



OPEN DATA · ARTIFICIAL INTELLIGENCE · PUBLIC INFRASTRUCTURE

TECHNICAL REPORT · TD-TR-2026-04

DamLevels-Data

From PDF bulletins to linked, AI-queryable open data on Tunisia's large dams, 2014–2025.

AUTHORS

Mohamed Hamza Nasfi · Soura Boulaares
Tarek Gasmi

PUBLISHED

October 2026

STATUS

Independently produced and published

VERSION · LICENSE

v1.0.1 · CC BY 4.0

DATASET AND SOFTWARE

[doi:10.5281/zenodo.23304326](https://doi.org/10.5281/zenodo.23304326)

REPOSITORY

[tanitdata/DamLevels-Data](#)

ABSTRACT

Every day the Tunisian dam administration (DGBGTH) publishes the state of the country's large dams as a three-page PDF bulletin: public, but locked away, with a drifting layout and column headings that do not mean what they seem. DamLevels-Data takes this source up the five levels of the open-data model, from a PDF on the web to linked data, and one step further, to a read-only Model Context Protocol (MCP) server for AI assistants. Deterministic engines read 11,927 pages without OCR and publish 129,390 daily storage values for 36 dams (2014–2025), each traced to its page fingerprint, column and printed characters. Five complementary checks audit the values; source errors are documented, never repaired. The DamLevels Ontology maps 84 printed columns to 33 concepts and 16 time rules on W3C and OGC vocabularies. The MCP server enforces the same rules at the point of use: it refuses dates outside the record, never fills a gap, and returns every number with its source page and caveats.

1. Introduction

Tunisia is a water-scarce country that depends on a network of large dams, most of them in the wetter north, to store winter rain for drinking water and irrigation. The **Direction Générale des Barrages et des Grands Travaux Hydrauliques (DGBGTH)** of the Ministry of Agriculture records their state every day, and the **Observatoire National de l'Agriculture (ONAGRI)** publishes it as a bulletin, *Situation des barrages* [12]: three PDF pages with the stock, inflow and release of every dam, its levels and capacities, and the national totals and transfers between dams. The figures are public and authoritative. They are also, for any computational use, effectively locked away.

There is no table and no API. Each day is a PDF whose layout has changed over the years, whose numbers use a French decimal comma, and whose dam names are printed in French and Arabic with varying spellings. Worse, several column headings do not mean what they seem to say: the value printed under “release of the day” is, in one of its two columns, the release on the same day *one year earlier*; flows printed on the bulletin of day D describe day $D - 1$; running totals restart on the bulletin of the 2nd of the month, not the 1st. A reader who copies these numbers into a spreadsheet produces values that are plausible, specific and wrong.

DamLevels-Data is the dataset that results from treating this as a data-engineering problem rather than a scraping exercise. This report organises the work along the **five-star open-data model** proposed by Tim Berners-Lee [1], because each star corresponds to a distinct engineering stage with its own evidence: the bulletin starts at one star, a PDF on the web, and each stage below lifts it by one level, up to linked data with an ontology. A final stage goes beyond the five stars: a read-only **Model Context Protocol (MCP)** server, added as the last step of the repository, through which an AI assistant can answer questions from the archive without being able to break its rules.

What was produced. The published record covers **2 July 2014 to 10 October 2025: 3,985 bulletin dates and 11,927 PDF pages**, from which **129,390 daily storage values for 36 dams and 47,556** qualified transfer facts were extracted. Every storage value carries its file, the SHA-256 fingerprint of that page, its column and its raw printed characters. The record is closed: no later bulletin is included and nothing is extrapolated beyond it. Dataset, code and server are published on GitHub [13] and archived together on Zenodo ([doi:10.5281/zenodo.23304326](https://doi.org/10.5281/zenodo.23304326)).

Where this comes from. The extraction, audit, watershed and climate work was carried out as a Licence capstone project at MUST University, within the DataDolt × MUST University incubator and for TanitData [2]. The public repository packages the accepted outputs of that project, adds the ontology view, recomputed quality indicators and the MCP server, and was hardened by an independent review before release. The contribution claimed here is not any single tool. It is the end-to-end discipline that lets a number travel from a printed PDF cell to an AI assistant’s sentence without losing the meaning, the date and the source that make it true.

2. Background

2.1 The five-star open-data model

The five-star scheme grades how usable published data is, each star adding to the previous ones (Table 1). It is deliberately simple, which makes it a good map for a data-engineering project: the hard work is not reaching a star but reaching it *without losing information on the way*. A CSV that drops the date semantics of its columns is three-star in format and worse than the PDF in meaning.

Level	Criterion [1]	DamLevels-Data stage (section)
★	On the web, under an open licence, in any format	The DGBGTH bulletin as published by ONAGRI: one PDF per day. Collected, fingerprinted and inventoried (§3).
★★	Structured, machine-readable data	Deterministic extraction of every page into tables, with cell-level lineage (§4).
★★★	Non-proprietary, open formats	Tidy CSV and Parquet tables, a Frictionless descriptor, a data dictionary, audited quality indicators and settled licences (§5).
★★★★	URIs to denote things	Stable identifiers for dams, observations, bulletins and pages; ontology IRIs; a Zenodo DOI (§6).
★★★★★	Linked to other data	The DamLevels Ontology (DLO 0.3), built on W3C and OGC vocabularies, with SPARQL queries over sample observations (§7).
Beyond	Machine-actionable for AI assistants	A read-only MCP server that answers from the published tables and enforces their rules (§8).

Table 1. The five-star open-data model and the stage of this project that implements each level. Section 10 states where a level is only partly reached: the ontology IRIs are not yet dereferenceable, and the links reach shared vocabularies but not yet other datasets.

