

DATASET CHARACTERISATION INSTRUMENT

Know Your Dataset (KYDS)

EXECUTIVE SUMMARY

Know Your Dataset (KYDS) is a characterisation instrument for any public-sector department or body that holds, generates, or shares a dataset. It records what a dataset is, how sensitive it is, where it lives, and the basis on which it may be shared or restricted — so that decisions about access, security, and disclosure rest on a documented, comparable record rather than on case-by-case recollection.

The form is structured around the two laws that govern every public dataset in India. The **DPDP Act 2023** applies wherever a dataset contains personal data, and asks what is owed to the individuals it describes. The **IT Act 2000**, read with the CERT-In Directions, applies to all electronic data without exception, and asks how securely it must be held and classified. Most datasets sit under both at once.

Each section opens with the guidance for that section, followed immediately by the questions it explains — so the reference material and the fields it governs sit together. Sections 0 to 7 cover the dataset itself; the optional Addendum at the end captures lawful basis and impact-assessment notes; Appendix A is the full glossary of data formats referenced in Section 1.

Complete one KYDS form per dataset. It is a self-assessment that records facts — it does not itself authorise or prohibit sharing, and it is not a legal determination. Confirm any lawful-basis entry with the department's legal team or Data Protection Officer before acting on it.

Instrument	Know Your Dataset (KYDS) — Dataset Characterisation Form
Contents	Executive summary · How this form works · Sections 0–7 · Addendum (optional) · Appendix A: Glossary of formats
Completed by	The data custodian for the dataset, with the department's DPO or legal team consulted on any lawful-basis entry
How to use	Open this PDF in Adobe Acrobat Reader. Every checkbox, option and answer box in Sections 0–7 and the Addendum is fillable on screen. Use File > Save As to keep your answers, then return the saved file.
Governing law	Digital Personal Data Protection Act 2023 · Information Technology Act 2000 read with the CERT-In Directions

How this form works

This is a characterisation instrument for any public-sector department or body that holds, generates, or shares a dataset. It records what a dataset is, how sensitive it is, where it lives, and the basis on which it may be shared or restricted. The aim is to give departments, oversight bodies, and prospective data-sharing partners a consistent, comparable record of each dataset, so that decisions about access, security, and disclosure can be made on a documented basis rather than case by case from memory.

It records facts about a dataset; it does not itself authorise or prohibit sharing, and it is not a legal determination. If you choose to fill the lawful basis section (refer to the Addendum), confirm with the department's legal team or Data Protection Officer before acting.

How to fill this form

Complete one KYDS form for each dataset.

If your department has a single dataset (for example, one table or file containing related information), fill out one form.

If a larger database contains multiple separate datasets, complete a separate form for each dataset. In each form, mention the name of the larger (parent) database.

Example: A Beneficiary Welfare Database contains three separate datasets: Ration Card Records; Bank Account Linkage Records; Benefit Disbursement Records. In this case, complete three KYDS forms — one for each dataset — and mention “Beneficiary Welfare Database” as the parent dataset in each form.

Tip: If the data can be managed, updated, or used as separate tables or files, treat each one as a separate dataset. If it is just one table with different columns, complete only one form.

The two laws this form covers

Every dataset is governed by two separate laws, and the form is structured around both. They ask different questions and neither substitutes for the other.

DPDP Act 2023 (Digital Personal Data Protection Act). Applies only where the dataset contains personal data — information that identifies or can identify a living individual (a “Data Principal”). It governs why the data was collected, the rights of the individuals it describes, and the obligations of whoever decided to collect and hold it (the “Data Fiduciary”). Section 2A of the form uses DPDP's own categories and terms.

IT Act 2000, read with the CERT-In Directions. Applies to all electronic data without exception — personal or not, sensitive or routine. It governs how data must be secured, how it is classified, and what happens if it is breached, including the CERT-In requirement to report qualifying incidents within six hours and retain logs for 180 days. Section 2B and the standing obligations in Section 7 of the form draw on this Act.

