EUDAT

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

Foreword

One of EUDAT's main ambitions is to bridge the gap between research infrastructures and e-Infrastructures through an active engagement strategy. EUDAT is privileged to have close and collaborative relations with its long-standing partner communities. At present, there are seven core communities – CLARIN, ELIXIR, ENES, EPOS, ICOS, LTER Europe and VPH in the consortium covering the Social Sciences & Humanities, Earth sciences, energy and environment, Biomedical and life sciences disciplines. They have helped us design and build our services and were instrumental in making two successful EC project proposals.

But we should see also that consortia with a fixed number of participants could easily become complacent with respect to the range of services offered and the research communities served. It is therefore essential that EUDAT regularly expands its horizons and renews and adds to the list of communities it works with. EUDAT invests in finding new communities and discusses their ideas and needs for research data management and supports them to solve their problems.

In this respect EUDAT is very satisfied that the last call for Data Pilots was very successful in bringing in 24 pilots from a large variety of disciplines and organizations. We were happy to accept all and are currently working together with them to define their implementation.

Originally we expected something around half that number, and although it will mean hard work for all involved, these pilots offer a unique chance to boost our involvement in new communities and projects and particularly to get input from new groups of researchers.

I hope the Data Pilots will find us open and sympathetic to their needs, and that we can work together on making EUDAT grow.

Finally we should also see such a large number of proposals as a reward and encouragement for our outreach and PR efforts that our image and reputation as a community friendly data management provider is successful. But let's not get complacent!

Daan Broeder, EUDAT Community Manager, Meertens Instituut & CLARIN-ERIC

Contents

Foreword 1
Contents 2
Introduction 4
Earth Sciences, Energy and Environment 4
- Core Community - ENES: European Network for Earth System Modelling 6
- Core Community - EPOS: European Plate Observing System 7
- Core Community -I COS: Integrated Carbon Observation System 8
- Core Community - LTER Europe: European Long Term Ecological Research Network 9
- Support to scientific research on seasonal-to-decadal climate and air quality modelling 10
- DataPublication@UPorto 12
- Unified Access to EISCAT radar data 14
- DATA SPHINX (DATA Storage and Preservation of HIgh resolutioN climate eXperiments) 16
- Public access to fine-grained city air quality data from roving sensors 18
- JADDS - Jülich Atmospheric Data Distribution Service 20
- Working towards an EUDAT Linked Data Service 22

Biomedical and Life Sciences 23
- Core Community – ELIXIR: A distributed infrastructure for life-science information 26
- Core Community - VPH: Virtual Physiological Human 27
- West Life Data Pilot 29
- IST DataRep 30
- Herbadrop 32
- The use of the EUDAT repository to store clinical trials in a secure and compliant way 34
- An EUDAT-based FAIR Data Approach for Data Interoperability 36

Physical Sciences and Engineering 38
- Tokamak data mirror for JET and MAST data – moving towards an open data repository for European nuclear fusion research 40
- Turbase DNS 42
- NFFA-EUROPE Information and Data Management Repository Platform for nano-science in Europe 44
- Direct simulation data of turbulent flows 46
- SIMCODE-DS 48

Social Sciences and Humanities 50
- Core Community - CLARIN 52
- Research data repository for students' own results 54
- Enriching Europeana Newspapers 56

Cloudy Culture: A study of EUDAT shared services to measure the potential of using cloud-like s ervices to improve the preservation of digital cultural heritage 58
Aalto data repository 60
Ancient OCR: Storing, Cataloguing, Relating, and Exposing OCR Objects from the Open Philology Project 62

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-Đend 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.

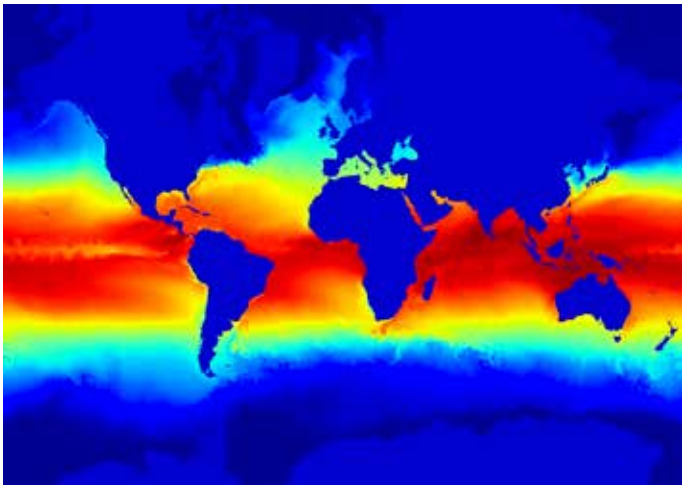
Why EUDAT?

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

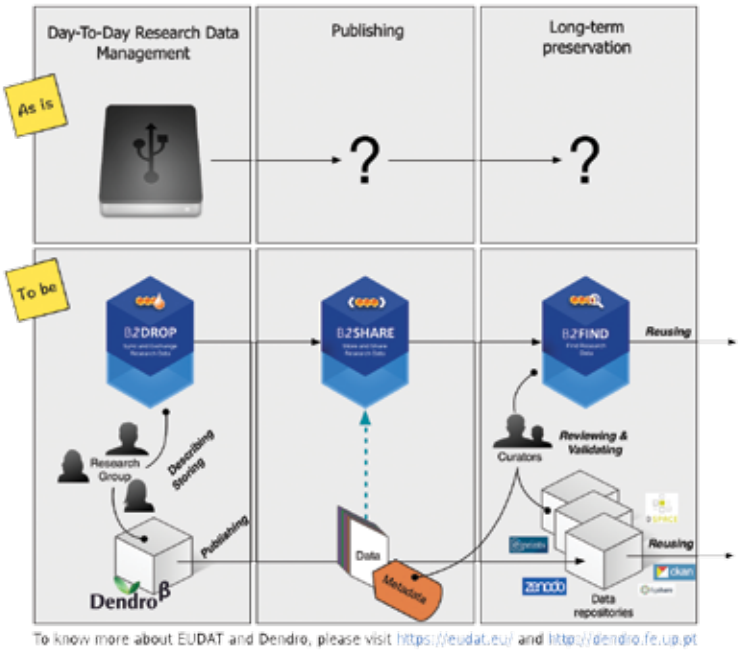
The DataPublication@UPorto pilot will close the loop on the RDM activities at the University of Porto, creating the conditions for the long-term deposit of datasets which have already been collected and for their re-use in the corresponding communities.

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

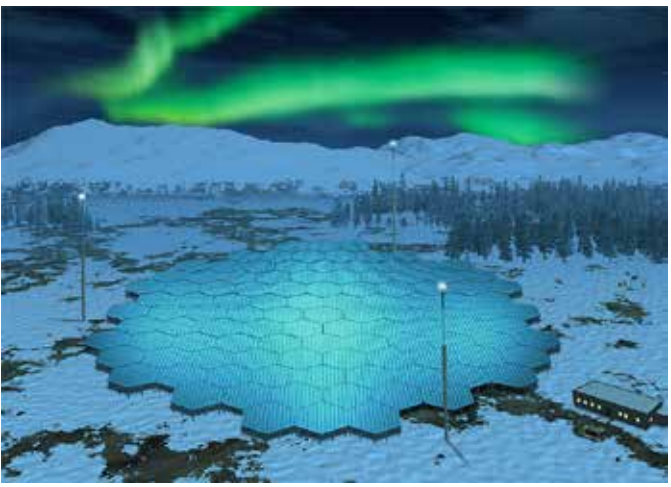
Accessible EISCAT data are divided into levels. EISCAT_3D data will be similar but data volumes will be considerably larger due to the volumetric data at high bandwidths. The data rates and volumes are expected to be in the order of magnitude as other large scientific experiments such as the LHC.

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 researches 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 HighResMIP 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?

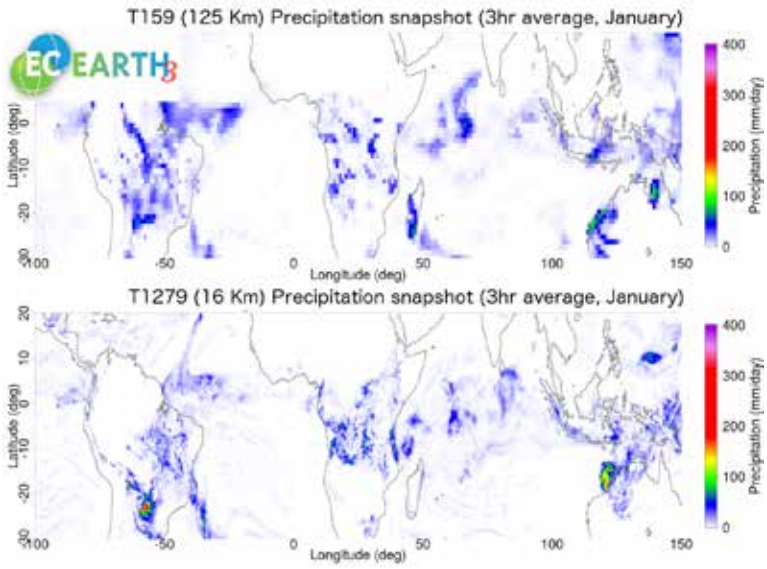
The pilot will expose stored data using an ESGF (Earth Science Grid Federation) node and a Thredds Data Server, deployed using the EUDAT "Service Hosting Framework". It will explore how to expose the ESGF instance through B2FIND for improving data discoverability. We will evaluate the possibility to 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, multiple-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

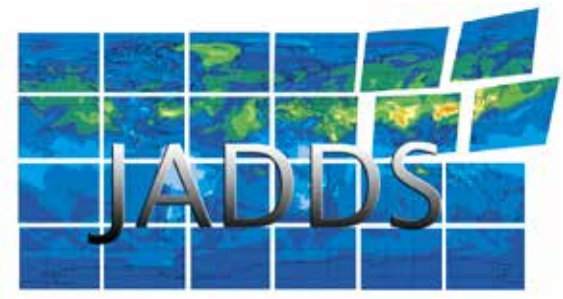
Global model data of atmospheric composition produced by the Copernicus Atmospheric Monitoring Service (CAMS) is collected since 2010 and serves as boundary condition for use by regional air quality modellers world-wide. An existing Web Coverage Service (WCS) for sharing these individually tailored model 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 expected 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.

