RESEARCH PAPER

The EnviDat Concept for an Institutional Environmental Data Portal

Ionuț Iosifescu Enescu[1], Gian-Kasper Plattner[1], Lucia Espona Pernas[1], Dominik Haas-Artho[1], Sandro Bischof[1], Michael Lehning[2,3] and Konrad Steffen[1,3,4]

[1] Swiss Federal Institute for Forest, Snow and Landscape WSL, CH

[2] WSL Institute for Snow and Avalanche Research SLF, CH

[3] School of Architecture, Civil and Environmental Engineering, EPFL, CH

[4] ETH Zurich, CH

Corresponding author: Ionuț Iosifescu Enescu (ionut.iosifescu@wsl.ch)

EnviDat is the environmental data portal developed by the Swiss Federal Institute for Forest, Snow and Landscape Research WSL. The strategic initiative EnviDat highlights the importance WSL lays on Research Data Management (RDM) at the institutional level and demonstrates the commitment to accessible research data in order to advance environmental science.

EnviDat focuses on registering and publishing environmental data sets and provides unified and efficient access to the WSL's comprehensive reservoir of environmental monitoring and research data. Research data management is organized in a decentralized manner where the responsibility to curate research data remains with the experts and the original data providers.

EnviDat supports data producers and data users in registration, documentation, storage, publication, search and retrieval of a wide range of heterogeneous data sets from the environmental domain. Innovative features include (i) a flexible, three-layer metadata schema, (ii) an additive data discovery model that considers spatial data and (iii) a DataCRediT mechanism designed for specifying data authorship. In addition, the overall user-friendly appearance in EnviDat provides an important opportunity for showcasing WSL research activities and results. The EnviDat portal builds on a conceptual system consisting of a core system, a set of guiding principles and a number of key services. Its development closely follows the conceptual framework, being guided by principles towards the ultimate goal of providing useful services for researchers.

1. Introduction
1.1. The Global Context of Environmental Data Management
Environmental information is becoming increasingly relevant for society as environmental change is proceeding at historically-unprecedented rates. Serious environmental risks were identified in the World Economic Forum's Global Risks Report 2018 (World Economic Forum, 2018), calling for increased efforts to monitor and understand these risks and to develop solutions for managing or mitigating them. At the same time, the amount and quality of environmental information (both data and model outputs) are rapidly increasing, providing a robust, science-based foundation for policy and decision making targeting a sustainable future. Yet, challenges remain as the availability and swift accessibility to environmental information are often inadequate, hindering full exploitation of the available information and thus slowing the progress in scientific understanding.

A number of important developments have been unfolding worldwide that might help to address these challenges. Trends towards more national and international collaboration have been reported (Adams, 2013). Increased sharing, improved accessibility or even open access to research data are being promoted at all levels, nationally and internationally, and across science fields (Perrino et al, 2013). Funding agencies now often ask for the data gathered during publicly funded research projects to be made accessible.

In Switzerland, for example, the Swiss National Science Foundation is following an open research data strategy highlighting the sharing of research data as a fundamental contribution to the impact, transparency and reproducibility of scientific research (Swiss National Science Foundation, 2018). Similarly, at European level, an open research data pilot has been launched for Horizon 2020, the EU Research and Innovation Programme for the period 2014 to 2020. The pilot aims to improve and maximize access to and re-use of research data generated by projects. Research data should be "as open as possible, as closed as necessary" (European Commission, 2016). Furthermore, many publishers of scientific journals nowadays explicitly encourage, expect or mandate data sharing, ultimately requiring the research data underlying publications to be published along with the journal articles. As a consequence, formal publication of research data with assignments of proper citation information using Digital Object Identifiers (DOIs) is becoming the standard across a wide area of research fields.

In some science areas, the research community has long moved ahead and agreed to sharing data at large and to develop shared data management standards and joint research data repositories. Examples include, e.g., the International Tree-Ring Data Bank (NCEI, 2018), the world's largest public archive of tree ring data from around the world, managed by the US National Centers for Environmental Information (NCEI), or the International Nucleotide Sequence Database Collaboration (Nakamura et al, 2013; NCBI, 2018), a collaboration among a number of national databases which has led to many beneficial projects that promise to proliferate in the molecular biology community. Furthermore, over the past years a number of environmental data centers have been established that are now widely being used in environmental science. Examples include the World Data Center PANGAEA (PANGAEA, 2018) or the US National Snow and Ice Data Center (NSIDC, 2018) which supports research into the Earth's cryosphere. Common to these data centers or web-based repositories is that they manage and distribute scientific data, create tools for data access, and support data producers and data users in providing and accessing the information. Some of these centers (and institutions) also perform scientific research and educate the public about the particular scientific fields. Furthermore, a wealth of Earth Observation data sets are federated by the Global Earth Observation System of Systems Platform (GEOSS) infrastructure (GEOSS, 2018) and its associated GEOSS Portal (ESA, 2018), while with the Research Data Alliance (RDA), an international research data management community has been established (RDA, 2018).

Nevertheless, many small institutional repositories are being established to complement existing community-wide data centers or fill the gap where such centers are missing. One advantage of institutional repositories is that the data remains close to the data producers and experts, which can help establish direct contacts between data users and producers and foster scientific collaborations. Initiatives such as re3data. org, operating under the auspices of DataCite (DataCite, 2018), help users to identify registered and certified data repositories, be they institutional, national or international, relevant to particular subject areas (Vierkant et al, 2015).

All these activities and developments are clear proof that research data management has become an integral part of scientific research. It is in this context that WSL is developing its institutional data portal for environmental research data. The goal is to offer easy and efficient access to WSL research data where possible and to provide a catalogue of existing research data information through metadata.

1.2. The EnviDat Institutional Data Portal

EnviDat (www.envidat.ch) is the institutional Environmental Data portal of WSL (Swiss Federal Institute for Forest, Snow and Landscape Research WSL, 2018a). WSL is part of the Swiss federal institutes of technology and research institutions (ETH Domain) dealing with terrestrial environmental systems and aiming for solutions that improve the quality of life in a healthy environment. WSL research focuses on five environmental areas, namely forest, landscape, biodiversity, natural hazards and snow and ice, with the latter being concentrated at the WSL Institute for Snow and Avalanche Research SLF. In addition, WSL is mandated by federal legislation to provide a range of national services, including running the Swiss National Forest Inventory, the Long-term Forest Ecosystem Research Program, the Avalanche Warning Service for Switzerland, the Swiss Forest Health Service, and the monitoring of the Swiss Natural Forest Reserves, as well as providing scientific-technical support for forest plant protection. WSL has therefore a long tradition in data collection, operating a comprehensive network for environmental research that includes more than six thousand observation sites for studying the terrestrial environment. The environmental data sets collected by WSL researchers include long-term monitoring data sets spanning over 130 years that cover most of Switzerland (**Figure 1**).

In addition to extensive expertise in the collection of research data, WSL also possesses substantial experience and expertise in research data management. Already in 2008 (Dawes et al, 2008), WSL researchers
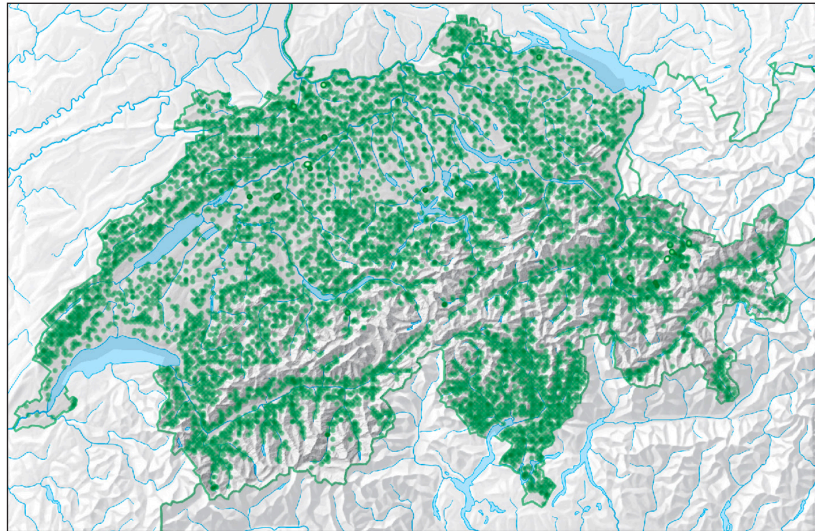
Figure 1: An overview (in green) of the spatial distribution of forest and snow data sets and experiments, known to be available at WSL, as identified by the EnviDat team in May 2018; background map based on data © 2007/2014/2016 swisstopo (JD100042).

at SLF initiated an integrated infrastructure for environmental data management: the Swiss Experiment (SwissEx) and its follow-up project the Open Support Platform for Environmental Research (OSPER) (Dawes et al, 2008, 2012). Both these activities focused on a multidisciplinary collaboration between environmental science and technology research projects across Switzerland for developing an integrated infrastructure for environmental data management, with a strong focus on sensor and GIS data (Dawes et al, 2012; Iosifescu Enescu et al, 2015; Jeung et al, 2010).

