ethord

Publishing Open Metadata for Open Research Data Projects of the ETH Domain

• Swiss Open Academic Data (SOAD) Day •

Lars Schöbitz
Global Health Engineering - ETH Zurich

Nicoló Massari
Global Health Engineering - ETH Zurich

Prof. Elizabeth Tilley
Global Health Engineering - ETH Zurich

September 10, 2025

ETH Domain Open Research Data (ORD) Program

✋ Hands-up ✋

Who has heard of the ETH Domain Open Research Data Program?

Measure 1: Calls for field specific actions

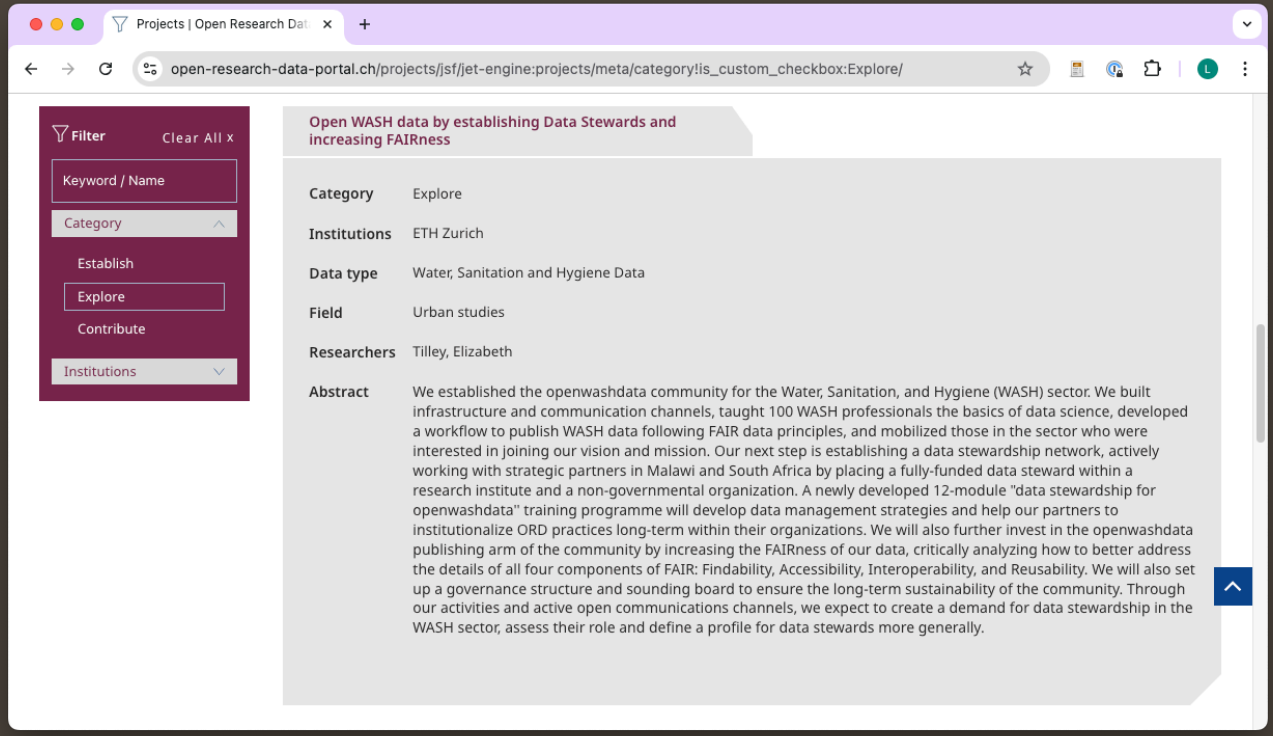
“The primary goal of the measure is to support ETH researchers to engage in, and develop ORD practices and to become ORD leaders in their fields”