Biomedical and Life Sciences

EUDAT has 2 core communities and 5 Data Pilots from the Biomedical and Life Sciences domain.

Core Community – ELIXIR: A distributed infrastructure for life-science information

ELIXIR builds a sustainable pan-European infrastructure for biological information, supporting life science research and its translation to medicine, agriculture, bioindustries and society.

ELIXIR unites Europe's leading life science organisations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research.

ELIXIR will provide the facilities necessary for life science researchers - from bench biologists to cheminformaticians - to make the most of a rapidly growing store of information about living systems, which is the foundation on which the understanding of life is built.

Areas of collaboration with EUDAT

ELIXIR is one of the core communities of EUDAT and actively contributes to providing end-user driving feedback to it as the goal of ELIXIR is to orchestrate the collection, quality control and archiving of large amounts of biological data produced by life science experiments.

Collaboration with EUDAT ensures that ELIXIR needs are taken into account in Europe's expanding HPC landscape, such as GEANT, the European Grid Infrastructure (EGI).

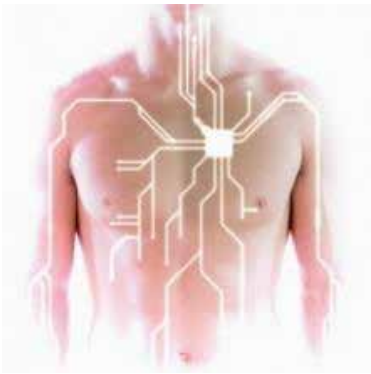
Main benefits for the community

Researchers targeted by ELIXIR will directly benefit from the collaboration. All the EUDAT's services can be of immediate and practical support to the daily work of hundreds of researchers across Europe and elsewhere.

For more information on ELIXIR see: https://www.elixir-europe.org/

Core Community - VPH: Virtual Physiological Human

The Virtual Physiological Human (VPH) project aims to provide digital representations of the entire human body, referred to as virtual humans. Instead of focusing on finding a specific cure for a specific disease, which is what currently happens in clinics, the VPH approach is to treat individual patients rather than treating diseases. The virtual humans are based on data collected from real patients, including biological, imaging, clinical and genomic data. As the data is unique to each patient, it will enable academic, clinical and industrial researchers to improve their understanding of human physiology and pathology, to derive predictive hypotheses and simulations, and to develop and test new therapies. The eventual outcome will be better disease diagnosis and treatment, along with improved prevention tools in healthcare.



Areas of collaboration with EUDAT

One of the major challenges faced by the VPH initiative is handling patient data and data generated from simulating patient reaction to certain treatments. Modelling, storing, sharing and processing large volumes of data, and the visualization of results, will play a central role in achieving VPH objectives, which clearly opens for relevant areas of collaboration with the EUDAT initiative.

Main benefits for the community

The VPH Research Community will be able to build on the generic data services provided by EUDAT to create rich, community-specific analysis platforms. The fact that many EUDAT partners are also large HPC centres participating in PRACE will make it easy for VPH researchers to co-locate their data with high performance computing resources.

For more information on VPH see: http://vph-portal.eu/

West Life Data Pilot

Overview of the pilot

West-Life will provide a VRE for structural biologists across Europe. Users will range from PhD students to professors. The raw data will be acquired at experimental facilities, and then a series of processing steps will create new data files, leading to the final PDB file. Larger experimental facilities already have arrangements for storing data, and this is the only possible approach where the technique produces large amounts of data. Smaller facilities will benefit from being able to use EUDAT services.

The scientific & technical challenge

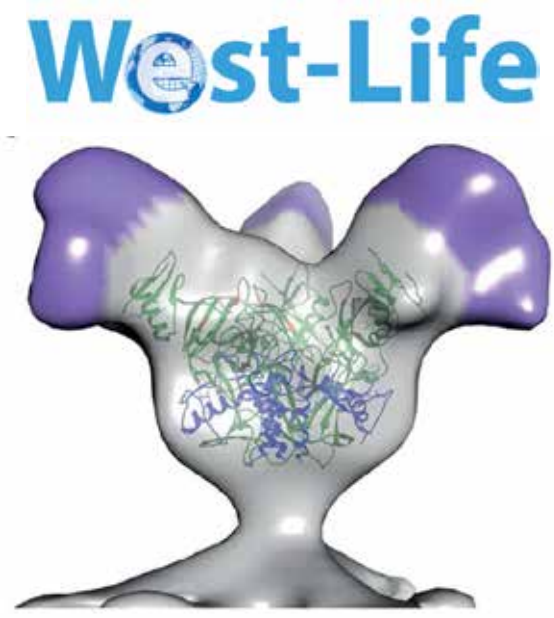
The use community consists of a few thousand scientists. Structural biologists used to identify themselves by their preferred technique (as "crystallographers", "electron microscopists" etc). Increasingly, they are targeting larger macromolecular complexes, so research projects now must combine several techniques, so data management and processing are becoming more complex.

There is a lot of value in being able to store metadata about provenance of data ("this file was created by processing those files, using that program with these keywords"). The standard ontology PROV-O will express most of what we need.

Expected outcomes

The benefit to facility providers is an interoperable way of providing support to users' data management needs, organised to conform to the new expectations of H2020, so that "doing the right thing" in sharing data becomes the path of least resistance.

The benefit to the end user is a seamless project overview of data and processing performed at facilities across Europe, by different members of a research collaboration.



Expected domain legacy

A future benefit of this VRE will be that it will create a context in which new processing pipelines can be developed and deployed. In particular, there are few algorithms that can balance evidence obtained by different experimental techniques to determine a consensus structure of a macromolecular complex. The provision and use of this VRE will pose the challenge of developing such algorithms in future.

IST DataRep

Overview of the pilot

IST DataRep is the institutional repository for publishing research output of IST Austria affiliates. IST DataRep was implemented to help scientists fulfil the requirements from funding bodies and to meet the growing impact of publishing research data. Therefore, the deposited data collections will be mainly open access.

The scientific & technical challenge

The repository is mainly designed for the demands of data publication. This was the main aspect we were focusing on regarding the data life cycle. Therefore each data collection is assigned a DOI to grant it's cite ability. But a DOI doesn't only enable citation it also facilitates persistence, which asks for longevity of the data collection. IST Austria has an internal back up strategy running but a truly safe is only guaranteed with offsite data storage.

Scientists at IST Austria are encouraged to deposit data at established subject repositories (I.e. Dryad, Gene Bank). For those domains – the long tail of science – which are not provided with internationally known and used subject repositories our institutional repository was designed for. IST DataRep is the institutional repository for a small scientific operation and even though the content is publicly accessible it needs to be indexed in international platforms/search engines to obtain sufficient visibility.

B2Safe and B2Find are planned to be additional services to guarantee long time archiving and visibility. Therefore the technical preconditions have to be fulfilled. On the one hand this is the capability of generating bundles (data collections + metadata) via a REST API and develop a workflow and technical features for the transfer to the EUDAT B2Safe service.

On the other hand the metadata has to be collected and indexed by EUDAT. Regarding B2Find we assume that the implementation of the service won't need any technical development because IST DataRep is an OAI-PMH compliant repository.

There are no indications in terms of the size of the potential audience to base on a predictive usage. Most likely it will be almost solely the respective scientific community.

Why EUDAT?

Most of this is answered in the previous section however one additional reason is that EUDAT offers various services, which could be of interest for the institute in the future.

Expected outcomes

Regarding the benefits for the end user, there is the ability to differentiate between depositing and accessing/reusing data. The obvious benefits are simultaneously the challenges already mentioned: Long-term archiving and visibility.

For users an additional benefit will be that they can search many repositories on an international level simultaneously, which increases the chance to find what you are looking for and/or to find more of what you are looking for.

Expected domain legacy

The main benefit of using EUDAT services is to guarantee the scientific inheritage of IST Austria.

Furthermore IST Austria is a multidisciplinary and interdisciplinary acting institution. The EUDAT services may comply with this approach and therefore constitute a sufficient service for all disciplines and cross-disciplines at our institute.

Herbadrop

Overview of the pilot

Herbadrop is both an archival service for long-term preservation of herbarium specimen images and a tool for extracting information by image analysis.

Developed by five institutes from Finland, France, Germany, Netherlands and Scotland it aims to be available to other herbaria in the future.

Making the specimen images and data available online from different institutes allows cross domain research and data analysis for botanists and researchers with diverse interests (e.g. ecology, social and cultural history, climate change)

The scientific & technical challenge

Herbaria hold large numbers of collections: approximately 22 million herbarium specimens exist as botanical reference objects in Germany, 20 million in France and about 500 million worldwide. High resolution images of these specimens require substantial bandwidth and disk space.

New methods of extracting information from the specimen labels have been developed using Optical Character Recognition (OCR) but using this technology for biological specimens is particularly complex due to the presence of biological material in the image with the text, the non-standard vocabularies, and the variable and ancient fonts.

Much of the information is only available only using handwritten text recognition or botanical pattern recognition which is less mature technology than OCR.

Why EUDAT?

The Herbadrop project intends to benefit from EUDAT services that are operated at CINES or in one other EUDAT partner. Technologies involved in the analysis and long-term preservation process can be gradually integrated since these steps are already running at CINES for different purposes.