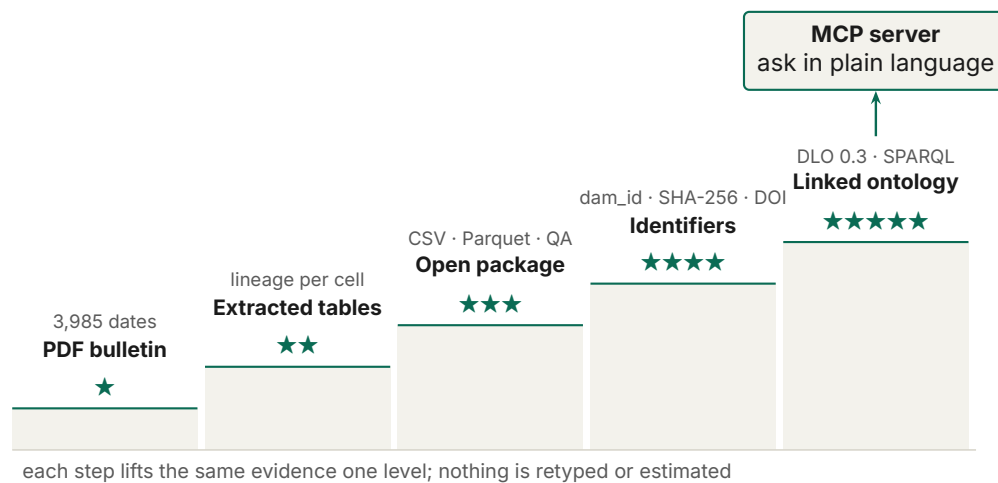


Figure 1. The staircase followed in this report. Each level is an engineering stage with its own artefacts and checks; the MCP server sits on top of the fifth level and only reads what the lower levels publish.

2.2 The bulletin as a three-page information system

Each bulletin has three pages with different roles (Figure 2 shows the first).

- **Page 1 — inflows, releases and stock, per dam.** Fifteen numeric fields per dam row, grouped by region with printed subtotals and a national total. Only two of the fifteen are the state on the bulletin day (stock and fill); the others are flows of the previous day, running totals and reference values from earlier years. The published storage table comes from this page.
- **Page 2 — the technical sheet.** Levels, capacities, sedimentation, rain at the dam and a second copy of storage, inflow and release, which serves as an independent cross-check.
- **Page 3 — national, regional and transfer information,** in Arabic, including the water transferred between dams through the northern network.

MINISTRE DE L'AGRICULTURE

APPORTS ET LACHERS DES BARRAGES

الجمهورية التونسية

وزارة الزراعة

المديرية العامة للسدود والتحكم الهيدروليكي

DIRECTION GENERALE DES BARRAGES ET DES GRANDS TRAVAUX HYDRAULIQUES

du 31-01-20 observée le 01-02-20

BARRAGES	APPORTS					LES LACHERS					المخزون بالسدود Stocks aux barrages					المستودع				
	du jour 31-01-20	du mois au 31-01-20	Moyenne du mois	Cumulés du 31-01-20	du 1-09-19 au 31-01-19	الفترة du jour 31-01-20	du jour 31-01-20	du mois au 31-01-20	Cumulés du 31-01-20	du 1-09-19 au 31-01-19	الفترة du jour 31-01-20	du jour 31-01-20	du mois au 31-01-20	Cumulés du 31-01-20	du 1-02-20 au 31-01-19					
																	%	المتغير	%	المتغير
MELLEQUE	0.000	1.809	9.454	33.943	136.915	97.091 p	0.086	0.924	2.678	33.209	128.019	50.0	25.556	31.804	ساق					
BEN METIR	0.038	1.138	9.882	6.595	34.137	20.229 T.p	0.099	0.087	2.564	10.876	15.115	67.1	40.544	59.696	بني مطير					
KASSEB	0.047	1.389	11.175	7.960	29.309	19.906 T	0.043	0.077	1.339	9.132	13.204	68.1	52.339	51.523	كاسب					
BARBARA	0.012	0.735	20.858	13.004	60.573	40.922 D	0.009	0.000	3.257	10.947	58.750	100.1	64.755	56.150	بربرة					
SIDI SALEM	0.604	20.520	93.870	114.380	222.608	155.591 p	1.038	0.043	8.809	140.736	39.703	78.6	456.345	385.734	سدي سلم					
BOU-HEURMA	0.076	4.115	22.697	19.879	66.091	33.057 E	0.128	5.203	2.081	16.510	23.355	49.6	55.592	86.614	بو هرتمة					
JOUMINE	0.100	4.592	20.710	21.062	63.116	34.217 E.p	0.182	0.440	3.651	30.309	21.668	51.2	60.843	88.650	جومين					
GHEZALA	0.016	0.689	2.781	2.195	7.251	3.996 E	0.005	0.022	0.060	0.826	0.577	79.4	7.943	10.079	غزالة					
MELAH	0.003	2.810	5.700	-	-	11.705	0.000	-	0.000	-	-	80.5	32.995	-	المايح					
SEJNANE	0.054	13.627	30.892	49.253	128.844	63.997 tr.p	0.294	3.726	10.105	45.570	79.821	67.4	90.281	131.305	سجنان					
Zarga	0.002	1.282	3.860	18.012	30.458	12.146	0.000	1.481	0.207	2.131	23.566	94.7	22.978	22.780	الزرق					
Kebir	0.008	1.564	11.004	9.466	-	19.554	0.000	0.000	1.492	9.187	-	9.3	5.970	5.149	الكبير					
Moula	0.001	1.349	7.572	7.214	-	18.855	0.000	0.690	1.073	6.609	-	97.6	25.705	23.665	المولى					
S. EL BARRAK	0.016	10.927	76.508	84.203	252.799	148.503	0.000	8.645	9.587	48.952	185.207	98.9	283.496	280.478	سدي البراق					
Zlatine	0.000	1.226	5.100	3.467	33.673	17.969 tr	0.004	1.436	0.183	9.267	22.992	97.7	32.533	32.238	الزلاطين					
Gangoum	0.001	0.719	2.486	12.179	15.260	3.916	0.000	0.507	0.000	1.061	9.075	96.0	17.565	17.235	الغانغوم					
SILIANA	0.004	0.350	5.662	5.374	22.258	18.247 E	0.036	0.028	0.793	2.999	0.447	66.8	22.338	30.289	سيلانة					
LAKHMESS	0.000	0.156	1.628	0.814	3.158	6.041 E	0.014	0.000	0.280	0.963	0.748	70.6	5.254	5.468	لكشمس					
RMIL	0.001	0.024	2.024	0.479	4.458	5.640 d	0.005	0.000	0.011	0.177	1.370	74.5	1.491	1.805	الرميل					
S/T Nord	0.986	69.020	343.862	409.478	1 110.907	731.581	1.944	23.309	48.150	379.461	623.615	74.7	1 304.524	1 320.661	مجموع سدود الشمال					
BIR M'CHERGA	0.010	0.288	4.859	4.652	22.280	14.680 E	0.002	0.000	0.047	1.355	0.670	84.5	35.146	32.773	بئر المشرقة					
RMEL	0.013	0.388	1.251	3.768	20.534	11.440 p	0.008	0.010	0.248	1.250	0.977	50.9	12.120	19.602	الرميل					
NEBHANA	0.012	0.570	2.488	3.053	16.400	12.318 E	0.009	0.021	1.145	2.755	2.796	24.6	14.480	17.083	نهبانة					
SIDI SAAD	0.000	0.539	1.487	5.581	57.721	30.247 E	0.036	0.139	0.550	3.580	10.149	43.3	57.692	76.415	سدي سعد					
EL HAOUAREB	0.002	0.019	1.377	2.400	10.267	9.890	0.000	0.009	0.166	0.785	0.490	0.4	0.300	3.894	هوارب					
Sidoua	0.000	0.009	13.261	0.047	3.187	17.969	0.000	0.000	0.000	0.000	0.200	20.2	1.557	2.346	سيديوة					
SIDI AICH	0.000	0.000	0.016	0.458	5.990	1.386	0.000	0.000	0.000	0.000	0.600	5.0	4.440	4.438	سدي عيش					
EL BREK	0.000	0.000	0.194	0.138	3.320	1.386	0.000	0.000	0.000	0.000	0.000	8.7	1.400	1.916	البرك					
S/T CENTRE	0.038	1.814	24.932	20.097	145.689	99.316	0.054	0.179	2.096	9.725	15.883	28.5	127.136	158.458	مجموع سدود الوسط					
BEZIRK	0.000	0.012	0.849	1.896	0.000	2.982	0.000	0.000	0.307	8.954	80.5	4.033	4.804	بزر						
CHIBA	0.001	0.044	1.091	1.542	17.750	4.051 E	0.007	0.010	0.204	1.249	8.013	25.7	1.296	4.610	شيبا					
MASRI	0.001	0.021	0.692	0.484	2.640	2.665	0.000	0.010	3.163	3.642	5.567	45.6	2.367	2.204	مصري					
LEBNA	0.013	0.241	3.760	18.511	30.554	11.100 E	0.003	0.003	0.058	5.906	10.122	95.3	22.211	22.482	لبنة					
HMA	0.001	0.043	0.433	0.322	2.420	2.976	0.000	0.000	0.000	0.182	0.013	15.5	1.990	2.772	الحمة					
ABID	0.000	0.026	0.734	3.206	10.407	3.324 E	0.006	0.002	0.026	0.773	3.707	81.8	7.864	9.451	العبد					
S/T CAP-BON	0.016	0.387	7.559	25.961	63.771	27.098	0.016	0.025	3.451	12.058	36.375	65.2	39.762	46.323	مجموع الوطان القبلي					
TOTAL	1.040	71.221	376.354	455.535	1320.37	857.996	2.014	23.513	53.697	401.244	675.873	65.3	1 471.421	1 525.441	المجموع العام					