In short: DPDP asks whether this dataset is about identifiable people and what is owed to them. The IT Act asks how securely any dataset — about people or not — must be held and classified. Most datasets sit under both at once.

HOW TO FILL THIS DOCUMENT

Every option and every answer box in Sections 0–7 and the Addendum is fillable on screen. Click a box to type; click an option to select it. Grey tinted panels are guidance and are not fillable.

Square boxes accept more than one answer — tick every one that applies. **Round options** accept one answer only — selecting a second replaces the first.

Fill the form in **Adobe Acrobat Reader** (free) and use **File > Save As** to keep your answers under a new name before returning the file. Some browser and phone PDF viewers save marks as annotations rather than as answers, which can be lost when the file is read.

GUIDANCE — read before answering

Section 0 records three things: the dataset's name, the department or ministry that holds it, and the date the characterisation was done.

Tips to name the dataset: Use the name the dataset is actually known by internally, not a descriptive title invented for the form. If the dataset has both a system name and a common name, give both, so that someone searching either will find this record.

The date matters because a KYDS record ages. Sensitivity, access level, and retention all change as a dataset grows or as its use changes. Treat the date as the point from which this characterisation can be relied on, and re-do the form if the dataset materially changes.

Description of the dataset. This is the only field in the form that explains what the dataset actually is. Everything else records a classification; this records the substance. Write it for a reader who has never seen the dataset and does not work in your department.

Cover three things: what the dataset holds, what one row or record represents, and what it is used for. For example: "Records of ration card holders in the district. One row per ration card, listing the card number, head of household, family members, category, and linked bank account. Used by the Supply Office to authorise monthly entitlements and to reconcile disbursement against distribution."

This description should be consistent with the rest of the form. If it describes named individuals but Section 2A is ticked "Open (no PII)", or describes one row per household but Section 4 is ticked at district level, one of the two is wrong. Treat any mismatch as a prompt to re-check both.

See the guidance above for term definitions (Fiduciary, Processor, composite dataset, aggregation, related datasets).

Dataset name**Description of the dataset**

In your own words, describe what this dataset contains, what a single row or record represents, and what it is used for.

Department / ministry**Date of characterisation**

DD / MM / YYYY

GUIDANCE — read before answering

Section 1 records the **modality** of the dataset — the shape and format the data takes. Tick every category present. Full definitions for every format under each category are in Appendix A — Glossary of Section 1 formats; use it to identify the specific format(s) for the optional precision line, or to check whether something belongs in a listed category before writing it under “Other”.

Tick the category (or categories) that apply. See Appendix A — Glossary of formats to identify the right modality.

Structured / tabular

Semi-structured

Binary serialisation / interchange

Database native

Geospatial

Documents and text

Scanned / image-as-document

Raster images

Audio

Video

Biometric

Logs and event data

Scientific / research

Communication / messaging

AI / model

Synthetic / derived

Mixed (tick the specific categories present)

Other (does not fit any category above — describe)**Specific format(s) within ticked categories**

Optional, for precision.

GUIDANCE — read before answering**How to spot personal data (PII) — practical indicators**

Use this list to decide which tiers in Section 2A apply. A dataset contains PII if it includes any of the following about a living individual, alone or in combination:

- **Direct identifiers** — name, father's/spouse's name, photograph, signature
- **Government-issued ID numbers** — Aadhaar, PAN, voter ID, passport, ration card, driving licence
- **Contact details** — address, phone number, email address
- **Biometric identifiers** — fingerprint, iris scan, facial geometry, voiceprint
- **Device or online identifiers** — IP address, device ID, cookie ID, MAC address
- **Financial identifiers** — bank account number, UPI ID, card number
- Date of birth or age, combined with other fields
- **Location data tied to a person** — home address, GPS traces, travel history
- Health, genetic, or biometric records linked to a named or identifiable person
- Any field that, combined with others in the same record, narrows down to one person (e.g. pin code + date of birth + gender)

If none of the above are present and the data cannot reasonably be linked back to an individual, the dataset is likely “Open (no PII)”. If unsure, treat the dataset as containing PII until the Data Protection Officer confirms otherwise.