Projects

15 mio CHF in funding
96 funded projects

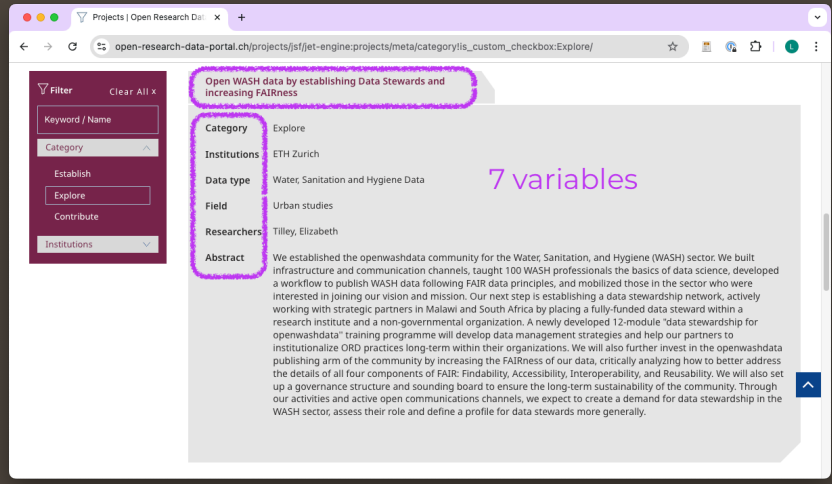

Metadata

Metadata: ORD portal



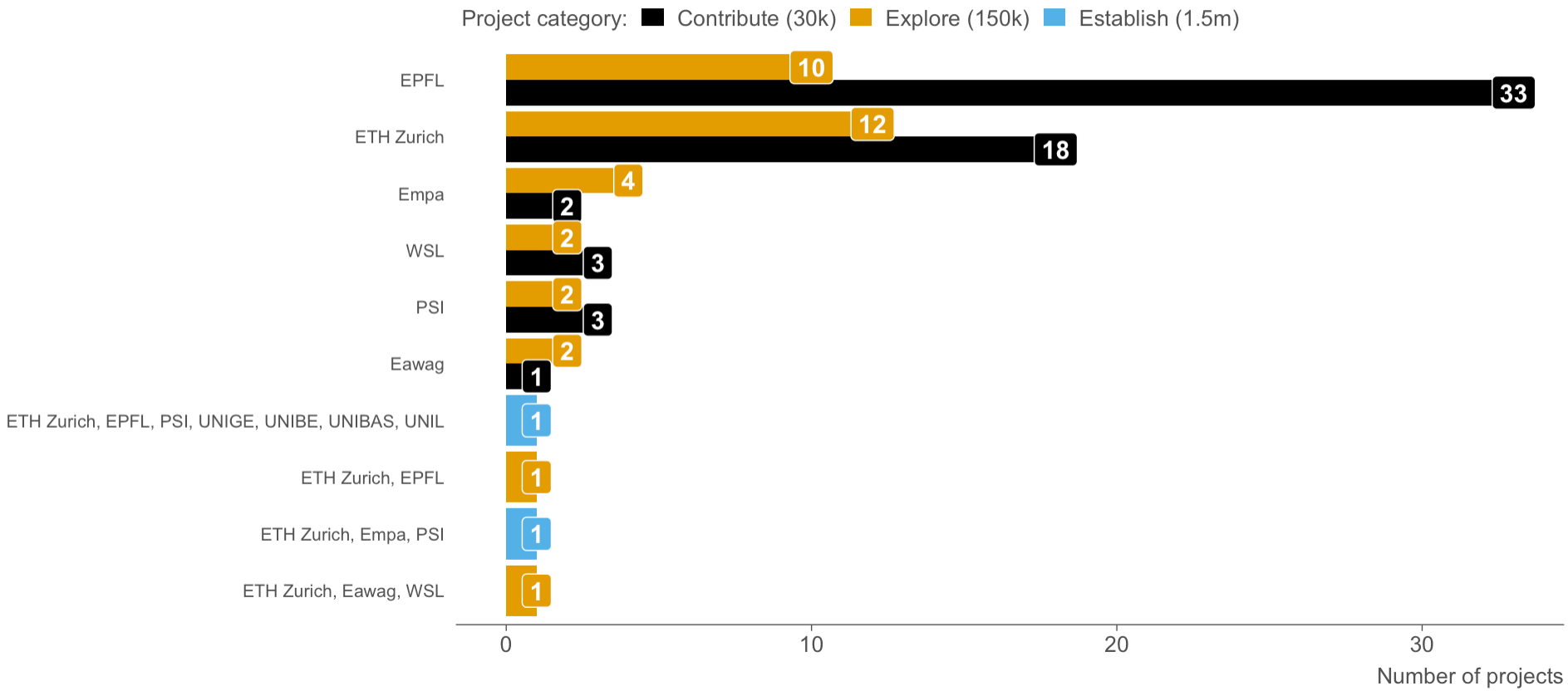
Metadata: ORD portal

- Portal shows titles, abstracts, institutions, applicant names
- No structured bulk data or programmatic access
- Limited visibility of reports and outputs
- No systematic tracking of project outcomes