The SwissEx/OSPER infrastructure was designed, implemented and supported by Swiss research institutes on the initiative of the CCES Competence Centre for Environment and Sustainability of the ETH Domain (CCES, 2018). WSL has played a central role in the development of the SwissEx/OSPER data platform, part of which is being integrated into EnviDat. The development of EnviDat has substantially profited from these valuable past experiences because the goals of EnviDat are largely complementary to the above-mentioned initiatives and share an emphasis on efficient access and availability of environmental research data.

1.3. EnviDat Strategy and Institutional Tie-in

EnviDat was launched in 2012 as a small project to explore possible solutions for a generic WSL-wide data portal, capable of satisfying requirements from researchers across the wide range of WSL research themes. The first EnviDat development phase was completed end of 2016 and the portal has been running in a "Beta"-version ready for testing operational use. As part of the strategic planning for 2017–2020, the WSL board of directors has selected EnviDat as one of three strategic initiatives of WSL. The development of Envi-Dat has since been intensified and EnviDat now has the status of a WSL program, formally cutting across the five WSL research areas and concerning all research units and central IT services.

EnviDat is set up as a service to researchers and data producers from WSL and SLF, although offering this service to other institutions within the ETH Domain is a mid- to long-term goal. EnviDat supports researchers in the publication of (i) key information about data sets collected, increasing the visibility, as well as (ii), if desired, the actual curated, publication-ready data sets. The basic idea behind this concept is that by sharing WSL data, individual researchers as well as the institution will benefit through new national and international collaborations. Currently, EnviDat has more than fifty registered researchers providing datasets for an estimated number of over a thousand data consumers. According to server logs, EnviDat users have caused more than 1 TB of network traffic in the month August 2018 alone. This exemplifies the importance WSL places on research data management at the institutional level and demonstrates WSL's commitment to enhance (long-term) access to its research data for the international research community as well as for the general public.

The development of EnviDat as a service to increase the visibility and facilitate access to research data, is sustained at the institutional level by the implementation of an overarching WSL data policy. The data policy defines general and overriding principles governing the handling of research data, provides instructions on the provision of research data and offers guidance on agreements between the WSL and the users of its research data. Core principles of the new policy are that (i) WSL will make its research data accessible within

two years after the completion of a research project or a program phase for long-term research programs and monitoring projects; (ii) all WSL research data will be registered in EnviDat through the inclusion of (detailed) meta data, and (iii) access to WSL data will not necessarily be "open access". In this context, access restrictions for certain data sets or resources is a core requirement by WSL data providers. EnviDat thus includes several accessibility layers and resources can be made public openly (under, e.g., an open source license) or restricted to registered users, members of an organization or even to a set of specifically defined users per data set. Consequently, even if the metadata is fully open and accessible, different data publication routes are possible, from fully open data to data that is made available only upon request.

The remainder of the paper is organized as follows: Section 2 introduces the EnviDat concept with the core system, its principles and services. Section 3 details the EnviDat architecture and technology stack. Section 4 highlights some of the differentiating aspects of EnviDat in the larger context of data portals. Finally, concluding remarks are presented in Section 5.

2. The EnviDat Concept

The EnviDat portal is being developed based on the conceptual framework shown in **Figure 2**. The Envi-Dat conceptual framework and its implementation into the data portal are regularly being revisited and, if necessary, adapted and extended. The overarching EnviDat concept includes several components: the EnviDat core system (**Figure 2**, center) and two supporting pillars (**Figure 2**, left/right), namely principles and services, with everything resting on experiences from, and exchanges with national, pan-European and international research data management initiatives.

The core system is focusing on the essential concepts related to the registration and integration of metadata and associated data sets. The principles-pillar consolidates the overarching guidelines, while the services-pillar specifies the most important high-level functionalities of the EnviDat portal. The entire EnviDat portal rests on a solid foundation of experiences and expertise gathered at national, pan-European and international levels, emphasizing that EnviDat is following best-practices from established and distinguished international Research Data Management (RDM) initiatives.

2.1. EnviDat Core System

The EnviDat core model focuses on the integration and publication of heterogeneous environmental data and the provision of accompanying metadata.

The first step for any data publication in EnviDat is the registration of the metadata describing the data. The purpose of metadata is twofold: as a documentation for the corresponding data and as an enabler of important portal functionalities such as facilitating searchability and accessibility by potential users. The quality of the metadata directly influences the findability of the data sets as well as their possible semantic cross-linking, or in other words, the ease with which the available data sets can be found. The metadata can be added directly, through a manual process for individual data sets, or indirectly, through automated bulk imports of metadata entries that are already available from other operational databases and information
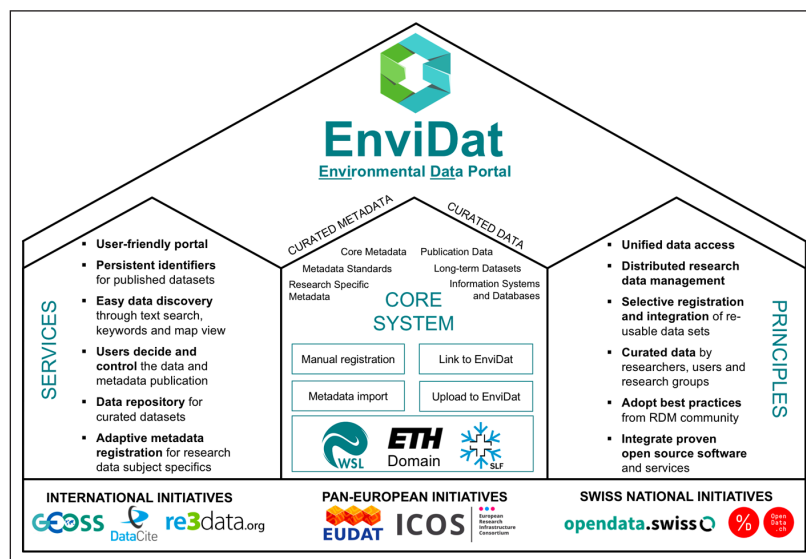

Figure 2: The EnviDat concept.

systems. The metadata registered in EnviDat ranges from core catalogue metadata to domain-specific research metadata. EnviDat has the capability to map such core metadata to other compatible metadata standards. More information about the EnviDat metadata schema is presented in Section 4.2.

In a second step, the curated, publication-ready data can then either be integrated by linking to existing information management systems and databases or uploaded to the EnviDat data repository. Both options represent core features of EnviDat. The upload of resources to an internal portal repository represents the "standard" for any research data management repository. The data sets integration through linking to existing information systems is an important feature, too. The latter allows data producers, including the long-term WSL programs Swiss Long-term Forests Ecosystem Research Program – LWF (Swiss Federal Institute for Forest, Snow and Landscape Research WSL, 2018b) and the Swiss National Forest Inventory – NFI (Swiss Federal Institute for Forest, Snow and Landscape Research WSL, 2018c), which already have in place well-designed and extensively used professional databases and operational web-based information systems, to use EnviDat without duplicating efforts. External metadata can be imported in bulk and, if the external information system requires authentication, an optional token mechanism (detailed in Section 3) allows authenticated EnviDat users to connect effortlessly to those systems. In addition, such information systems have implemented specific requirements adapted to their target audience, such as data selection based on domain-specific themes or data types, that are costly to implement and maintain and therefore not necessary to be replicated in a generic environmental data repository such as EnviDat.

The EnviDat core system described above is steered by a set of principles and services that support and define the EnviDat value proposition, namely a unified access to environmentally-relevant metadata and curated publication data available at WSL. The principles reflect the guiding philosophy behind the EnviDat system, the Services define the main functionalities provided by the portal to users.

2.2. EnviDat Guiding Principles
The EnviDat concept defines six EnviDat principles as shown in the right pillar of **Figure 2**. The first four principles, namely (P1) unified data access, (P2) distributed research data management, (P3) selective registration and integration of data sets and (P4) curated data, can be summarized under the motto "unified access, distributed curation".