1) B2SAFE will be used in the first step of the ingestion process. Existing images of herbarium specimens along with the associated data are transmitted to the CINES repository using a Data synchronisation & exchange service.
2) The ingestion into B2SAFE will always be carried out in accordance with the centralized persistent identifiers (PID) management system used in EUDAT (e.g. EPIC handle);
3) The discovery, sharing and visualization of the data objects can be performed with the EUDAT B2FIND service.

Expected outcomes

Online data storage and image processing are not the main skills of Natural History Collection institutes. By bringing together the knowledge of each institute on herbarium specimen images and the experience of CINES on long term digital preservation we plan to build an infrastructure both powerful and easy to use. The system will provide the best OCR technology adapted to the requirements of herbarium specimen images and will require minimal installation in each institution.

Expected domain legacy

Safeguarding long-term data storage is an important precondition for reliable access to herbarium specimen information. Thanks to this pilot, it is possible to envisage a long term storage for herbarium specimen images.

Moreover, the specimens will be discoverable by the entire scientific community. Thus, undescribed species stored in herbaria can be examined by experts to aid identification and discovery of new species. Distribution information for species over time can be evaluated and these data could provide evidence of the point of time when an invasive species first occurred in a certain area. Historians could analyze herbarium data to create itineraries for historical characters. The data can be used to calibrate predictive models of the oncoming changes in biodiversity patterns under global threats. This diverse information will be useful for a wide user community including conservationists, Policy makers, and politicians.

The use of the EUDAT repository to store clinical trials in a secure and compliant way

Overview of the pilot

The EUDAT repository will be combined with EUDAT services for the secure, GCP (Good Clinical Practice) compliant and transparent storage of clinical trials data. For such a safe and accessible storage of clinical trials, an authentication service (AAS) manages the access rights for users; a PID service refers users, digital objects and associated documents to each other. In addition, the linking to metadata and the data type registry is necessary. The leading European clinical researchers are centred in the European Clinical Research Infrastructures Network (ECRIN), whose members and other researchers will access the repository to analyse data of different European trials.

The scientific & technical challenge

Randomized clinical trials (RCT) are the important step to bring treatments from preclinical development to the patient. But many scientific and technical challenges exist for clinical trials, like the need for innovation in trial design and for more objective interpretations of trial outcome data. There exists still a gap in the translation of basic scientific discoveries into clinical trials and of clinical trials into medical practice. Although, biomedical sciences has provided an unprecedented supply of information for improving human health, clinical trials data do not participate in the activities of the research environment in an important way. After the conclusion of a clinical trial, most raw data is withdrawn and archived inaccessible in archives and only statistical summary results are published. What is missing is a repository for clinical trial data (raw data in anonymised form) that may become the first step for the provision of this data to the research community for analysis.

The course of clinical trials is determined by a detailed study protocol; patient data is collected by many investigators at different sites using electronic Case Report Forms (eCRF). Increasingly data from biobanks, nutritional and genetic data and data from electronic health records (EHR) are involved and exist in different formats (Fig. 1). After the end of the study, data is analysed using statistical software and study results may be published. Nonetheless, the clinical trials raw data is stored in isolated archives without metadata enrichment and without links and references to preclinical data, trial documents, publications, analysis results, contents of trial registries and without the possibility of access by the research community.

Why EUDAT?

For the safe and accessible storage of clinical trials, an authentication service (AAS) will manage access rights for different user groups; links and references to metadata and data type registries will provide the searchability of the trials data and a PID service will refer users, digital objects and associated documents to each other ensuring transparency. One

possibility is to develop a suitable data warehouse from scratch for the storage of clinical trials data. But we decided against this solution and in favour of employing EUDAT services for several reasons. In our experience, different EU projects often developed similar solutions with limited reach and usability for other projects. Thus, a more generic approach is needed to open clinical trials data for the research environment and the joint analysis with life science, genetic, and nutritional data. By employing EUDAT services we can build on the experiences of other research groups, use common standards and tools for access control and data protection; but most importantly we can integrate trials data and metadata into the generic EUDAT service layer that is being developed for all kind of research, including climate, oceanographic and earth sciences. In addition, being part of such a large and trusted infrastructure will encourage clinical researchers to provide their trials data to the EUDAT repository.

Expected outcomes

Raw clinical data of several clinical trials will be stored in a standard-based, secure and compliant way in the EUDAT repository. After appropriate authorisation users will be able to access and analyse the stored clinical trials data. The data is characterised by accompanying metadata and data type specifications and linked to each other and to study results, documents (like data management plan, the statistical analysis plan) and publications by PIDs. In this way, researchers and investigators can get access to raw clinical study data and documents and can analyse trials on the individual patient level and by using cross-database examinations.

Expected domain legacy

Once the EUDAT repository is being filled with data from many different clinical trials, users will be able to access and analyse clinical data on the individual patient level and conduct meta-analysis between different trials, including trials that not only were properly finished and published, but also trials that were aborted or trials with a negative result that never were published as well as trials where analysis procedures were only insufficiently described in their publications (underreporting). In this way, more reliable results will be obtained. Users will have to add and evaluate different analysis tools and processes for the repository, but in general, accessing the EUDAT repository investigators and researchers will be able to compare restricted access clinical data with open access data available in a multitude of different biomedical and genetic databases, a precondition for the improvement clinical trial design and of medical treatments and especially for the successful application of personalised medicine.

An EUDAT-based FAIR Data Approach for Data Interoperability

Overview of the pilot

In order to achieve data publication in a FAIR manner and foster their findability, accessibility, interoperability and reusability, a set of (FAIR) data tools are being developed, including the FAIR Data Point (FDP), which is a software layer on top of datasets to expose them as FAIR (inter-linkable) data. The FDP provides information about the available datasets in terms of their metadata as well as the actual access to the data in an interoperable format. Our pilot will evaluate whether current EUDAT services could be extended to behave as FDPs or a new FDP-based service should be proposed.

The scientific & technical challenge

Although the FAIR Data Point service will be useful to any research community facing massive data management and interoperability issues, the proposed pilot will specifically target the life science community. Due to the complex nature of biology, life science data arguably represent one of the most heterogeneous, diverse and challenging type of research data. Exposing new and existing datasets following the FAIR data principles will facilitate the improvement of our ability to interpret and combine these data.

A FDP service built on the EUDAT infrastructure offers advantages to individual researchers, as well as research groups and consortia. Existing, small-scale semantic data repositories are frequently managed by the researchers themselves and are notoriously difficult to maintain, resulting in frequent unavailability and short repository life spans. Therefore, one of the benefits of the service is to be able to completely remove this burden from the researcher. We would like to emphasize the novelty of such a service since to date; no Semantic Web-enabled repository services are available to the general research community. Moreover, FDPs are designed to enable data citation and maintain statistics about data accesses, which means that impact will be measurable for any FDP deployment.

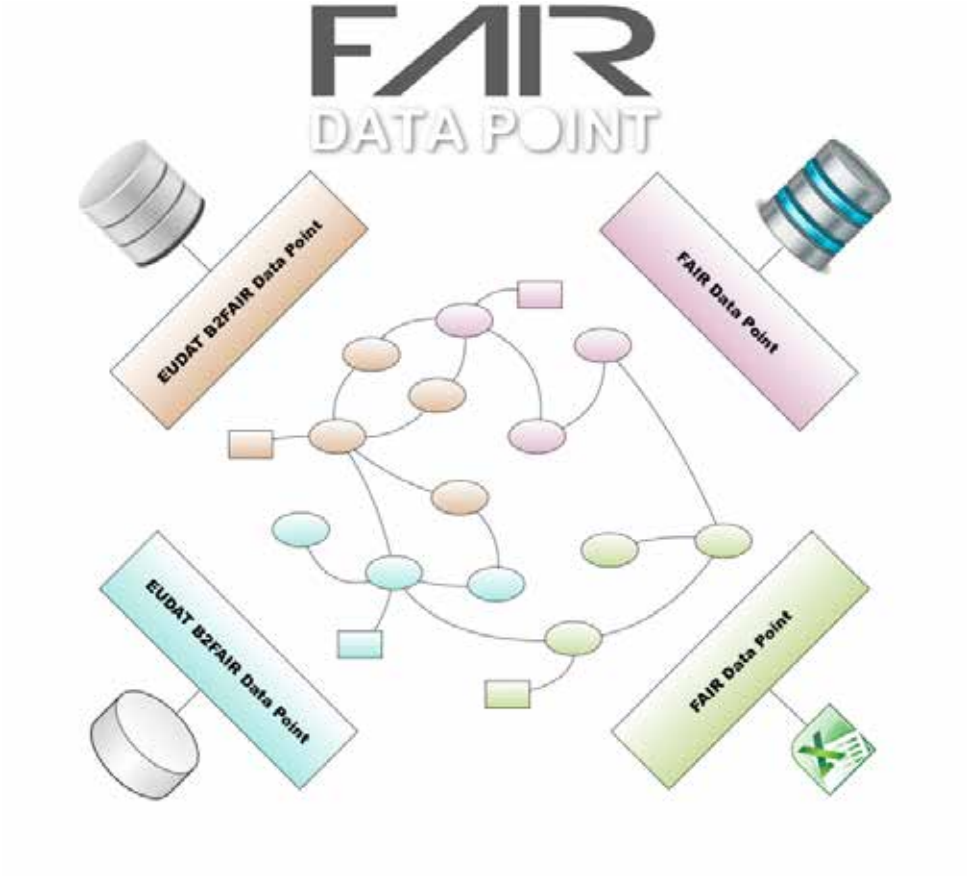
Within this pilot, we aim to implement and deploy a FDP using a combination of existing Semantic Web standards and frameworks for the front-end, and (existing or new) EUDAT services for the back-end. A FDP provides access to the data and metadata using REST-APIs conforming to the W3C Linked Data Platform specification.

