

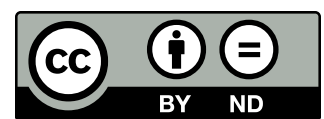
Swiss Research Data Day 2020

#SRDD2020

22 OCTOBER 2020

(Meta)Data Quality and Logistics

The FAIR Data Publication Workflow at Eawag and WSL



Harald von Waldow (Eawag) & Ionuț Iosifescu Enescu
(WSL)

RDM CHALLENGES AT WSL AND EAWAG

Very **heterogeneous** research topics

=> heterogeneous data characteristics & use cases

- climate simulations
- GIS data & remote sensing
- personal data, surveys
- biodiversity data
- hydrogeological data
- monitoring data
- high resolution mass spectrometry
- gen-, transcript-, proteomics
- sensor networks
- software
-
- rapidly increasing **data size** and **complexity**
- researchers' need for **specific RDM consulting**
-- way before data publication --

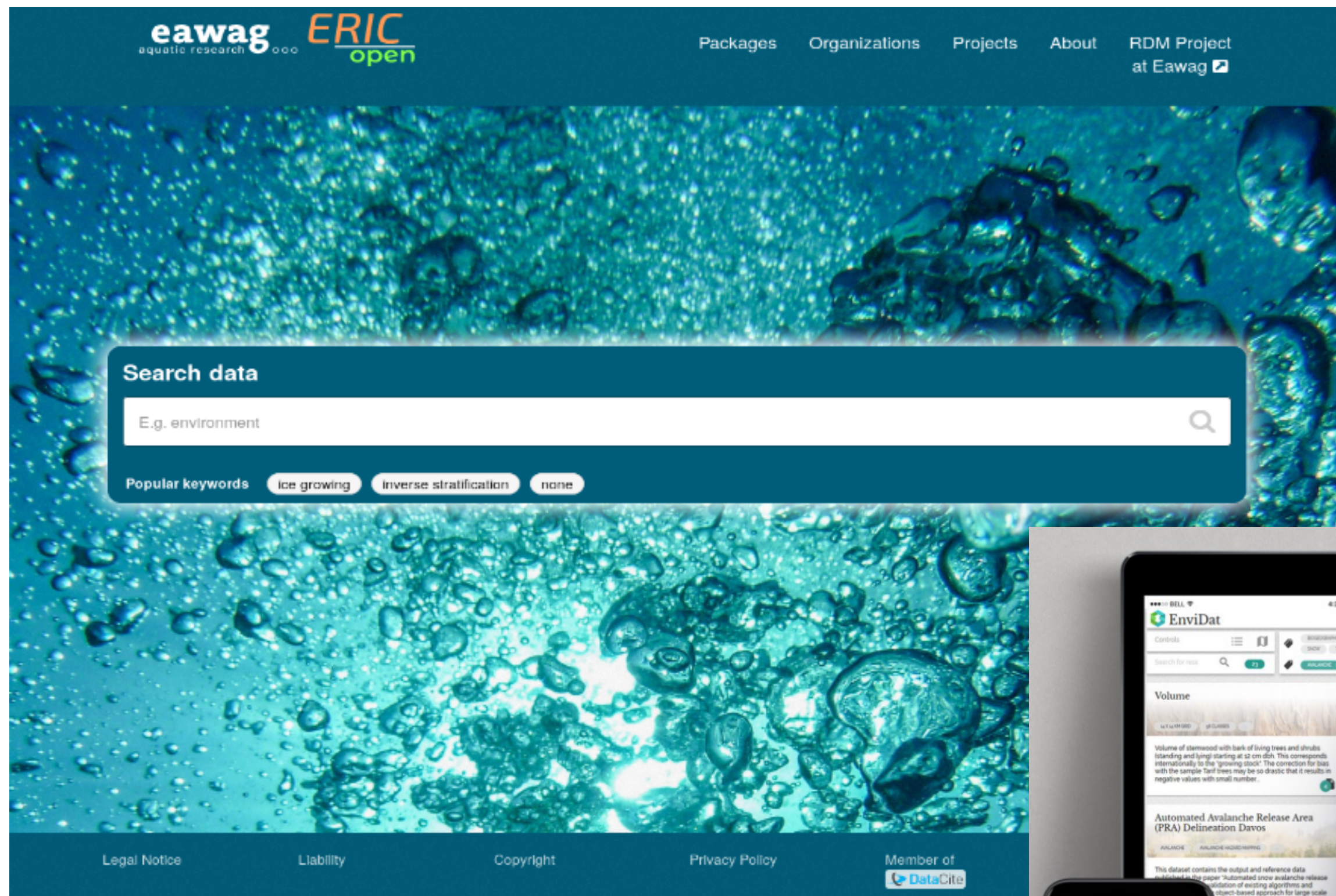
ENVIDAT & ERIC: THE INSTITUTIONAL RESEARCH DATA REPOSITORIES AT WSL END EAWAG

ERIC/open:

`opendata.eawag.ch`

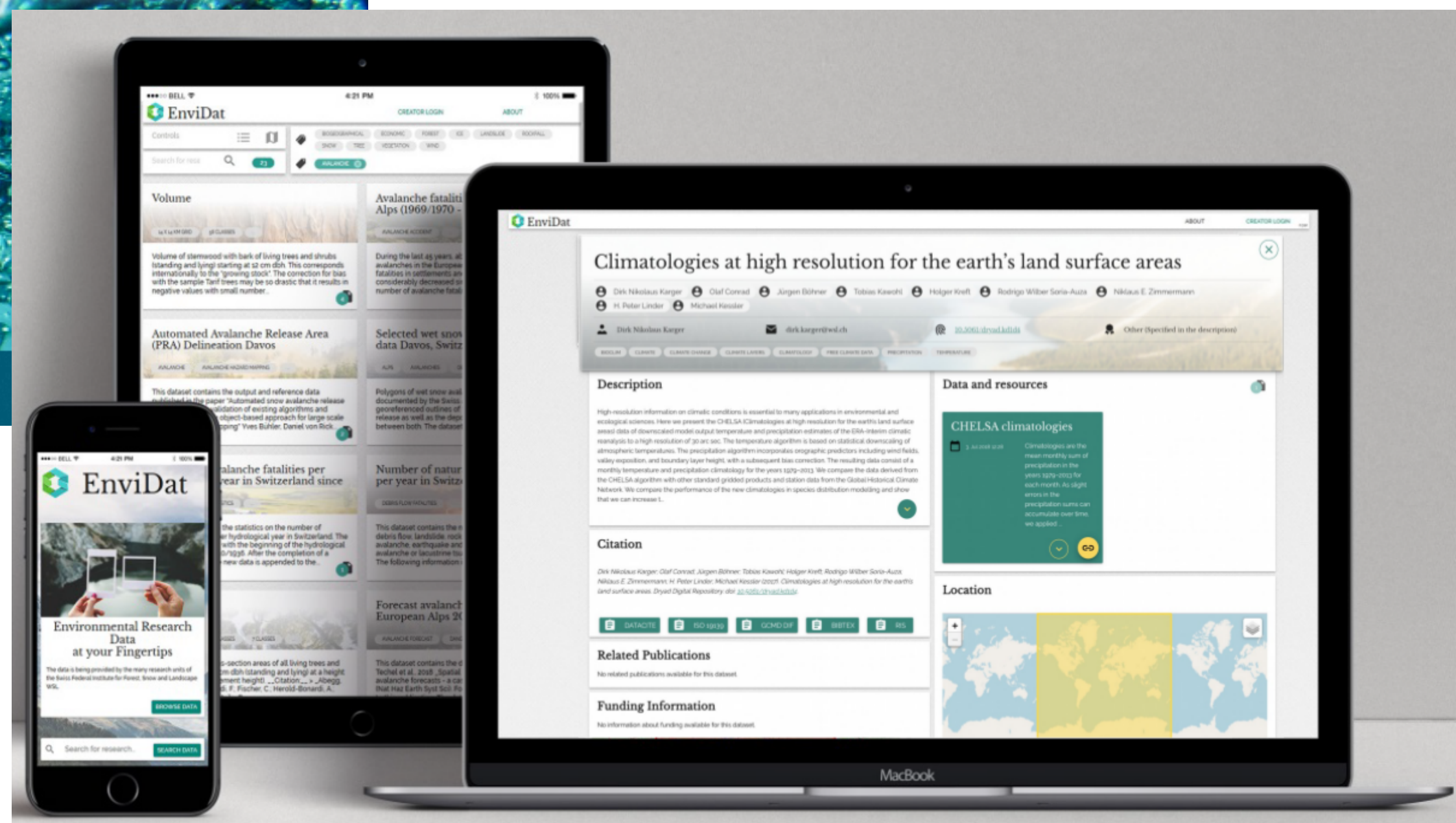
ERIC/internal:

`data.eawag.ch`



EnviDat:

`envidat.ch`



THE EVOLUTION OF ENVIDAT

Initial purpose: public **metadata-portal**

- **2013** ⇨ Exploratory project to explore solutions for a WSL data portal.
- End of **2016** ⇨ Custom prototype (Java) for a metadata portal.
- Mid **2017** ⇨ **Evolved purpose: full Research Data Repository** (CKAN based)
Cross-cutting WSL program with a dedicated technical team.
- March **2018** ⇨ Operational www.envidat.ch portal launched.
- May **2018** ⇨ WSL Data Policy comes into effect (**mandatory "Open Data"**).
- September **2018** ⇨ Launch of updated, modern frontend (Vue.js).
- Mid **2020** ⇨ Development of ancillary *Next Generation Cloud Repository*.

EnviDat is a long-term commitment of WSL (financed at least until 2024).

ERIC/internal Initial purpose: internal data repository

- 2016 ⇨ Start current RDM management "project" at Eawag: 1 FTE
- 2017 ⇨ Requirements engineering, understanding CKAN
- April 2018 ⇨ "public" beta
- January 2019 ⇨ operational as core service
- January 2019 ⇨ **Policy: Directive on the archiving of research data at Eawag (mandatory internal record)**

ERIC/open Evolved purpose: Open Data Repository

- November 1, 2019 ⇨ operational.
- Same system. But completely independent.
- No ingress, no users.
- In the process of being populated (~80 packages backlog).
- Every package has a DOI (DataCite).

ENVIDAT \rightleftharpoons ERIC : DIFFERENCES, PARALLELS, SYNERGIES

↳ Different initial focus


↳ Convergence towards similar capabilities & use-cases

Differences, e.g.

- **two systems** \leftrightarrow **one system**
- metadata-only ok \leftrightarrow keeps all data
- organization & **resources**
(IT **Services Dept** \leftrightarrow GIS group,
Research Unit)
- **mandatory** \leftrightarrow **voluntary** publication

Parallels

 common technological basis

 code sharing

 reciprocal consultation

joint development of concepts

DOI WORKFLOW

IMPORTANCE AND DIFFICULTIES

Why DOIs are **important**

- well recognized **persistent identifier**: The "F" in FAIR
- incentive to publish data in the first place.

