Workflow for an Improved FAIR Environmental Data Publication in EnviDat

Ionuț Iosifescu Enescu
Programme EnviDat
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland
ionut.iosifescu@wsl.ch

Gian-Kasper Plattner
Programme EnviDat
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland
gian-kasper.plattner@wsl.ch

Lucia Espona Pernas
Programme EnviDat
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland
lucia.espona@wsl.ch

Dominik Haas-Artho
Programme EnviDat
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland
dominik.haas@wsl.ch

Rebecca Kurup Buchholz
Programme EnviDat
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland
rebecca.kurup@wsl.ch

David Hanimann
Programme EnviDat
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland
david.hanimann@wsl.ch

Abstract—*The Swiss Federal Institute WSL strives to increasethe fraction of environmental data that is easily available for reuse. With the Environmental Data portal EnviDat, WSL facilitates the publication of FAIR (Findable, Accessible, Interoperable, Reusable) and high-quality environmental research datasets by providing: A) a formal data publication process for the data producers, B) a technical workflow for improving data-quality with automatic validation, interactive quality checks, and iterative improvement of (meta-)data quality in support of the formal publication process and C) a DataCRediT mechanism for declaration of data authorship roles.*

Keywords—EnviDat, FAIR, RDM, (meta-)data quality, DataCRediT, environmental data publication

I. INTRODUCTION

EnviDat is the institutional data portal of the Swiss Federal Research Institute WSL, dedicated to hosting and publishing environmental datasets from forest, landscape, biodiversity, natural hazards and snow and ice research. As graphically summarized in Fig.1, EnviDat offers a range of functionalities and services for publishing data, software, and documentation in support of best practices in Research Data Management (RDM) and Open Science (Iosifescu et al., 2018; Iosifescu et al., 2019). Through its capabilities to host and publish data sets, EnviDat provides unified and managed access to WSL's comprehensive reservoir of environmental research data, and thus actively contributes to the goal of increasing the fraction of environmental data that is easily accessible for reuse by researchers and the public.

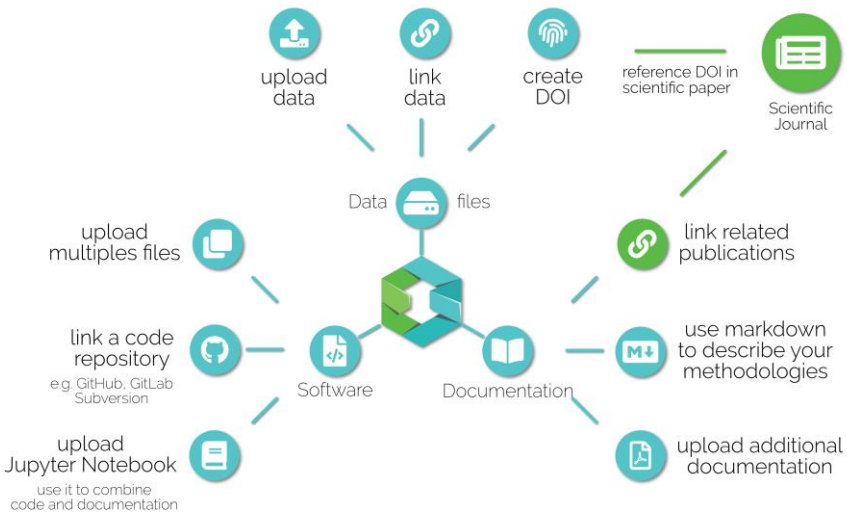

Fig. 1. EnviDat's core functionalities in a nutshell

II. ENVIDAT AS A FAIR PORTAL AND REPOSITORY

EnviDat actively implements the FAIR principles – Findability, Accessibility, Interoperability and Reusability – for scientific data management (Wilkinson et al. 2016). First, an essential requirement for a findable dataset is the assignment of a persistent identifier. The datasets published in EnviDat are assigned globally unique and persistent identifiers (PIDs). All open datasets are also assigned Digital Object Identifiers (DOIs). The minted DOIs are kept in a separate database that is managed independently from the EnviDat main database, in order to protect the persistence of the DOIs in case of technical failures. Moreover, datasets are described by comprehensive metadata records that explicitly include the persistent identifier(s) of the dataset, and present to users a formal citation for the corresponding dataset that includes the assigned DOI. Finally, the metadata of the EnviDat datasets are indexed and made searchable through www.envidat.ch, as well through DataCite Search, ESA's geoportal.org, NASA EarthData Global Change Master Directory (GCMD) and, in the near future, also through the Swiss public administration's central portal for open government data opendata.swiss.

