

# EUDAT Community Engagement

Core communities and Data Pilots



EUDAT works directly with a wide range of research communities to deliver common data services to support and resolve their research data management challenges. To be successful in this ambitious initiative, EUDAT uses novel methods to involve all the stakeholders, both in the discussions to determine the required services, and in the process of designing, developing and implementing those services. These methods include involving communities in the core Research, Innovation & Development activities, known as EUDAT Core Communities, as well as collaborating with communities through specific data Pilots. This booklet gives an overview of EUDAT's 7 core communities and 24 data Pilots currently running. For more information see [www.eudat.eu](http://www.eudat.eu)



Introduction

# Earth Sciences, Energy and Environment

EUDAT has 4 core communities and 7 Data Pilots from the Earth Sciences, Energy and Environment domain.





# Core Community - ENES: European Network for Earth System Modelling



A major challenge for the climate research community is the development of comprehensive Earth system models capable of simulating natural climate variability and human-induced climate changes. These models also need to capture complex nonlinear interactions between different components of the Earth system and assess how these interactions can be perturbed as a result of human activities.

The European Network for Earth System Modeling (ENES) is developing a common climate and Earth system modeling distributed research infrastructure in Europe. This integrates the European community on Earth's climate System Models (ESMs) and their hardware, software, and data environments.

The overarching goals of ENES are to, further integrate the European climate modeling community; ease the development of full ESMs; foster the execution and exploitation of high-End simulations; support the dissemination of model results and the interaction with the climate change impact community.

## Areas of collaboration with EUDAT

The ENES Partners, several institutions including university departments, research centres, meteorological services, computer centres, and industrial partners, agreed to create ENES with the purpose of working together and cooperating towards the development and maintenance of a European network for Earth system modelling, which is synergistic with the EUDAT effort.

The ENES community, one of the current EUDAT core communities, will add to EUDAT's existing services and long-term archived data for interdisciplinary applications. ENES is also working on scalability aspects of the federation, on workflow engines and web services, data curation and preservation, authentication and policy rules as well as interfaces to data archives in a federated environment.

## Main benefits for the community

ENES is already being hit by the 'data tsunami', and this volume of data will just continue to grow. By collaborating with EUDAT and being exposed also to other scientific disciplines experiencing similar data challenges, ENES can adapt the architecture of its own federation of data servers to meet this new reality. The climate research community will also benefit from easier access to data from such other disciplines, because climate researchers, and especially those working on evaluating the impact of climate change, require data from multiple scientific fields to perform their research effectively.

For more Information on ENES see: <https://verc.enes.org/>

# Core Community - EPOS: European Plate Observing System



The European Plate Observing System (EPOS) is the integrated solid Earth Sciences research infrastructure approved by the European Strategy Forum on Research Infrastructures (ESFRI) and included in the ESFRI Roadmap in December 2008. EPOS is a long-term integration plan of national existing Research Infrastructures (RIs).



The establishment of EPOS will foster worldwide interoperability in Earth Sciences and provide services to a broad community of users. EPOS aims to be an effective coordinated European-scale monitoring facility for solid Earth dynamics taking full advantage of new e-science opportunities.

## Areas of collaboration with EUDAT

The real challenge for EPOS is to successfully coordinate- and provide access to- the data infrastructures for solid Earth Science in Europe. This requires strengthening the European capability to create high quality data, both observed and simulated, and to facilitate access to data products, completely aligned with EUDAT's overall scope.

EPOS acknowledges, with interest the developments of EUDAT since it is designing and building its own e-infrastructure, and, ultimately, EPOS can provide IT solutions that would be difficult for the solid Earth sciences community to provide on its own.

## Main benefits for the community

EPOS aims to provide all researchers of its community with basic e-science services relevant to solid Earth science, and to exploit the “core services” provided by EUDAT to build a robust e-infrastructure that uses state-of-the-art technologies for tasks as diverse as data staging and data replication, the implementation of B2ACCESS - AAI procedures and adoption of metadata and persistent identifiers. The adoption of EUDAT's IT solutions is important for EPOS as it will ensure optimum standardization across the participating sub-communities within the solid Earth sciences.

For more Information on EPOS see: <http://www.epos-eu.org/> -



# Core Community - ICOS: Integrated Carbon Observation System



ICOS (Integrated Carbon Observation System) is a new European research infrastructure (RI) with the mission to enable research to understand the greenhouse gas (GHG) budgets and perturbations in Europe and adjacent regions. ICOS is based on the collection of high-quality observational data by measurement stations operated long-term (15+ years) as national networks in the RI member states. The data is quality controlled and

processed at common Thematic Centers (TCs) by experts on Atmospheric, Ecosystem and Marine data streams.