The unified data access principle (P1) emphasizes the main mission and vision of the EnviDat portal as a unique gateway to WSL data sets, irrespective of where they are physically stored and curated. The second guiding principle (P2) is specifically emboldening a system for supporting decentralized processes in research data management, in contrast to a centralized system such as an institutional database for all research-related data. A centralized system for all research data of an institution might have advantages in terms of consistent data management and curation, but at the cost of a separation of research data from domain experts and the creation of redundancies with little or no value added. This principle however, does not exclude the use of a data warehouse (or data cache) that is important for the implementation of specific use cases such as interactive visualizations or efficient real-time extraction of data subsets. The third and the fourth principles tackle the selective (P3) registration and integration of quality-controlled (P4) data sets, both being features which help EnviDat stand out in terms of data and metadata management compared to other data portals. Finally, the last two principles are related to leveraging the considerable know-how already available in the research data management community by adopting best practices (P5) and by integrating proven open source software and services (P6) into EnviDat. A strong example of leveraging the community contributions is the integration of CKAN – Comprehensive Knowledge Archive Network (CKAN, 2012) into the EnviDat system architecture, an open source repository standard supported by the Open Knowledge foundation (Open Knowledge International, 2018). More information about the technology and architecture of EnviDat is provided in Section 3.

2.3. EnviDat Services
Services are the second important pillar supporting the core EnviDat system. Services define the portal's functionalities offered to users of EnviDat, both data producers and data consumers. To this end, EnviDat provides generic services for researchers that include (S1) registration, (S2) data repository, (S3) controllable publication process, (S4) easy data discovery, (S5) provision of persistent identifiers and (S6) an overall user-friendly experience **Figure 2**, left).

The main service of the EnviDat portal, influencing all subsequent services, is the adaptive metadata registration (S1) that is based on a layered metadata model with three layers (core, optional and domain-specific research metadata). More details about the EnviDat layered metadata model is presented in Section 4.2.

The EnviDat repository (S2) for curated, publication-ready data offers the possibility to directly link metadata to the corresponding data files and other resources hosted in EnviDat. EnviDat accepts many types of resources, i.e., from time-series data and numerical models to images and software, as required by researchers. All data sets are curated and prepared for publication by the researchers and/or data providers. The choice of data file formats is entirely left to the data providers (in accordance with the domain-relevant de-facto or de-jure community standards), yet EnviDat offers guidance and advice on data formats suitable for long-term preservation. EnviDat currently accepts individual data file uploads over the HTTP protocol up to several GBs. The self-assessed upper feasibility limit for the EnviDat repository is of around 50–100 GB per file, depending mainly on the available network bandwidth and the reliability of the HTTP protocol for supporting large file downloads to data consumers. Furthermore, significant effort is being invested for drastically increasing this limit in the future, based on the ScienceDMZ design pattern (ESnet, 2018).

The full control over the data publication process (S3) is another important EnviDat service available to data producers. Newly created metadata records are by default marked and saved as "draft" and kept as "unpublished" even after all the mandatory metadata fields have been completed. A data provider must thus actively change the status to "public" in order to publish a data set. With this approach, we provide the data producers full control over the data publication process and ensure that they have sufficient time to carefully review, select, clean or curate the (meta-)data sets before publication.

Easy data discovery (S4) is a distinct feature for any environmental data portal. The additive data discovery mechanism, as further detailed in Section 4.3, represents the foundation for this service in EnviDat. The data discovery and findability are greatly facilitated by the availability of spatial information, as the geometry of the spatial extent is a mandatory part of the core EnviDat metadata schema. Similarly, links or semantic connections between data sets based on the Linked Data paradigm could be added in the future in order to improve the search results and their ranking, as described in Section 4.4.

In terms of data publication, EnviDat offers its users persistent identifiers (S5) as well as a process for generating and assigning ("minting") Document Object Identifiers (DOIs) to their published data sets though the DOI-Desk of ETH Zurich (ETH Library, 2018). The creation and assignment of permanent identifiers for data sets is becoming more and more important in research in general and for data portals in particular (research data catalogues, citations etc.)

Finally, the user-friendly access (S6) to the available data sets published in EnviDat is paramount for data users. It results from the combination of many small development steps dedicated to increasing the overall user-friendliness. Some examples are the inclusion of the geographical search option implemented through a map-based interface (further detailed in Section 4.3), the graphical display of the spatial extent close to the detailed metadata record or the possibility to draw the geometry of the spatial extent in a user-friendly manner (instead of writing a geometry string) when registering a data set (**Figure 3**).

Figure 3: Example of increasing the user-friendliness of the EnviDat portal through the inclusion of several map-based graphical user interfaces.

In order to further improve the user-friendliness of EnviDat, a new web-interface for data consumers is currently being designed and developed. Consideration of both data user and provider perspectives requires developing separate systems for data provisioning and data consumption. This new EnviDat interface is the starting point for adapting to the requirements of the users of WSL research data, which differ substantially from the needs of WSL data providers, focused mainly on the publication of their data sets.

3. EnviDat System Architecture and Technology Stack
The availability of the generic EnviDat Services, as well as any future portal interface and functionality improvements, are strictly interrelated with the EnviDat system architecture and technology stack. Consequently, EnviDat implements the functional requirements for data publishing derived from the Services part of the conceptual model. In line with our principles, EnviDat adopted best practices and standards in data sharing by integrating prominent technologies for data repositories that are available from the wider research data management community. The EnviDat portal for data providers is leveraging community software such as CKAN, Apache Solr and PostgreSQL. Additional custom software is being developed for automatic DOI minting and for a new graphical user interface for data consumers.

Non-functional requirements such as reliability, security and maintainability are considered through a multi-server system architecture (**Figure 4**). The EnviDat three-tier system architecture separates the presentation, application and data layers, with productive servers from each layer having additional hot-spare servers that can take over the duties of the primary servers, as a failover mechanism in order to safeguard the overall availability of the EnviDat portal.

The data layer bundles the PostgreSQL databases, the extensible virtual storage for the file repository, and the infrastructure mechanisms for ensuring the safety of the metadata and data such as backup and mirroring. The application logic layer contains the majority of computational components, such the core CKAN (CKAN, 2012) and the Solr (The Apache Software Foundation, 2017) search server, as well as the middleware needed for the communication between the data layer (back-end) and the presentation layer (front-end). The third layer, namely the presentation layer is dedicated to serving graphical user interfaces (GUIs) built using Web standards such as HTML (W3C, 2018a), CSS (W3C, 2018b) and JavaScript (W3C, 2018c).

CKAN is the main open source software component powering EnviDat. It is used in a number of prominent Swiss open data projects such as, e.g., opendata.swiss (Swiss Confederation, 2018) or the open data portal of the City of Zurich (City of Zurich, 2018). Furthermore, the Swiss Federal Institute of Aquatic Science and Technology Eawag, another federal research institute of the ETH Domain, is also implementing an institutional repository for research data based on CKAN (Eawag, 2016), which ensures continuous exchanges and mutual support.
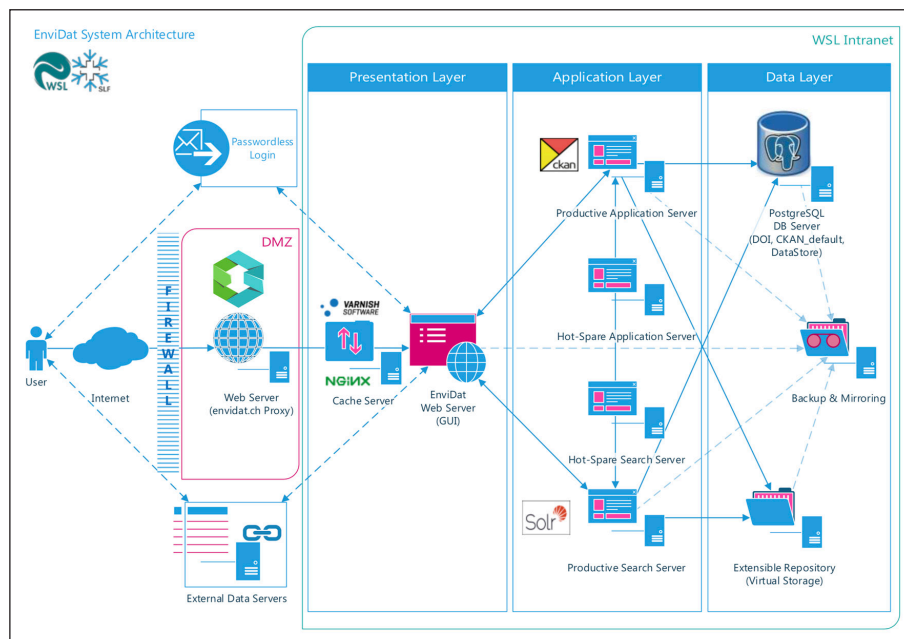
Figure 4: EnviDat three-tier system architecture containing clearly defined data management, application and presentation layers.