Pompes et transfert (Mm3)	du jour	Cumulé à ce jour du jour précédente	Cumulé à ce jour du jour précédente	المخزون du jour du jour précédente
Barbara (zultina)	0.000	0.578	0.020	بربرة
Sidi el Barrak	0.000	15.689	14.332	سدي البراق
Lacher vers Ichkeul	0.009	11.064	1.404	بحيرة الشكّل

Côte	Volumen lachés (Mm3)			المخزون
	Jour	Mois	Saison	
R.S. (g/l)	1.800	35.80	0.000	45.134
Régie El amrouse	0.00	0.00	0.00	0.00
Small Totesse (Po)	0.00	0.00	0.00	0.00
K13 sur Melleque	0.50	0.00	0.00	0.00

D : Déversement E : Exploitation T : Transfert a : déstockage pour abaissement du plan d'eau, d : déversement, T: Turbinage, P: suite de drainage

Figure 2. Page 1 of the bulletin of 1 February 2020, one of the two sample bulletins redistributed in the repository. The header reads *du 31-01-20 observée le 01-02-20*: flows describe the previous day, stocks the bulletin day. Source: DGBGTH, published by ONAGRI.

2.3 The traps that shape every later decision

- **Two dates in one row.** Stock, level and fill printed on bulletin *D* are the state on day *D*; inflow, release, transfer and evaporation are amounts over day *D* – 1.

- **A misleading heading.** Under *Lâchers du jour* two values are printed. The first is the release over day $D - 1$; the second, extracted as `p1_release_today`, is the release on the *same day one year earlier*.
- **Counters with a one-day lag.** Month-to-date totals run from the first day of the month of $D - 1$, so they restart on the bulletin of the **2nd**; season totals start on 1 September and restart on the bulletin of 2 September.
- **Layout drift.** In July 2014 Page 1 printed fewer dams and no fill column; dams join the roster later (Sarrat only from April 2021); one Page 2 column changes meaning on 27 April 2024.
- **Blanks are not zeros**, and a printed zero is not a blank. Neither may be filled.
- **Names are not identifiers.** The same dam appears under several French spellings and an Arabic name; one Page 2 column (R.S.) is not documented by the publisher at all.

3. Level ★: Collect and fingerprint the source

The starting point is a PDF on the web. ONAGRI's listing page links each day's bulletin; the national open-data licence (*Licence Nationale de Données Publiques Ouvertes*) is the default for ONAGRI's open data, although the bulletin page itself states no licence. The first engineering stage does not change a single value. It establishes *which* files the dataset is built from, so that anyone can obtain the same bytes.

- **Fingerprinting.** Every file is hashed with SHA-256 and split into its three pages; each page is hashed again. Published values refer to the hash of the exact page they were read from.
- **A logical manifest** (`sources/manifest.csv`) has one row per bulletin date: 3,985 dates and 11,927 pages, of which 23 dates have an incomplete set of pages. For 3,979 dates the original ONAGRI URL is recorded with the evidence that matched it (a page-hash chain, or the listing date plus the server filename); six dates have no matching listing entry.
- **A physical inventory** (`sources/file_manifest.csv`, 31,843 rows) keeps every file found, including duplicates, backups and files whose name disagrees with their content. Two Page 3 files of September 2019 were re-dated from their content and fingerprint rather than from their filename.

The bulletin corpus itself is not redistributed; two complete sample bulletins are included so a reader can check the extraction against the original without downloading anything.