The finalized observational data products are then distributed via the ICOS Carbon Portal (CP). In addition, various “elaborated data products”, i.e. outputs of modelling activities based on ICOS observations, will be distributed by the CP. A main role of the CP is to provide human users with services that enable them to discover, download and visualize ICOS data products. Selected ICOS data layers will also be made available for e.g. visualization at other data portals.

## Areas of collaboration with EUDAT

Given the primary goals of ICOS, immediate areas of collaboration with EUDAT are the use of trusted repositories for sensor data; PID services (at all data product levels); safe, long-term data product storage with fast access (via ICOS data portal). Also, simplification of the usage of ICOS data “at source” for modelers (high-volume users) is an area of collaboration. Finally, user authorization, authentication and identification service (beyond current “federations”) is of great relevance and potential positive impact.

## Main benefits for the community

The growing ICOS community will immediately benefit from usage of the B2DROP, B2FIND, B2SAFE, B2SHARE, and B2STAGE services. ICOS acknowledges also the fact that synergies are necessary in the area of data infrastructures in order to overcome ICOS's formidable challenges of being able to tackle them on its own.

For more Information on ICOS see: <http://www.icos-ri.eu/>

# Core Community - LTER Europe: European Long Term Ecological Research Network



Long-Term Ecosystem Research (LTER) is an essential component of world-wide efforts to better understand ecosystems. This comprises their structure, functions, and long-term response to environmental, societal and economic drivers. LTER contributes to the knowledge base informing policy and to the development of management options in response to the Grand Challenges under Global Change.



From the beginning (around 2003) the design of LTER-Europe has focussed on the integration of natural sciences and ecosystem research approaches, including the human dimension. LTER-Europe was heavily involved in conceptualizing socio-ecological research (LTSER). As well as LTER Sites, LTER-Europe features LTSER Platforms, acting as test infrastructures for a new generation of ecosystem research across European environmental and socio-economic gradients.

## Areas of collaboration with EUDAT

The uptake plan of LTER is still evolving. However, although not finished, it already contains two strategic aspects that from (Sept 2015) be addressed with EUDAT technology, which are:

- the use of B2SAFE for distributing, archiving virtual machines (including the data)
- the integration of DEIMS with B2SHARE, using B2SHARE as an deposition/archiving option from DEIMS

## Main benefits for the community

LTER's community shall directly benefit from practical usage of the EUDAT services, and, indirectly, it will also be positively impacted, as far as data infrastructure issues are concerned, from the cross-fertilisation with the other disciplines that are encompassed in the EUDAT perimeter.

For more Information on LTER Europe see: <http://www.lter-europe.net/>



# Support to scientific research on seasonal-to-decadal climate and air quality modelling

## Overview of the pilot

This pilot aims at “better simulate” climate change, at seasonal to decadal time scale and forecast air quality using both existing and locally developed models EC-Earth (global circulation model, GCM) and NMMB-BSC (air quality model). By “better simulate,” we mean making a better use of the huge amount of raw data generated by these models. That includes data transfer between the different research institutes using the data that are disseminated all over the world, but also curation, and data discovery on portals where different projects store their data.

## The scientific & technical challenge

In the latest version of the climate models, that will be used in the next Coupled Model Inter-comparison Project (CMIP6) between other projects, the resolution used has increased up to 25km in the ocean with 75 vertical levels and 40km in the atmosphere (T511/ORCA025) and the trend is to go to T1279/ORCA012, doubling the resolution. The time frequency at which the outputs are saved is also increasing and the size of the outputs consequently explodes: for example, one year of a typical experiment can occupy 1TB, knowing that in a climate experiment, hundreds of years are simulated for each experiment. Once this raw data is produced by a local institute, it needs to be shared among the whole community. We can estimate that a community of several hundreds of scientists disseminated in more than 30 research institutes around the world will use this raw data. The sharing and (multiple) transfers of such an amount of data is one of the first technical obstacles we have to cope with. The other technical challenge is how to get meaningful information from this huge amount of data for climate scientists but also downstream communities such as health (impact of climate and aerosols on health) and climate services (renewable energy industry, policy makers). Simple diagnostics such as time means or calculations of indices along the time series can become almost impossible (or at least extremely time consuming) if one needs to explore the whole dataset, retrieve the data and compute the output needed. A more technical challenge that climate scientists are facing with the increase in volume but also in data sources (satellites, observations, many instances of models) is the data discovery and indexing part. The Earth System Grid Federation (ESGF) is an example of web portal that serves this kind of data.