DPDP sensitivity tiers

Tier	What it means
Open (no PII)	Contains no personal data about any identifiable person.
Aggregated from personal data	Derived from PII; re-identification risk depends on cell size and linkability.
Quasi-identifiable	Identifies someone only in combination (e.g. pin code plus date of birth plus gender).
PII	Directly identifies a Data Principal.
Special-category PII	Protection comes mainly from sectoral and constitutional law, not DPDP alone. Sub-types and their governing law are listed below.

Special-category sub-types and their governing law

Sub-type	Governing law
Health / medical records	Sectoral health law; Mental Healthcare Act 2017 for mental health
HIV status	HIV/AIDS Act 2017
Biometric incl. voiceprint	Aadhaar Act 2016; constitutional privacy
Genetic data	Constitutional privacy
Caste / tribe	Articles 14 and 15; SC/ST (Prevention of Atrocities) Act 1989
Religion	Articles 14, 15 and 25
Sexual orientation	Constitutional privacy
Children's data (under 18)	DPDP Section 9
Financial data	Banking confidentiality; RBI Master Directions
Disability	Rights of Persons with Disabilities Act 2016
Trade union membership	Constitutional privacy
Political opinion / affiliation	Articles 19 and 21
Location traces over time	Constitutional privacy; Telecommunications Act 2023

IT Act classification levels

Level	What it means
Top Secret	National-interest — disclosure could cause exceptionally grave damage to national security.
Secret	National-interest — disclosure could cause serious damage to national security or interests.
Confidential	National-interest — disclosure could cause damage to national security or interests, below the threshold for Secret.
Restricted	National-interest — for official use only.
Unclassified	National-interest — no national-security protection required; select only to record that the national-interest scheme was considered and found not to apply.

A. Personal-data sensitivity (DPDP Act)

Tick every tier present in the dataset. If unsure which tier applies, see the guidance above.

Open (no PII)

Aggregated from personal data

Quasi-identifiable

PII

Special-category PII (tick sub-types below)

SPECIAL-CATEGORY SUB-TYPES

Health / medical records

HIV status

Biometric incl. voiceprint

Genetic data

Caste / tribe

Religion

Sexual orientation

Children's data (under 18)

Financial data

Disability

Trade union membership

Political opinion / affiliation

Location traces over time

Degree per tier (proportion of records / which fields)

--

B. Information / dataset classification (IT Act)

Select one level in each scheme that applies. If unsure which level applies, see the guidance above.

ORGANISATIONAL CLASSIFICATION

Confidential

Restricted

Internal use

Unclassified

NATIONAL-INTEREST CLASSIFICATION (ONLY IF APPLICABLE)

Top Secret

Secret

Confidential

Restricted

Unclassified

Level	What it means
Open data	Published, anyone, no gate.
Statutorily mandated disclosure	Public by law, e.g. MCA filings, RTI Section 4, electoral data.
Registered access	Identified users who accept terms.
Controlled access	Case-by-case approval for defined purposes.
Restricted (inter-agency / MOU)	Carries a lawful-basis claim that must be substantiated. An MOU on its own is not a lawful basis — cite the actual statute or exemption in the Addendum.
Embargoed	Held now, released on a trigger or date.
Internal only	Never leaves the holding department.

Pick the most restrictive level that governs the dataset. See the guidance above for definitions, including why an MOU alone is not a lawful basis.

Statutorily mandated disclosure

Controlled access

Embargoed

If a more open subset is published separately, what and at what aggregation

GUIDANCE — read before answering

Tick the finest level present in the dataset.

Note: If this dataset is itself an aggregation of more granular data held elsewhere, for example a national dataset built from state-level records, record that relationship in Section 7 rather than trying to capture both granularities here.