Why EUDAT?

The reliability of the EUDAT data infrastructure would prevent the loss of valuable research data, while the functionality provided by the FDP improves the discoverability, interoperability and reusability of semantically rich research data. In our pilot we target primarily the EUDAT B2Safe and B2Share services. The ultimate goal is to have these services also

complying with the FAIR Data Point behaviours and, therefore, adhering to the FAIR Data principles by offering metadata and data in a FAIR manner.

To accomplish this goal we plan to first investigate the EUDAT services specifications in order to verify what adjustments need to be made for them to expose the behaviours of the FAIR Data Point. Then we will develop a prototype of these necessary adjustments to demonstrate the feasibility of the FAIR-complying EUDAT services and discuss with EUDAT organisation a plan to realise the extend services in the production environment.

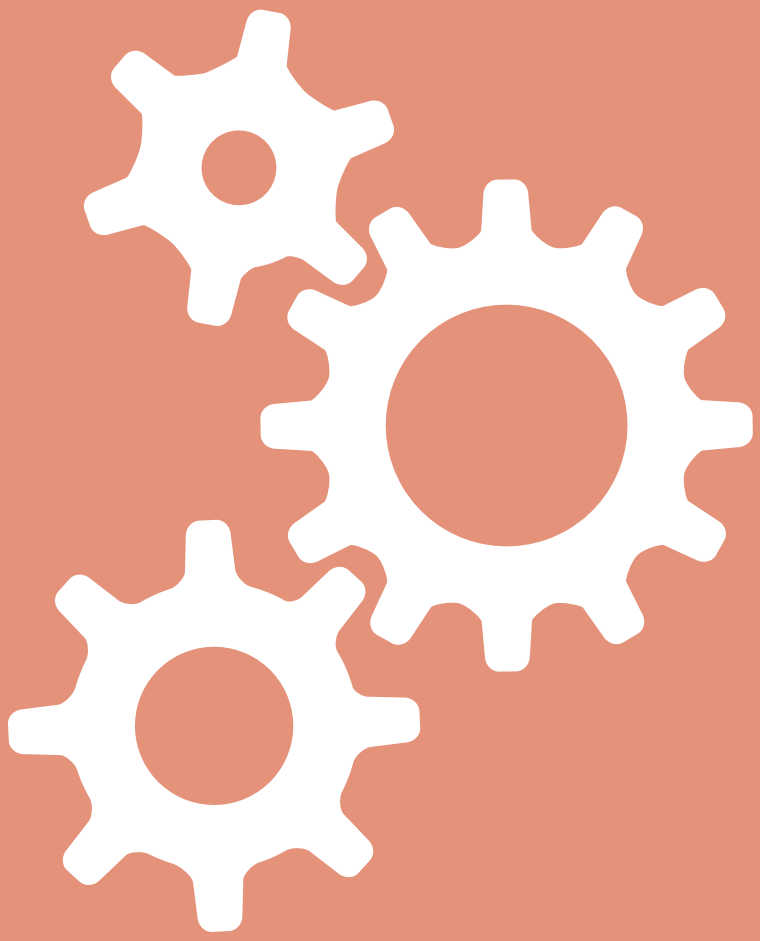


Expected outcomes & Expected domain legacy

By the end of our pilot, we expect to have demonstrated the feasibility and benefits of having a large scale data repository service provider such as EUDAT allowing the published datasets to be exposed in a FAIR manner. For the community, the benefits will be to have a reliable way of publishing the datasets in a way that promotes data findability, accessibility, interoperability and reuse and, therefore, fulfils part of a good data stewardship plan, which is being increasingly demanded by funding agencies as part of their research grants. This infrastructure is also part of the preparation for the upcoming requirements of the European Open Science Cloud.

Physical Sciences and Engineering

EUDAT has 5 Data Pilots from the Physical Sciences and Engineering domain.



Tokamak data mirror for JET and MAST data – moving towards an open data repository for European nuclear fusion research.

Overview of the pilot

Our data pilot will provide a mirror of experimental data from two magnetic confinement nuclear fusion devices (Tokamaks) at the Culham Centre for Fusion Energy (CCFE): the Joint European Torus (JET) and the Mega Amp Spherical Tokamak (MAST).

The research community will be plasma physics and fusion researchers, engineers and technologists from the 29 members of the EUROfusion consortium and around 100 associated organisations, including those delivering the next generation nuclear fusion device (ITER) in southern France, namely ITER-IO (France) and Fusion 4 Energy (Spain).

The scientific & technical challenge

Data from the JET and MAST experiments has been collected over many years (JET has been operating since 1984). It is hosted at CCFE and made available via bespoke APIs and visualisation tools.

We would like to make more use of cloud-based data infrastructure including object-storage platforms. The challenges in making use of a third party platform include:

- Maintaining the native data versioning and validation status information
- Maintaining the link to local identifiers for data items
- Not losing information from the native hierarchical structure of the data
- Complying with UK government and EU policies on hosting and access restrictions
- Keeping mirrored data in sync as new versions of individual data items supersede old ones

There is scope for EUROfusion members to make more use of each other's data. We intend to make it simpler to access JET and MAST data remotely.

Data volumes are ever increasing - both the total per experiment and the size of individual signals such as high-resolution camera data. It's necessary to plan ahead and evolve our data infrastructure to cope with this continued growth.

We are also keen to develop and pilot data management approaches for the next generation nuclear fusion device, ITER, which is currently being constructed in southern France. ITER's individual experimental runs will have a much longer duration than the current generation of tokamaks and will generate up to 0.4PB of data per day.

There is lots of potential for researchers to make more use of HPC facilities and we aim to provide more convenient ways to make data available for this purpose.

We estimate that several hundred users might initially make use of the EUDAT data mirror once it's fully tested and publicised.

Why EUDAT?

The EUDAT platform appeals because the general approach and the services available match well with our own ideas about the future of our data management infrastructure. The Europe-wide nature of our research community is a good fit with EUDAT's scope.

B2SHARE will be used to provide on-demand access to individual data items via APIs. We will collaborate with EUDAT developers to address some of the challenges around data structure, versioning and access controls.

B2FIND will be used for data discovery. We aim to provide improved meta-data such as aliases or tags for commonly used signals to help users who aren't familiar with the machine-specific signal names.

B2SAFE will be used for resilient data storage. This will improve the redundancy of our data management infrastructure and allow bundles of data to be downloaded for particular purposes.

B2STAGE will be used to test shipping data sets between EUDAT storage sites and HPC clusters at Harwell (UK) and CINECA (Italy). This will reduce the need to create and move data bundles manually which can be difficult to manage and break the provenance chain.

Expected outcomes

The EUDAT data pilot provides us with a chance explore how our data systems can best be integrated with cloud-based data management infrastructure. Because it is separate from our existing systems, there will be more freedom to come up with the best solution without having to address total backward-compatibility from day one.

The project will enable users from the EUROfusion community to access data from the JET and MAST experiments more conveniently without complicated remote access arrangements. We intend to extend the number of researchers using the data by making it more easily discoverable and providing access in more convenient ways.

The ability to ship datasets to HPC clusters for processing should encourage more use of these facilities and improve the convenience and traceability of the workflow.

Expected domain legacy

This study will be a proof of concept for delivering nuclear fusion data on EUDAT services. If successful we would like to make it the primary route for remote access to our data and continue to improve the meta-data and access interfaces.

Use of EUDAT could be a means of ensuring the continued availability of the data beyond the lifetime of the current experiments. In the longer term we could aim gradually to increase the scope of the data hosted to include more of the raw data from JET in addition to the more commonly used processed data.

If the pilot is successful we hope it will grow into a shared repository for data from other nuclear fusion experiments across Europe. This could be a step towards more common tools and interfaces, shared between the various experiments. We are keen to develop and pilot data management approaches for ITER, the next generation nuclear fusion device, along with our colleagues in other organisations.

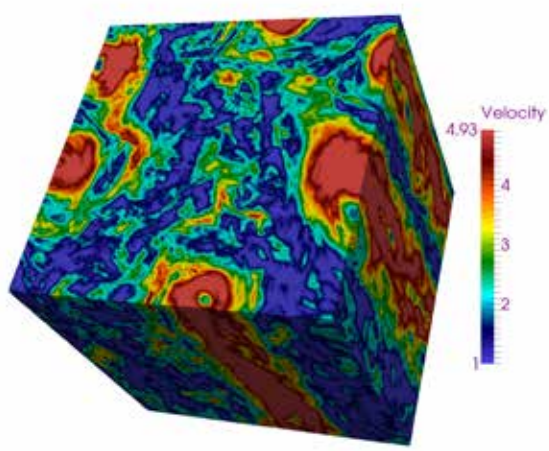
Turbase DNS

Overview of the pilot

Our project is meant to preserve and standardize a first set of state-of-the-art numerical datasets in computational fluid dynamics, concerning: (i) fully homogeneous and isotropic turbulence evolved on a fractal Fourier set, (ii) a world record simulation of a turbulent flow with rotation at 40963 collocation points (iii) multi-component microfluidics in complex geometries. Data-sets include both Eulerian and Lagrangian data, i.e. snapshot of the velocity field and trajectories of particles affected by the flow. All data are of potential interest for a vast community of researchers, mostly in Europe and in the USA, in the fields of theoretical physics, geophysics, meteorology, chemical and bio-engineering.