## Why EUDAT?

As explained before, one of the Big Data challenges in Earth Sciences is how to gather the data from all the different centres to a centralized place and then eventually back to the individual centre to do the calculations on the data. In this sense, B2DROP could be of great interest to increase the velocity of the transfers. In the case where CPU intensive computa-

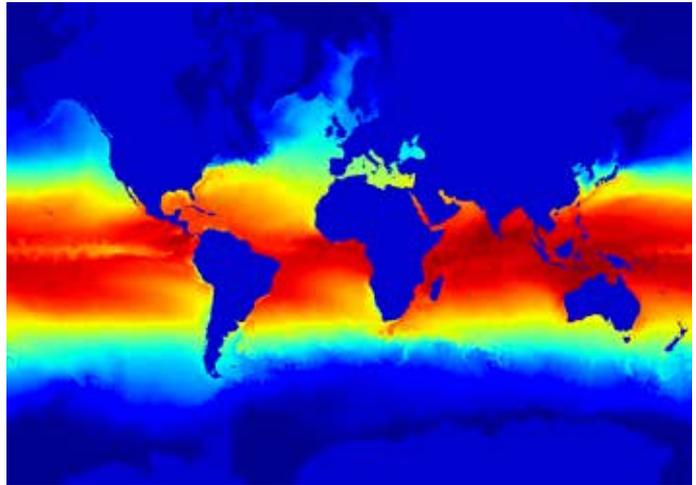


tions are required to get the diagnostics from data (this is very often the case with data set that are very big and disseminated in many files), these offline diagnostics could be done directly at the HPC where the data has been produced and

B2STAGE could be used to improve this part of the work. When all the data is centralized in a same portal (which is the case of the ESGF), people need to explore easily the metadata and how the files are organized to be able to know what is there before using the data (to find, for example the temperature at 2 meters for all models initialized in a given year). This is where B2FIND can be useful.

## Expected outcomes

The first benefit that is expected from the use of the EUDAT tools, if the project is successful, is a benefit in time: making the transfers and the calculations faster would of course be beneficial but would also allow scientists to do things that were unthinkable because of operations duration they could not afford (or tools that didn't allow the operations). From the point of view of our data, the success of the pilot would clearly improve their visibility and, in addition to the technical improvements in terms of research on data, it would allow the scientists to perform more advanced scientific diagnostics on the data, improving the research.



## Expected domain legacy

In the Earth Sciences and particularly the climate and air quality modelling communities, open data and open access to model developments are key. Therefore, it is almost impossible to distinguish between legacy for our scientific domain in general and expected outcomes for us, given that the developments brought by our pilot would ideally benefit as much to us as to the rest of the community and “our data” (produced at the centre), will be accessible to all the community. The same idea also applies to the software developments done within the EUDAT data pilot.



# DataPublication@UPorto

## Overview of the pilot

The DataPublication@UPorto pilot gathers experiments where Dendro, a prototype Research Data Management platform, is used as a gateway to EUDAT. Dendro provides an ontology-based environment for dataset description and publication for the long tail of research. It is built as a multi-disciplinary platform and its preliminary evaluation was carried out with a panel of research groups from the University of Porto. In the scope of the pilot, researchers from several domains within the University of Porto will be asked to follow the steps of a prescribed workflow and organize, describe and deposit datasets created in the scope of their projects.

## The scientific & technical challenge

The main scientific challenge in the RDM research line where DataPublication@UPorto fits is the definition of diverse metadata models and their joint use in the Dendro platform. This has led to the use of recommendation techniques in Dendro, to help users in each domain pick the appropriate descriptors for their data.

The second challenge concerns the data management workflows. We intend to build on previous small-scale experiments covering the definition of metadata models, and their use in Dendro to describe datasets, expanding the pilot to a larger multi-domain community.

The main technical challenge in the DataPublication@UPorto pilot is the use of EUDAT as a long-term repository for the University of Porto. Besides this, the pilot will also consider the data staging services of EUDAT and assess their features, in order to compare them with those already available in Dendro. Given the diversity of research domains in the pilot, we expect that this will result in some solutions being more appropriate for some research groups than others. The extension of the panel will provide more evidence of the effectiveness of the Dendro platform, while in other cases an all-EUDAT solution may prove more effective. Another possibility to be considered is a hybrid approach, where Dendro is used in the early stages of RDM, providing a data storage, description and deposit environment to researchers, similarly to what B2Drop and B2Share already do, but long-term deposit and retrieval will be handled by the EUDAT platform.

Our platform has so far been tested with a panel of 11 research groups, which we expect to extend to 50 groups during the pilot.

## Why EUDAT?

The DataPublication@UPorto pilot explores the B2Share service of EUDAT to store and provide access to data generated by research groups in several domains in the University of Porto. The pilot will test B2Share on two perspectives: technical and operational. On the technical part, the most relevant aspects are the features of the API, the interoperability with external systems (namely Dendro) and the robustness of the storage infrastructure. On



the operational part, B2Share will be assessed with respect to the features provided to the managers at the University services and to the end users.