Second, the EnviDat datasets are made accessible via their corresponding metadata landing pages on www.envidat.ch, which are linked to their assigned PIDs/DOIs through the HTTP(S) protocol. EnviDat uses the Open Knowledge Foundation (OKFn) Comprehensive Knowledge Archive Network (CKAN) as a backend metadata repository. CKAN includes a rich Application Programming Interface (API) which allows developers to write code that interacts with CKAN sites and their hosted datasets. The dataset resources are stored using established protocols and providing the HTTPS GET interface for a straightforward download of the data files. Furthermore, even if the metadata are open and accessible by default, the uploaded files and resources can be restricted. In case of restricted resources, interested data users can trigger an access request that will ask for the permission of the data owner, which the data owner must approve before users are allowed to download such restricted datasets. For these cases, EnviDat implemented passwordless authentication procedures, where a unique one-time login token is sent directly to a user's mailbox, thus completely eliminating the need for storing and securing user passwords. Finally, even if data files may be removed at the request of the depositor, the metadata will be kept and the DOIs will continue to point to "tombstone" landing pages containing a modified description that explains the withdrawal reasons. Valid reasons for withdrawal are: violations of WSL research integrity guidelines, proven copyright violation or plagiarism, or legal requirements. Therefore, any metadata registered in EnviDat will persist even if the data should no longer be available, thus avoiding broken links from scientific citations.

Third, the metadata records in EnviDat are organized according to a three-layer schema model (with core, optional and domain-specific research metadata), in order to meet current and future domain-relevant community standards and to ensure interoperability (Iosifescu et. al, 2018). At the core of the EnviDat metadata schema there exist a number of mandatory metadata fields: title, description, keywords, author(s), affiliation, license, publisher, publication year, contact information and the geometry of the spatial extent. Furthermore, a unique EnviDat PID and a DOI are automatically added by the system during the publication workflow as an integral part of the dataset's core metadata record. The EnviDat core metadata schema is designed to make maximal use of the latest DataCite metadata-schema (DataCite, 2019) for DOIs and exploits its ability to store spatial information about environmental measurements. In the EnviDat metadata schema there are also two optional metadata fields, namely "Related Publications" and "Related Datasets", designed to document associated scientific article(s) and other dataset(s) through qualified references/citations. These optional metadata fields improve the documentation of data provenance, since they record and link to additional information influencing the data of interest. The metadata captured with the EnviDat metadata schema is fully interoperable with (and exportable to) various standards such as: Dublin Core, the latest DataCite Metadata Schema (4.3), ISO 19139 or GCMD DIF 10.2 (current operational version). Yet, the interoperability of the data files is a constant concern. Even though EnviDat does not impose any restrictions regarding the format of the data itself, we advise, encourage and even support the users to publish their data in file formats that adhere to existing community standards. EnviDat also fosters new standards, one example of this effort being the Non-Binary Environmental Archive Data (NEAD). This format is being developed as a "generic and intuitive format that combines the self-documenting features of NetCDF with human readable and writeable features of CSV" and it is being specifically "designed for exchange and preservation of time series data in environmental data repositories" (Iosifescu et al., 2020).

Fourth and final, in order to support the reusability of the dataset, the data publishers need to select or define a license that documents the terms of use, thus defining rights, permissions and restrictions of use for the dataset. WSL makes its research data available to users in accordance with the WSL Data Policy, to the extent permitted by the relevant laws, ordinances and contracts with third parties. Any exemptions from the obligation to share research data with users after a clearly specified period must be substantiated and approved by the WSL Directorate. Consequently, EnviDat does not impose any restrictions on "free and open access conditions" conditions – depositors may freely choose to release their data under CC0 (Creative Commons "No Rights Reserved") or

equivalent license. In practice however, the most open licenses preferred by the dataset authors are ODC-ODbL (Open Data Commons Open Database License) or CC-BY-4.0 (Creative Commons Attribution 4.0 International), which both require an appropriate citation/attribution of the dataset.

III. IMPROVING FAIR

EnviDat takes the implementation of the FAIR principles seriously. Unfortunately, however, the FAIR principles do not address an important aspect, namely the quality of the published (meta-)data contents, because publishing a dataset by respecting the FAIR principles is not necessarily correlated to (meta-)data quality. EnviDat thus offers guidance and support to researchers throughout the entire data publication process and strives to achieve the highest quality for environmental research data with an improved FAIR publication workflow.