The identifiability flags exist because granularity alone can understate risk. A district-level dataset restricted to one sub-population, or linked to a persistent identifier like Aadhaar, can be far more identifying than its granularity label suggests.

What each level means — individual level

Level	What it means
Individual / Record level	One row per person or per named entity. The row describes that subject's attributes or current state. This is the finest granularity the form recognises and carries the highest identifiability.
Event / Transaction level	One row per occurrence rather than per subject: a payment, a visit, a grievance, an application, a sensor reading. The same person may appear in many rows. Do not read this as less identifying than record level. Timing, sequence, and frequency are themselves identifying, and a run of dated events builds a behavioural trail that a single record does not.
Household / Family	One row per household or family unit. Not anonymous. Household composition plus an address usually resolves to named individuals, and a household row often carries attributes of every member.

Local level

Level	What it means
Building / Property / Parcel / Survey Number	Tied to one specific physical property. Where a property has a single owner or occupant, this is individual level in practice, whatever the label says.
Street / Locality / Village	A named street, urban locality, or revenue village. In sparsely populated areas, village level can approach household level.
Gram Panchayat / Municipal Ward	The lowest elected administrative unit. Population per unit varies enormously between and within states, so the same label can mean very different things for identifiability.
Community / Neighbourhood	An informally bounded area such as a settlement, colony, or basti. Boundaries are usually not standardised, which makes these units hard to compare across datasets and easy to mis-join.
Block / Taluk / Tehsil / Mandal	The sub-district administrative unit. The name differs by state; all four sit at the same tier.
District	The standard administrative unit below the state.

Regional level

Level	What it means
City / Municipality / Urban Local Body	The jurisdiction of a single urban local body.
Metropolitan Region	A planning or development region that may span several local bodies and cross district lines.
State / UT	State or union territory.
Regional / Zonal Office	An administrative catchment defined by the holding department itself rather than by standard geography. These boundaries often do not map onto districts or states, so note what the zone actually covers.
National	Countrywide.

Statistical level

Level	What it means
Aggregated / Statistical only	Counts, totals, rates, or indices, with no underlying records held in this dataset. If the record-level source is held elsewhere, that belongs in Section 7 under aggregation.
Mixed granularities	Different parts of the dataset sit at different levels, for example record-level rows for one scheme and district totals for another. Tick this and say in Section 7 whether it should really be characterised as a composite dataset instead.

Identifiability flags

Answer these separately from the granularity tick. Each can raise real-world identifiability well above what the granularity level suggests.

Restricted to a sub-population? The dataset covers only a defined group, for example only disability certificate holders, only one caste category, or only patients on a particular treatment. Membership of the dataset is itself a disclosure. A district-level count stops being neutral when everyone counted shares a sensitive attribute.

Linked to a persistent identifier? Aadhaar, PAN, a beneficiary ID, a case number, or any key that stays stable across systems. A persistent identifier lets this dataset be joined to others, so its effective sensitivity becomes the sensitivity of everything it can be joined to, not just what it holds on its own.

Longitudinal? The same subjects are observed repeatedly over time. Repeat observations narrow identity quickly; a small number of dated points is often enough to single out one person even where no direct identifier is present.

Select the finest level present — one option across all four groups below. See the guidance above for why the identifiability flags matter even at coarse granularity.

INDIVIDUAL LEVEL

Individual / Record level

Event / Transaction level

Household / Family

LOCAL LEVEL

Building / Property / Parcel / Survey Number

Street / Locality / Village

Gram Panchayat / Municipal Ward

Community / Neighbourhood

Block / Taluk / Tehsil / Mandal

District

REGIONAL LEVEL

City / Municipality / Urban Local Body

Metropolitan Region

State / UT

Regional / Zonal Office

National

STATISTICAL LEVEL

Aggregated / Statistical only

Mixed granularities

IDENTIFIABILITY FLAGS

Restricted to a sub-population?

No

Yes

Linked to a persistent identifier?