As a second line, the pilot also intends to explore the B2DROP service and compare it to Dendro, the in-house platform for data organisation and description. This comparison takes into account the requirements of the different stakeholders from a variety of research domains.

## Expected outcomes

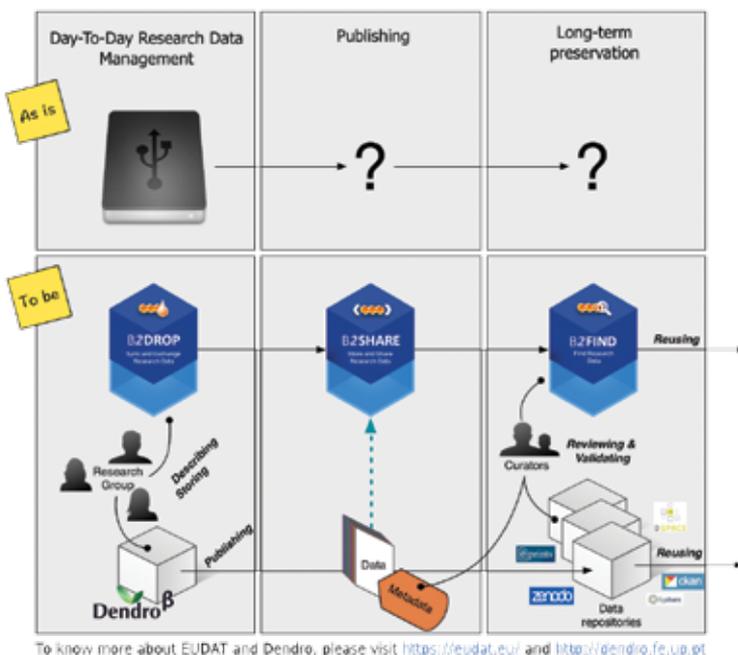
The DataPublication@UPorto pilot will close the loop on the RDM activities at the University of Porto, creating the conditions for the long-term deposit of datasets which have already been collected and for their re-use in the corresponding communities.

The use of an international platform such as EUDAT is compared with the local solution consisting of a University-wide repository. The catalogue of EUDAT services, and the existence of other data pilots, is used to show researchers the expected results of their work, and hopefully engage them in the full cycle of RDM actions.

Upon completion of the pilot, the University will have a collection of use cases which can be used to showcase salient research projects and groups.

## Expected domain legacy

The DataPublication@UPorto pilot runs in the context of a generic RDM service for the University of Porto, not as a disciplinary endeavour. For the domains considered in the participating research groups, the investment in the pilot is expected to kick-start the RDM efforts in these areas. The success in the pilot also means that RDM for the long tail can be handled with generic tools combined with domain-dependent metadata models, which can evolve based on their use by researchers.





# Unified Access to EISCAT radar data

## Overview of the pilot

The European Incoherent Scatter Scientific Association (EISCAT) operates three incoherent scatter radars, instruments probing the Earth's ionosphere. Two are located in northern Fennoscandia and one on Svalbard. EISCAT are now developing the next generation radar, EISCAT\_3D, consisting of antenna arrays in northern Norway, Sweden and Finland.

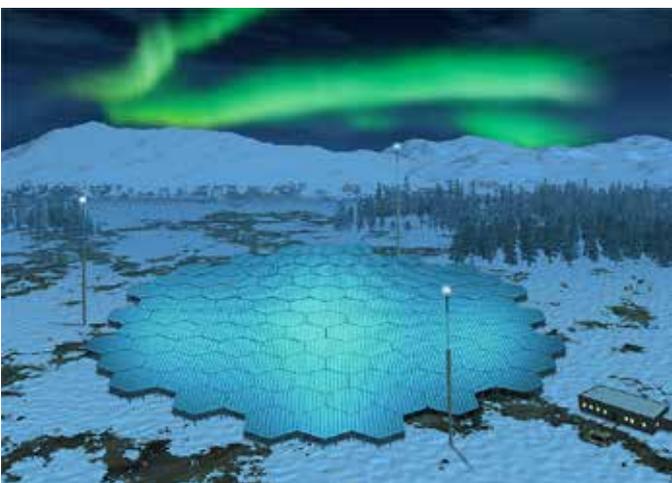
The purpose of this data pilot is to use EUDAT services to establish a unified archival and data search system for the existing EISCAT incoherent scatter radars. The outcome will be used to explore whether and how EUDAT services can be customised for data archival and discovery for the future EISCAT\_3D radar system.