4. Level ★★: Read the pages into structured data

4.1 Deterministic engines, no OCR

The bulletins are born-digital: their characters and table rulings exist in the PDF itself. Three frozen engines read them directly, Pages 1 and 2 with pdfplumber and Page 3 with PyMuPDF. No image is rendered and **no OCR is used anywhere**, so running an engine

twice on the same file yields the same output. Figure 3 shows the whole path from bulletin to published tables.

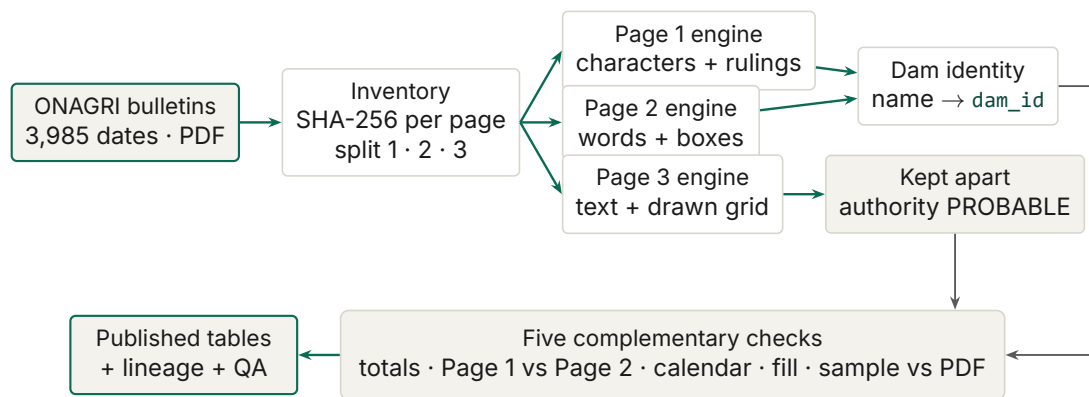


Figure 3. Extraction and packaging. Each page type has its own engine; rows that break a rule leave through a quarantine exit instead of being silently fixed; Page 3 facts are kept apart with authority **PROBABLE**. Checks compare values but never create or overwrite one.

4.2 How one page is read

The Page 1 engine first fingerprints the file, then locates the table from its vertical rulings. The number of columns identifies the layout (the layout with a fill column has sixteen); an unknown layout goes to quarantine rather than being guessed. Each printed dam name is matched to a governed roster of 36 dams by exact alias, never by fuzzy similarity; an unmatched name is quarantined. Each character is assigned to the cell that contains its centre, and the cell's characters are parsed with the French decimal comma: 283,496 becomes 283.496, a blank stays blank. Finally the dams of each region are summed and compared with the printed subtotal. The difference is recorded as evidence; it is never used to correct a cell.

Across the corpus, Page 1 yields 129,391 dam-day observations, Page 2 129,472 and Page 3 274,183 facts. The Page 1 lineage scan traced 1,896,094 cells.

4.3 Every value keeps its passport

A number without its origin cannot be checked. Every published storage value therefore carries its provenance in the row, and its cell-level evidence in `provenance/storage_lineage.csv`. Figure 4 follows one value.

\$ EL BARRAK	0,016	10,927	76,508	84,203	252,799	148,503	0,000	8,645	9,567	48,952	185,207	98,9	283,496	280,478	سیدی البراق
Column	Value														Meaning
dam_id, date	sidielbarrak, 2020-02-01														stable identifier; stock is a state on the bulletin day
value, unit	283.496, Mm3														converted from the printed 283,496
source_file, page	01-02-2020_1.pdf, 1														the page it was read from
source_sha256	0bb40123..6032c5														fingerprint of that exact page
source_field	p1_stock_today_mm3														the printed column
engine_version	page1-authoritative-0.2.3														the reader that produced it
lineage record	column 14, bounding box, raw text 283,496, NUMERIC_PARSED														the cell itself, and the check against the printed regional subtotal

Figure 4. From a printed cell to a row of data: Sidi El Barrak on the bulletin of 1 February 2020. If a number looks wrong, a reader can open the original page, go to the cell and see what was printed.

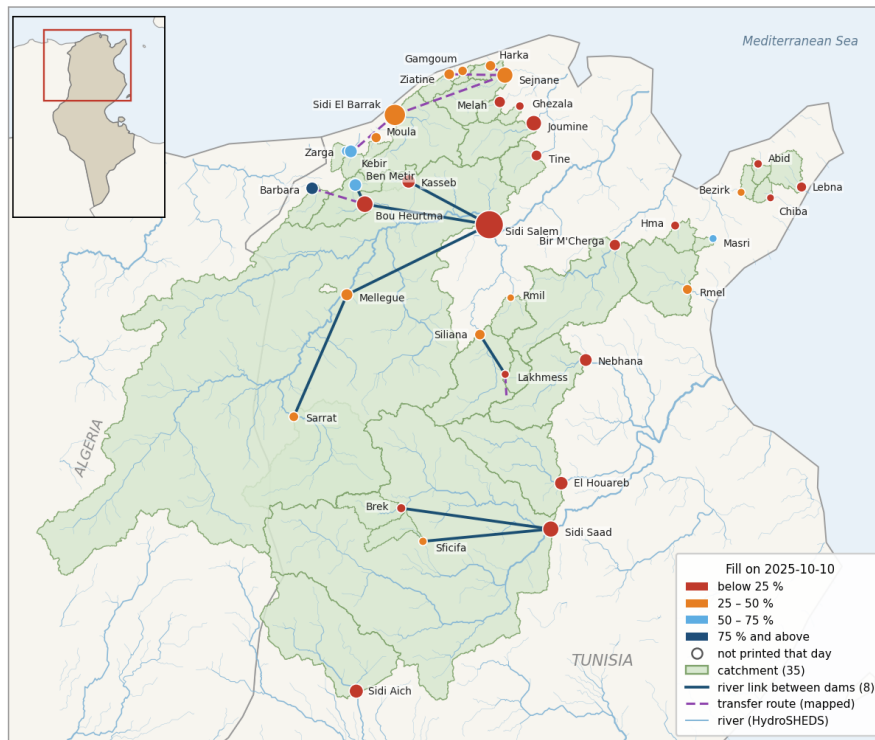
5. Level ★★★: An open, documented and audited package

5.1 Tables that explain themselves

The third star asks for open formats. The package goes further, because a CSV without meanings, units and exceptions is open but not usable; the FAIR principles [7] ask for meanings, identifiers, provenance and licences as well. Four things travel together: the data, a dictionary that says what every field means, the provenance that leads back to the PDF, and the audit with its exceptions.