No

Yes

Longitudinal?

No

Yes

GUIDANCE — read before answering**Update frequency**

Record the cadence at which the dataset actually changes, not the cadence policy says it should. Frequency matters for two reasons: it tells a user how current the data is, and it indicates how long a stale copy stays wrong once shared.

Most options are self-explanatory. The ones that get confused:

- **Streaming / Continuous** — data arrives as a constant flow with no batch boundary. Distinct from real-time, which describes latency rather than shape.
- **Real-time, Near real-time, Hourly, Daily** and the rest — these describe how quickly a change in the world appears in the dataset. Pick the observed interval.
- **Event-triggered** — updates when something happens, not on a clock. State the trigger if it is not obvious from the dataset name.
- **Irregular / Ad-hoc** — updated, but on no defined schedule.
- **Historical / Archival** — closed. No further updates are expected.

Retention

Retention records how long the dataset is kept before it is deleted or moved to archive, and on whose authority. This is often the least documented field in a department and the one most likely to be challenged, so state the actual position rather than the intended one.

- **Indefinite** — kept with no end date. This is a decision, not a default. For a dataset containing PII, expect to justify why no end point applies.
- **Defined statutory period** — a law, rule, or records-retention schedule sets the period. Cite the instrument and the period in the citation line, for example “Rule X of Y Rules, 8 years from closure.”
- **Until purpose served** — kept only while the purpose for which it was collected is live, then erased. This is the default expectation for personal data under DPDP. Write the purpose into the citation line, so that “served” is testable by someone other than the team that collected it.
- **Until specific event** — a defined trigger ends retention: case closure, scheme wind-up, contract expiry, appeal period lapse. Name the event.
- **Operational discretion** — no fixed rule; the holding team decides when to delete. The weakest position on the list. Workable for routine non-personal operational data, difficult to defend for PII.
- **Not yet determined** — an honest placeholder for a new dataset. It should not stay selected. Treat it as an open action with an owner, not as an answer.

UPDATE FREQUENCY

Select the observed cadence.

Streaming / Continuous	Real-time (< 1 min)
Near real-time (< 1 hour)	Hourly
Daily	Weekly
Fortnightly	Monthly
Bi-monthly	Quarterly
Half-yearly	Annual
Multi-year	Event-triggered
Irregular / Ad-hoc	Historical / Archival

RETENTION

Indefinite	Defined statutory period (cite below)
Until purpose served	Until specific event
Operational discretion	Not yet determined

Retention citation / purpose / event

--

GUIDANCE — read before answering

This records the physical and technical location of the dataset — not who may access it (that is Section 3), but where it actually resides. Storage location matters because it determines which security duties under the IT Act 2000 and CERT-In Directions apply (network exposure, the 180-day log retention requirement, and the six-hour breach reporting window), and because data held offline, on-premise, or with a cloud provider each carries a different threat profile, jurisdiction, and recovery path in the event of an incident.

Tick every location where this dataset resides. See the guidance above for why storage location matters.

OFFLINE

Local file server

Storage area network (SAN)

Airgapped / offline storage

Network-attached storage (NAS)

On-premise data warehouse

Physical media

ONLINE

Cloud-based storage

GUIDANCE — read before answering**Standalone or composite**

- **Standalone dataset** — one data product with consistent characteristics throughout.
- **Composite dataset** — one named resource holding several sub-datasets with different characteristics. Name the parent and the specific sub-dataset: complete a separate form for each sub-dataset.

Example of a composite dataset: a “Beneficiary Welfare Database” that bundles ration card records (structured, Restricted), bank account linkage records (structured, PII, Controlled access), and disbursement logs (event data, Internal only). Each sub-dataset here has different sensitivity, format, and access needs, so each gets its own KYDS form, naming the parent dataset each time.

Provenance

Record whether the department collected this data itself or received it from another body. If received, name the body and state how the transfer was authorised. An MOU or a data-sharing arrangement is a mechanism, not an authorisation; the lawful basis behind it belongs in the Addendum.

