

```
nome,indirizzo,comune,tipologia,posti_auto,posti_disabili,a_pagamento,tarif
fa_oraria_euro
Parcheggio Demo,Via degli Esempi 4,Comune Demo,a raso,120,6,si,1.00
```

Contributi erogati (in forma aggregata)

Per i contributi vale l'avvertenza della Sezione 7: mai i nominativi dei beneficiari persone fisiche, ma dati aggregati per categoria.

```
anno,tipologia_contributo,settore,numero_beneficiari,importo_totale_euro
2025,contributo demo,associazionismo,14,25000
```

Patrimonio immobiliare

```
denominazione,indirizzo,comune,categoria,superficie_mq,stato_utilizzo,anno_
costruzione
Immobile Demo,Via degli Esempi 5,Comune Demo,ufficio,850,in uso,1975
```

> **Nota:** questi schemi sono un punto di partenza. Se per la tua tipologia di dato esiste uno standard nazionale o un'indicazione della Regione, quella prevale: verifica sempre nella documentazione ufficiale del portale.

10. Checklist: il mio CSV è pronto?

Prima di passare alla pubblicazione, verifica che il tuo file rispetti questi punti.

Struttura della tabella

- ☐ Una riga per ogni record, una sola cosa per riga
- ☐ Una colonna per ogni tipo di informazione, un solo valore per cella
- ☐ Eventuali raggruppamenti trasformati in colonne (es. `provincia`), non in righe-titolo
- ☐ Prima riga con intestazioni chiare, minuscole, senza spazi né simboli
- ☐ Niente celle unite, righe vuote, righe di totale, loghi o titoli

Contenuto

- ☐ Date tutte nello stesso formato (consigliato `AAAA-MM-GG`)
- ☐ Numeri senza unità di misura e con separatori coerenti

- Valori uniformi dentro ogni colonna (sempre `Lecce`, non anche `LECCE`)
- Celle vuote dove il dato manca, davvero vuote (nessun valore inventato, nessun `N/D` o `0`)
- Valori riportati fedelmente, anche quelli “strani” (semmai segnalati, non corretti)

Formato, nome e privacy

- Salvato in **CSV UTF-8**, separatore **virgola**
- Accenti corretti (verificati aprendo il file)
- Nome del file chiaro, minuscolo, senza spazi né accenti (es. `farmacie_2024.csv`)
- Nessun dato personale (o dati aggregati/anonimizzati previa valutazione)

Se tutte le caselle sono spuntate, il dato è pronto per la pubblicazione descritta nel **Documento 3, Pubblicare su dati.puglia.it**.

Riferimento ufficiale sui requisiti tecnici per l’apertura dei dati: Linee Guida AgID sull’apertura dei dati e il riutiizzo dell’informazione del settore pubblico.

Documento formativo e operativo introduttivo: per i requisiti tecnici specifici del portale (metadati obbligatori, formati ammessi) fanno fede la documentazione ufficiale di dati.puglia.it e le Linee Guida AgID. Gli esempi contenuti in questo documento sono fittizi e hanno scopo puramente dimostrativo.