- **Long ("tidy") tables in CSV and Parquet:** `storage` (129,390 rows, one per `dam_id` × date), `dams` (36), `capacity_history` (60 dated capacity regimes) and `transfers` (47,556 Page 3 facts: 35,667 physical transfer edges and 11,889 accounting lines).
- **A Frictionless Data descriptor [11]** (`datapackage.json`) declares 15 resources with their schemas and per-resource licences; every one of the 168 schema fields has a description.
- **Explicit missingness.** A printed cell without a usable value produces no row, and nothing is interpolated, forward-filled or set to zero. The 44,772 such cells, 44,771 printed-fill cells and one storage cell (Barbara, 27 January 2017, printed #N/A), are listed in `provenance/storage_missingness.csv`; a dam not yet on the printed roster simply has no rows before its first appearance. Capacity is looked up from the regime in force on the date and is reported `UNAVAILABLE` or `AMBIGUOUS` rather than filled.
- **GIS layers** (GeoJSON, EPSG:4326): the 36 dams, 35 accepted catchments, 34 incremental catchments, 8 natural river links and 10 transfer routes, with a QGIS project.
- **Licences settled per part:** source values under the national open-data licence with the DGBGTH/ONAGRI attribution; project tables, documentation and ontology under CC BY 4.0; catchment layers derived from FABDEM and MERIT Hydro under CC BY-NC-SA 4.0; code under MIT.

Tunisia's 36 large dams



Marker size: reference capacity. Fill = storage / capacity in force on that date. Data: DGBGTH bulletins via ONAGRI (Licence Nationale de Données Publiques Ouvertes); catchments CC BY-NC-SA 4.0 (FABDEM, MERIT Hydro); outlines Natural Earth; rivers derived from HydroSHEDS (Lehner et al. 2008).

Figure 5. The 36 dams on the last bulletin date of the record, 10 October 2025, coloured by fill with the capacity in force that day. Together they held 649 million m³, 27 % of their capacity; 18 dams were below a quarter full and only Barbara was above 75 %. Map and catchments: CC BY-NC-SA 4.0, derived from FABDEM v1-2 [9] and MERIT Hydro [10]. FABDEM is produced using Copernicus WorldDEM-30 © DLR e.V. 2010-2014 and © Airbus Defence and Space GmbH 2014-2018 provided under COPERNICUS by the European Union and ESA; all rights reserved.

5.2 Five complementary checks

No value is trusted on its own. Five checks compare values with something independent; each applies only where its comparison exists, so each has its own denominator (Table 2). The bulletin prints totals, duplicates some values across pages and repeats running counters, which gives the audit three internal witnesses; the PDF itself is the fourth, and the fill percentage the fifth.

Check	Rule	Result
Printed totals (K12)	the dams of a bulletin add up to the printed national total, within the rounding of every printed value	3,965 / 3,967 complete bulletins
Other page (K09a)	storage printed on Page 1 and Page 2 for the same dam and day is equal	128,995 / 128,996 pairs
Calendar (K10a)	today's month-to-date inflow minus yesterday's equals today's inflow ($\pm 0.0015 \text{ Mm}^3$), restarting where day $D - 1$ opens a new month	122,317 / 122,957 steps (99.48 %)
Fill (K11)	printed fill equals storage / capacity, allowing for the rounding of all three numbers	84,470 / 84,471 rows
Original PDF (K15)	a random sample of cells re-read directly from the page	2,392 / 2,392 sampled cells (800 + 800 + 792)

Table 2. The five checks, from `qa/kpi/results/KPI_RESULTS.csv`. All results and every exception row can be recomputed with `python qa/kpi/run_kpis.py`. The PDF check is a sample, not a census; 44,770 storage values have no printed fill to compare with.

WHEN THE SCORE WAS WRONG, THE RULE WAS WRONG

An earlier automated QA scored the inflow counters at 96.26 % and the fill check at 95.67 %. Both rules were at fault, not the data. The counter rule restarted the month on the bulletin of the 1st, which still reports the last day of the previous month; restarting on the bulletin of the 2nd, as the time rules require, raises the pass rate to 99.48 %. The fill rule ignored that storage and capacity are themselves rounded; with all three roundings, 84,470 of 84,471 rows agree. Storage above the printed useful capacity, counted as an error before, is physically possible during a flood and is now a diagnostic (1,488 rows; every one of the 836 with a printed fill shows a fill of at least 100 % in the bulletin itself). The lesson: validate a rule against what the source means before trusting a score.

5.3 Errors in the source are documented, never repaired

Some problems are in the bulletins, not the extraction. They are published as printed and listed in `docs/known_issues.md`: the season-to-date inflows of Ziatine and Gamgoum are printed in each other's rows from 2017 to 2021; one fill value does not match (Masri, 9 February 2022: printed 44.9 %, computed 43.87 %); two printed national totals do not add up (30 December 2017 and 9 February 2022); and the Page 2 column R.S. is kept as printed but not interpreted. Correcting any of these would make the dataset disagree with its own source and destroy the audit trail.

6. Level ★★★★★: Identifiers for every thing

The fourth star asks that things be named by identifiers rather than by labels. The package uses four kinds:

- **Dams:** a stable `dam_id` (`sidielbarrak`, `sidisalem`, ...) to which a governed alias register maps every printed French and Arabic spelling. Tables are joined on `dam_id`, never on names.

- **Observations and evidence:** `observation_id`, `lineage_id`, `bulletin_id` and `transfer_observation_id` link each value to its cell record, its bulletin and the manifest.
- **Content identifiers:** the SHA-256 of each page identifies the exact bytes a value was read from, independently of where the file is stored.
- **Web identifiers:** the ontology names its classes, concepts, time rules, dams, reservoirs and catchments with IRIs under <https://w3id.org/damlevels/>, and the whole release carries a Zenodo DOI.

The fourth star is reached in structure but not yet on the web: the w3id.org/damlevels namespace is not registered, so these IRIs cannot yet be looked up. Registering it is the first item of future work (§11).

7. Level ★★★★★: Linked data and the DamLevels Ontology

7.1 An ontology is a label on every number

A spreadsheet can hold 8,645, but it cannot say what the number is. The **DamLevels Ontology (DLO 0.3)** is the vocabulary that says it: *what* was measured, *where*, *for which* day, *in which* unit, and *which* bulletin page it came from. In the Sidi El Barrak row of 1 February 2020, 8,645 becomes: released water volume · Sidi El Barrak reservoir · **31 January 2019** · million m³ · a reference value from one year earlier, not today's release. The rule "a flow printed on day D belongs to day $D - 1$ " is written once in the model and applied to every value.

