

Package: washopenresearch (via r-universe)

October 5, 2026

Title Dataset about open research data information in Water, Sanitation, and Hygiene

Version 0.5.0

Description The goal of washopenresearch is to provide an overview of open research data related to Water Sanitation and Hygiene (WASH). The package provides access to seven datasets: `washdev`, `ws`, `jwh`, `aqua`, `uncnewsletter`, `ploswater`, and `datapapers`. Each dataset collects information on scientific articles about (1) article metadata (e.g. title, first author, correspondence author), (2) supplementary material information, (3) data availability statement, and (4) semantic information (e.g. keywords).

License CC BY 4.0

Encoding UTF-8

Roxygen list(markdown = TRUE)

Depends R (>= 3.5)

Imports dplyr

Suggests europepmc, rcrossref, testthat (>= 3.0.0)

LazyData true

Config/testthat/edition 3

Config/Needs/website rmarkdown

Date 2026-10-02

URL <https://github.com/openwashdata/washopenresearch>

BugReports <https://github.com/openwashdata/washopenresearch/issues>

Config/roxygen2/version 8.0.0

X-schema.org-keywords open-data, open-research-data, open-science, openwashdata, sanitation, wash

Repository <https://openwashdata.r-universe.dev>

Date/Publication 2026-10-02 10:05:46 UTC

RemoteUrl <https://github.com/openwashdata/washopenresearch>

RemoteRef HEAD
RemoteSha 466e6033ec97b2707eae53e1826ef983fb8efe7c

Contents

aqua	2
das_in_paper_support	3
datapapers	4
jwh	5
ploswater	7
score_fair	8
uncnewsletter	10
washdev	11
ws	12
Index	14

aqua	<i>Dataset about data availability in the journal AQUA</i>
------	--

Description

AQUA - Water Infrastructure, Ecosystems and Society is an IWA journal. Coverage starts in 1998, the first year in the journal’s online archive. The journal changed its title during this period; the dataset uses the current title for every year. The dataset is updated monthly. Each data release adds the journal issues published since the last one. Scraped from iwaponline.com by data-raw/iwa_scraping.R (#33).

Usage

aqua

Format

- aqua:
- paperid** ID number of the paper on the journal website
 - volume** Volume number of the journal
 - issue** Issue number of the journal, as text because combined issues (for example "1-2") occur
 - paper_url** Official website url of the paper
 - journal** Full name of the journal
 - title** Title of the paper
 - published_year** Year of publication
 - is_supp** Whether the paper has supplementary materials
 - num_supp** Number of supplementary material files
 - supp_file_type** File types of the supplementary materials, separated by "; " when there are multiple

supp_url Website urls of the supplementary materials, separated by "; " when there are multiple

num_authors Number of the authors

first_author_name Name of the first author

first_author_affiliation Academic affiliation of the first author

first_author_affiliation_country Country of the first author parsed from first_author_affiliation, encoded with United Nations names

first_author_orcid ORCID of the first author

correspondence_author_name Name of the correspondence author

correspondence_author_affiliation Academic affiliation of the correspondence author

correspondence_author_affiliation_country Country of the correspondence author parsed from correspondence_author_affiliation, encoded with United Nations names

correspondence_author_orcid ORCID of the correspondence author

has_das Whether the paper has a data availability statement

das Original data availability statement of the paper. NA if it does not have a data availability statement.

das_type Type of the data availability statement including in paper(data in full paper scope like supplementary material or appendix or main content) on request(data available on request to the authors) available in online repository(data is shared in a public online repository) not shareable(data is not shareable). NA if it does not have a data availability statement.

das_repo_url Website urls of the data if the relevant data of the paper is shared on a public repository, separated by "; " when there are multiple

keywords Keywords of the paper, separated by "; "

url_source Publisher website of the paper

doi DOI of the paper. Collected by the R scraper; populated for every row

das_in_paper_support *Classify how a "data in paper" claim is backed by the supplement fields*

Description

The stock statement "all relevant data are included in the paper or its supplementary information" is the modal data availability statement across the IWA journal snapshots. As a single category it is nearly uninformative: it covers papers whose complete dataset genuinely fits in the printed tables as well as papers whose tables hold only summary statistics (issue #47). This function scores the claim jointly with the supplement fields already recorded, so "in paper plus xlsx supplement" is distinguished from "in paper, nothing attached".