## The scientific & technical challenge

Accessible EISCAT data are divided into levels. EISCAT\_3D data will be similar but data volumes will be considerably larger due to the volumetric data at high bandwidths. The data rates and volumes are expected to be in the order of magnitude as other large scientific experiments such as the LHC.

The present archives of EISCAT data at level 2 and 3 are completely separate and use different systems for access. The proposed pilot is intended to unify the access to data at these two levels. The project can be divided into the following tasks:

- (1) Archive, index and stage data.
- (2) Data discovery and search. Data at levels 2 and 3 will have to be connected to each other. Several versions of level 3 data corresponding to the same level 2 data may exist, depending on analysis methods and parameters of the analysis algorithms.
- (3) Data visualization. E.g., browse level 3 data visually for occurrence of aurora, and download data from these events. EISCAT\_3D will also require volume rendering of level 3 data in four dimensions (time development of parameters in a volume) or seven dimensions (time development of a velocity vector field).
- (4) Access control and user authentication. Different access rules apply to EISCAT data at different levels. There is currently no fine-grained access control in the EISCAT data access systems. An authentication system must be implemented in order to grant access to EISCAT users following the data policy, regardless of their geographical location at the moment of data download.





## Why EUDAT?

A system controlled by the EISCAT community sends a level 2 dataset to B2SAFE. The dataset is registered and replicated according to the data management policy of the EISCAT community. B2SAFE extracts and associates technical metadata from the level 2 dataset with the dataset making it harvestable by B2FIND.

In B2FIND, a researcher queries datasets by giving a set of conditions. B2FIND returns a list of datasets fulfilling the search conditions. The researcher assesses the list and selects a number of datasets for further analysis. The researcher assesses the datasets by reviewing the dataset description.

The entitled researcher can reuse the dataset. The researcher stages data from B2SAFE to a computing cluster through the use of PID shown by B2FIND. After the datasets has been processed and analysed, the researcher sends a resulting level 3 dataset to B2SHARE. Alternatively, the resulting level 3 dataset can be sent to B2SHARE by an automatic process through the B2SHARE API.

## Expected outcomes

With the development of a functional archive for the EISCAT data, the EUDAT pilot will make a foundation for new discoveries and significant scientific breakthroughs.

The system will be robust and allow refinements and further developments of the access of data. Important is also the training of the users, with valuable feedback, making the updated system ready for wider use. The system is also expected to lay a foundation for the development of a data archive for EISCAT\_3D.

## Expected domain legacy

The design of the next generation incoherent scatter radar system, EISCAT\_3D, opens up opportunities for physicists to explore many new research fields. On the other hand, it also introduces significant challenges in handling large-scale experimental data which will be massively generated at great speeds and volumes. This challenge is typically referred to as a big data problem and requires solutions from beyond the capabilities of conventional database technologies.

The overall ambition is to provide the users of incoherent scatter radar with tools that improves opportunities for scientific discovery. This competence centre is also important for the build-up towards EISCAT\_3D and the tools developed will form a base for further development.



# DATA SPHINX (DATA Storage and Preservation of High resolution climate eXperiments)

## Overview of the pilot

This pilot will allow long-term storage and sharing among a wide scientific user community of high-resolution climate model output data. It aims at building a repository serving the climate change impact modelling community, providing selected variables at high temporal and spatial resolution, with a focus on climate extremes and the hydrological cycle in areas with complex orography. Potential users include researchers studying the impacts of climate on ecosystems, floods, landslides, fires. The archive will contain high-resolution data from the PRACE project Climate SPHINX and will later be extended with simulations from the projects PRIMAVERA, CRESCENDO and HighResMIP.

## The scientific & technical challenge

An open issue which is currently being actively investigated is the sensitivity of climate simulations to model resolution and determining if very high resolution is useful for a realistic representation of the main features of climate variability. Also the advantage of sub-grid parameterizations capable of capturing small-scale variability, such as stochastic parameterizations, has to be determined. To this end extremely high resolution climate integrations are necessary and they are being performed or planned in the framework of several initiatives (Climate SPHINX, HighResMIP, CRESCENDO, PRIMAVERA).

In a first stage the EC-Earth Earth-System model is being used to explore the impact of Stochastic Physics in long climate integrations as a function both of model resolution (from 80km to 16km for the atmosphere). This research will for the first time investigate extensively and systematically the impact of resolution and stochastic parameterisations for climate simulations. As a result, we estimate data storage needs around 50-300 TB in this first stage.

In a second stage, the archive will be further expanded with high-resolution coupled simulations performed mainly with the EC-Earth model in the framework of the CMIP6 High-ResMIP initiative and of the PRIMAVERA and CRESCENDO H2020 projects. For this second phase we estimate storage needs around 300-700 TB.