A first step towards the implementation of this EnviDat future architecture was implemented by separating three critical architectural components from the main CKAN server, namely the PostgreSQL (The PostgreSQL Global Development Group, 2018) database (for metadata storage), the CKAN repository (for data storage) and the Solr search server (delivering search results to CKAN) on separate virtual servers. As a result, the current system is now using a productive, regularly backed up PostgreSQL database of the IT Services in Birmensdorf (for safe metadata storage), a 2 TB network share that is regularly backed-up and mirrored to a secondary site at the SLF in Davos (for safe data storage) and a dedicated Solr virtual server (powering the CKAN metadata search).

As the visualization of time-series and GIS data are also envisioned to become part of EnviDat in the long-term, the EnviDat technology stack will also be considerably expanded by GIS-related technologies. For example, a possible solution for implementing future support for GIS-data in EnviDat is to assimilate the technology stack already available from the geodata4edu.ch and GeoVITe (Geodata Versatile Information Transfer environment) projects (geodata4edu.ch, 2018; Iosifescu-Enescu, 2016). As a result, the EnviDat technology stack that currently includes CKAN, Apache Solr and PostgreSQL could be expanded with GIS-related technologies such as QGIS Server (QGIS, 2018), GDAL/OGR (GDAL, 2018) or GeoTools (GeoTools, 2018).

Furthermore, as the HTTP protocol for supporting successful large file transfers from the EnviDat repository to data consumers has its limitations, additional architectural changes are needed in order to implement the ScienceDMZ design pattern (ESnet, 2018). More specifically, Data Transfer Nodes (DTNs), which are additional servers dedicated to supporting large file transfers, will need to be placed outside the Firewall (**Figure 4**). Finally, a passwordless login rounds up the architecture and enables logged in users to be seamlessly linked to external systems, through an optional token mechanism based on the hashed user ID and a shared secret between EnviDat and the external system.

Finally, as the EnviDat development workflow will gradually shift to Agile software development principles (Beck et al, 2001), this change will impact the EnviDat System Architecture. Consequently, a continuous architectural redesign of the EnviDat system architecture is necessary in order to support the proper application of the Continuous Delivery (Beck et al, 2001) concept through an agile architecture that, according to (Waterman et al, 2015) "is being able to be easily modified in response to changing requirements, is tolerant of change, and is incrementally and iteratively designed – the product of an agile development process".

## 4. Differentiating Aspects of EnviDat in the Context of Data Portals

The philosophy behind EnviDat is revolving around developing solutions for efficient, unified and managed access to the WSL's comprehensive reservoir of monitoring and research data, in accordance with the data policy and the strategy. Consequently, EnviDat has several specific aspects that differentiate us from other data portals. These differentiating aspects include fostering the publication of curated, quality-controlled data sets, with a clear mechanism designed for specifying data authorship (D1), a flexible and adaptive three-layer metadata schema (D2), an additive data discovery mechanism that considers spatial data and semantics (D3), intrinsic support for Open Science (D4), and integrated visualization of the hosted environmental data (D5).

4.1. Curated, Quality-controlled Data (D1)

The first aspect of EnviDat to be highlighted is that it requires data providers to register curated, quality-controlled, publication-ready data sets, that are ideally conforming to the FAIR (Findability, Accessibility, Interoperability and Reusability) principles (Wilkinson et al, 2016). The FAIR conformance for (meta)data sets requires them to be described with a formal metadata language (promoting Interoperability) containing a plurality of rich and relevant metadata attributes that meet domain-relevant community standards (fostering Findability and Reusability) and released with a clear usage license even in the absence of the data (safeguarding Accessibility and Reusability).

In order to improve the quality-control for research data sets, we aim for introducing an approach that is similar to the regular peer-review process applied for scholarly articles, a process currently often missing in the data publication process. For this reason, the EnviDat team encourages nomination of EnviDat data administrators for every data provider organization. Data administrators can review the (meta)data with regard to completeness and reusability. The responsibility to curate the research data and metadata remains with the domain specialists, namely the original data providers and their research units or groups at WSL, thus associating data sets with their detailed provenance in order to further increase the reusability of (meta)data. It is important to note that curated, quality-controlled, publication-ready data sets only make up a small percentage of the total amount of data accumulated throughout a research project. Restricting the
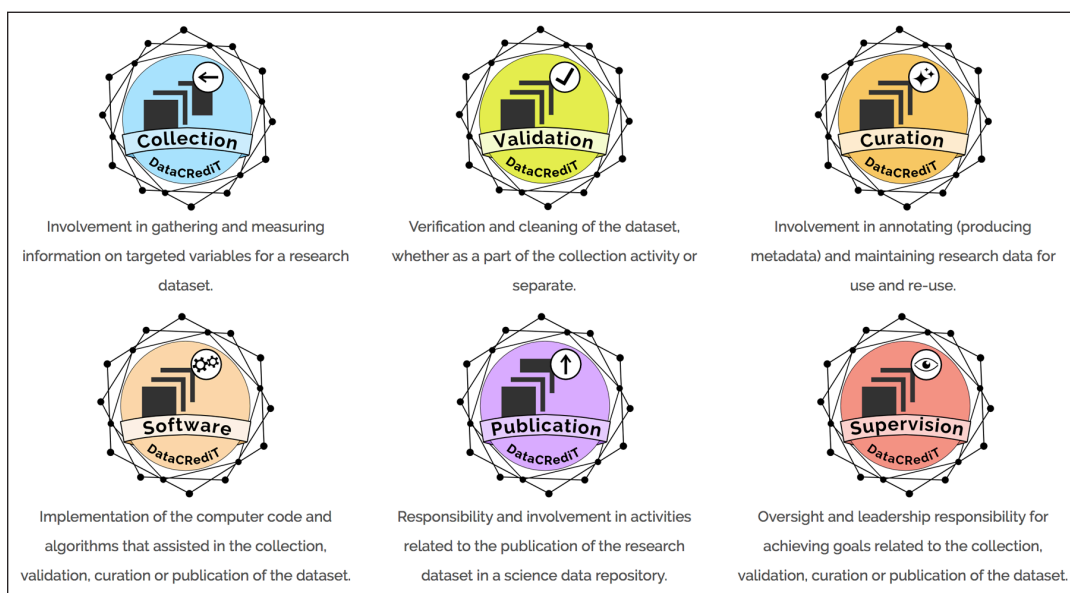
Figure 5: The Data Contributor Roles (DataCRediT) taxonomy of EnviDat.

repository to only such data sets not only avoids the cluttering of the repository and improves the findability of relevant data, but also promotes only the data sets that have the highest potential for being re-used by others. A notable exception to this is represented by offering streamed data from automatic measurement stations, such as weather stations, due to the fact that the immediate availability of such data can speed up research. Nevertheless, such streamed data are of limited temporal coverage and long-term time series need to be validated and curated by the corresponding data providers.

Ultimately, we do envision that such reviews of metadata information and data sets could also be conducted by users of the data sets and issues identified in the data might be reported back to the data originators. Yet, organizations in EnviDat are strongly encouraged perform their due-diligence, too, by identifying data administrators tasked with checking compliance of the data sets published in EnviDat. A topic of interest for EnviDat in this context is the possibility to specify individual author contributions to the publication of a particular data set. We have defined a DataCRediT (WSL, 2018) mechanism for data authorship specification, inspired by the Contributor Roles Taxonomy (CASRAI, 2018) included in the DataCite metadata schema (DataCite Metadata Working Group, 2016). The DataCRediT mechanism is adapted and enhanced to cover a number of contributor roles of people involved in research data publishing. Currently we propose six DataCRediT contributor roles: DataCRediT/Collection, DataCRediT/Validation, DataCRediT/Curation, DataCRediT/Software, DataCRediT/Publication and DataCRediT/Supervision (**Figure 5**). We do however plan to regularly review, adapt and extend the DataCRediT taxonomy in the future based on user feedback.

We believe that such a DataCRediT taxonomy can become an important tool in supporting the principle of publishing curated, well-documented data sets. The taxonomy applied supports transparency in contributions to research data sets published in EnviDat and it provides an improved system of attribution, credit, and accountability for scientific data publication, thus encouraging the application of the FAIR data principles by the individual researchers and data providers.