More precisely, EnviDat enables the publication of highquality environmental research datasets by providing (A) a formal data publication process for the data producers, (B) a technical workflow for improving data-quality in support of the formal publication process and (C) a DataCRediT mechanism for the declaration of data authorship roles.

A. Formal Data Publication Process

The formal data publication process has the role of making the researchers aware of their responsibilities and accountabilities when publishing a research dataset in EnviDat. This process, depicted graphically in Fig. 2, contains the following six main steps:

1. Login. An initial passwordless login in EnviDat registers the email of the researcher. Then the researcher is directed to inform their group leader. The group leader can either direct the researcher to the group's or unit's data manager (if available) or confirms the researchers request to publish data in EnviDat.
2. Receiving editing rights. If the group leader approves, a data manager or the EnviDat support team will grant the necessary rights for data publication and points them to the portal's guidelines and policies.
3. Creation of a "New Dataset" in EnviDat and registration of the necessary metadata according to the EnviDat metadata schema explained in the previous section.
4. Upload of the research data and further resources (e.g., images, software, supporting files etc.). The upload of large files is supported though the provision of individual FTP accounts or, for multi-TB datasets, the provision of individual object store access keys.
5. Publishing the dataset. After finalizing the metadata registration and uploading the research data and possibly other resources, the "Publish" button will become available. This step includes an important workflow for improving data-quality with interactive quality checks and iterative improvement of (meta-) data quality in support of the formal publication process, that will be explained and detailed in the next heading.
6. Curation of the research dataset. Since the researcher's responsibility does not end with the data publication, the data owners are encouraged to periodically revisit and improve the published dataset and the associated metadata. For example, if the published research data is linked to a scientific publication that is in review, then the researchers are asked to enter the final publication in the "Related Publications" field of the EnviDat metadata form. Also, in EnviDat the researchers have the possibility to perform corrections and additions to the published data by uploading new versions.

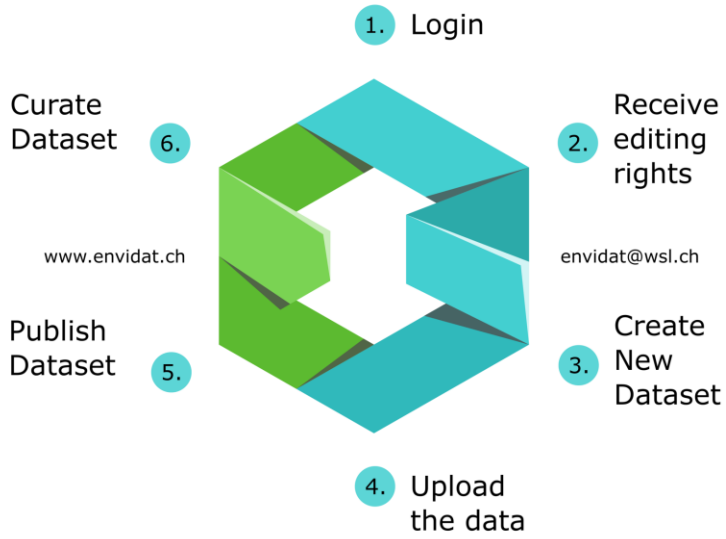

Fig. 2. Formal research data publication process in EnviDa

On one hand, EnviDat's formal publication process verifies that the group leaders are informed of the data publishing efforts by their supervised researchers. On the other hand, the researchers will get in touch with the EnviDat support team, which points them to important EnviDat guidelines and policies before receiving editing rights. For example, the researchers are made aware that the published metadata (but not necessarily the data files) will become public domain, therefore it can be re-used in any medium for any purpose and without prior permission. (although, in the EnviDat policies we request metadata harvesters, if technically feasible, to provide a link to the original EnviDat metadata record).

Furthermore, by reading the EnviDat guidelines and policies, researchers are also made aware that the validity, authenticity and quality of the content of submissions is their responsibility, hence they should only submit metadata and content items for which they have the necessary permissions and rights for distribution and publication. Copyright violations related to the submission of metadata and content items to EnviDat are the responsibility of the depositors.

Finally, the other steps of the process pertaining to how create a new EnviDat dataset record, how to upload the data, and how to publish and curate the dataset are detailed in the EnviDat's guidelines for data publication, with the most upto- date version available on www.envidat.ch.