7.2 Built on shared vocabularies

DLO reuses W3C and OGC standards rather than inventing terms: SOSA/SSN for observations [3], OWL-Time for days and intervals [4], QUDT for units, PROV-O for lineage [5], SKOS for code lists, I-ADOPT for the decomposition of variables [6], HY_Features and ENVO for water bodies and catchments, and DQV for quality measurements. One printed cell becomes one `dlo:BulletinObservation`, linked to the page it was read from, its time, its result, the column mapping that read it and the feature it describes (Figure 6). The dam (the wall), its reservoir (the water) and its catchment (the land) are separate features, because storage, capacity and erosion belong to different things.

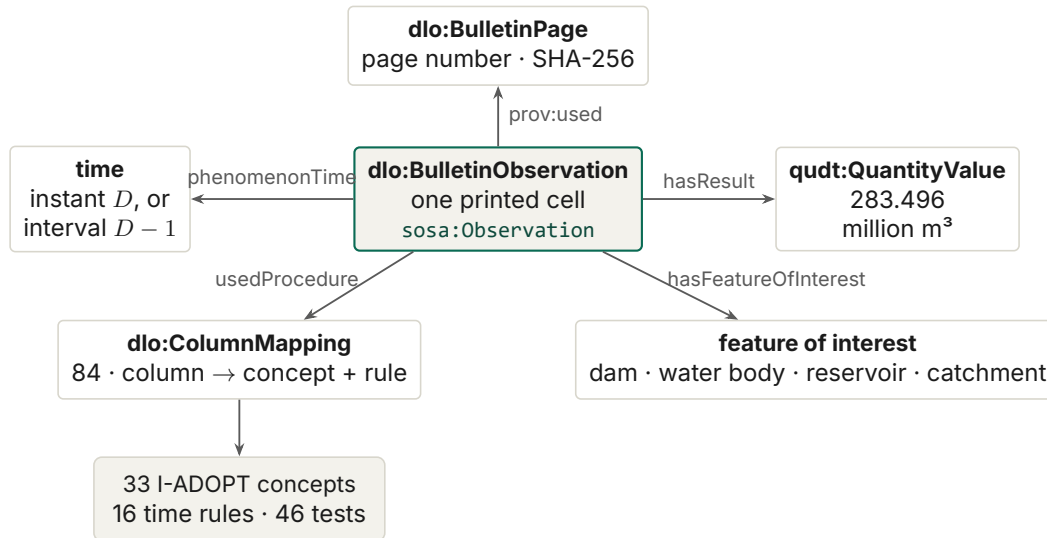


Figure 6. Core of DLO 0.3, simplified. Each printed cell is an observation that points to its source page, its phenomenon time, its result, the column mapping that read it, and the feature it describes. The mapping carries the concept and the time rule, so the rule is stated once for every value of that column.

7.3 From 84 columns to 33 concepts and 16 time rules

Across the three pages, **84 printed columns** map to **33 concepts** and **16 time rules** (state at the bulletin date; amount over the day before; month-to-date; season-to-date; same date one year earlier; three-year means; and others). Inflow volume, for example, is fed by columns from all three pages under different headings; it now has one name. The undocumented R.S. column is mapped but given no concept: the model is explicit about what it does not know.

The time rules were tested rather than assumed: 46 falsifiable hypotheses about column meanings were checked on the whole corpus, with alternatives and negative controls. One result was new: around leap years, the publisher's "one year earlier" is the same day of the year, not the same calendar date, so DLO keeps a reference date for each bulletin. In the project build [2], the converter turned the 11,927 pages into 3,243,357 observations and 38.6 million triples (values reprinted unchanged are stored as run-length observations); the graph answered 13 of 13 competency questions, reproduced 10 of 10 published counts by SPARQL alone, and its SHACL shapes rejected 18 of 18 deliberately injected mistakes.

7.4 What the repository ships

The public repository ships a **view** of DLO 0.3 in Turtle and RDF/XML (8,083 triples): the schema, code lists, the 36 dams, reservoirs and catchments, and 200 sample observations, each linked to its page by SHA-256, with three example SPARQL queries. The full instance graph is not included. Querying the view for Sidi El Barrak gives each number in one printed row its own day:

inflow volume	0.016	2020-01-31/P1D	amount over the day before the bulletin
released water volume	0.0	2020-01-31/P1D	amount over the day before the bulletin
stored water volume	280.478	2019-02-01	state on the same date one year earlier
stored water volume	283.496	2020-02-01	state at the bulletin date

The test suite checks that the Turtle and OWL files carry the same statements (graph isomorphism, blank nodes included), that the documented counts hold, that every observation has its feature, concept, time, mapping and source page, that the 20 sample storage values equal `data/storage.csv`, and that OWL RL reasoning finds no contradiction; each of the two equivalence and consistency checks has a self-test that must fail on a deliberately broken graph.

8. Beyond five stars: an MCP server for AI assistants

8.1 Why a server, and why last

Five-star data still requires a reader who knows SPARQL or pandas, and who remembers that a flow is dated $D - 1$. Most people who want to know how full Sidi Salem was last October will ask an AI assistant instead, and an assistant that reads a PDF or a CSV on its own can repeat every trap of §2.3 with confidence. The last step of the repository is therefore a read-only server implementing the **Model Context Protocol (MCP)** [8], the open standard by which assistants such as Claude Desktop and Claude Code call external tools. It was added last on purpose: it adds no data and no new interpretation. It reads only the published tables and the ontology view, so everything it says can be checked against levels ★★ to ★★★★★.

8.2 The tool suite

The server (`mcp_server/`, built on the official SDK's FastMCP framework) runs locally over stdio, opens no network listener, and exposes ten read-only tools and three document resources (the data dictionary, the known issues and the licence text). All query logic lives in one module that can be tested without MCP; the server only registers it.