4.2. Adaptive Metadata Schema (D2)

The FAIR data principles must be extensively supported not only by the data providers but also by data repositories such as EnviDat, in order to allow the data providers to describe the data sets with rich metadata meeting current and future domain-relevant community standards. Consequently, we have developed an adaptive metadata schema model with three main layers (core, optional and domain-specific research metadata) depicted in **Figure 6**.

At the core of the EnviDat metadata schema there are a minimal number of mandatory metadata fields for cataloguing and documenting the data set: title, description, keywords, author(s), organization, publisher, publication year, contact information and the geometry of the spatial extent. A unique identifier is automatically added by the system. Core metadata can also be mapped to various standards such as Dublin Core (DCMI, 2018), DataCite metadata schema (DataCite Metadata Working Group, 2016) or ISO 19139 (ISO, 2007).
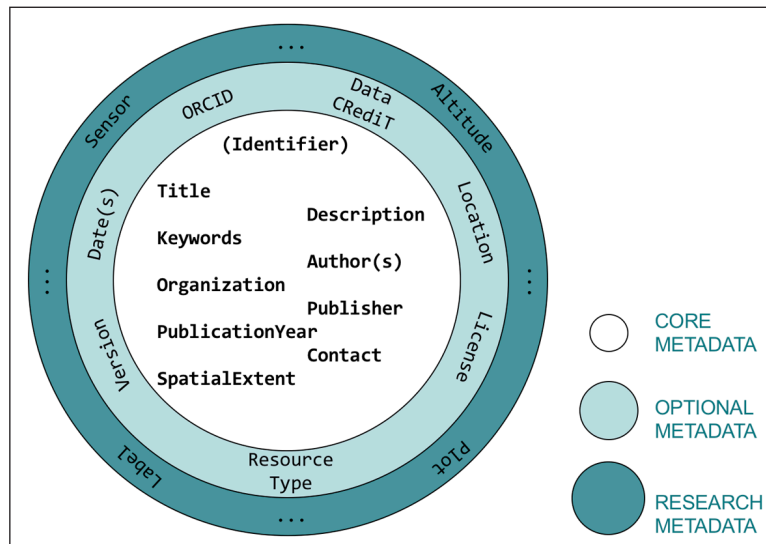
Figure 6: EnviDat three-layer metadata schema model.

A second layer of optional metadata is enhancing the mandatory fields with additional information, such as (1) additional standard identifiers for authors, (2) a selection of date fields (e.g., created, collected, updated), (3) an opportunity to specify the license, (4) the type of the resources (e.g., data set, software, etc.), (5) the current version of the data set, (6) the toponym or (7) a Data Contributor Roles (DataCRediT) taxonomy for data set authors (as detailed previously). Supported standard identifiers for authors include ORCID IDs (ORCID, 2018), ISNI – International Standard Name Identifier (ISNI International Agency, 2018), ResearcherID (Thomson Reuters, 2018), ResearchGate (ResearchGate, 2018) or any other future standard, and data collection dates and ranges may become mandatory in the future in order to support temporal search.

The third layer of domain-specific research metadata is needed by various WSL databases and information systems that have specific metadata considered important for efficient searching, retrieving and correct ranking of their relevant data sets. For this reason, any domain-specific research metadata can be recorded in EnviDat in a custom metadata field following a key-value approach, where the key represents the name and the value represents the content of the custom metadata field. Furthermore, such extended metadata can hold information for cross-linking between data sets. For the future, the methods described by the ISO 19115-1:2014 standard (ISO, 2014) for extending metadata to fit particular needs are also under consideration, especially if complex hierarchical metadata will become a requirement. Finally, all these additional metadata fields are automatically indexed and therefore covered by the EnviDat additive data discovery mechanism as detailed further in Section 4.3.

4.3. Additive Data Discovery Mechanism (D3)
The additive data discovery mechanism is another distinct feature for an environmental data portal. The main idea is to consider, in an additive and commutative manner, the different dimensions of the metadata, in order to supplement the well-known free-text search.

Currently, we employ three mutually connected search functionalities: geospatial search option, keywords selection and the free-text metadata search (**Figure 7**). The most important element, the full indexing of metadata content is important for full-text searches on all available metadata fields. In addition to the full-text metadata search, the spatial geometry, a mandatory metadata field in EnviDat, is central to search and discovery functionalities based on geospatial information characteristics. Finally, the relevant keywords defined by the data providers, can be used either to list all data sets that are tagged with the corresponding keywords or to further filter the results given by the full-text search.

4.4. Extending Data Discovery with Semantics (D3)
The additive data discovery mechanism introduced in Section 4.3 could additionally include semantic connections between data sets, in order to augment the findability of related data sets and to improve the ranking of search results. Linked Data (Bizer et al, 2009) could be used as a new paradigm to organize and represent both metadata and data. In an earlier test, we mapped the DataCite Metadata Schema (DataCite
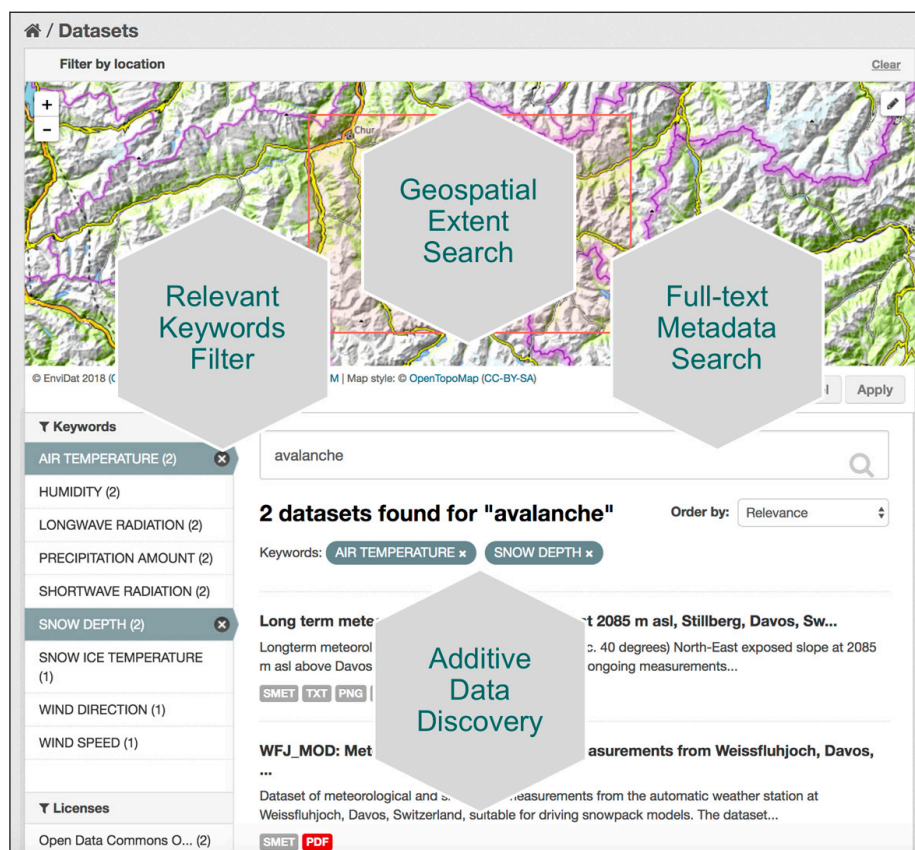
Figure 7: Additive data discovery as implemented in EnviDat, combining relevant keywords, geospatial
extent and full-text metadata search.

Metadata Working Group, 2016) with the help of the DataCite Ontology (Shotton and Peroni, 2018) to
Resource Description Framework – RDF (W3C, 2014). RDF is allowing to define the structure of Linked Data
and thus represents one of the fundamental technologies enabling the Semantic Web. The Semantic Web
represents a data Web, providing a common framework for integrating across different machine-readable
data on the Web (Bizer et al, 2009).

The future integration of the concept of Linked Data in EnviDat might help to further improve efficient
search and data discovery since Linked Data has the potential to increase the re-usability of data sets by
exposing, sharing, and interconnecting pieces of data. Linked Data has, in the context of a data portal, two
notable characteristics: a) a graph-based data model that is easily expandable even beyond system bounda-
ries and b) data that can be broken up to an atomic granularity. Here, two use cases were particularly inter-
esting to investigate, namely (i) to extend core metadata and (ii) to describe the structure of data sets for
enabling cross-linking between them. The first use case was to extend the information about the data sets
by linking to an author ID system like ISNI which offers a Linked Data entry point, which potentially allows
us to find more information about data set contributors. Furthermore, the geographic location of data sets
can be further extended with DBpedia (DBpedia, 2018), which offers Wikipedia (Wikipedia, 2018) content as
Linked Data, therefore linking to information that transcends the system boundaries.