B. Workflow for an impoved (meta-)data quality

The publishing of the dataset represents an important step in EnviDat. In order to improve the quality-assurance for data sets, we aim for introducing an approach that is similar to the peer-review process applied for scholarly articles, a process that is currently missing for research data publication. For this reason, the EnviDat team encourages nomination of EnviDat data managers by every data provider organization. Data managers can peer-review the (meta)data with regard to quality characteristics such as accuracy, completeness, reliability, relevance, and timeliness. EnviDat has extended this data publishing process step to be more than just a simple assignment of a DOI by chaining together researcher input, automatic validation, interactive quality checks, and iterative improvement of (meta-)data quality. During the quality assurance workflow, the request for a DOI is simply only one substep of the workflow which brings the dataset through an approval process with a double-checking principle. The publication workflow ends with the submission of the metadata-record to DataCite and the final publication of the metadata record in EnviDat

During this workflow, as depicted in Fig. 3, the dataset itself moves between different states, from "Unpublished" towards "Published", "Pending" and the "Approved" states.

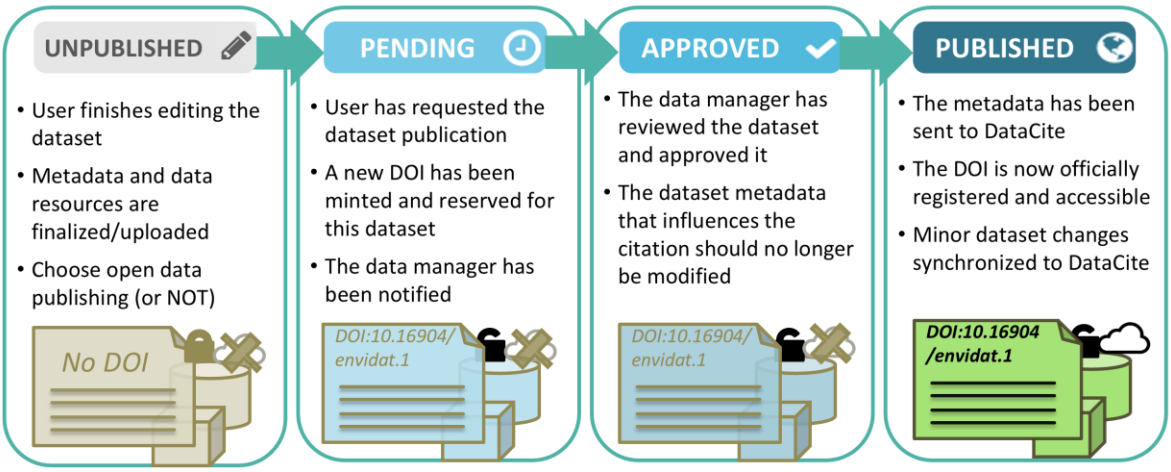

Fig. 3. EnviDat's quality assurance workflow

The workflow can be viewed as a decentralized peerreview and quality improvement process for safeguarding the quality of published environmental datasets. This workflow is being further developed and refined together with partner institutions within the ETH Domain on regular basis, with an especially strong cooperation regarding concepts and software existing between WSL and the Swiss Federal Research Institute for Aquatic Science and Technology Eawag (von Waldow and Iosifescu, 2020).

C. DataCRediT

The overall workflow for an improved FAIR data publication in EnviDat is further improved by increasing the transparency for the range of contributions that a dataset author's make to the published data. Therefore, the EnviDat metadata schema has fields designed to capture and document the individual author contributions to the publication of a particular dataset and related contents such as the software that was used to process or generate the data set. These fields are documented in The Data Authorship Contributor Roles Taxonomy – DataCRediT (WSL, 2018), a mechanism for data authorship specification inspired by and adapted from the Contributor Roles Taxonomy (CRediT) for scientific scholarly output developed by the Consortia Advancing Standards in Research Administration Information (CASRAI, 2018). DataCRediT currently covers six contributor roles: Collection, Validation, Curation, Software, Publication, and Supervision, as detailed in Fig. 4.

The taxonomy supports transparency of contributions to published research data sets by providing an improved system of attribution, credit, and accountability for scientific data publication, thus further encouraging the vigilant application of the FAIR data principles by the individual researchers.