Aggregation of more granular data

Some datasets held centrally are aggregations of record-level data that originates at, and continues to be held at, a more granular level. For example, a central dataset built from state-level civil registration records, where the state retains the underlying record-level data separately. If this applies, note the source granularity and the authority holding it, and flag whether that source dataset has, or needs, its own KYDS form. This keeps the two datasets — the granular state-held original and the aggregated central version — separately and correctly documented, rather than one form trying to describe both.

Related datasets

This is different from a composite dataset. A composite dataset is one resource bundling sub-datasets internally; “related datasets” is for noting links to other, separately held datasets. For example, a dataset that is derived from another, duplicates another, or is regularly joined to another via a common identifier (such as Aadhaar, PAN, or a case ID). Use this to flag dependencies an access decision-maker would want to know about, even if those other datasets aren't governed by this same form.

STANDALONE OR COMPOSITE?

Standalone dataset

Composite dataset (name parent and sub-dataset below)

Parent dataset / sub-dataset (if composite)

PROVENANCE

Originally collected by this department

Received from another body

If received: from whom and how authorised

AGGREGATION OF MORE GRANULAR DATA?

No

Yes — aggregates more granular data held elsewhere

If yes — source granularity

If yes — authority holding the source data

If yes — does the source have its own KYDS form? (Y / N / reference)

RELATED DATASETS

None known

Related to other dataset(s) — complete below

Related dataset name(s)

Nature of relationship

Source/copy, aggregation, linked key, feeds into, other.

GUIDANCE — read before answering

If access is Restricted: cite the lawful basis — an express statute, DPDP Section 7(b) or 7(c), an established Section 17 exemption other than 17(2)(a), or Section 3(c)(ii) for data that is already publicly available. An MOU between agencies is not, by itself, a lawful basis.

If Special-category PII is present: cite the sectoral statute or constitutional protection that applies, from the table in the Section 2 guidance above.

Standing obligations apply regardless of any answer elsewhere in the form: every electronic dataset carries IT Act 2000 security duties and the CERT-In Directions (incident reporting within six hours, 180-day log retention). Personal data also carries DPDP Section 8(5) security obligations and the rights of Data Principals.

See the guidance above for which lawful basis or statute to cite.

If Restricted access — lawful basis cited

If Special-category — sectoral / constitutional basis cited

Data Protection Impact Assessment done?

Yes

In progress

No

Not required

Notes

Standing obligations: IT Act / CERT-In security duties apply to every electronic dataset. DPDP Section 8(5) and Data Principal rights apply wherever personal data is present.

ABOUT THIS GLOSSARY

Full definitions for every format under each Section 1 category, grouped the same way. The form only asks you to tick the category; use this to identify the specific format(s) if you want to note them in the optional precision line in Section 1 — Modality, or to check whether something belongs in a listed category before writing it under “Other”.

Structured / tabular

CSV	comma separated text table, opens in Excel
Excel (.xlsx / .xls)	Microsoft spreadsheet
Feather / Arrow	fast memory based table format
Fixed-width flat files	text where each field has a set column width
ODS	LibreOffice / OpenOffice spreadsheet
ORC	compressed columnar format (Hadoop)
Parquet	compressed columnar table for big data
R data (.rds / .rdata)	R statistics data file
SAS (.sas7bdat)	statistics software data file
SPSS (.sav)	statistics software data file
Stata (.dta)	statistics software data file
TSV	tab separated text table

Semi-structured

EDI (EDIFACT / X12)	standard for business documents between systems
HTML (as data)	web page markup used to carry data
JSON	nested key and value data, common in web and APIs
JSON Lines (NDJSON)	one JSON record per line
SDMX	standard for exchanging statistical data
TOML	simple config file format
XML	tagged text for structured data
YAML	readable config or data format