Open Research Data Program of the ETH Board

Number of funded projects per institution and project category




Metadata (not public)

Proposals & Reports & List of Outputs

- What is their scientific role (e.g. Professor, Post Doc, PhD, etc.)?
- How were budgets distributed among their cost categories?
- How many publications are derived from these projects?
- How many ORD datasets have been published?
- etc.

Limitation

- proposals, scientific reports, and lists of outputs are not available as
  - open
  - structured
  - machine-readable
  - data

The Real Limitation

- proposals, scientific reports, and lists of outputs are:
  - confidential
  - only available to EPFL Research Office
  - protected as intellectual property by researchers

Consequences

- Limits discoverability and impact assessment
- Reviewers have privileged access while public cannot evaluate program effectiveness
- Contradicts goal of helping researchers become ORD leaders

Solution

- Ask all applicants for permission to extract metadata from project documentation
- We are experts in FAIR data sharing principles, so let’s do it!
- Publish FAIR-compliant, DOI-assigned metadata dataset


ethord R data package

ethord

- https://global-health-engineering.github.io/ethord/ (Massari, Schöbitz, and Tilley 2025)
- R data package with four tables
- open source development (GitHub)
- permissive license (CC-BY)
- assigned DOI (Zenodo)

Resource: `docs_proposal`

variable_name	variable_type	description
project_id	numeric	A unique identifier for each project, represented as a numerical value.
cost_personnel_senior_staff_fr	logical	The cost of personnel for senior staff in Swiss Francs (CHF), expected to be a numerical value.
cost_personnel_postdocs_fr	numeric	The cost of personnel for postdoctoral researchers in Swiss Francs (CHF), represented as a numerical value.
cost_personnel_other_fr	numeric	The cost of personnel for other staff members in Swiss Francs (CHF), represented as a numerical value.
cost_personnel_students_fr	logical	The cost of personnel for students in Swiss Francs (CHF), expected to be a numerical value.
cost_travel_fr	numeric	The cost of travel expenses in Swiss Francs (CHF), represented as a numerical value.
cost_equipment_fr	numeric	The cost of equipment in Swiss Francs (CHF), represented as a numerical value.
cost_publication_fr	logical	The cost of publication expenses in Swiss Francs (CHF), expected to be a numerical value.
cost_social_fr	numeric	The cost of social expenses in Swiss Francs (CHF), represented as a numerical value.

variable_name	variable_type	description
cost_other_fr	numeric	The cost of other expenses in Swiss Francs (CHF), represented as a numerical value.
cost_subcontracting_fr	numeric	The cost of subcontracting in Swiss Francs (CHF), represented as a numerical value.

Question: How much money was spent in each cost category?

```r
library(ethord)
library(dplyr)
library(tidyr)

# data wrangling to get sub-categories

docs_proposal_long <- docs_proposal |>
  pivot_longer(cols = !project_id,
               names_to = "cost_category",
               values_to = "CHF") |>
  mutate(cost_sub_category = case_when(
    str_detect(cost_category, "personnel") ~ "personnel",
    str_detect(cost_category, "travel") ~ "travel",
    str_detect(cost_category, "equipment") ~ "equipment",
    str_detect(cost_category, "publication") ~ "publication",
    str_detect(cost_category, "social") ~ "social",
    str_detect(cost_category, "subcontracting") ~ "subcontracting",
    .default = "other"
  ))

# data summary to display table output
docs_proposal_long |>
  filter(!is.na(CHF)) |>
  group_by(cost_sub_category) |>
```

