EUDAT

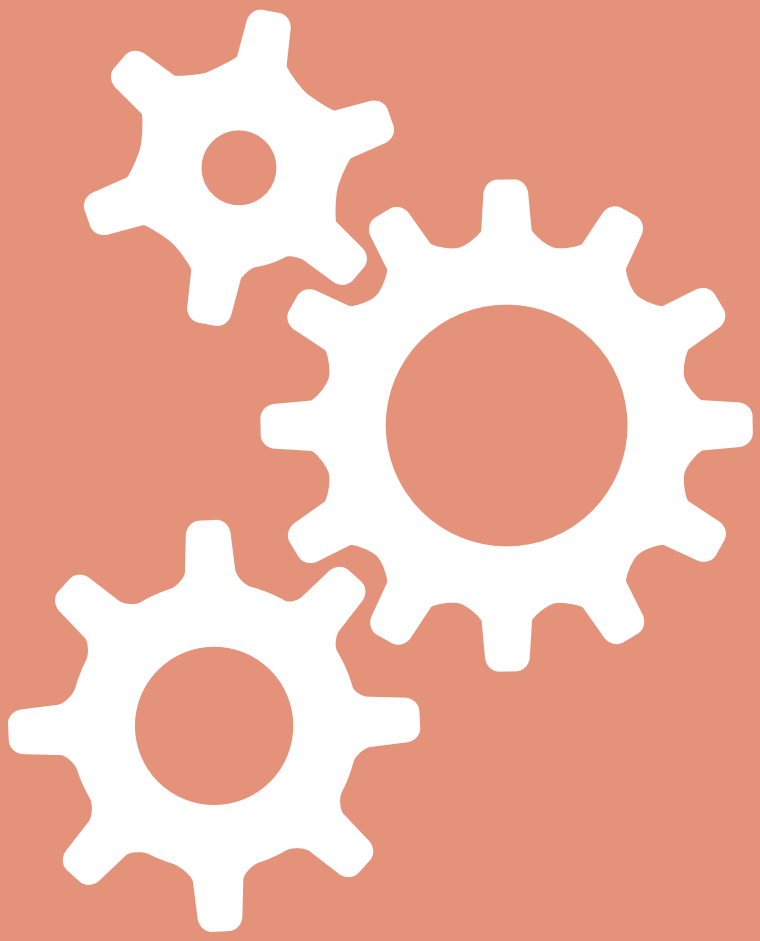
EUDAT Community Engagement

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

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.

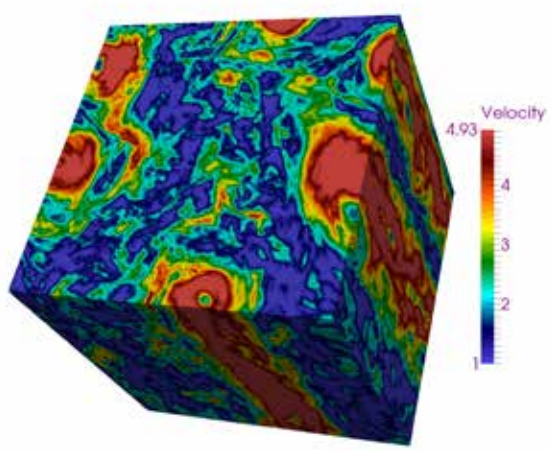
Turbase DNS

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.