Binary serialisation / interchange

Avro	compact binary with schema
Protobuf	Google's compact binary format
Thrift	cross language binary format
MessagePack	compact binary JSON
FlatBuffers	fast access binary format
Cap'n Proto	very fast binary format
CBOR	compact binary object format

Database native

Relational SQL	tables of rows and columns, related to one another through shared key fields rather than stored as one flat table (e.g. PostgreSQL, MySQL, Oracle, SQLite). This is what “relational” refers to in a database context — not a relationship between separate datasets (see “Related datasets” under Section 7 for that).
NoSQL document	flexible JSON like records (MongoDB, Firestore, CouchDB)
Columnar	column oriented store for scale (Cassandra, HBase, BigQuery)
Key-value	simple lookup by key (Redis, DynamoDB)
Graph	nodes and relationships (Neo4j, Neptune)
Time-series	timestamped measurements (InfluxDB, TimescaleDB)
SQLite file (.db)	single file database

Geospatial

Shapefile (.shp)	common GIS map layer of points, lines or areas
GeoJSON	map features in JSON
KML / KMZ	Google Earth map data
GeoTIFF	image with map coordinates, e.g. satellite
GeoPackage (.gpkg)	single file GIS database
DXF / DWG	CAD drawing files
WKT / WKB	text or binary representation of shapes
TopoJSON	compact map format that keeps shared borders
GML	XML for geographic data
NetCDF	gridded scientific or geo data

Documents and text

PDF	fixed layout document
Word (.docx / .doc)	Microsoft document
ODT	LibreOffice document
Plain text (.txt)	unformatted text
Markdown	lightly formatted text
RTF	rich text document
HTML pages	web pages
LaTeX	typesetting source for documents
EPUB	e-book format

Scanned / image-as-document

BMP	uncompressed bitmap
DICOM	medical imaging (X-ray, MRI, CT)
GIF	simple or animated image
JPEG	compressed photo
JPEG / PNG of documents	photos or scans of documents
PDF (image-based)	scanned PDF that cannot be searched as text
PNG	lossless image with transparency

Raster images

RAW	unprocessed camera image
TIFF (scanned)	high quality scanned pages
TIFF	high quality image
WebP	modern compressed web image

Audio

AAC	compressed audio (Apple)
AMR	speech or voice recording (mobile)
FLAC	lossless compressed audio
M4A	compressed audio
MP3	compressed audio
OGG	open compressed audio
WAV	uncompressed audio

Video

AVI	older video container
MKV	flexible video container
MOV	Apple video
MP4	common compressed video
MPEG-2	broadcast or DVD video
WebM	open web video

Biometric

ANSI/NIST-ITL	standard for fingerprint, face and iris exchange
CBEFF	common biometric exchange format
FingerMinutiae	fingerprint feature points
ISO 19794	international biometric data standard
WSQ	compressed fingerprint image

Logs and event data

Apache / Nginx access logs	web server request records
CDR	call detail records, phone call metadata
CEF	common event format (security)
JSON logs	structured log records
LEEF	log event format (security)
PCAP	captured network traffic
Plain text logs (.log)	system or application activity records
Syslog	standard system log format

Scientific / research

NetCDF	gridded scientific data
HDF5	large hierarchical scientific data
FITS	astronomy data and images
MATLAB (.mat)	MATLAB data file
NumPy (.npy / .npz)	Python array data
Zarr	chunked array storage for big data

Communication / messaging

MIME / EML	a single email message
PST / OST	Outlook mailbox archive
SMS / MMS archives	exported text or multimedia messages
WhatsApp export (.txt / .zip)	exported chat history

AI / model

ONNX	portable machine-learning model
SafeTensors	safe model weights format
Pickle (.pkl)	Python serialised object (caution: can run code)
TfRecord	TensorFlow training data
JSONL	one JSON record per line (training data)
Hugging Face datasets	ML dataset format

Synthetic / derived

Synthetic datasets	artificially generated, no real individuals
Aggregated summary tables	totals and counts, not individual records
Anonymised / pseudonymised variants	identifiers removed or masked