A further advantage of Linked Data is the atomic granularity with which (meta)data can be represented.
Therefore, the second use case in the context of a data portal is to use the Linked Data concept to describe
the structure of the data sets. The Linked Data could contain a generic description of data sets analogue
to the RDF Data Cube Vocabulary (W3C, 2018d) implementing a (hyper)cube of space, time and attributes
as described in (Ott and Swiaczny, 2012) or (Iosifescu Enescu et al, 2015). In this way, the semantics and
therefore the meaning helps increasing the value of a data set. Furthermore, this granularity could be the
starting point of linking and recombination between different data sets or to extract data subsets if so
desired by the data users. Unfortunately, the new technologies like Linked Data are not yet as mature as for
example relational databases, and therefore more effort is required to transfer them into a productive mode.
Nevertheless, the achievements of the Semantic Web and the paradigm of Linked Data has a lot of potential
for linking and expanding the metadata already existing in data portals.

4.5. Support for Open Science (D4)

Open Access (OA) publishing is gaining widespread traction in the global academic community (Frosio, 2014) and as a result, there is an increasing availability of open research papers. Moreover, as detailed in section 4.1, EnviDat requires data providers to register curated, quality-controlled, publication-ready data sets, that are conforming as much as possible to the FAIR principles described in (Wilkinson et al, 2016). Hence, one of EnviDat's distinctive features is to support Open Science as described in (Fecher and Friesike, 2014; Finnish Open Science and Research Initiative, 2014; OECD, 2015), by making research data accessible as a means for "accelerating research, […] enhancing transparency and collaboration, and fostering innovation" (OECD, 2015). However, in order to truly support Open Science, scientists still have the problem of sharing and reproducing the computations needed to process and visualize the (open) data according to the methodology presented in a research paper.

Fortunately, Jupyter Notebooks (Project Jupyter, 2018), emerged as a one possible solution for allowing scientists to share detailed and understandable descriptions of their code, making it easy for reviewers and other scientists to explore the code and thus follow the corresponding "computational narrative" (Project Jupyter, 2015; Shen, 2014). They allow to combine live code with explanations as narrative text, as well as equations, visualizations and other media (Project Jupyter, 2018). Furthermore, Jupyter Notebooks support a wide range of programming languages often used by environmental researchers, such as Python, R, Octave, Scilab, Matlab, C, Java or Scala.

Although the EnviDat repository already allows scientists to save any kind of resources in digital form, EnviDat aims for a deeper integration of user-uploaded Jupyter Notebooks. The long-term ambition is to offer a solution for accessing tools and code from anywhere and providing a computational infrastructure for running the Web-based notebooks interactively on EnviDat. In spite of limitations related to the integration of proprietary software libraries and the reduced computing infrastructure resources that are available to EnviDat, developing a solution for bridging the gap between Research Data Management and Open Science is another distinctive forward-looking feature of EnviDat.

4.6. Visualization of Environmental Data (D5)

The WSL researchers are collecting, validating, curating and publishing environmental data in a variety of digital formats. Furthermore, the data availability in the environmental domain appears to be increasing overall, as proven for example by the growing number of Earth Observation data sets listed in the GEOSS infrastructure (GEOSS, 2018).

Curated metadata are at the heart of EnviDat, as they describe complex and large data sets. But only linking and hosting the data is not enough. Although the EnviDat repository is allowing the publication of data sets in any digital form (such as CSV or links to operational information systems and databases), currently there is no possibility to preview the data before downloading it, or in other words the data consumers need to first download the data content in order to assess its suitability for their purpose.

The redesign of a new EnviDat frontend for data consumers takes a user centric design approach (Nielsen, 1999). Its main goal is to provide an effortless and intuitive user experience when browsing, finding and downloading the data. This is expected by current (Web) design practice; if a website provides a great experience and helps users to achieve their task, they will use it with pleasure (Fintech, 2017; Norman, 2003). Taking a modern approach on the way data is described means to include visual tools for the data creators to effectively communicate what their data is about.

On a larger scale, it is well recognized that information interpretation efficiency is not advancing to the same extent as data acquisition (Dam et al, 2000; Johnson, 2004) and that attractive things work better (Norman, 2003). Furthermore, (Slocum et al, 2008) confirms that map-based visualization is an effective way of presenting spatial and statistical data. For example, it is both more effective and more user-friendly to (i) look at a graphical representation of numerical data from a time-series rather than looking at raw or tabular data and (ii) to preview Geographic Information Systems (GIS) data in a map-based interface. Consequently, the visualization of environmental data it is not only more user-friendly but it can also contribute to improving data interpretation and therefore assisting in understanding complex environmental processes by reducing the gap between the data acquisition and the speed of data interpretation. Moreover, it can provide a synergistic effect with semantics (as described in Section 4.4), as visualization can also profit from the existence of generic data descriptions. As a result, an innovative feature currently under consideration for the future development of EnviDat is the web-based visualization of data on maps and charts. As a first proof of concept, we have performed a small test to allow the preview of the latest data from the automatic weather stations of the Greenland Climate Network – GC-Net (Steffen Research Group, 2018). Different numerical

parameters such as atmospheric pressure, temperature, wind speed or wind direction are visualized on a chart, thus allowing an immediate interpretation of the latest atmospheric data from Greenland. Giving the data producers such tools that would allow to present their data automatically and effortlessly, will in turn improve the overall experience of the EnviDat portal itself and sets EnviDat apart from other data portals which rely on a standard textual description of complex data sets.

In this context, EnviDat is also intended to be showcasing the possible convergence between repositories dedicated to data publication, traditionally developed by the research data management community, and geoportals, traditionally developed by the geospatial community. Since geoportals are web-based systems designed for discovering, presenting and retrieving geodata as a specific subset of data from earth system research (Kellenberger et al, 2016), geoportals lack EnviDat's focus on publishing and archiving. Yet, EnviDat is enabling geoportals to federate its contents to European and international communities by implementing support of geospatial data and ISO 19139 metadata. Consequently, after becoming a GEOSS data provider, EnviDat is now featured as one of the community portals in the GEOSS Portal (ESA, 2018).

5. Discussion and Conclusions

The EnviDat portal, builds on a solid conceptual framework encompassing the core system, principles and services, as a well a three-tier system architecture. For data providers, the (meta)data publication and the availability of a data repository are the top most important services that EnviDat offers. The metadata publication is implemented with a custom metadata schema for EnviDat served through the CKAN repository standard. Yet, the metadata recorded by individual data producers greatly vary in length and detail. This is due to substantial differences in complexity between metadata records, e.g., between metadata resulting from established national research programs such as LWF and NFI, and the (meta)data collected as part of a master thesis. To account for this complexity EnviDat offers a three-layer model for metadata handling. Data sets registered in EnviDat require only a minimum of core catalogue information (including title, author(s), keywords), though a more extensive metadata schema is supported and encouraged. In addition, EnviDat offers the opportunity to specify author contributions to data sets through a newly developed DataCRediT taxonomy, one of the differentiating features of EnviDat that we believe to gradually become important also for the wider research data management community. Active collaboration and contribution to the wider RDM community as well as connecting to successful RDM initiatives will continue to be important for EnviDat in the near future. The coordination between different data portal initiatives is also important in order to mitigate the proliferation of unconnected data portals operating in the area of environmental science.

The EnviDat data repository is a vital complement to the metadata publication, while EnviDat also allows linking existing external operational systems. Both options are core to the usability of a research data management portal for the environmental domain. The EnviDat repository is exclusively focused on curated, publication-ready data for which full copyright ownership exists. Furthermore, in terms of data users, the EnviDat focus will be on further increasing usability and user-friendliness by, e.g., introducing map-based interfaces. The user-friendly access to data through EnviDat ultimately fosters further advances in environmental science, since long-term data sets are particularly valuable towards obtaining an integrated view of the Earth System.