Technical issues to be solved include the implementation appropriate tools for the distributing and searching the data, for post-processing and data extraction and for comparing them with available observations from other archives. To this end the integration of standard tools from the climate research community (such as ESGF nodes) will be explored.

This pilot will be used to demonstrate the integration of existing solutions, still under development, with relevant EUDAT services. The size of the potential user base can be estimated as hundreds of scientists in the climate change and climate impact fields.



## Why EUDAT?

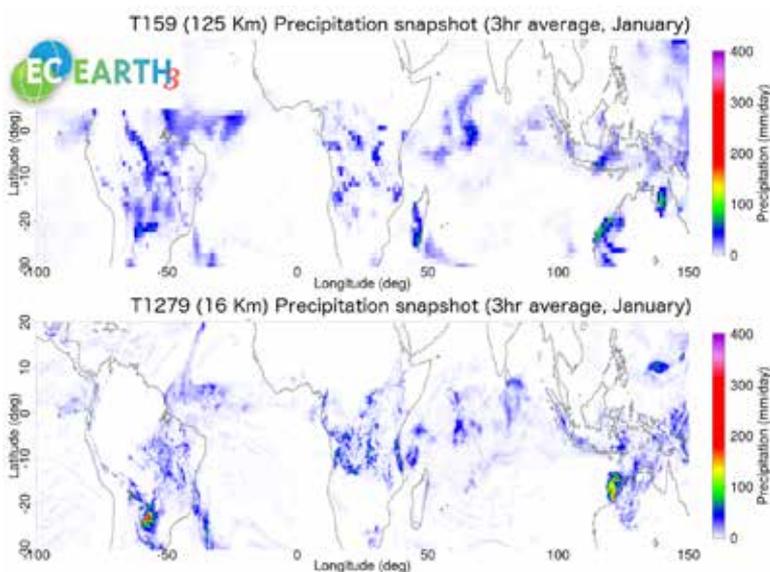
The pilot will expose stored data using an ESGF (Earth Science Grid Federation) node and a Thredds Data Server, deployed using the EUDAT “Service Hosting Framework”. It will explore how to expose the ESGF instance through B2FIND for improving data discoverability. We will evaluate the possibility to register the data sets either through the DOI or the PID. The use of B2SHARE as catalogue where to store meta-data records only will be evaluated. Specific EUDAT services involved include data repository, (long tail) data sharing and data staging for analysis and processing.

## Expected outcomes

The data repository, data sharing and staging services offered by the pilot will be crucial to allow a wide user base to have access to a set of climate variables at high temporal resolution and at extremely high spatial resolutions, not commonly available at this time. These services will represent one important source of very high resolution simulation data in preparation for following international efforts (such as HighResMIP and current and future H2020 projects), to perform preliminary studies following the work programme of these projects and to develop further data analysis, diagnostic and visualization tools.

## Expected domain legacy

The pilot will provide a platform for medium term storage and to facilitate the access and discovery of state-of-the-art high-resolution climate simulations. The EUDAT services will be used to allow easy and fast access, sharing and analysing efficiently selected variables from extremely high resolution datasets (particularly storage intensive), with a particular focus on climate extremes and the hydrological cycle. This will facilitate scientific collaboration and will foster research facilitating data analysis and post-processing. The services offered by this pilot will be made available to participants of different climate research communities or participants in national and international research projects.





# Public access to fine-grained city air quality data from roving sensors

## Overview of the pilot

The project LIFE+RESPIRA has a network of 50 air pollution sensors carried around by volunteer cyclists during their urban commutes within Pamplona, Spain, a fairly average European city. Contaminant gasses and particles are recorded at very fine spatial and temporal resolution, and transmitted in near-real time for processing. A huge volume of data is being produced and, after heavy processing, serves to feed an air quality model allowing prediction of best routes for city dwellers. The pilot wants to ensure that citizens and researchers alike can fully access the pre- and post-processed data for any scientific, social, or policy purpose.

## The scientific & technical challenge

Our sensor suites (up to 50) are reading each up to 10 environmental, geo-located, multi-data parameters at a rate of 5 Hz. Despite heavy internal processing and averaging, we are still producing cumulative, stored data at a high rate. Although these have a limited life for any immediate purpose (e.g. what is NOW the level of this contaminant HERE), air quality models are extremely sensitive to many variables: time, weather, climate, urban structure, winds, etc.; only a large, distributed, nearly-continuous dataset can account for the parametrization of the models—that is, ALL “past” data are useful to understand how air quality is, was, and will be under current and future conditions. Thus, we need to build and store a multi-million-record dataset at a raw resolution better than 10 meters and 10 seconds.