Usage

```
das_in_paper_support(data)
```

Arguments

data A data frame with the columns `das_type` and `supp_file_type`. `is_supp`, `num_supp`, and `supp_url` are used when present to detect supplements with an unknown format. The `washdev`, `uncnewsletter`, and `ploswater` datasets and the IWA journal snapshots in `data-row/` all carry these.

Details

A paper makes the in-paper claim when `das_type` is the normalized value "in paper" (`washdev` style) or contains the stock phrasing "included in the paper" / "included in the article" (the IWA journal snapshots keep the raw sentence in `das_type`). Papers without the claim get NA.

Claims are classified by the strongest attachment that could carry the data, using the same format tiers as `score_fair()`:

- "open supplement": an open machine-readable format is attached (csv, txt, tsv, json, xml).
- "structured supplement": structured but proprietary formats only (xlsx, xls, docx, doc, sav, dta, rds, parquet).
- "unstructured supplement": a supplement exists but only as pdf or images, or its format is unknown.
- "no supplement": nothing is attached; the claim rests entirely on the printed tables and figures.

The classification is structural: it says where the claimed data could be, not whether it is actually there. Verifying the content of printed tables and supplement files is the follow-up work in issue #47.

Value

data with one character column added: `das_in_paper_support`. NA for papers that do not make the in-paper claim.

Examples

```
classified <- das_in_paper_support(washdev)
table(classified$das_in_paper_support, useNA = "ifany")
```

datapapers

Dataset about WASH data papers published in dedicated data journals

Description

Candidate WASH-related data papers harvested from Crossref and Europe PMC for seven data journals (Scientific Data, Data in Brief, Gates Open Research, F1000Research, GigaScience, GigaByte, and Data), screened for relevance with a committed decision sheet. Data papers describe a shared dataset, so the repository link takes the role that the data availability statement variables play in `washdev` and `uncnewsletter`.

Format

datapapers:

paperid ID number of the paper within this dataset

doi DOI of the data paper

paper_url Official url of the paper (DOI resolver link)

url_source Publisher website of the paper

journal Full name of the journal

title Title of the paper

published_year Year of publication

num_authors Number of the authors

first_author_name Name of the first author

first_author_affiliation Academic affiliation of the first author

first_author_affiliation_country Country of the first author parsed from first_author_affiliation variable encoded with United Nations names

data_repo_url Website urls of the repository holding the dataset the paper describes, separated by "; " when there are multiple

data_repo Name of the data repository (e.g. Zenodo, Dryad, Figshare, OSF, Dataverse) parsed from data_repo_url

license License url of the paper from Crossref metadata

related_paper_doi DOI of a linked research article, if any (e.g. Data in Brief co-submissions), separated by "; " when there are multiple

abstract Abstract of the paper as provided by the metadata source

query_term WASH search term(s) that retrieved the paper, separated by "; "

retrieval_date Date the paper metadata was harvested from the API

Details

This dataset is frozen. It was harvested once, on 23 July 2026, and its screening is closed. It is still built with every release but no longer updated.

Source

Crossref (<https://api.crossref.org>) and Europe PMC (<https://europepmc.org>); see data-raw/README.md for the pipeline.

jwh

Dataset about data availability in the Journal of Water and Health

Description

The Journal of Water and Health is an IWA journal covering water quality and health. Coverage starts in 2003. The dataset is updated monthly. Each data release adds the journal issues published since the last one. Scraped from iwaponline.com by data-raw/iwa_scraping.R (#32-#34).

Usage

jwh

Format

jwh:

paperid ID number of the paper on the journal website

volume Volume number of the journal

issue Issue number of the journal, as text because combined issues (for example "1-2") occur

paper_url Official website url of the paper

journal Full name of the journal

title Title of the paper

published_year Year of publication

is_supp Whether the paper has supplementary materials

num_supp Number of supplementary material files

supp_file_type File types of the supplementary materials, separated by "; " when there are multiple

supp_url Website urls of the supplementary materials, separated by "; " when there are multiple

num_authors Number of the authors

first_author_name Name of the first author

first_author_affiliation Academic affiliation of the first author

first_author_affiliation_country Country of the first author parsed from first_author_affiliation, encoded with United Nations names

first_author_orcid ORCID of the first author

correspondence_author_name Name of the correspondence author

correspondence_author_affiliation Academic affiliation of the correspondence author

correspondence_author_affiliation_country Country of the correspondence author parsed from correspondence_author_affiliation, encoded with United Nations names

correspondence_author_orcid ORCID of the correspondence author

has_das Whether the paper has a data availability statement

das Original data availability statement of the paper. NA if it does not have a data availability statement.