Tool	What it returns
<code>dataset_info</code>	coverage, counts, units, time rules, source and licence
<code>list_dams</code>	the 36 dams with identifiers, names, coordinates, reference capacity and reporting period
<code>find_dam</code>	a <code>dam_id</code> from a French or Arabic name or spelling variant, with the match type; candidates when ambiguous
<code>get_storage</code>	daily values (up to 400) or monthly summaries; missing bulletin dates listed, bulletins without a Page 1 counted apart from days without any bulletin
<code>get_fill</code>	storage, the capacity in force that day, the fill, and the fill printed in the bulletin
<code>national_storage</code>	the total of the dams printed that day, the fill over dams with a known capacity (numerator, denominator and excluded dams), and the printed national total with its rounding bound and any known disagreement
<code>compare_years</code>	the same calendar day in every year, for one dam or for all
<code>get_source</code>	for a storage value: the ONAGRI URL, Page 1 file and SHA-256, column, cell position and printed text
<code>get_transfers</code>	Page 3 transfer facts with both endpoint mapping statuses, their own source fields and the time rule of the measure; candidate crosswalks listed separately
<code>describe_variable</code>	meaning, unit, labels and time rule of a term or printed column, from DLO; all candidates when a label is ambiguous

Table 3. The ten tools. Discovery resolves a dam or a term; retrieval returns values with their provenance; every answer carries the record's coverage, the source attribution and its caveats.

8.3 Rules the assistant cannot break

The rules of §2.3 are built into the server, so they hold whatever the assistant asks:

- **No estimate outside the record.** A date after 10 October 2025, or before 2 July 2014, is refused with `OUTSIDE_RECORD`.
- **No silent empty answer.** Malformed input is refused with an error code (`BAD_DATE` for anything but `YYYY-MM-DD` or an impossible day such as 30 February; `BAD_RANGE`; `BAD_LIMIT`; `UNKNOWN_DAM`), so a typing error cannot read as "no data".
- **No filling.** A missing day is reported as missing, with the nearest dates given but never substituted. A bulletin that exists without a Page 1 is reported as such, not as a day without a bulletin.
- **No guessing between meanings.** When a name or a label matches several things, all candidates are returned and none is chosen. Two Arabic labels in the bulletins each belong to two concepts with different time rules; the server returns both and asks for the printed column, so the answer cannot change from one run to the next.
- **Time and uncertainty travel with the number.** Flows are dated $D - 1$; fill uses the capacity in force; transfers are marked `PROBABLE`; and Page 3 nodes whose link to a dam is only a candidate are never promoted to that dam.

8.4 What an answer looks like

Asked how full Sidi Salem was on 10 October 2025 and where the number comes from, the server returns a storage of **137.196 million m³**, a capacity in force of 580.35 million m³ and a fill of **23.6 %**, equal to the 23.6 % printed in the bulletin; `get_source` adds the bulletin’s ONAGRI address, the Page 1 file `10-10-2025_1.pdf` and the printed text `137,196`. Asked for the national storage on 10 October of every year, it returns (million m³, sum of the dams printed that day):

Year	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025
Mm ³	1,179	1,139	660	586	896	1,393	966	669	736	579	501	649
Dams	26	26	26	30	32	33	34	36	36	36	36	36

Table 4. Total storage on 10 October of each year, as returned by `compare_years`. The roster grew from 26 to 36 dams, so early years are not fully comparable; the server says so in its notes.

9. Evaluation and quality assurance

9.1 Automated tests

The repository’s test suite (110 tests, run with `pytest` and needing nothing outside the repository) checks the contracts of the published tables, the capacity lookup, the GIS packaging, the ontology view and the MCP server. The 55 MCP tests compare tool outputs with the published CSV tables (daily storage, monthly means, fill, national and leap-day totals, source cells, transfer amounts and their mapping statuses), check every rule of §8.3, and run real stdio sessions in which the server is started as a subprocess, including the ambiguous Arabic labels under three different hash seeds. Contract validation and a Frictionless check (15 of 15 resources valid) run alongside.

9.2 An independent adversarial review

Before release, the repository was reviewed by a separate AI agent instructed to distrust the project’s own claims. It had read-only access, rebuilt the environment from a fresh clone, recomputed the published numbers, and drove the MCP server over real stdio with 322 tool calls and 14 additional probes. It reported 18 findings: 3 critical, 9 major and 6 minor. Normal numeric checks all passed; the findings were about meaning and disclosure at the edges:

- an ambiguous Arabic label could silently take a different meaning after a restart;
- a bulletin that existed without a Page 1 was reported as a day without a bulletin;
- local machine paths survived in an executed notebook and in project metadata;
- transfer answers dropped one endpoint’s mapping status, hid candidate endpoints when filtering by dam, and gave no route back to the Page 3 cell;
- malformed dates and a negative row limit returned successful-looking answers;
- two tests could not detect the failures their names claimed to cover.

All eighteen were fixed in release 1.1.1, each with a test, and the git history was rewritten to remove the local paths. The review's lesson matches the one behind the audit of §5: the values were right, but the *meaning* did not always reach the consumer. A check that asks "does the caveat arrive with the number?" finds errors that a check of the number cannot.

10. Limitations

- **Page 3 is the weakest source.** Transfer facts are published with authority [PROBABLE](#). Two endpoints (Bouhertma → Bou Heurtma, Zayatin → Ziatine) are only candidate matches. The stored date of every transfer is the bulletin date with the rule [SAME_DAY](#), while the ontology dates daily transfers to $D - 1$; the conflict is disclosed, not resolved (only 80.3 % of date columns match the DLO phenomenon time, all storage rows included).
- **Stars four and five are partial.** The [w3id.org/damlevels](#) namespace is not registered, so the IRIs cannot be looked up; the ontology links to shared vocabularies but not yet to other datasets (for example Wikidata entries for the dams); and the full instance graph is not published.
- **Verification against the PDF is a sample.** 2,392 cells were re-read from the original pages, computer-assisted; other checks are internal consistency checks.
- **Coverage gaps are real.** Capacity is unknown for three documented periods; Mel-lègue has no accepted catchment because much of its basin lies in Algeria; basin names are given for 6 dams and Arabic names for part of the roster.
- **Licence of the source.** The national open-data licence is the default for ONAGRI's open data, but the bulletin page itself does not state a licence.
- **Not official statistics and no forecast.** This is an independent reconstruction from public bulletins. The wider project prepared watershed climate data (CHIRPS, ERA5-Land, SMAP) and a feature panel for future forecasting [2]; these are not part of this release, and no forecasting model was trained.
- **Local server only.** The MCP server runs on the user's machine from a clone of the repository; there is no hosted endpoint.