The scientific & technical challenge

The Computational Fluid Dynamics (CFD) community is facing more and more the problem of data preservation, data standardization and data analysis (by both the data owners and by third parties). It is therefore mandatory to develop user friendly supports and optimal interfaces to make the data available and useful for a long period of time. Besides the obvious scientific interests of the owner groups, the availability of these large data sets is potentially crucial for a much wider audience of theoretical and applied scientists working in different cross-disciplinary domains, who do not intend --or cannot do—numerical simulations on their own. Moreover, it is important to mention that these accurate and high-resolution (both in space and time) datasets cannot be obtained by any commercial CFD software because of the very strict requirements about the precision, error control, statistical accuracy etc. Systematic analysis and classification of huge datasets is a challenge for both the needed man power and for the storage requirements. Our research group is involved in many collaborations all over Europe and worldwide, including the International Collaboration for Turbulence Research (ICTR) and the European High-Performance Infrastructures in Turbulence (EuHIT) project, two initiatives that count more than 100 scientists in the domain of numerical, theoretical and experimental fluid mechanics.



Why EUDAT?

In a first stage of the pilot we are mainly interested to the EUDAT B2STAGE service to maintain, analyse and standardize the data produced by several PRACE projects about Rotating Turbulence, Turbulence under Shear, Turbulence at

changing the dimension of the embedding Fourier dynamics and micro-control of droplet formation in T- and Y-junctions. Typical storage requirement for each of the above applications is about 50 TB. Moreover, most of the data analysis requires applications developed for high performance computing on massively parallel architectures. During the EUDAT pilot we intend to complete the data analysis and to develop new data comparisons among the different data-sets. It is therefore crucial to have all the data on the same storage. In a second step of the pilot, the standardization and classification of the data-sets will produce useful metadata that can be used by B2FIND for correctly/easily browse the data, which will promote the dataset usage among a wider community.

Expected outcomes

Our group will greatly benefit from B2SHARE service for two main reasons. First, the comparison and standardization will allow us to reach new scientific targets concerning the common physical problems shared by all our numerical applications. The second important outcome is to create strongly coherent datasets to be preserved and refined with valuable metadata information, such that a subset of each dataset can be used by the potential community of end users as a reference of well-established and trustable numerical data-sets.

Expected domain legacy

There exist already in Europe the EuHIT project meant to standardize and preserve a series of experimental turbulent data collected from various laboratories in Europe. Our EUDAT pilot intends to complement the previous database with a set of prototypical numerical data such as to exploit synergies among the different data type. Accurate understanding and modeling of fundamental fluid dynamical properties benefit from the realization of benchmark measurements to be used as solid reference: new ideas, algorithms and probe techniques can be tested against common well-documented case studies. These benefits can be foreseen only using a service as the one offered by EUDAT: that is a user-friendly, reliable and trustworthy way for researchers, scientific communities and citizen scientists to store and share small-scale research data from diverse contexts. In particular this is crucial for scientists in the CFD community that do not have adequate facilities for storing data with metadata, and that cannot guarantee long-term availability of their locally-stored data, and/or do not have adequate facilities to easily share data, results or ideas with colleagues. A snapshot of the velocity modulus obtained from Direct Numerical Simulations of a three-dimensional homogeneous and anisotropic flow, under strong rotation.

NFFA-EUROPE Information and Data Management Repository Platform for nano-science in Europe

Overview of the pilot

Another challenge is the need for the integrated Information and Data management Repository Platform (IDRP) to cover the full research lifecycle for the user community. It will involve automated acquisition of key metadata into a data repository for future data access, also defining a data policy including the need to address the IPR issues.

The efficient data archiving for nanoscience community is another challenge, i.e. harvesting from open-access scientific Data Repositories (DR) that could support sample/material preparation protocols with absolute metrology, and adequate metadata for the characterization and scientific investigations

It is fundamental that that existing standards, recommendations and evolving best practices of data management are incorporated, as well as sensible reuse of existing e-infrastructures where applicable rather than building own e-infrastructure for nanoscience from scratch. So NFFA, on one hand, will consider using the existing and emerging EUDAT services for data archiving, sharing and discovery, and on the other hand, will contribute to testing EUDAT services in-the-field and provide feedback for their tuning or extension.

Why EUDAT?

To address the above challenges a very close cooperation with EUDAT and the adoption of their results, whenever possible, is of great significance.

Thus in this data pilot, we propose to use EUDAT data services to:

- Focus on developing a data service around the NFFA-EUROPE IDRP rather than developing yet another e-infrastructure
- Provide data with clear identity as many NFFA partners do not mint persistent identifiers for data and some EUDAT services offer the data identifiers functionality out-of-box
- Support sharing of long tail experimental data (via B2SHARE)
- Provide easy and flexible discovery of data in a central location (via B2FIND)
- Explore opportunities for scalable and trusted storage and replication of raw experimental data (via B2SAFE)

These services will be integrated into the NFFA-EUROPE IDRP, with due consideration to the actual data management policies and technology maturity of the project partners.

Expected outcomes

The integration of B2SHARE into the NFFA-EUROPE IDRP infrastructure will allow users the chance to access and share their long tail data. If successful the use of B2SHARE may become a recommended "mainstream" solution for long tail data in nanoscience

Regarding B2FIND, the mapping of the nanoscience metadata standard that NFFA is currently developing to the B2FIND metadata might open the nanoscience data to a huge number of new communities, and the registration of published datasets could provide an immediate benefit on both sides.

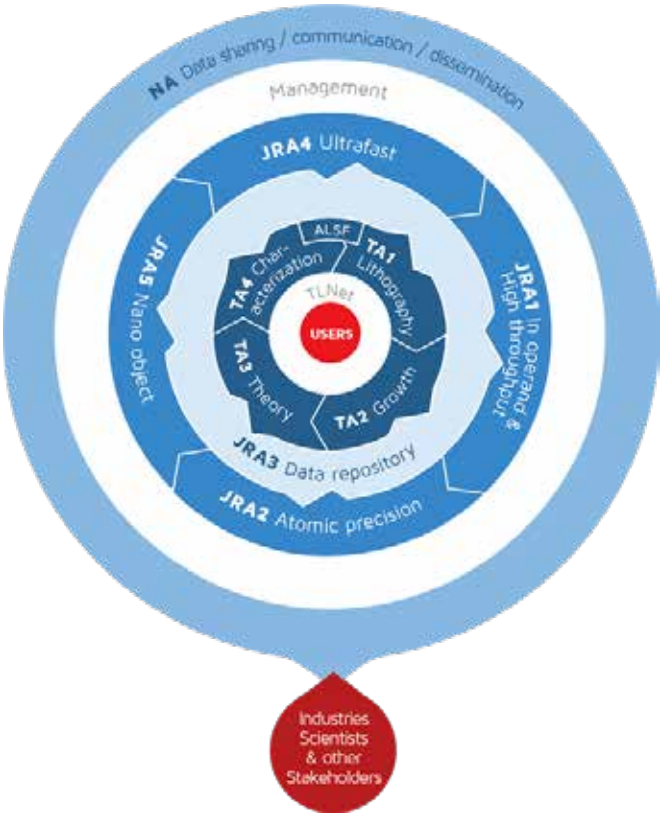
The active use of B2SAFE is currently considered a more experimental activity compared to B2SHARE or B2FIND. NFFA-EUROOPE hopes to identify suitable partners and compelling use cases where B2SAFE is going to bring benefits to the nanoscience community without breach of data management policies or unreasonable stress of the network infrastructure.

Supplying the nanoscience data with persistent identifiers can be seen as a valuable by-product of using EUDAT services.

Expected domain legacy

A major innovation potential is related to developing a sensible data access and data reuse policy for the NFFA-EUROPE IDRP that supports common cases for intellectual property management like giving academic and innovation credits to researchers who collect and process experimental data. Sustainable supply of persistent data identifiers should support this common business case of data reuse and intellectual property management.

The EUDAT services will help us to develop our IDRP and advanced data services around it that are required by the NFFA EUROPE-wide community. The NFFA-EUROPE will provide training for data practitioners from the industry or other user communities, with EUDAT services and the cases for their use in NFFA-EUROPE IDRP as illustrative examples of a modern e-infrastructure which applicability may extend beyond the lifespan of a particular EU project.



Direct simulation data of turbulent flows

Overview of the pilot

Turbulence is a relevant field in science and engineering nowadays, as countless industrial and technical applications rely on fundamental research for better performance and efficiency. A worldwide community in universities and research centres have direct numerical simulations as main research tool and DNS data analysis is of great importance for experimentalist and industrial model tuning. Our group has been using DNS of turbulence for 30 years, and probably owns at the moment the largest public data base of turbulence data. New challenges arise as the amount of data to store, preserve and share increase with larger simulations.

The scientific & technical challenge

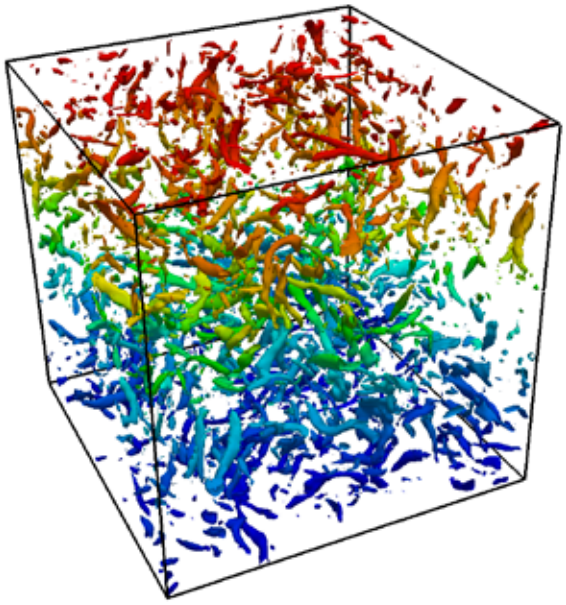
Direct simulation of turbulent flows produces large amounts of data that can be used for multiple purposes and multiple researchers besides those originating them. After initial publication, those data can be shared freely worldwide. The community has come to expect that to be done, since large simulations are too expensive to be repeated by everybody, but there have been up to now few programs specifically dedicated to data preservation and sharing. DNS data comes in two types: 1) Summary statistics, which are processed data sets of size (<= 1GB) that can easily be shared from a departmental web page, but which have very long-term value (>20 years) and, 2) raw fields that require some kind of processing before they are useful, have typical individual file size of 50-100 GB. A data set should include several hundred individual flow fields to create good statistics, and a typical complete data set is 50-100 TB. The useful life of raw data is typically 10 years. Sharing these data is limited by the lack of an agreed standard format, even for meta-data, and by lack of resources. Since data generation is typically linked to a research grant, data tend to die after that grant ends. The community of established researchers on turbulence may be estimated as a few thousand worldwide. They use data in a variety of ways, going from fundamental research to model testing or tuning. Because experiments are difficult, and only observe a limited number of variables, DNS simulations have revolutionised the field. However it involves managing large amount of resources and computational time for post-processing and storage which are the main limitation for research groups.