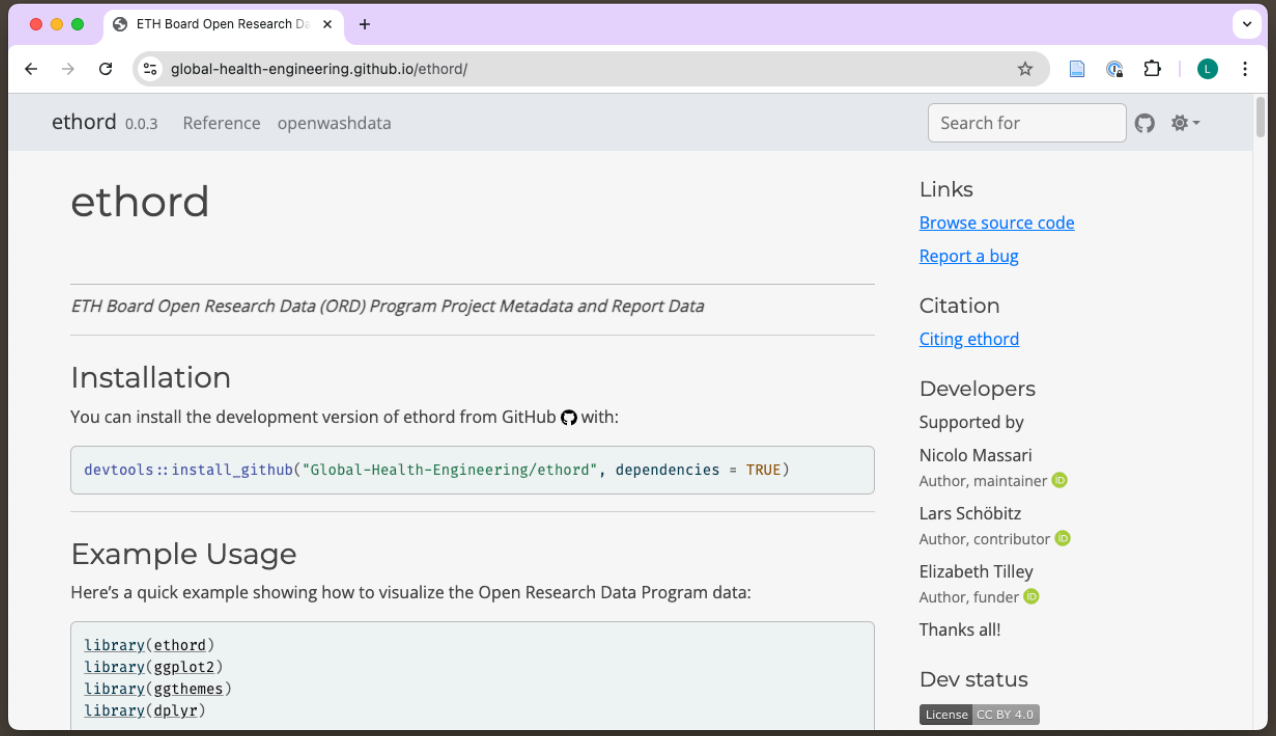
Question: How much money was spent in each cost category?

