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

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).

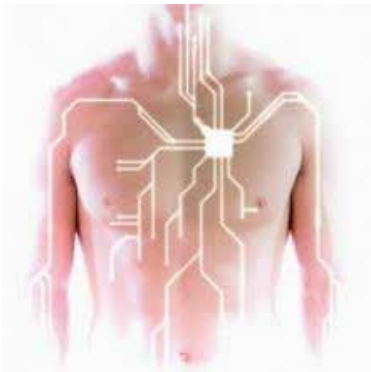
Main benefits for the community

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

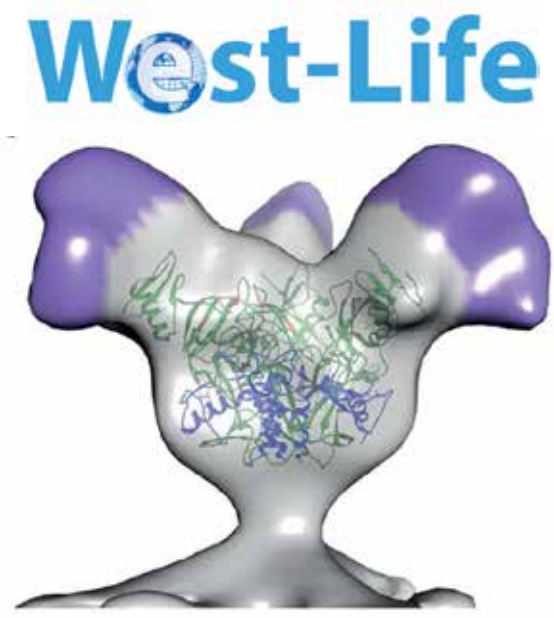
The use community consists of a few thousand scientists. Structural biologists used to identify themselves by their preferred technique (as "crystallographers", "electron microscopists" etc). Increasingly, they are targeting larger macromolecular complexes, so research projects now must combine several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

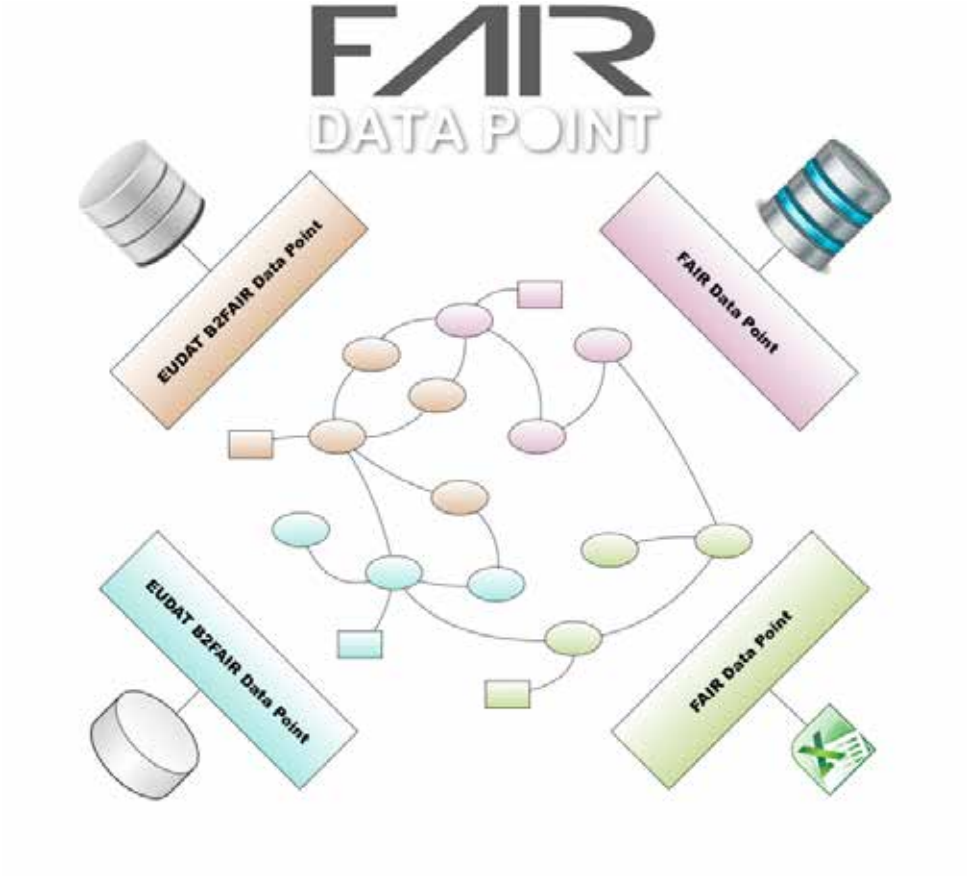
Within this pilot, we aim to implement and deploy a FDP using a combination of existing Semantic Web standards and frameworks for the front-end, and (existing or new) EUDAT services for the back-end. A FDP provides access to the data and metadata using REST-APIs conforming to the W3C Linked Data Platform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