These data may prove invaluable to analyse how air pollution evolves in a city—not only as an overall parameter, but at a human scale. The dataset could thus be used to build models that go beyond the statistical average for an area, down to what the individual can experience during his or her daily walk or ride. Corrective measures could be applied when and where they matter most.

Research that we haven’t even figured could be undertaken on the data, and we want to ensure that that research is possible. Within the LIFE+RESPIRA consortium there are several research subjects that need to filter, select, and group the data according to specific needs—and therefore the project will be the main user of the data at first. But at a larger scale, we want these data to be made available to all: other scientists,

officers, technicians, policy makers, and ordinary citizens that may also require selecting and combining the data as they see fit.





## Why EUDAT?

We may well be using all five EUDAT services within the pilot although under different intensities. Within the project, B2DROP will substitute, at an advantage, our current sharing of the main repository so all teams access the latest updates, while either B2SHARE or B2SAFE, according to the scale of the dataset, would be preferred to store a copy of both the raw data and the results to be disseminated among the interested public: researchers and citizen researchers wanting to access the raw or processed data. Thus, either B2SHARE or B2SAFE would act as final, permanent, post-project repositories of the project's data, and could become the primary repositories for any project extension.

Although we envisage selection and downloading of data through common database-interfaced tools (especially within the project), we would still document the dataset finely enough to allow B2FIND to select processed data and products by other scientists outside our team.

At this moment our workflow does not yet need staging to large supercomputing facilities and therefore we do not think we'll be using B2STAGE—but this may eventually become a need, so within the pilot we plan to explore B2STAGE for eventual integration within our workflow.

## Expected outcomes

We envisage EUDAT contribution as a solution to two problems. Firstly, to ensure that our teams have ready access to all data as they are produced, in a way that does not depend from, or detract of, limited resources available to the project or from external services not specifically tailored for scientific research. Second, we believe that EUDAT would be invaluable in ensuring: (a) the permanent availability of all collected and processed data after the project ends, by releasing them to an external, permanent facility; and (b) to be able to retrieve the data through a common access point for all interested parties. Researchers working on atmospheric sciences or urban planners, to name a few, could then tap on the raw and modelled data and apply the shared knowledge to their own projects.

## Expected domain legacy

Our “sharing model” could be used as a “model to share” within the urban air quality community. If we can successfully release our raw and processed datasets to the community, documenting it in a way that facilitates retrieval of relevant data (e.g., making it possible to avoid having to download large sections of mostly unnecessary data that need to be filtered afterwards), we could provide the urban air quality community with a source of data for additional studies that we haven't even yet thought of. We also hope our participation in EUDAT could act as a catalyst to promote open access to shared environmental data. Finally, we want to stress that the goal of the project is to improve the quality of life of citizens—therefore, EUDAT would be linked to the outreach of current science to the general public.



# JADDS - Jülich Atmospheric Data Distribution Service

## Overview of the pilot

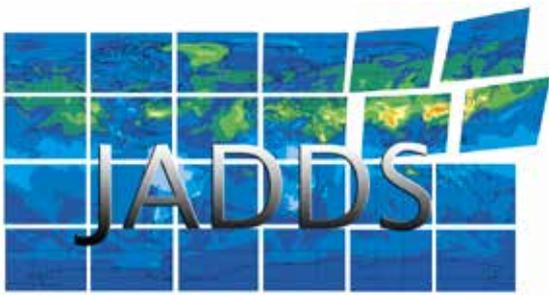
Global model data of atmospheric composition produced by the Copernicus Atmospheric Monitoring Service (CAMS) is collected since 2010 and serves as boundary condition for use by regional air quality modellers world-wide. An existing Web Coverage Service (WCS) for sharing these individually tailored model results shall be re-engineered to make use of a modern, scalable database technology in order to improve performance, enhance flexibility, and allow the operation of catalogue services. The WCS protocol shall be upgraded to WCS2.0 and the metadata shall be interfaced with the EUDAT service structure. In effect the current self-written WCS software package shall be replaced by a modernized and more efficient out-of-the-box solution.

## The scientific & technical challenge

The Jülich Atmospheric Data Distribution Service (JADDS) is aimed primarily at regional atmospheric air quality modelling groups from all over the world. Regional Air Quality (RAQ) models need time-resolved meteorological as well as chemical lateral boundary conditions for their individual model domains. While the meteorological data usually come from well-established global forecast systems, the chemical boundary conditions are not always well defined. In the past, many models used 'climatic' boundary conditions for the tracer concentrations, which can lead to significant concentration biases, particularly for tracers with longer lifetimes which can be transported over long distances (e.g. over the whole northern hemisphere) with the mean wind. The Copernicus approach utilizes extensive near-real time data assimilation of atmospheric composition data observed from space which gives additional reliability to the global modelling data and is well received by the RAQ communities.