ORD Program Budget Distribution

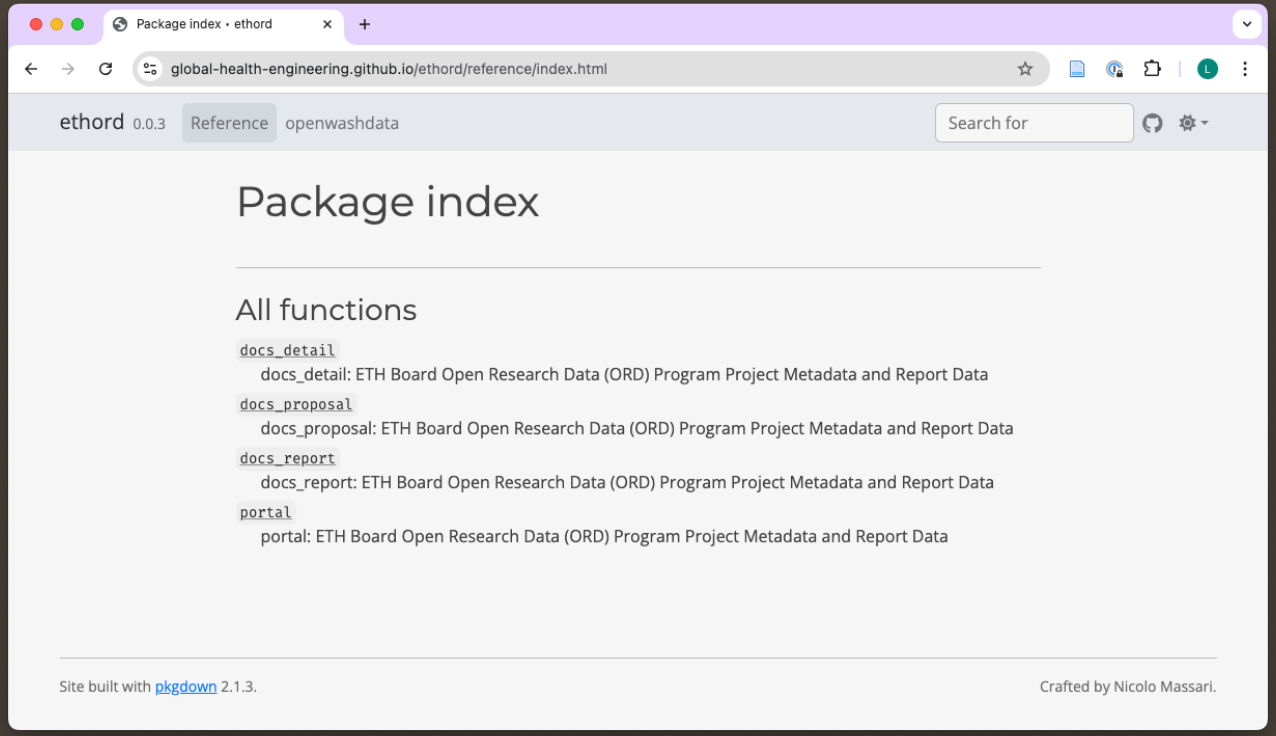
Cost breakdown across 4 funded projects

Cost Category	Total (CHF)	Percentage
personnel	281,835	78.3%
subcontracting	36,000	10.0%
travel	22,570	6.3%
social	16,900	4.7%
other	2,000	0.6%
equipment	600	0.2%

Product: A public website



Product: Documentation for each dataset



Future Possibilities

How could we make such administrative data “open by default” in the future?

Swiss Open by Default Policy Framework

Federal Foundation: Open Government Data Strategy 2019-2023 (Swiss Federal Council 2019) established “open by default” principle for all federal agencies

Legal Mandate: Federal Act EMBAG Article 10 (Swiss Federal Assembly 2024) legally requires open data publication unless restricted by privacy or security

Implementation: OGD Masterplan 2024-2027 (Federal Statistical Office 2024a) operationalizes through:

- Progressive data opening with documented exceptions
- Quality standards aligned with FAIR principles (Wilkinson et al. 2016)
- Centralized coordination via opendata.swiss (Federal Statistical Office 2024b)
- Standardized metadata using DCAT-AP CH (Federal Statistical Office 2023)

Thanks!

Slides created via revealjs and Quarto: https://quarto.org/docs/presentations/revealjs/

Data (Nicoló Massari) available at: Massari, Schöbitz, and Tilley (2025)

Access slides as PDF on GitHub

References

Federal Statistical Office. 2023. “DCAT Application Profile for Data Portals in Switzerland (DCAT-AP CH).” https://www.dcat-ap.ch/.

———. 2024a. “Open Government Data Masterplan 2024-2027.” https://www.bfs.admin.ch/bfs/en/home/services/ogd.html.

———. 2024b. “Opendata.swiss: Swiss Open Government Data Portal.” https://opendata.swiss/.

Massari, Nicolo, Lars Schöbitz, and Elizabeth Tilley. 2025. “Ethord: ETH Board Open Research Data (ORD) Program Project Metadata and Report Data.” https://doi.org/10.5281/zenodo.16563064.

Swiss Federal Assembly. 2024. *Federal Act on the Use of Electronic Means for the Fulfilment of Government Tasks (EMBAG)*. https://www.fedlex.admin.ch/eli/cc/2023/682/en.

Swiss Federal Council. 2019. “Open Government Data Strategy Switzerland 2019-2023.” https://www.admin.ch/gov/en/start/documentation/media-releases.msg-id-74641.html.

Wilkinson, Mark D et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data*. https://doi.org/10.1038/sdata.2016.18.