13.11.2025

Datasets

Jan Bárta

barta
.legal

Data

Distinguish **dataset** vs. **database** vs. **individual items**

Know when copyright and sui generis database right bite

Choose & apply open data licences (CC, Open Data Commons)

Spot & mitigate GDPR / trade secret / licence risks

Understand AI training constraints (TDM exceptions; EU AI Act duties)

barta
.legal

What is a “dataset”? What is a “database”?

Dataset (practice): any structured collection used for analysis/model training

Database (legal): collection of independent materials arranged systematically or methodically and individually accessible (Directive 96/9/EC, Art. 1(2)).

Individual items: facts/data points may be unprotected by copyright; protection (if any) sits in the selection/arrangement or in the database right.

barta
.legal

Types of database

Copyright in **databases** (if original selection/arrangement), not in mere contents.

Sui generis database right (SGDR) protects substantial investment in obtaining/verifying/presenting contents; restricts extraction/re-utilisation of substantial parts and repeated insubstantial parts.

barta
.legal

What Is *Not* Protected

Raw facts/data are not protected by copyright

Non-original compilations lacking creativity and substantial investment

Yet: GDPR, trade secrets, **terms of use** (ToU), and **access restrictions** can still apply

Publication ≠ waiver of trade secrets, privacy, or contractual limits

barta
.legal

Licensing Options

Proprietary / B2B licenses (closed or paid datasets; negotiated terms)

Open licenses: Creative Commons (CC0, CC BY, CC BY-SA), Open Data Commons (PDDL, ODC-By, ODbL)

Strategy: align license with intended reuse, compatibility, and compliance

barta
.legal

Proprietary Datasets: Contract Structure

Scope: purpose/territory/term, TDM & AI training, sublicensing

Warranties & provenance, audit rights, notice-and-takedown, termination & deletion

Define Derived Database vs. Produced Works; attribution obligations

barta
.legal

CC BY 4.0 License

Free use of the database – copy, share, modify, and use commercially

Applies to the database as a whole or in part

Can be combined with other data sources

Conditions

Attribution required – name of creator, title, license link, and source

Indicate if changes were made

No additional restrictions – you can’t impose new legal or technical limits

Example attribution

“Data © 2025 John Doe, licensed under CC BY 4.0“

barta
.legal

CC BY-SA 4.0 License

Free use – copy, share, adapt, remix, and use commercially

Applies to the database as a whole or in part

Can be combined with other materials, as long as license terms are respected

Conditions

Attribution required – name, title, license link, and source

Share-Alike (Copyleft):

If you modify or build upon the database, you must distribute your version under the same license (CC BY-SA 4.0)

Indicate changes if modifications were made

Example attribution

“Data © 2025 Jane Doe, licensed under CC BY-SA 4.0.”

barta
.legal

CC0 1.0

Free use without any restrictions – copy, share, modify, distribute, and use commercially

No need to give credit or indicate changes

Applies to the database and its contents (where legally possible)

The creator waives all rights (copyright and database rights) to the extent allowed by law. If full waiver isn't possible, the license grants unconditional permission for any use

Example notice

"Data released under CC0 1.0 – no rights reserved."

barta .legal

Open Database License (ODbL) 1.0

Free use of the database – copy, share, modify, and use commercially

Create derived databases and produce works from the data (may be under another license but must attribute the original database)

Conditions

Attribution – credit the creator and include a link to the license

Share-Alike (Copyleft) – if you modify or build upon the database, you must distribute your version under ODbL 1.0

Keep open data accessible – if you publicly use a modified database, you must also make your modified version available under the same license

No restriction on data extraction, but derivative databases trigger the same obligations

Example attribution

“Contains data © 2025 OpenData Initiative, licensed under the ODbL 1.0”

barta
.legal

ODC-By 1.0 – Open Data Commons Attribution License

Free use of the database – copy, share, modify, and use commercially
Create derivative databases or combine with other data sources

Conditions
Attribution required
Credit the creator and include a link to the license
Indicate if any changes were made
No “Share-Alike” or “Non-Commercial” limits – only attribution is required

Must not misrepresent the original data source or author

Example attribution
“Contains data © 2025 OpenData Initiative, licensed under the ODC-By 1.0”

barta
.legal

PDDL 1.0 – Public Domain Dedication and Licence

Unrestricted use – copy, modify, share, distribute, and use commercially
Applies to the database and its contents (where legally possible)

No attribution required – though voluntary credit is encouraged

Full compatibility with open data and public domain projects

Legal meaning

The author waives all rights (copyright and database rights) to the extent allowed by law

If complete waiver isn't possible, PDDL grants unconditional permission for all uses

Designed for maximum openness and reuse of data

Example notice

"Data released under the PDDL 1.0 – no rights reserved."

barta
.legal

Combining Licenses

CC0 × anything → typically compatible (still consider GDPR/ToU)

CC BY 4.0 × ODbL → keep layers separate to avoid triggering SA on the database

CC BY-SA × proprietary → often incompatible for combined distributions

barta
.legal

Text & Data Mining (DSM)

Research organizations

Need to state the source

Non-profit

Relevant part of the database

Requires lawful access to content; no opt-out by rightholders

internal citation methodology + archiving of evidence of purpose status

barta
.legal

Text & Data Mining (DSM)

Common and appropriate use

Not substantial part of database

Public database

Not systematical or repeatedly

Cannot infringe the rights of the owner

barta
.legal

Text & Data Mining (DSM)

OK examples: one-time queries, occasional citations of small snippets, test extractions

Risky: script/governor that periodically downloads small parts → can be “repeated and systematic”

Pay attention to the ToU of the website/API and rights to individual items (photos, texts)

Evidence: log the purpose and scope to be able to show proportionality

barta
.legal

Data Act (EU) 2023/2854)

Application phases from 9/2025; aims at fair access to IoT/connected product data

SGDR cannot be used to block access/portability of in-scope IoT data

barta
.legal

GDPR in Datasets

Personal data: legal basis, purpose limitation, data minimization

Anonymization vs. pseudonymization — different compliance outcomes

barta
.legal

Common Mistakes

“Free to use” without an actual license; missing or wrong attribution

Ignoring TDM opt-outs and source ToU

Confusing pseudonymization with anonymization;

Unclear scope of license

barta
.legal

Jan Bárta

jan.barta@barta.legal

+420 776 725 597

barta
.legal