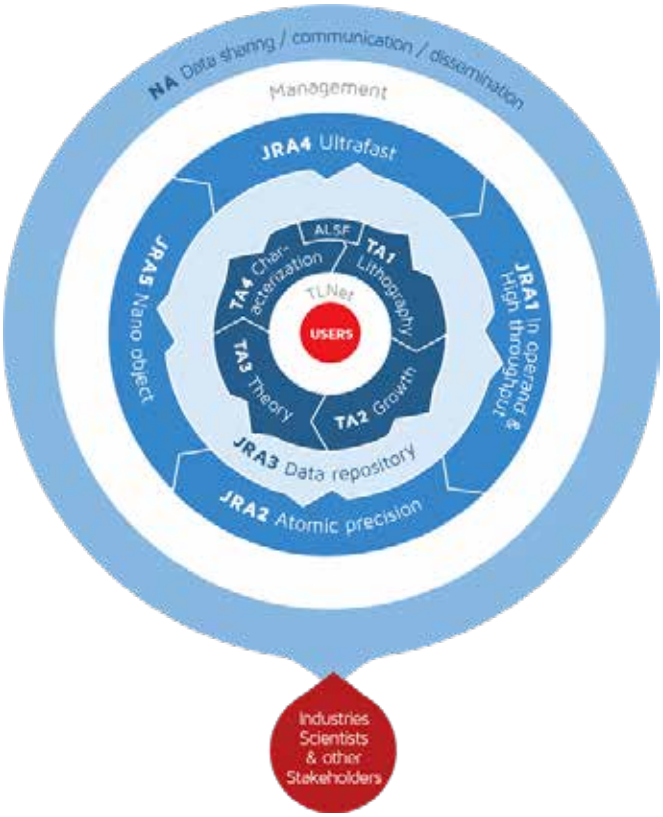
The active use of B2SAFE is currently considered a more experimental activity compared to B2SHARE or B2FIND. NFFA-EUROOPE hopes to identify suitable partners and compelling use cases where B2SAFE is going to bring benefits to the nanoscience community without breach of data management policies or unreasonable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

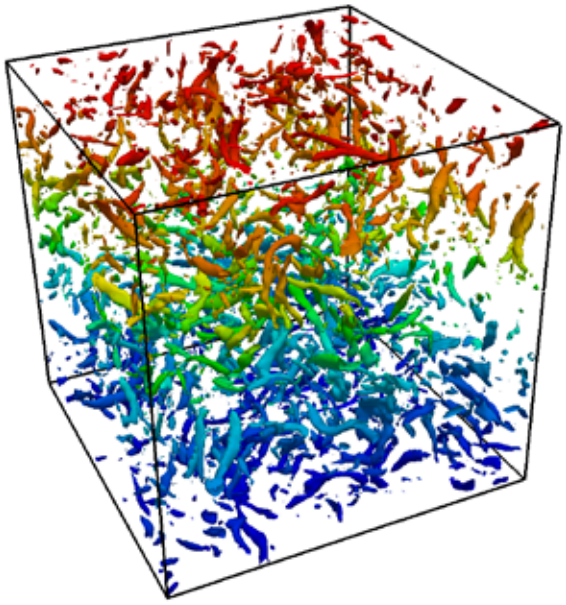
Direct simulation of turbulent flows produces large amounts of data that can be used for multiple purposes and multiple researchers besides those originating them. After initial publication, those data can be shared freely worldwide. The community has come to expect that to be done, since large simulations 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

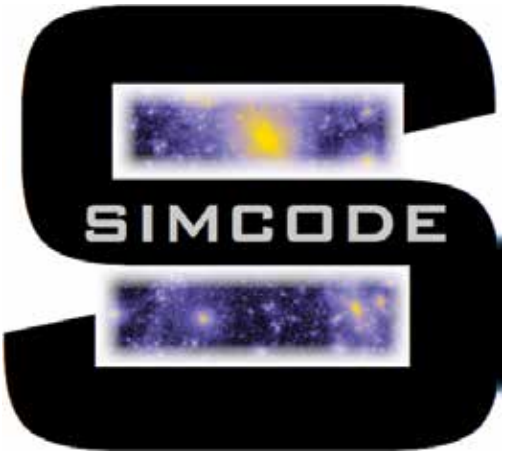
Cosmological simulations are an essential ingredient for the success of the next decade of "Precision Cosmology" observations, including also large and costly space missions as e.g. the Euclid satellite. Since the required precision and the need to test for statistical anomalies, astrophysical contamination, 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.