However, developing an institutional portal for environmental research data also forces us to reflect upon the future. Creating a zoo of unconnected web-based data portals worldwide will limit their usability and, in fact, might even reduce their usefulness and service to the science community and the public. Therefore, we consider it important to coordinate and connect between the various initiatives in order to avoid fragmented parallel developments. While a coordinated, central portal dedicated to environmental data does not currently exist in Switzerland, EnviDat is working towards collaboration with national portals such as the opendata.swiss (Swiss Confederation, 2018), the Swiss authorities' portal for open data, and international initiatives such as the GEOSS Portal, which federates a wealth of heterogeneous collections of databases and other portals, including EnviDat (ESA, 2018). Such cooperation will also promote the dissemination of WSL environmental research data at European and international level.

Acknowledgements

Competing Interests
The authors have no competing interests to declare.

Author Contribution
Contributor Roles/Writing – original draft: Original draft written by Ionuţ Iosifescu Enescu and Gian-Kasper Plattner with contributions from Sandro Bischof and Michael Lehning, based on the abstract submitted to the Göttigen-CODATA RDM Symposium 2018.

Contributor Roles/Writing – review & editing: All authors have reviewed and commented the original draft.

Contributor Roles/Conceptualization: The EnviDat differentiating features and the conceptual framework emerged from the common work of Ionuţ Iosifescu Enescu, Sandro Bischof, Lucia Espona Pernas and Gian-Kasper Plattner.

Contributor Roles/Visualization: All figures were produced by Ionuţ Iosifescu Enescu. The EnviDat logo and the DataCRediT badges used in Figures 2, 4, and 5 were contributed by Dominik Haas-Artho.

Contributor Roles/Software: The current EnviDat portal is the technical work of Lucia Espona Pernas and Ionuţ Iosifescu Enescu.

Contributor Roles/Supervision: Gian-Kasper Plattner heads the EnviDat program, Konrad Steffen and Michael Lehning share the strategic oversight, and Ionuţ Iosifescu Enescu is technical coordinator of the EnviDat program.

References
Adams, J. 2013. Collaborations: The fourth age of research. *Nature*, 497(7451): 557–560. DOI: https://doi.org/10.1038/497557a

Beck, K, Beedle, M, van Bennekum, A, Cockburn, A, Cunningham, W, Fowler, M, Grenning, J, Highsmith, J, Hunt, A, Jeffries, R, Kern, J, Marick, B, Martin, RC, Mallor, S, Shwaber, K and **Sutherland, J.** 2001. *The Agile Manifesto.* The Agile Alliance.

Bizer, C, Heath, T and **Berners-Lee, T.** 2009. Linked Data – the story so far. *International Journal on Semantic Web and Information Systems*, 5(3): 1–22. DOI: https://doi.org/10.4018/jswis.2009081901

CASRAI. 2018. CRediT – CASRAI. Available at: http://docs.casrai.org/CRediT [Last accessed 8 May 2018].

CCES. 2018. CCES Research platforms. Available at: http://www.cces.ethz.ch/research/platforms.html [Last accessed 8 May 2018].

City of Zurich. 2018. Open Data Zürich – Stadt Zürich. Available at: https://data.stadt-zuerich.ch/ [Last accessed 9 May 2018].

CKAN. 2012. About CKAN. *ckan.* Available at: https://ckan.org/developers/about-ckan/ [Last accessed 9 May 2018].

DataCite. 2018. Welcome to DataCite. Available at: https://www.datacite.org/ [Last accessed 14 June 2018].

DataCite Metadata Working Group. 2016. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.0. *DataCite Schema.* [Last accessed 8 May 2018]. DOI: https://doi.org/10.5438/0012

Dawes, N, Kumar, KA, Michel, S, Aberer, K and **Lehning, M.** 2008. Sensor Metadata Management and Its Application in Collaborative Environmental Research. In: *2008 IEEE Fourth International Conference on EScience*, 143–150. DOI: https://doi.org/10.1109/eScience.2008.27

Dawes, NM, Lehning, M, Bavay, M, Sarni, S, Iosifescu, I, Gwadera, R, Scipion, DE, Blanchet, J, Davison, A, Berne, A, Hurni, L, Parlange, MB and **Aberer, K.** 2012. Open Support Platform for Environmental Research (OSPER) – tools for the discovery and exploitation of environmental data. *AGU Fall Meeting Abstracts*, 51: IN51B-1691.

DBpedia. 2018. DBpedia. Available at: http://wiki.dbpedia.org/ [Last accessed 8 May 2018].

DCMI. 2018. DCMI Specifications. Available at: http://dublincore.org/specifications/ [Last accessed 8 May 2018].

Eawag. 2016. eawag-ckan: Our customized version of CKAN (ckan.org). *Eawag – Research Data Management.* Available at: https://github.com/eawag-rdm/eawag-ckan [Last accessed 9 May 2018].

ESA. 2018. GEOSS Portal. Available at: http://www.geoportal.org [Last accessed 14 June 2018].

ESnet. 2018. Science DMZ. Available at: https://es.net/science-dmz/ [Last accessed 9 May 2018].

ETH Library. 2018. ETH Zurich DOI Desk. Available at: http://www.library.ethz.ch/en/Dienstleistungen/Publizieren-registrieren-verwalten/DOI-Desk-der-ETH-Zuerich.

European Commission. 2016. Guidelines on FAIR Data Management in Horizon 2020. Available at: http://
ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-
mgt_en.pdf [Last accessed 14 June 2018].

Fecher, B and **Friesike, S.** 2014. Open Science: One Term, Five Schools of Thought. In: *Opening Science*,
17–47. Springer, Cham. DOI: https://doi.org/10.1007/978-3-319-00026-8_2

Finnish Open Science and Research Initiative. 2014. Open Science and Research Handbook. Available at:
https://openscience.fi/handbook [Last accessed 15 May 2018].

Frosio, G. 2014. Open Access Publishing: A Literature Review (SSRN Scholarly Paper No. ID 2697412). Roch-
ester, NY: Social Science Research Network. Available at: https://papers.ssrn.com/abstract=2697412
[Last accessed 15 May 2018].

GDAL. 2018. GDAL: GDAL – Geospatial Data Abstraction Library. Available at: http://www.gdal.org/ [Last
accessed 9 May 2018].

geodata4edu.ch. 2018. Project info. *geodata4edu.ch*. Available at: https://www.geodata4edu.ch/en/about-
geodata4edu-ch/project-info/ [Last accessed 9 May 2018].

GEOSS. 2018. Available at: https://www.earthobservations.org/geoss.php [Last accessed 17 May 2018].

GeoTools. 2018. GeoTools The Open Source Java GIS Toolkit – GeoTools. Available at: http://www.geotools.
org/ [Last accessed 9 May 2018].

Iosifescu-Enescu, I. 2016. Geodata Download Service Technology Overview, 20.

Iosifescu Enescu, I, Vescoukis, V, Iosifescu Enescu, CM, Müller, F, Panchaud, NH and **Hurni, L.** 2015.
Hypercube-Based Visualization Architecture for Web-Based Environmental Geospatial Information Sys-
tems. *The Cartographic Journal*, 52(2): 137–148. DOI: https://doi.org/10.1080/00087041.2015.1119469

ISNI International Agency. 2018. ISNI|How ISNI Works. Available at: http://www.isni.org/how-isni-works
[Last accessed 8 May 2018].

ISO. 2007. ISO/TS 19139:2007 – Geographic information – Metadata – XML schema implementation. Avail-
able at: https://www.iso.org/standard/32557.html [Last accessed 8 May 2018].

ISO. 2014. ISO 19115-1:2014 – Geographic information – Metadata – Part 1: Fundamentals. Available at:
https://www.iso.org/standard/53798.html [Last accessed 14 June 2018].

Jeung, H, Sarni, S, Paparrizos, I, Sathe, S, Aberer, K, Dawes, N, Papaioannou, TG and **Lehning, M.** 2010.
Effective Metadata Management in Federated Sensor Networks. In: *2010 IEEE International Conference
on Sensor Networks, Ubiquitous, and Trustworthy Computing*, 107–114. DOI: https://doi.org/10.1109/
SUTC.2010.29

Johnson, C. 2004. Top scientific visualization research problems. *IEEE Computer Graphics and Applications*,
24(4): 13–17. DOI: https://doi.org/10.1109/MCG.2004.20

**Kellenberger, B, Iosifescu Enescu, I, Nicola, R, Iosifescu Enescu, CM, Panchaud, NH, Walt, R,
Hotea, M, Piguet, A** and **Hurni, L.** 2016. The wheel of design: Assessing and refining the usability of
geoportals. *International Journal of Cartography*, 2(1): 95–112. DOI: https://doi.org/10.1080/23729
333.2016.1184552

Nakamura, Y, Cochrane, G and **Karsch-Mizrachi, I.** 2013. The International Nucleotide Sequence Database
Collaboration. *Nucleic Acids Research*, 41(D1): D21–D24. DOI: https://doi.org/10.1093/nar/gks1084

NCBI. 2018. International Nucleotide Sequence Database Collaboration. Available at: https://www.ncbi.
nlm.nih.gov/genbank/collab/ [Last accessed 14 June 2018].