11. Future work

- **Complete the fourth and fifth stars:** register the [w3id.org/damlevels](#) namespace, publish the full instance graph with dereferenceable IRIs, and link dams and catchments to external datasets.
- **Publish the climate layer** — watershed rainfall, meteorology and soil moisture — with the same provenance discipline, so that forecasting work starts from a shared, audited foundation.
- **Resolve the Page 3 questions** that only the publisher can settle: the two candidate crosswalks, the effective date of transfers and the meaning of R.S.
- **Lower the barrier to use:** a packaged or hosted, still read-only, MCP endpoint.

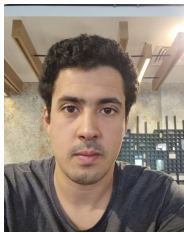
FOR PUBLISHERS OF PUBLIC DATA

Most of the effort in this project was spent recovering what a table next to each PDF would have given for free. Five low-cost actions would make the bulletins usable directly: publish the bulletin as a machine-readable table alongside the PDF; document what each column means and when a layout changes; state the licence on the bulletin page; give each dam a stable public identifier; and treat data publication as part of the monitoring work, with time and budget.

12. Conclusion

DamLevels-Data takes a public but unusable source, a daily PDF, up the five levels of open data: fingerprinted files, deterministically extracted tables with a passport for every value, an open and audited package, identifiers for every thing, and an ontology that writes the meaning and the date of each printed column once and applies it everywhere. Its last step, an MCP server, does not add data; it carries the same rules to the point where people now ask their questions, so that an AI assistant answering from the archive cannot fill a gap, mistake last year's release for today's, or quote a number without its source. The limitations are stated plainly, and the guiding principle holds at every level: every value should lead back to its evidence.

About the Authors



Mohamed Hamza Nasfi

AI RESEARCHER, DATADOIT × MUST UNIVERSITY

Mohamed Hamza Nasfi is an AI Researcher working within the DataDolt × MUST University technological incubator and collaborating with the TanitData research team. His research focuses on building innovative solutions powered by Artificial Intelligence and Data Engineering.



Dr. Soura Boulaares

PROFESSOR, MUST UNIVERSITY

Dr. Soura Boulaares is a Computer Science Professor and Researcher at the Maghreb University of Science and Technology (MUST University). Specialized in Artificial Intelligence, Data Science, and Machine Learning, she actively collaborates with the TanitData research team to supervise innovative technological projects and co-author high-impact international scientific publications.



Tarek Gasmi

FOUNDER, TANITDATA

Tarek Gasmi is Founder of tanitdata and Head of Data and AI at datadoit, with an academic affiliation at the University of Manouba. His work focuses on AI governance, digital infrastructure, sustainability, and technology policy, with particular attention to transitional economies and the geopolitical economy of AI systems.

Citation and License

Suggested citation. Nasfi, M. H., Boulaares, S., & Gasmi, T. (2026). *DamLevels-Data: From PDF Bulletins to Linked, AI-Queryable Open Data on Tunisia's Large Dams, 2014–2025*. Technical report TD-TR-2026-04. tanitdata, Tunis.

Dataset and software. Nasfi, M. H., Boulaares, S., & Gasmi, T. (2026). *DamLevels-Data: daily storage of Tunisia's 36 large dams, 2014–2025*. Zenodo. [doi:10.5281/zenodo.23304326](https://doi.org/10.5281/zenodo.23304326).

License. This report is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license; the map in Figure 5 is CC BY-NC-SA 4.0, as stated in its caption. The DamLevels-Data tables are CC BY 4.0, the catchment layers CC BY-NC-SA 4.0 and the code MIT. The source values are produced by the Direction Générale des Barrages et des Grands Travaux Hydrauliques (DGBGTH), Ministère de l'Agriculture, des Ressources Hydrauliques et de la Pêche, Tunisie, published by ONAGRI, and reused under the Licence Nationale de Données Publiques Ouvertes. This is an independent, unofficial project and is not affiliated with or endorsed by the DGBGTH, ONAGRI or the Ministry.

Version. v1.0.1, October 2026 (v1.0.1 corrects the spelling of an author name); describes DamLevels-Data release 1.1.2.

References

- [1] Berners-Lee, T. (2006, five-star scheme added 2010). *Linked Data — Design Issues*. W3C. [w3.org/DesignIssues/LinkedData.html](https://www.w3.org/DesignIssues/LinkedData.html); see also 5stardata.info.
- [2] Nasfi, M. H. (2026). *Reconstruction, Preprocessing and Auditing of an Open Hydrological Data Pipeline for Tunisian Dams*. Licence report, MUST University; supervised by S. Boulaares, co-supervised by A. Souai, company supervisor T. Gasmi.
- [3] Haller, A., Janowicz, K., Cox, S. J. D., et al. (2019). The modular SSN ontology: A joint W3C and OGC standard specifying the semantics of sensors, observations, sampling, and actuation. *Semantic Web*, 10(1), 9–32. W3C Recommendation *Semantic Sensor Network Ontology*, 2017.
- [4] Cox, S., & Little, C. (eds.) (2017). *Time Ontology in OWL*. W3C Recommendation.
- [5] Lebo, T., Sahoo, S., & McGuinness, D. (eds.) (2013). *PROV-O: The PROV Ontology*. W3C Recommendation.
- [6] RDA Interoperable Descriptions of Observable Property Terminology (I-ADOPT) Working Group. *I-ADOPT Framework ontology*. [w3id.org/iadopt](https://www.w3id.org/iadopt).
- [7] Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.
- [8] Model Context Protocol — open standard specification and SDKs. modelcontextprotocol.io.
- [9] Hawker, L., Uhe, P., Paulo, L., et al. (2022). A 30 m global map of elevation with forests and

buildings removed. *Environmental Research Letters*, 17, 024016. FABDEM v1-2, University of Bristol, [doi:10.5523/bris.s5hqmjcdj8yo2ibzi9b4ew3sn](https://doi.org/10.5523/bris.s5hqmjcdj8yo2ibzi9b4ew3sn).

[10] Yamazaki, D., Ikeshima, D., Sosa, J., et al. (2019). MERIT Hydro: A high-resolution global hydrography map based on latest topography datasets. *Water Resources Research*, 55, 5053–5073.

[11] Frictionless Data — Data Package standard. frictionlessdata.io.

[12] DGBGTH / ONAGRI. *Situation des barrages* (daily bulletins). onagri.nat.tn/Situation-des-barrages.

[13] DamLevels-Data source repository, github.com/tanitdata/DamLevels-Data (tables, provenance, QA, ontology, MCP server and tests).