Why EUDAT?

The EUDAT services that we are interested in are B2store and B2Share. First, we require a reliable, safe and robust way to preserve and access our turbulence database now and in the future. Also to guarantee long-term persistence of data is an important feature of B2store services as will allow the data to be accessible and available for a longer time beyond research periods. Second, B2share services will allow us to enlarge the pool of users that we already have and to provide standardized meta-data extensions and user-friendly interfaces to increase the impact of our activity in the turbulence research community.

Expected outcomes

The idea of this pilot is to acquire new resources that give us the possibility to archive raw data and share it in a more standardised and stable way for public access. Since many of these data can only be processed on a supercomputer, it is also interesting to use EUDAT pilot together with time in Prace or similar computer resources to explore post-processing options. EUDAT pilot offers us the opportunity to store and improve our wide database features. Designing the meta-data for the community is important to reach a standard that allows easy and fluent data exchange between research groups in our community. Optimized and fast access for end users is also expected, which will surely improve our existing open access service to our database. We also expect to benefit from EUDAT as new data intensive simulations are planned in the near future and storage capacity is already a limitation at present.



Expected domain legacy

We expect our database and also data generated in the future to be stored and organized in a reliable, accessible and safe way. Also standardized procedures to read this data will make it easier for researcher in turbulence community to benefit from it. EUDAT also offers services and tool to store and preserve data and make it publicly reachable beyond the research grant period in which it is generated.

SIMCODE-DS

Overview of the pilot

The SIMCODE-DS project deals with the need of high resolution simulations in view of the advent of what is known as the epoch of "Precision Cosmology". The latter term indicates the huge quality leap in the accuracy of observational data expected for the next decade (mostly through large galaxy surveys as the European satellite mission Euclid) that will allow tests of the cosmological model to percent precision. As a robust interpretation of such high-quality data will require a large number of cosmological simulations, the community will face in the next years a serious issue of big data storage and sharing.

The scientific & technical challenge

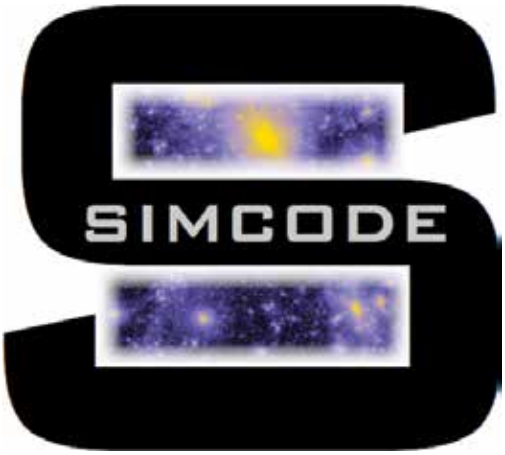
Cosmological simulations are an essential ingredient for the success of the next decade of "Precision Cosmology" observations, including also large and costly space missions as e.g. the Euclid satellite. Since the required precision and the need to test for statistical anomalies, astrophysical contamination, parameter degeneracies, etc will require a large number of such simulations, the community is about to face the issue of storing and sharing big amounts of simulated data through a Europe-wide collaboration. In fact, cosmological simulations are getting progressively cheaper as computing power increases, and even for the exquisite accuracy and the huge dynamical range that are required for Precision Cosmology, the main limitation will be determined by data handling rather than by computational resources. Also, while large simulations can now be run in a relatively short time taking advantage of highly optimised parallelisation strategies and of top-ranked supercomputing facilities, their information content might require years of post-processing work to be fully exploited. A typical example is given by the Millennium Simulation (Springel et al. 2005) that is now more than 10 years old but is still employed for scientific applications. The present Pilot aims at testing possible strategies to make large amounts of simulations data available to the whole cosmological community and to store the data for a timescale comparable with the duration of collaboration such as Euclid (~10 years). The main idea behind the project is that various types of simulations (differing by size, dynamical range, physical models implemented, astrophysical recipes, etc) can be safely stored on a central long-term repository and their content made easily available through metadata and indexing procedures to the community at large, which can range from a small group of collaborators to the whole Euclid Consortium (> 1000 people) depending on the specific nature of the stored simulations.

Why EUDAT?

The EUDAT Pilot Call promises to provide dedicated infrastructures for long-term data storage, data handling and indexing, and data sharing over broad communities of users. This

represents a valuable opportunity for the community of cosmological simulators that are presently facing the difficulties of fully exploiting their numerical products. Also, the storage provided by supercomputing infrastructures is getting progressively less suitable for this purpose as data size increases since the scratch areas of computing clusters need to be periodically cleaned up to leave room for running applications. This makes it always a struggle to "park" finished simulations in a safe place where they can reside for a sufficiently long time to allow a full exploitation and a thorough post-processing analysis.

Expected outcomes



As a cosmological simulator I have access to several computing facilities located in various parts of Europe. In some case the access is granted based on collaborative works with the hosting institutions, in some case it comes as a result of a competitive call (as for the case of PRACE or DECI accounts) and in some case it is granted by individual affiliation. In any case, all these accounts are generally limited in time and require removing the data from the machine at the end of the allocation period, which is normally shorter than the time after which simulations become obsolete. This means that a single research group can produce simulations data on different machines and then spend a significant fraction of its time in struggling to move data from one place to another trying to save scientifically useful data from deletion. From the present EUDAT pilot I expect to have finally a single centralised storage location where to move all the finished simulations performed on different computing facilities to allow for long-term collaborations relying on an easy access to simulated data.

Expected domain legacy

In the field of computational cosmology it has often been cheaper and more convenient to re-run a large simulation that had been previously carried out on a remote machine rather than trying to move the data. This is clearly a waste of computational resources, a waste of time, and a useless duplication of work. Furthermore, it won't be feasible for the size and dynamical range of the simulations required in the next decade. Building a stable infrastructure for sharing simulations data and for allowing an easy browsing of large datasets would represent a significant improvement for the field. This would allow easier collaborations and a more efficient planning of HPC allocations.

Social Sciences and Humanities

There is a one 1 community member and 5 data pilots represented for the Social Sciences and Humanities

Core Community - CLARIN

The CLARIN project is a large-scale pan-European collaborative effort aimed at making language resources and technology readily available for the whole European Humanities (and Social Sciences) community. This includes coordinating the development of appropriate resources. Amongst other things, CLARIN will offer scholars tools for computer-aided language processing. In more detail, CLARIN offers:

- Comprehensive services to the humanities disciplines with respect to language resources and technology.
- Technology for overcoming the many barriers created by institutional, structural and semantic interoperability problems and fragmenting the resources and tools landscape.
- Tools and resources that will be interoperable across languages and domains, thus addressing the issue of preserving and supporting the multilingual and multicultural European heritage.
- Comprehensive training and education programs that include university education in the different member states.
- Improvement and extension of web-based collaborations, i.e. creating virtual working groups breaking the discipline boundaries.
- Development or improvement of standards for language resource maintenance.
- A persistent and stable infrastructure that researchers can rely on decades to come.

Areas of collaboration with EUDAT

CLARIN has been one of EUDAT's core communities since 2011 and the service that it has been most involved with, up to now, is B2SAFE. Various CLARIN centres are using it to perform safe replication of the language data they are hosting. The University of Tübingen (Eberhard Karls Universität Tübingen), the LINDAT/CLARIN Centre for Language Research Infrastructure in the Czech Republic (usually known as LINDAT/CLARIN) and the Max Planck Institute for Psycholinguistics in the Netherlands are all using B2SAFE. There are a further eight centres ready to use B2SAFE and tailored uptake plans are under development to be deployed over the coming months. CLARIN will harvest the B2SHARE metadata related to language material and make it accessible via their search portal, the Virtual Language Observatory. Additionally, EUDAT's B2DROP service has been tested and is used for internal data exchange and sharing. CLARIN community metadata has been integrated into B2FIND, EUDAT's metadata portal.

Main benefits for the community

EUDAT has a strong focus on facilitating the uptake of its services by research communities through specific uptake plans driven by the active involvement of community experts. This collaboration between EUDAT & CLARIN makes it possible for "homeless" researchers and

citizen scientists to deposit their resources into a good data repository with long-term preservation. There are other services – for example, the metadata integration into the infrastructure, or safe replication – that are all the kinds of services on which CLARIN and other research communities can build. It makes sense to pool these resources and make sure there are good, reliable and stable services that can be used by everyone as, in the end that will benefit all the research communities.

For more information on CLARIN see: https://www.clarin.eu/

Research data repository for students' own results

Overview of the pilot

In the Department of Physics of University of Helsinki, the masters level training of physicists; have traditionally included extensive laboratory experiments, their documentation and reporting. This pilot is for including data publication and curation of the experiment results in the laboratory courses: storing the observations, together with relevant metadata into a repository, where the course assistants would have access. The students would then learn to publish and document their data as a normal part of scientific workflow. Naturally it would be needed also to include methods to "cite" the data sets using a PID offered by the system.