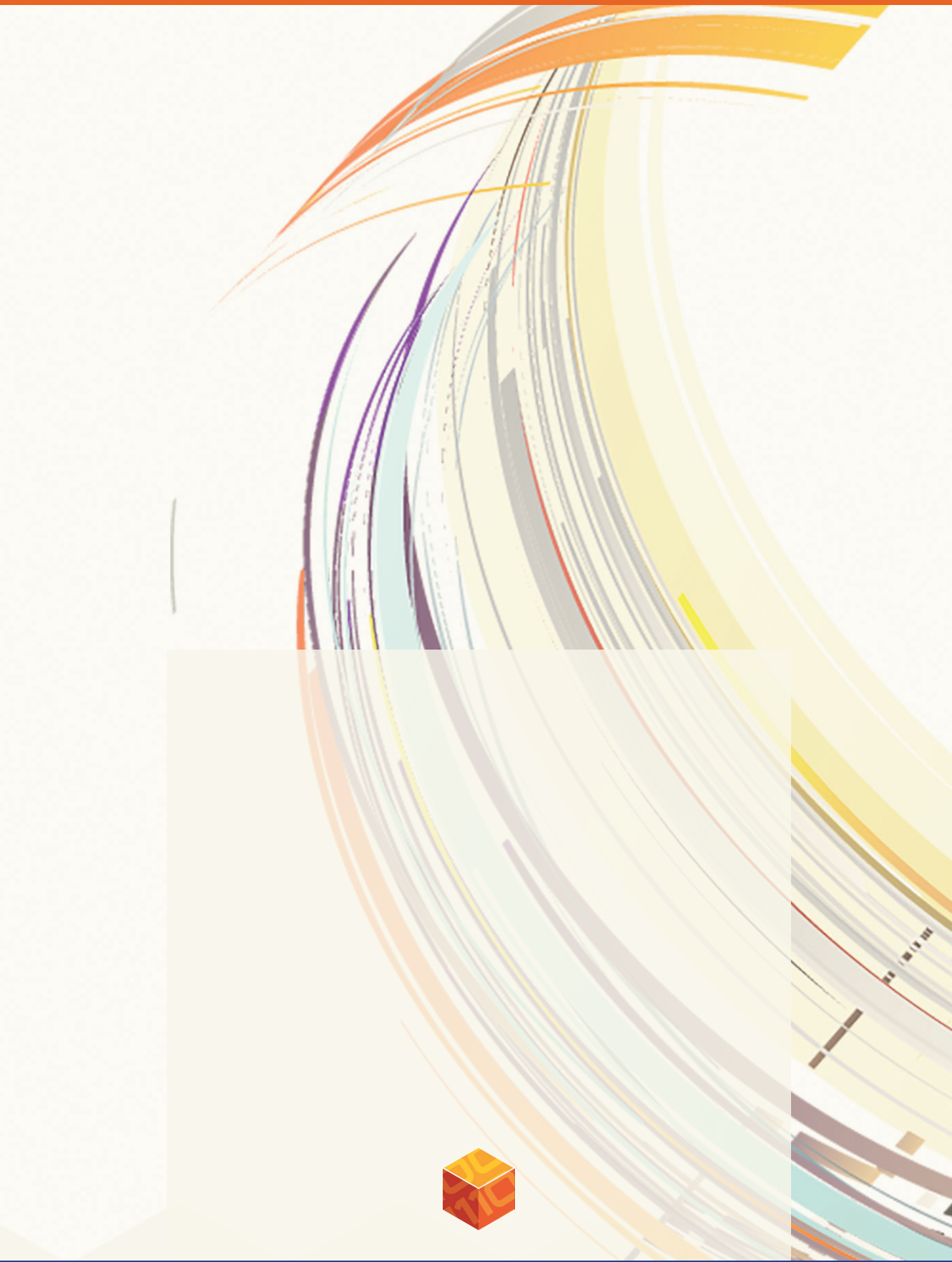
Expected outcomes



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.



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