das_type Type of the data availability statement including in paper(data in full paper scope like supplementary material or appendix or main content) on request(data available on request to the authors) available in online repository(data is shared in a public online repository) not shareable(data is not shareable). NA if it does not have a data availability statement.

das_repo_url Website urls of the data if the relevant data of the paper is shared on a public repository, separated by "; " when there are multiple

keywords Keywords of the paper, separated by "; "

url_source Publisher website of the paper

doi DOI of the paper. Collected by the R scraper; populated for every row

ploswater

Dataset about data availability in PLOS Water

Description

Article metadata for all articles published in PLOS Water since its first volume (2022), collected through the public PLOS search API and each article's JATS XML. Data availability statements are mandatory at PLOS, so `has_das` is TRUE for research articles throughout; the interesting variation lies in `das_type`, `das_repo_url`, and `das_repo_name`. All article types are included; use `article_type` to restrict to research articles.

Usage

ploswater

Format

ploswater:

paperid DOI of the paper, identical to the `doi` variable

volume Volume number of the journal; volume 1 is 2022

issue Issue number of the journal

paper_url Official website url of the paper

journal Full name of the journal

title Title of the paper

published_year Year of publication

is_supp Whether the paper has supplementary materials

num_supp Number of supplementary material files

supp_file_type File types of the supplementary materials, separated by "; " when there are multiple

supp_url Website urls of the supplementary materials, separated by "; " when there are multiple. These links are stable download endpoints, they do not expire.

num_authors Number of the authors

first_author_name Name of the first author

first_author_affiliation Academic affiliation of the first author

first_author_affiliation_country Country of the first author parsed from `first_author_affiliation`, encoded with United Nations names

first_author_orcid ORCID of the first author

correspondence_author_name Name of the correspondence author

correspondence_author_affiliation Academic affiliation of the correspondence author

correspondence_author_affiliation_country Country of the correspondence author parsed from `correspondence_author_affiliation`, encoded with United Nations names

correspondence_author_orcid ORCID of the correspondence author

has_das Whether the paper has a data availability statement

- das** Original data availability statement of the paper. NA if it does not have a data availability statement.
- das_type** Type of the data availability statement including available in online repository(data is shared in a public online repository) in paper(data in full paper scope like supplementary material or appendix or main content) on request(data available on request to the authors) not shareable(data is not shareable) no data generated(the study produced no datasets). NA if it does not have a data availability statement or no classification rule matched.
- das_repo_url** Website urls and dataset DOIs mentioned in the data availability statement, separated by "; " when there are multiple
- das_repo_name** Recognized data repositories behind das_repo_url (e.g. zenodo, dryad, figshare, osf, github, dataverse), separated by "; " when there are multiple
- keywords** Subject terms of the paper from the PLOS search API, separated by "; ". PLOS Water articles carry no author keywords in their XML.
- url_source** Publisher website of the paper
- doi** DOI of the paper
- article_type** Article type, e.g. "Research Article", "Opinion", "Review"
- publication_date** Date of publication (ISO 8601)

Details

The dataset is updated monthly. Each data release adds the articles published since the last one.

score_fair

Score a paper's data management against the FAIR principles

Description

Scores each paper on the four FAIR dimensions (Findable, Accessible, Interoperable, Reusable) from the data-sharing fields the package already records. Each dimension is scored 0, 1, or 2, and `fair_total` is their sum (0 to 8). The rubric is deliberately simple and documented so the scoring is reproducible and open to criticism (issue #19); it is a screening instrument, not a certified FAIR assessment.

Usage

```
score_fair(data)
```

Arguments

<code>data</code>	A data frame with the columns <code>has_das</code> , <code>das_type</code> , <code>das_repo_url</code> , <code>das_repo_name</code> , and <code>supp_file_type</code> . The <code>washdev</code> , <code>uncnewsletter</code> , and <code>ploswater</code> datasets all carry these. <code>das_repo_name</code> is optional; when absent, the repository signal falls back to <code>das_repo_url</code> .
-------------------	---

Details

The four dimensions are scored as follows.