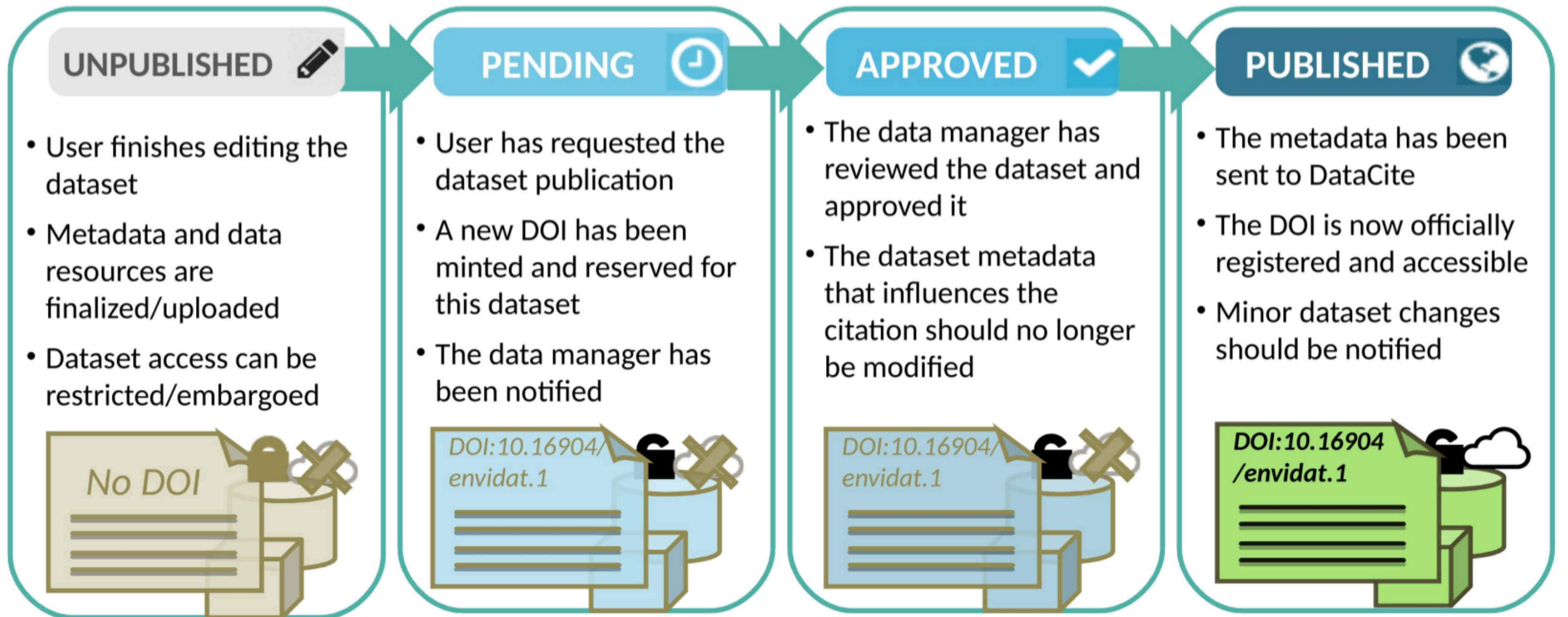
 **DataCite** also serves as:

- metadata schema supplier
- citation formatter
- DOI database with REST-API
- **metadata distributor**
- ...

Why DOIs are **difficult**

- **not designed** for datasets
- require **immutability** (citeability)
- **metadata-updates** must be propagated
- *relatedIdentifiers* need updating
 - versions
 - ORCIDs
 - related articles
 - ...
- **accessibility of data** needs to be ensured