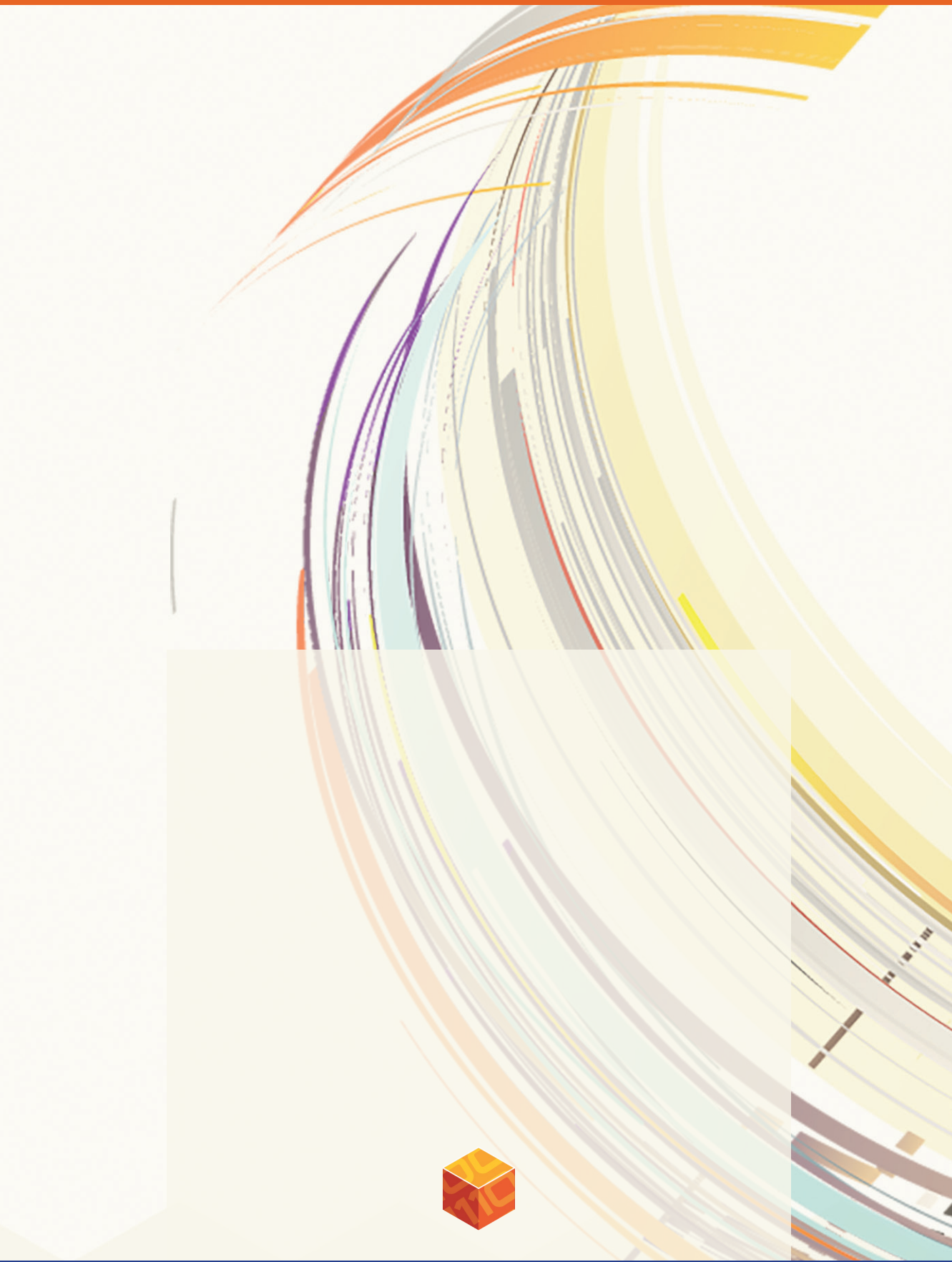
complying with the FAIR Data Point behaviours and, therefore, adhering to the FAIR Data principles by offering metadata and data in a FAIR manner.

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.



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