Findable (persistent identifier and registered location):

- 2: the data is in a registered repository (das_repo_name is set) and a repository link or dataset DOI is present.
- 1: a data location is stated (das_type is "available in online repository", or a das_repo_url is present) but without a recognised repository, or the data is in the paper or its supplement.
- 0: no data location (no DAS, "on request", or "not shareable").

Accessible (can a reader get the data without a barrier):

- 2: a repository or supplement link is present (das_repo_url or supp_url), so the data is directly retrievable.
- 1: the data is stated to be in the paper or supplement but no link is recorded.
- 0: "on request", "not shareable", or no DAS.

Interoperable (open, machine-readable shared formats), from supp_file_type:

- 2: any open machine-readable format (csv, txt, tsv, json, xml).
- 1: structured but proprietary formats only (xlsx, docx, sav, dta).
- 0: unstructured only (pdf, images), or no shared files.

Reusable (license and repository metadata). Because a license column is not yet collected for these datasets, this dimension is scored from the repository signal as a lower bound:

- 2: data in a recognised repository (das_repo_name set), which normally carries a license and rich metadata.
- 1: a data location is stated but not in a recognised repository.
- 0: no shared data.

When a license column is added (see issue #19), raise this dimension to use it directly.

Note that Accessible scores the shared files, not the authors' intent. A paper whose das_type is "on request" or "not shareable" can still score 2 on Accessible if it ships a supplement with a supp_url, because that supplement is directly retrievable. The das_type value stays visible alongside the score, so a restricted-data paper that still shares a supplement is distinguishable from a fully open one.

Value

data with five integer columns added: fair_findable, fair_accessible, fair_interoperable, fair_reusable, and fair_total.

Examples

```
scored <- score_fair(ploswater)
table(scored$fair_total)
```

uncnewsletter

*Dataset about data availability in the UNC Water Newsletter***Description**

A curated list of articles from the Research section of the newsletter North Carolina Water News, 2020 to 2023.

Usage

uncnewsletter

Format

uncnewsletter:

paperid ID number of the paper on the journal website**issue_url** Volume number of the journal**paper_url** Official website url of the paper**url_source** Publisher website of the paper**journal** Full name of the journal**title** Title of the paper**published_year** Year of publication**is_supp** Whether the paper has supplementary materials**num_supp** Number of supplementary material files**supp_file_type** File types of the supplementary materials, separated by "; " when there are multiple**supp_url** Website urls of the supplementary materials, separated by "; " when there are multiple**num_authors** Number of the authors**first_author_name** Name of the first author**first_author_affiliation** Academic affiliation of the first author**first_author_affiliation_country** Country of the first author directly parsed from first_author_affiliation variable encoded with United Nation names**first_author_orcid** ORCID of the first author**correspondence_author_name** Name of the correspondence author**correspondence_author_affiliation** Academic affiliation of the correspondence author**correspondence_author_affiliation_country** Country or region of the correspondence author directly parsed from correspondence_author_affiliation variable encoded with United Nation names**correspondence_author_orcid** ORCID of the correspondence author**has_das** Whether the paper has a data availability statement**das** Original data availability statement of the paper. NA if it does not have a data availability statement.

- das_type** Type of the data availability statement including in paper(data in full paper scope like supplementary material or appendix or main content) on request(data available on request to the authors) available in online repository(data is shared in a public online repository) not shareable(data is not shareable). NA if it does not have a data availability statement.
- das_repo_url** Website urls of the data if the relevant data of the paper is shared on a public repository, separated by "; " when there are multiple
- citations** Number of citations of the paper, as entered by the annotators during the manual collection in January 2024 or earlier; the source of the count is not documented. NA where no value was entered.
- keywords** Keywords of the paper, separated by "; "
- doi** DOI of the paper, backfilled via a Crossref title search (issue #20); NA where no match cleared the title-similarity threshold.

Details

This dataset is frozen. The newsletter ceased publication in May 2024, so the dataset is still built with every release but no longer updated.

washdev	<i>Dataset about data availability in the Journal of Water, Sanitation and Hygiene for Development</i>
---------	--

Description

All articles of the Journal of Water, Sanitation and Hygiene for Development, an IWA journal, from its first volume in 2011. The dataset is updated monthly. Each data release adds the journal issues published since the last one. Scraped from iwaponline.com by data-raw/iwa_scraping.R.