The scientific & technical challenge

One of the biggest issues in data sharing is a cultural one. Current research paradigm does not necessarily consider data sharing and curation as a part of normal scientific process. Teaching this as a part of normal course work is the way to get the message across to new scientists, thus creating possibility of new generation of scientists who do not need to be taught and forced to publish their data – no more than they are to publish their results. Taking data publication as a normal part of course work is the key.

The data publication should however be realistic, easy and flexible. The system should be similar which are used in long-tail of the research data applications and the overall system should also serve the overall course work. For this reason, the system should be capable of authentication, team work, controlled evaluation and to be completely citable. The annual number of students in the initial phase is less than hundred and the data amounts are small.

Why EUDAT?

From the EUDAT side the system will be realised using a version of EUDAT B2SHARE platform, with some local variations (access rights, template). The access to results would (at least in default) be for the student (data originator) and the course assistants (for control). The students would have realistic, but practical user experience for publishing

small datasets. The overall size of the datasets is not large. Crucial issue is a good interface (including somewhat specified template) and ease of use. The template will be developed by UHEL together with EUDAT2020 team. There should be an unique identifier for each data set for realistic inclusion to the students' reports

Two optional additions can also be considered:

- In long run, the repeated experiments could be of interest to e.g. pedagogical research. Thus possibility of anonymization of data set producers and completely open access to results after a grace period could be optional goal.
- Another optional goal could be direct commenting of the data sets (in this case by course assistants), and if corrections are needed - new versions of data sets.

Expected outcomes

New physicists graduating from the Department of Physics will take data publication and curation as a natural part of their scientific work. This will increase their employability and significance of their research. The education will also give secondary benefits when part of these students will continue as PhD students in the research teams. The education of data publication and curation is also an important transferrable skill. This system can also be then generalized on different courses throughout the University of Helsinki and Finnish education sector – all important users of EUDAT services.

Expected domain legacy

Most important factor is to foster the culture of open research. Teaching students early on that data publication is crucial and natural part of the scientific process is one of the key ways to make this change. Experiences on metadata inclusion with from a large pool of students can be useful to find practical bottlenecks of such systems.

Enriching Europeana Newspapers

Overview of the pilot


Le Petit Parisien

C'EST UNE PAGE UNIQUE DES ANNALES DU CRIME

COMMENT BONNOT FUT PRIS ET TUÉ

BONNOT

DUBOIS

Enriching Europeana Newspapers aims to expose the full text aggregated as part of the Europeana Newspapers project. It contains over 11 million pages of full text of historic newspapers (mainly but not all 19th century), drawn from national and research libraries across Europe. A portal is already in place at http://www.theeuropeanlibrary.org/tel4/newspapers. This pilot aims to expose and improve the text for more data driven usage (ie large scale data analysis of the whole corpus)

The scientific & technical challenge

The key scientific challenges are these:

- Creating best practice guidelines for the publication, citation and impact measurement of cultural heritage data (ie the newspapers in question). Standards for citing and judging the impact of open cultural data are still far from being established.
- Enriching the newspapers corpus, via the automatic extraction of topics and named entities; the current corpus is only searchable via free text searches
- Showcasing the value of the enrichment by a quantitative analysis of the occurrence of topics/entities over time and across borders.

A particular challenge will be the extraction of topics across texts in multiple languages (over 40 languages are featured in the corpus from French to Yiddish to Estonian) and variable quality of the digitised text.

Digital humanities scholars will be interested in the raw OCR texts; the number of these is likely to be in the 100s rather than 1000s. We also suspect that others in linguistics, economics, information science and computer science can make use of the datasets

If successful the enriched texts could also be placed in the current Newspapers interface (http://www.theeuropeanlibrary.org/tel4/newspapers). This received over 1.4m page impressions in 2015, around 5 to 6,000 users a month. Better search facilities will help improve these numbers

Why EUDAT?

We will be using the B2SAFE and B2FIND services. These will help us undertake the enrichment of the datasets and, more generally, expose them for re-use by other academics,

particularly those outside the digital humanities. At present, users of the service tend to be 'traditional' historians who are familiar with the search and browse possibilities of the portal – connecting with the tools and, just as importantly, the EU-DAT community.

Expected outcomes

We expect to meet to three significant use cases:

1. to have a better understanding of the topics, themes and subjects featured in the newspapers, allowing researchers richer understanding of the how certain issues and ideas were phrased in the corpus under question
2. the extraction of topics will also assist with resource discovery – allowing users of The European Library portal to search not just for free text words but topics. (Note: the re-integration of the enriched dataset into the existing newspapers portal is not foreseen in this piece of work)
3. exposure of the datasets via EUDAT will allow for much greater discovery and reuse of the newspapers corpus as a whole

Expected domain legacy

Meeting the use cases described above will help the study and understanding of historic newspapers as source material. The use of news[papers has been standard within many disciplines in the humanities and social sciences for a while, but their availability as a 'big data corpus' opens up many methodological avenues currently unexplored. It enables new research questions on language, communication as well as any topic featured in the newspapers.

Creating and exposing a corpus drawn from so many different countries also has benefits in developing transnational history, ie exploring themes and relationships between different European countries or the continent as a whole .

Cloudy Culture: A study of EUDAT shared services to measure the potential of using cloud-like services to improve the preservation of digital cultural heritage

Overview of the pilot

For the benefit of cultural organisations the National Library of Scotland, working with Edinburgh Parallel Computing Centre (EPCC) and with the support of the National Galleries of Scotland and the Digital Preservation Coalition will explore the potential of EUDAT cloud-like services to preserve European digital cultural heritage. The pilot will inform practitioners in digital preservation, curation and archiving and will test 3 elements of the EUDAT platform:

- The online transfer of large amounts of data to EUDAT (c 100TB/20 million files)
- Safe storage of data over time
- The use of high performance computing to accelerate preservation actions

The scientific & technical challenge

Cultural organisations need to preserve access to an increasing amount of digital content that they are creating and acquiring. For example the National Library of Scotland expects its data to grow 10 times over 10 years. This growth increases the strain on the core preservation requirements to store data in multiple geographic locations (cost of setting up more data centres), and to check if data changes over time (costs of increased computing power/time). High level studies suggest that traditional cloud services offer no net benefit for large volumes of data (100s of TB) that require on-going access to undertake preservation actions. There is little openly published information that describes or quantifies the practical limits and costs of using cloud-like services. For example how long will it take to transfer data? Is transfer and data monitoring scalable? What additional tools and services are required to automate the process?

Cloudy Culture will use EUDAT to hold a safe preservation copy of data to allow locally held access copies to be repaired if they change over time. For this reason access to the copy at EUDAT will be restricted to those few people who are undertaking preservation actions on the data. However the local copy of the data, mainly digitised collections, is freely and openly available via www.nls.uk where the audience size is millions of visitor sessions per year.

Why EUDAT?

Running in parallel with the growth of digital cultural heritage is the development of large data centres with a focus on science data. The European Commission sponsored 2014 Digital Cultural Heritage Roadmap for Preservation identifies existing e-Infrastructures as a solution to this problem, connecting these facilities with digital cultural heritage to ensure

our heritage remains accessible and usable long term. The Cloudy Culture pilot, supported by EUDAT, wishes to exploit this same synergy.

Cloudy Culture will integrate local tools with those developed by EUDAT and use B2SAFE, B2STAGE, iRODS, storage and computing power at the EUDAT facility at the Edinburgh Parallel Computing Centre (EPCC) to:

- automate the managed transfer of digital content and metadata between the National Library of Scotland and EUDAT
- automate and report on preservation actions undertaken in the EUDAT environment such as fixity checking and file format characterisation, accelerated by the availability of large amounts of computing power

Expected outcomes

During the 18 month pilot the EUDAT services will enable a third copy of high value digital cultural heritage at the National Library of Scotland to be stored in a different location and on different technology than the other two copies kept locally. This reduces the risk of losing the same data from all copy locations and so improves preservation. Because the EUDAT copy is monitored for changes through fixity checking any unwanted changes can be identified and repaired using intact local copies and vice versa. In addition the Cloudy Culture team will:

- improve local and EUDAT tools and workflows and the automation of data transfer and preservation actions to reduce human resource requirements and make preservation more sustainable
- gain an improved understanding of using EUDAT services such as transfer and compute times, transfer stability, scalability and costs
- understand the potential to use EUDAT and other cloud-like services beyond the pilot phase

Expected domain legacy

Cloudy Culture partners, in particular the Digital Preservation Coalition, will help to disseminate the results of the pilot for the benefit of the wider digital preservation, curation, archiving and cultural heritage domains. By doing this Cloudy Culture will increase community understanding of using cloud-like services and improve the communities' decision making. Cloudy Culture will:

- share information about the costs and benefits of preservation storage using EUDAT that can be transposed to other cloud-like services
- understand the viability and limiting factors of cloud-like services for large amounts of data
- improve tools for automated data management and share these with accompanying documentation so they are useful to others
- expand EUDAT's potential to act as a digital preservation option for European digital cultural heritage

Aalto data repository

Overview of the pilot

We are creating a central online location for data sharing for all Aalto University researchers. This will host both data and metadata: the name, description, ownership, source, and information on usage. Other dataset hosting sites exist, so our main target use case expanding EUDAT scope is intra-Aalto University interaction. Researchers with data analysis skills will be able to find data related to their work, as well as the domain experts responsible for that data. Furthermore the solutions should be tightly integrated to existing computing resources and Big Data platforms available nationally.

The scientific & technical challenge

In Aalto University, Big Data and Data Science have been recognized as key areas of ICT and digitalization at all levels of rapidly developing socio-economic societies. These systems generate ever-increasing amounts digital data, which can in unprecedented ways serve as a gold mine for researcher of various disciplines to study as well as enable the private sector players and public sector to develop their services, processes and technologies. Hence there is need to respond and find solutions to this data deluge, which is also reflected in the Aalto University application for profiling of Finnish Universities.