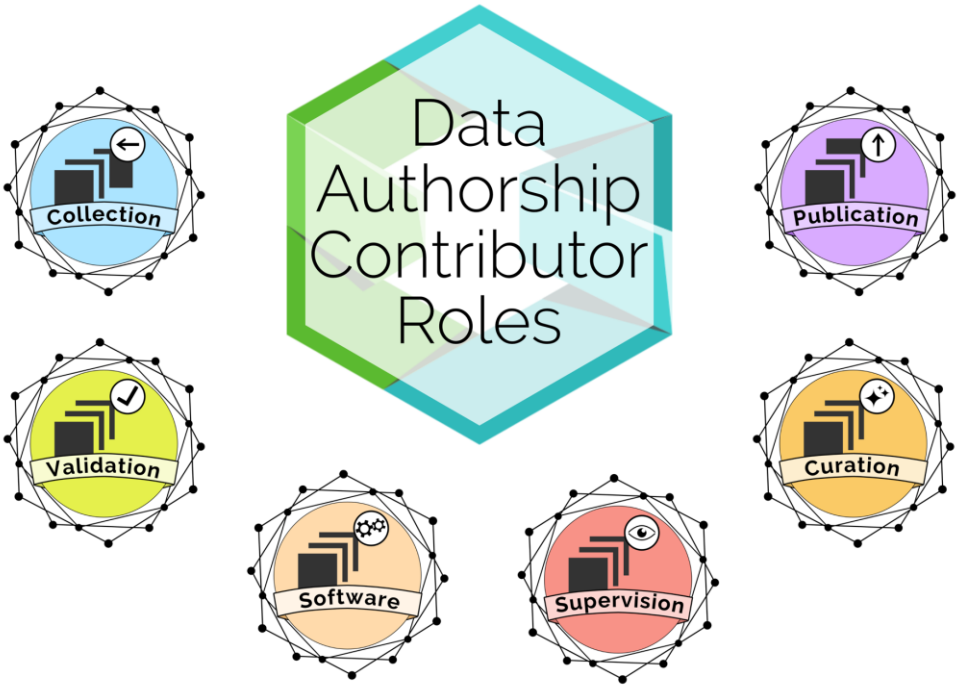

Fig. 4. The Data Authorship Contributor Roles Taxonomy (DataCRediT)

IV. CONCLUSION

The FAIR data publication workflow can be greatlyimproved by implementing basic quality assurance. Since publishing data comes with significant restrictions – like not being allowed to delete any parts of the already published data – it is important to ensure that the original data is of the highest possible quality from the beginning. This can be achieved with: A) a formal data publication process, B) iterative improvement of (meta-)data quality through an approval workflow with a double-checking principle, and C) an improved system of attribution, credit, and accountability forscientific data publication.

DEDICATION AND ACKNOWLEDGMENTS

The Environmental Data Portal EnviDat was initiated by Prof. Dr. Konrad Steffen, the former WSL director, who died in 2020 during field work in Greenland. We dedicate this article to him, to acknowledge and honor his crucial role for EnviDat. His vision was that EnviDat will facilitate the work of researchers by supporting them with the publication of their data, thus creating new collaboration opportunities within WSL, the ETH Domain and beyond.

REFERENCE

CASRAI (2018). CRediT – Contributor Roles Taxonomy. http://docs.casrai.org/CRediT – last accessed on February 15, 2021

DataCite Metadata Working Group (2019). DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.3. DataCite e.V. https://doi.org/10.14454/f2wp-s162

Iosifescu Enescu, I., Plattner, G. K., Bont, L., Fraefel, M., Meile, R., Kramer, T., Pernas, L. E., Haas-Artho, D., Hägeli, M. and Steffen, K. (2019). Open science, knowledge sharing and reproducibility as drivers for the adoption of FOSS4G in environmental research. In M. A. Brovelli & A. F. Marin (Eds.), International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences: Vol. XLII-4/W14. FOSS4G 2019 – Academic Track (pp. 107-110).https://doi.org/10.5194/isprs-archives-XLII-4-W14-107-2019

Iosifescu Enescu, I., Plattner, G. K., Pernas, L. E., Haas- Artho, D., Bischof, S., Lehning, M., and Steffen, K. (2018). The EnviDat concept for an institutional environmental data portal. Data Science Journal, 17, 28 (17 pp.). https://doi.org/10.5334/dsj-2018-028

Iosifescu Enescu, I., Bavay, M. and Mankoff, K. (2020). Non-Binary Environmental Data Archive (NEAD) format. EnviDat. https://doi.org/10.16904/envidat.187

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

WSL (2018) DataCRediT - Contributor Roles Taxonomy for Data. https://www.wsl.ch/datacredit/#feat – last accessed on February 15, 2021

von Waldow, H. and Iosifescu Enescu, I. (2020). (Meta)Data Quality & Logistics: the FAIR Data Publication Workflow at Eawag and WSL. Presentation at the Swiss Research Data Day 2020 (online), Switzerland. https://mediaserver.unige.ch/play/137206 – last accessed on February 15, 2021