Usage

washdev

Format

- washdev:
- paperid** ID number of the paper on the journal website
 - volume** Volume number of the journal
 - issue** Issue number of the journal
 - paper_url** Official website url of the paper
 - journal** Full name of the journal
 - title** Title of the paper
 - published_year** Year of publication
 - is_supp** Whether the paper has supplementary materials
 - num_supp** Number of supplementary material files

supp_file_type File types of the supplementary materials, separated by "; " when there are multiple

supp_url Website urls of the supplementary materials, separated by "; " when there are multiple

num_authors Number of the authors

first_author_name Name of the first author

first_author_affiliation Academic affiliation of the first author

first_author_affiliation_country Country of the first author parsed from first_author_affiliation, encoded with United Nations names

first_author_orcid ORCID of the first author

correspondence_author_name Name of the correspondence author

correspondence_author_affiliation Academic affiliation of the correspondence author

correspondence_author_affiliation_country Country of the correspondence author parsed from correspondence_author_affiliation, encoded with United Nations names

correspondence_author_orcid ORCID of the correspondence author

has_das Whether the paper has a data availability statement

das Original data availability statement of the paper. NA if it does not have a data availability statement.

das_type Type of the data availability statement including in paper(data in full paper scope like supplementary material or appendix or main content) on request(data available on request to the authors) available in online repository(data is shared in a public online repository) not shareable(data is not shareable). NA if it does not have a data availability statement.

das_repo_url Website urls of the data if the relevant data of the paper is shared on a public repository, separated by "; " when there are multiple

keywords Keywords of the paper, separated by "; "

url_source Publisher website of the paper

doi DOI of the paper. Collected by the R scraper for recent articles and backfilled via Crossref for legacy rows (issue #20); NA where no Crossref match was found.

WS

Dataset about data availability in the journal Water Supply

Description

Water Supply is an IWA journal covering drinking water supply, treatment and distribution. Coverage starts in 2001. The dataset is updated monthly. Each data release adds the journal issues published since the last one. Scraped from iwaponline.com by data-raw/iwa_scraping.R (#32-#34).

Usage

WS

Format

WS:

paperid ID number of the paper on the journal website

volume Volume number of the journal

issue Issue number of the journal, as text because combined issues (for example "1-2") occur

paper_url Official website url of the paper

journal Full name of the journal

title Title of the paper

published_year Year of publication

is_supp Whether the paper has supplementary materials

num_supp Number of supplementary material files

supp_file_type File types of the supplementary materials, separated by "; " when there are multiple

supp_url Website urls of the supplementary materials, separated by "; " when there are multiple

num_authors Number of the authors

first_author_name Name of the first author

first_author_affiliation Academic affiliation of the first author

first_author_affiliation_country Country of the first author parsed from first_author_affiliation, encoded with United Nations names

first_author_orcid ORCID of the first author

correspondence_author_name Name of the correspondence author

correspondence_author_affiliation Academic affiliation of the correspondence author

correspondence_author_affiliation_country Country of the correspondence author parsed from correspondence_author_affiliation, encoded with United Nations names

correspondence_author_orcid ORCID of the correspondence author

has_das Whether the paper has a data availability statement

das Original data availability statement of the paper. NA if it does not have a data availability statement.

das_type Type of the data availability statement including in paper(data in full paper scope like supplementary material or appendix or main content) on request(data available on request to the authors) available in online repository(data is shared in a public online repository) not shareable(data is not shareable). NA if it does not have a data availability statement.

das_repo_url Website urls of the data if the relevant data of the paper is shared on a public repository, separated by "; " when there are multiple

keywords Keywords of the paper, separated by "; "

url_source Publisher website of the paper

doi DOI of the paper. Collected by the R scraper; populated for every row

Index

* datasets

- aqua, [2](#)
- datapapers, [4](#)
- jwh, [5](#)
- ploswater, [7](#)
- uncnewsletter, [10](#)
- washdev, [11](#)
- ws, [12](#)

aqua, [2](#)

das_in_paper_support, [3](#)
datapapers, [4](#)

jwh, [5](#)

ploswater, [7](#)

score_fair, [8](#)
score_fair(), [4](#)

uncnewsletter, [10](#)

washdev, [11](#)
ws, [12](#)