NCEI. 2018. Tree Ring|National Centers for Environmental Information (NCEI) formerly known as National
Climatic Data Center (NCDC). Available at: https://www.ncdc.noaa.gov/data-access/paleoclimatology-
data/datasets/tree-ring [Last accessed 14 June 2018].

Nielsen, J. 1999. *Designing Web Usability: The Practice of Simplicity*. Thousand Oaks, CA, USA: New Riders
Publishing.

Norman, DA. 2003. *Emotional Design*. Australia Books.

NSIDC. 2018. National Snow and Ice Data Center. Available at: http://nsidc.org/ [Last accessed 14 June
2018].

OECD. 2015. Making Open Science a Reality. *OECD Science, Technology and Industry Policy Papers No. 25*.
DOI: https://doi.org/10.1787/5jrs2f963zs1-en

Open Knowledge International. 2018. Open Knowledge International. Available at: https://okfn.org/
about [Last accessed 9 May 2018].

ORCID. 2018. Structure of the ORCID Identifier – Feedback & support for ORCID. Available at: https://sup-
port.orcid.org/knowledgebase/articles/116780 [Last accessed 8 May 2018].

Ott, T and **Swiaczny, F.** 2012. *Time-Integrative Geographic Information Systems: Management and Analysis of Spatio-Temporal Data.* Springer Science & Business Media.

PANGAEA. 2018. PANGAEA. Data Publisher for Earth & Environmental Science. Available at: https://www.pangaea.de/ [Last accessed 14 June 2018].

Perrino, T, Howe, G, Sperling, A, Beardslee, W, Sandler, I, Shern, D, Pantin, H, Kaupert, S, Cano, N, Cruden, G, Bandiera, F and **Brown, CH.** 2013. Advancing Science Through Collaborative Data Sharing and Synthesis. *Perspectives on Psychological Science,* 8(4): 433–444. DOI: https://doi.org/10.1177/1745691613491579

Project Jupyter. 2015. Project Jupyter: Computational Narratives as the Engine of Collaborative Data Science. *Jupyter Blog.* Available at: https://blog.jupyter.org/project-jupyter-computational-narratives-as-the-engine-of-collaborative-data-science-2b5fb94c3c58 [Last accessed 15 May 2018].

Project Jupyter. 2018. Project Jupyter. Available at: http://www.jupyter.org [Last accessed 15 May 2018].

QGIS. 2018. QGIS as OGC Data Server. Available at: https://docs.qgis.org/2.8/en/docs/user_manual/working_with_ogc/ogc_server_support.html [Last accessed 9 May 2018].

RDA. 2018. RDA|Research Data Sharing without barriers. Available at: https://www.rd-alliance.org/ [Last accessed 14 June 2018].

ResearchGate. 2018. About us|ResearchGate, the professional network for scientists. *ResearchGate.* Available at: https://www.researchgate.net/about [Last accessed 8 May 2018].

Shen, H. 2014. Interactive notebooks: Sharing the code. *Nature News,* 515(7525): 151. DOI: https://doi.org/10.1038/515151a

Shotton, D and **Peroni, S.** 2018. The DataCite Ontology. Available at: http://purl.org/spar/datacite [Last accessed 8 May 2018].

Slocum, TA, McMaster, RB, Kessler, FC and **Howard, HH.** 2008. *Thematic Cartography and Geovisualization, 3rd Edition.* Upper Saddle River, NJ: Pearson.

Steffen Research Group. 2018. Greenland Climate Network (GC-Net). Available at: http://cires1.colorado.edu/steffen/gcnet/ [Last accessed 17 May 2018].

Swiss Confederation. 2018. About – opendata.swiss. Available at: https://opendata.swiss/en/about/ [Last accessed 9 May 2018].

Swiss Federal Institute for Forest, Snow and Landscape Research WSL. 2018a. EnviDat: the Environmental Data Portal. Available at: https://www.wsl.ch/en/about-wsl/programmes-and-initiatives/envidat.html [Last accessed 9 May 2018].

Swiss Federal Institute for Forest, Snow and Landscape Research WSL. 2018b. Long-term Forest Ecosystem Research (LWF). Available at: https://www.wsl.ch/en/forest/forest-development-and-monitoring/long-term-forest-ecosystem-research-lwf.html [Last accessed 9 May 2018].

Swiss Federal Institute for Forest, Snow and Landscape Research WSL. 2018c. Swiss National Forest Inventory NFI. Available at: https://www.wsl.ch/en/forest/forest-development-and-monitoring/nfi.html [Last accessed 9 May 2018].

Swiss National Science Foundation. 2018. SNF. Swiss National Science Foundation. Available at: http://www.snf.ch [Last accessed 14 June 2018].

The Apache Software Foundation. 2017. Apache Solr. Available at: http://lucene.apache.org/solr/ [Last accessed 9 May 2018].

The PostgreSQL Global Development Group. 2018. PostgreSQL: The world's most advanced open source database. Available at: https://www.postgresql.org/ [Last accessed 9 May 2018].

Thomson Reuters. 2018. ResearcherID.com. Available at: http://www.researcherid.com [Last accessed 8 May 2018].

UX Design Agency for Banking and Fintech. 2017. TOP 10 Most Interesting UX Design Case Studies To Inspire Your Service Reinvention In 2018. *Muzli-Design Inspiration.* Available at: https://medium.muz.li/top-10-most-interesting-ux-design-case-studies-to-inspire-your-service-reinvention-in-2018-ea2309e4104b [Last accessed 17 May 2018].

van Dam, A, Forsberg, AS, Laidlaw, DH, LaViola, JJ and **Simpson, RM.** 2000. Immersive VR for scientific visualization: A progress report. *IEEE Computer Graphics and Applications,* 20(6): 26–52. DOI: https://doi.org/10.1109/38.888006

Vierkant, P, Pampel, H, Kindling, M, Witt, M and **Scholze, F.** 2015. Re3Data.Org – Registry Of Research Data Repositories. DOI: https://doi.org/10.5281/zenodo.30782

W3C. 2014. Resource Description Framework (RDF). Available at: https://www.w3.org/RDF/ [Last accessed 8 May 2018].

W3C. 2018a. W3C HTML. Available at: https://www.w3.org/html/ [Last accessed 9 May 2018].

W3C. 2018b. Cascading Style Sheets. Available at: https://www.w3.org/Style/CSS/ [Last accessed 9 May 2018].

W3C. 2018c. JavaScript Web APIs – W3C. Available at: https://www.w3.org/standards/webdesign/script [Last accessed 9 May 2018].

W3C. 2018d. The RDF Data Cube Vocabulary. Available at: https://www.w3.org/TR/vocab-data-cube/ [Last accessed 8 May 2018].

Waterman, M, Noble, J and **Allan, G.** 2015. How Much Up-front?: A Grounded Theory of Agile Architecture. In: *Proceedings of the 37th International Conference on Software Engineering*, 1: 347–357. Piscataway, NJ, USA: IEEE Press. Available at: http://dl.acm.org/citation.cfm?id=2818754.2818799 [Last accessed 15 May 2018].

Wikipedia. 2018. Wikipedia. Available at: https://www.wikipedia.org/ [Last accessed 8 May 2018].

Wilkinson, MD, Dumontier, M, Aalbersberg, IjJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, J-W, Santos, LB da S, Bourne, PE, Bouwman, J, Brookes, AJ, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, CT, Finkers, R, Gonzalez-Beltran, A, Gray, AJG, Groth, P, Goble, C, Grethe, JS, Heringa, J, 't Hoen, PAC, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, SJ, Martone, ME, Mons, A, Packer, AL, Persson, B, Rocca-Serra, P, Roos, M, van Schaik, R, Sansone, S-A, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, MA, Thompson, M, van der Lei, J, van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J and **Mons, B.** 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. DOI: https://doi.org/10.1038/sdata.2016.18

World Economic Forum. 2018. *The Global Risks Report 2018* (No. 2018). Geneva: World Economic Forum. Available at: https://www.weforum.org/reports/the-global-risks-report-2018/ [Last accessed 14 June 2018].

WSL. 2018. DataCRediT – Contributor Roles Taxonomy for Data. Available at: https://www.wsl.ch/datacredit/#feat [Last accessed 8 May 2018].