COMMON **QUALITY** ASSURANCE WORKFLOW



relevant differences ERIC ↔ EnviDat

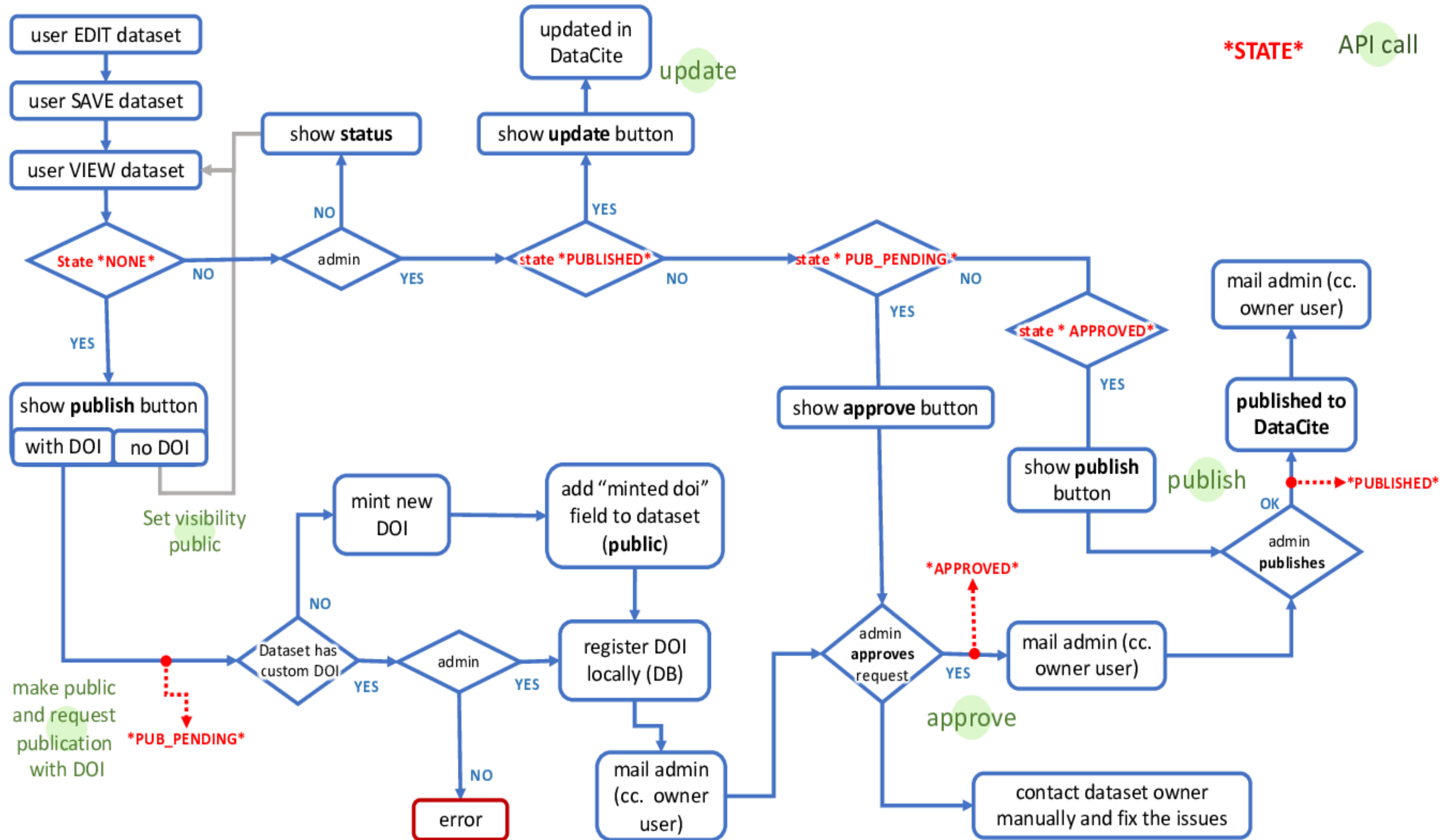
ERIC

- no DOI \Leftrightarrow **internal** archive only
- DOI \Leftrightarrow **Open Research Data**

EnviDat

- no DOI \Leftrightarrow **meta-data** only
- DOI \Leftrightarrow **accessible**

COMMON **QUALITY** ASSURANCE WORKFLOW



VALIDATION AND Q/C

automatic validation & Q/C (CKAN validators)

- mandatory fields
- minimum content (e.g. min. 5 keywords)
- choices (e.g. only ERIC-users can be selected as contact person)
- auto-fill based on article reference
- ...

manual checks EnviDat

- title **descriptive** but short
- description long and **informative**
- keywords appear **appropriate**
- contact person exists, **email is valid**
- correct **geo-reference**
- files **can be opened**

manual checks ERIC

- some **form-** and **consistency** checks
- "usage contact" is permanent staff
- **filenames** are sane
- there is a **README-file**

TECHNICAL AND CONCEPTUAL CHALLENGES

ERIC: **two** systems

- synchronization
- configuration-drift
- redundancies
(storage space)

EnviDat: **one** system

- publish with/without DOI
- access restrictions
- degrees of "openness"
- security

BOTH:

- embargo handling
- versioning
- DOIs on file-level (or not)

CONCLUSIONS

- Workflow assures **metadata-quality** to a large extent
- Publication of "meaningful" data is supported (but **not assured!**) by this workflow.
- **Curation** is very important, cannot be automatized, requires a **lot more (human) work** than presented here.
- Bottleneck is the **manual process**. Need for **cultural & organizational changes** (data managers, data champions, incentives, recognition, training, ...)

A DEMO!