For the solution we have identified a set of design requirements that may pose also as a technical challenge. A few such are the requirements of 1) the metadata of datasets is full text searchable, 2) published datasets are assigned a persistent identifiers, 3) there are no restrictions to the type of uploaded data, 4) the datasets in the system can be made public for all the world to see, 5) the tool imposes no restrictions to the type of research data stored, 6) the metadata templates offered by the system satisfy the needs of different fields of science and also national requirements, 7) the system should be integrated to the already existing user management system.

In the first phase the system will have tens of users. If found to be successful, the solution will be scaled first within the University and possibly, even beyond to a national level. The possible user base may increase tenfold or even further in these cases. As Aalto users are working with large data sets the data volumes can already in the pilot phase potentially extend to tens of terabytes.

Why EUDAT?

We engaged into discussion with Centre for Scientific computing (CSC) that coordinates EUDAT operations in Finland. Based on these discussions we have assessed especially B2SHARE & B2DROP functionalities based on the requirements.

The results look quite promising and from this analysis we can see that the B2SHARE service would probably be the most suitable tool for addressing both our publishing and data management requirements. As a part of the pilot a suitable metadata templates can be provided

in B2SHARE and customized for Aalto University. Further, as federated authentication via B2ACCESS is supported in B2SHARE is should be possible to authenticate to the national identity service within EUDAT services. Also interfacing to national metadata and storage solutions is of interest to us in the EUDAT pilot.

Expected outcomes

The resulting data platform enables system and method level development, resulting in research innovations but it also opens up possibilities for educational and training purposes on Data Science and Big Data. In the initial phase there is a need to concentrate on data policy and repository practices for contributing to increased research effectiveness and generating wider goals of data sharing and open data in alignment with CSC's planned Big Data platform, but also to start widening and enhancing the skill base for data related research. We see that EUDAT solutions can paly a major role in supporting the implementation of the ambitious research data management goals set by Aalto University.

Expected domain legacy

This approach can enhance the system and process level understanding of any area of research, basic and applied, from science to engineering and humanities. It will strengthen multi and cross disciplinary research and step up product development, as well as facilitate novel innovations and services. At the same time it lowers disciplinary boundaries both in the public and private sectors, thus adding to innovativeness and new breakthroughs in R&D, and covering all areas of society and science from discovery and security to general competitiveness. The data repository will be initially used mainly for research purposes, but yet at the same time it will serve as a pilot and education platform for methodological development, training, and data repository technology.

Ancient OCR: Storing, Cataloguing, Relating, and Exposing OCR Objects from the Open Philology Project

Overview of the pilot

The Open Philology Project at the University of Leipzig has developed a modular, multi-threaded OCR pipeline to reach our goal of digitizing 100,000 books in the next three years. This pilot project gives us a way to store, catalogue, and expose the results of this pipeline, from original image to final OCR results. The users of the EUDAT system will be at the University of Leipzig and Tufts University (USA). The users of the data would be the same as those of the Perseus Digital Library, i.e., researchers and students in classical languages worldwide.

The scientific & technical challenge

The Open Philology Project produces many different data types according to many different data standards, all of which refer to the same real object, e.g., page images, HOCR data, XML text, syntactic treebank data, GIS data, etc. That is, the data that we publish should be thought of as a possibly ever-expanding vertical collection of objects. Perhaps our greatest challenge right now is to discover how we can get these different types of objects to interact not only within our own collections but also with collections outside of our own that might transform portions of our data or vice versa.

To do this, we need persistent identification of text and related data object

- that is stable throughout the creation, curation, publication, and post-publication lifecycle
- that can be leveraged easily across projects
- which can allow for a wide variety of PID schemes without requiring code changes for each one
- that is supported by institutional infrastructure
- and still allows for domain and project specific PID schemes

We also need a means to formalize and express details about the data types we are working with without coupling them tightly to the identifier schemes used to identify data objects that adhere to them. And, finally, we need support for multiple different models of collections (e.g., both horizontal and vertical) and a well-defined CRUD API for working with them.

The potential size of the user group is the same as that of the Perseus Digital Library, i.e., currently about 500,000 unique users

Why EUDAT?

We believe that the Data Types Registry and the PID Types API are both relevant to this use case. The Data Types Registry provides a data model for formally expressing data types, an API for CRUD and Query operations on a data type registry, and fulfillment of a dependency for the PID Types API. The PID Types API provides a conceptual model for a PID record and a CRUD API for interacting with PID records which can abstract the differences between PID implementations. We would like to begin to formalize the data types we are referencing here, possibly through use of the Data Types Registry. And we would like to explore use of the PID Types API to abstract the differences between different identifier types, as ultimately we expect this to be integrated with other components of the Perseus infrastructure, which uses a variety of different identifier types.

Expected outcomes

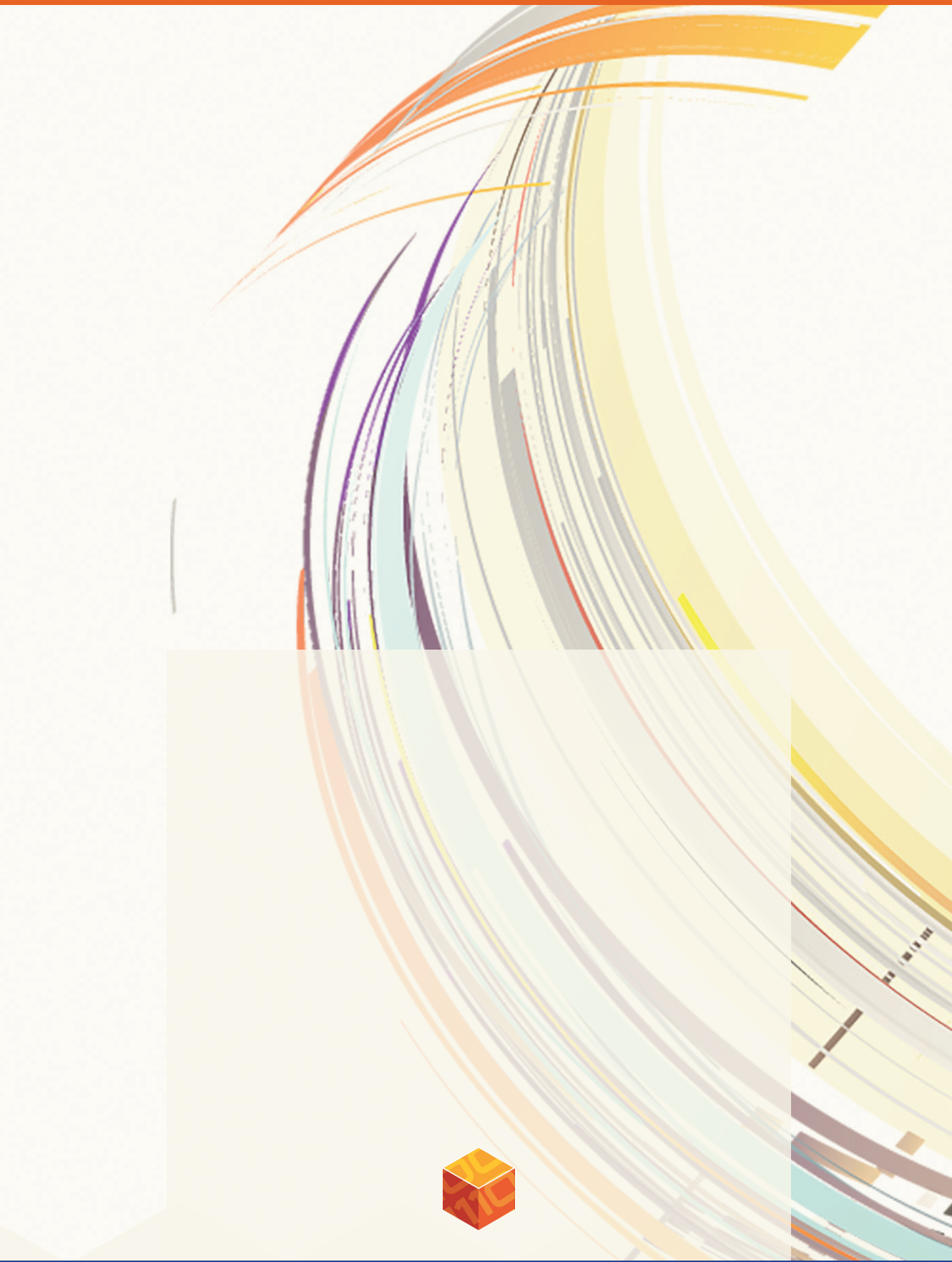
The concrete benefits of using EUDAT would be that it gives our data types domain neutrality, it improves our data management practices, it enhances the ability for our data to be reused by others because the data would be more clearly and thoughtfully expressed and because EUDAT provides a documented API for interacting with it. It would also give us the ability to scale more rapidly to support new PID types.

EUDAT could also provide the potential benefits of

- greater sustainability
  - if implementations of the APIs are built and maintained by diverse communities and projects
- greater bi-directional interoperability
  - if other projects with whom we want to share data find the same benefit and implementations
- Institutional support and long-term preservation
  - if our libraries and institutional repositories also deploy and implement the solutions

Expected domain legacy

Our goal at the Open Philology Project is to produce high quality, digital editions of every ancient work. EUDAT will help us to reach this goal by providing a way for us to store and reference our data, making it available to users on the outside. This availability will not only be to use the data but also to manipulate and improve it. In the realm of OCR, the best example of this will be the ability to access a page image and its resultant OCR results through the OCR proof reader we have designed, allowing them to check and correct the OCR results. Without EUDAT, we would be able to do this only on a very small scale. EUDAT will make it possible for any user of the Perseus Digital Library to engage in this vital activity of textual enhancement.



European Commission | EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065.