The Jülich server adheres to the Web Coverage Service WCS standard as defined by the Open Geospatial Consortium OGC. This enables the user groups to flexibly define datasets they need by selecting a subset of chemical species or restricting geographical boundaries or the length of the time series. The data is made available in the form of different catalogues stored locally on our server. In addition, the Jülich OWS Interface (JOIN) provides interoperable web services allowing for easy download and visualization of datasets delivered from WCS servers via the internet.

So far, the WCS server has been hosted on a local workstation in the atmospheric sciences institute and it is based on a deprecated WCS version (1.1.2). The performance of the service is limited by the server hardware and by the file-based data storage. The outdated WCS version prevents automatic harvesting of metadata for web catalogue services such as those offered by B2FIND.



## Why EUDAT?

We aim to better connect the CAMS boundary condition service with EUDAT concepts and services and to implement the following improvements: 1) the server shall migrate to the central storage site of the Jülich Supercomputing Centre (JSC) at FZ Jülich, where the ex-

pected growing data amount (about 25 TB/yr) can be optimally handled and made available for fast data access by external users, 2) the WCS protocol shall be upgraded to WCS2.0 and adapt the EUDAT service structure, and 3) the current self-written WCS software package shall be replaced by a modernized and more efficient out of the box solution in order to improve efficiency and connectivity.

The WCS 2.0 service shall be harvested by the EUDAT B2FIND service. Selected data products that are frequently requested shall be stored in B2SHARE and made available in the WCS infrastructure and through the JOIN web interface. In addition users of JOIN have the opportunity to store their data selections with a PID in B2SHARE for referencing them in publications.

## Expected outcomes

Currently, there are about a dozen users who regularly access data from the Jülich server and many more who occasionally browse the data or download specific parts for their analysis. With operational data delivery in CAMS the number of users is expected to grow, if easy and fast data extraction can be guaranteed. This will further broaden the acceptance of CAMS data products and its use in the RAQ communities. As additional value CAMS data stored on the modernized JADDS server are available much longer than at the originating centre ECMWF and can thus be used for scientific analyses such as interpretation of field campaign data or model inter-comparison projects. Through the web interface Jülich OWS Interface (JOIN; <https://join.fz-juelich.de>) they can also be interactively visualized and compared to observational data. While the CAMS data are the focus of this project, JADDS can later be expanded to distribute other datasets with similar properties, such as meteorological reanalyses or satellite data products.

## Expected domain legacy

The envisaged JADDS data distribution server will help to strengthen the links between global and regional modelling communities, particularly for less developed countries where RAQ modelling is often lacking reliable and easy to use boundary conditions. Moreover, it will help to further disseminate CAMS products for atmospheric composition and thus foster collaboration on air quality monitoring in and beyond Europe.





# Working towards an EUDAT Linked Data Service

## Overview of the pilot

The EUDAT semantic annotation service aims to look at the technical options for providing a linked data service to EUDAT participants and stakeholders. The pilot will build on a previous data pilot b2note that was looking at providing

linked data of metadata mobilised by EUDAT. The EUDAT semantic annotation service extends the integration of metadata to select data use cases from the Long-Term Ecological Research (LTER) community



## The scientific & technical challenge

Within the LTER community a lot of data providers expose their data as simply structured spreadsheets or CSV files. To facilitate correct reuse of those data it is necessary to annotate them with concepts from controlled vocabularies which describe the meaning of data, their provenance, their features of interest and further details.

Services providing information on the meaning and disambiguation of the data (like LD services) will facilitate reuse of data. The output of this data pilot project will be of interest for other communities and could be further developed as an EUDAT service.

## Why EUDAT?

Within EUDAT first prototypes to annotate the metadata have been done. This data project shall extend those prototypes with the possibility to annotate the data.

This pilot project complements core current core EUDAT services. In order to do so the data have to be deconstructed to their elements so that elements can undergo a first automated analysis for found keywords which are equal or similar to concepts in the controlled vocabulary used. The data provider, of course has to review the results of such an automated process, but once he/she has done so, data are ready for exposure via data services.

Those services will be compliant to existing standards (e.g. OGC WFS, SOS and/or W3C linkedData service and geoSPARQL endpoints).

## Expected outcomes

The expected outcome will be a case study for implementing linked data provision from LTER Data. Linked data principles and flexibility can apply too other communities served by EUDAT but also across these communities



## Expected domain legacy

The pilot results will be used to inform future developments of an LTER Research Infrastructure. Using EUDAT will increase impact of the pilot and applicability for other communities.



