



DICE

Data Infrastructure Capacity for EOSC

Deliverable D5.1

Pilots description and validation

Responsible Partner:	Barcelona Supercomputing Center
Status-Version:	Final version- v1.0
Date:	29/09/2021
Distribution level (CO, PU):	PU

Project Number:	GA 101017207
Project Title:	DICE: Data infrastructure capacity for EOSC

Title of Deliverable:	Pilots description and validation
Due Date of Delivery to the EC	30.09.2021
Actual Date of Delivery to the EC	29.09.2021

Work package responsible for the Deliverable:	WP5 – Integration with Community platforms
Editor(s):	Nadia Tonello - BSC
Contributor(s):	Mika, A. – ASTRON Holties, H. - ASTRON Zarrabi, N. – SURF Vermeulen, A. - ULUND
Reviewer(s):	Testi, D. - CINECA Pursula, A. - CSC
Recommended/mandatory readers:	WP2, WP4

Abstract:	This deliverable describes the use cases pilot design for integration of the platforms selected, and validation tests in progress or planned to demonstrate their impact.
Keyword List:	Use case, pilot, services, integration
Disclaimer	This document reflects only the author's views and neither Agency nor the Commission are responsible for any use that may be made of the information contained therein



Document Description

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v0.1	19.08.2021	First draft version	BSC
v0.2	01.09.2021	Comments and suggestions received by WP5 partners	ALL
V0.3	10.09.2021	Last draft version for review	ALL
v1.0	28.09.2021	Final version addressing reviewers comments	BSC



Table of Contents

Table of Contents	4
List of Figures	4
List of Tables.....	5
Terms and abbreviations.....	6
Executive Summary.....	7
1 Introduction	8
1.1 About this deliverable.....	8
1.2 Document structure.....	8
2 CompBioMed pilot	9
2.1 The CompBioMed community.....	9
2.2 Integration use case design	9
2.3 Validation tests	11
2.4 Status and timing of integration tasks.....	12
2.5 Expected impact for the community	12
3 LOFAR pilot.....	13
3.1 The LOFAR community	13
3.2 LOFAR Integration use case design.....	13
3.3 LOFAR use case validation tests.....	14
3.4 Status and timing.....	15
3.5 Expected impact for the community	16
4 ICOS pilot.....	17
4.1 The ICOS community	17
4.2 Integration use case design	17
4.3 Validation tests	18
4.4 Status and timing.....	18
4.5 Expected impact for the community	19
5 Conclusions	20

List of Figures

FIGURE 1 COMPBIOMED WORKFLOW	10
FIGURE 2 LOFAR INSTRUMENT DATA IS CAPTURED IN THE CENTRAL PROCESSING FACILITY HOSTED BY THE UNIVERSITY OF GRONINGEN AND ARCHIVED AT ONE OF THE LOFAR LONG TERM ARCHIVE DATA CENTERS. ADVANCED DATA PRODUCTS ARE GENERATED, PRIMARILY ON COMPUTE INFRASTRUCTURE CO-LOCATED WITH ARCHIVE STORAGE, AND DEPOSITED IN THE SURF HOSTED SCIENCE DATA REPOSITORY IN AMSTERDAM. PUBLISHED DATA COLLECTIONS ARE REGISTERED IN THE VIRTUAL OBSERVATORY. PUBLISHED DATA COLLECTIONS ARE HARVESTED BY THE B2FIND SERVICE HOSTED BY DKRZ.....	15



FIGURE 3- SAFE STORAGE OF THE RESEARCH DATA WILL BE GRANTED BY B2SAFE AT FJZ AND CSC. SURF IS PROVIDING PIDS FOR DATASETS. B2SHARE AT CSC WILL GIVE A GENERIC PUBLICATION PLATFORM FOR DATASETS FROM RESEARCH DATA ANALYSIS, AS AN ALTERNATIVE TO THE ICOS PORTAL, SITED AT UNIVERSITY OF LUND. 18

List of Tables

TABLE 1: DICE SERVICES FOR COMBIO MED USE CASE.....	10
TABLE 2: COMBIO MED USE CASE VALIDATION TESTS.....	11
TABLE 3 COMBIO MED USE CASE STATUS AND TIMING OF THE INTEGRATION TASKS	12
TABLE 5 LOFAR USE CASE VALIDATION TESTS.....	14
TABLE 6 LOFAR USE CASE INTEGRATION STATUS AND TIMING	15
TABLE 7 DICE SERVICES FOR ICOS USE CASE	17
TABLE 8 ICOS USE CASE VALIDATION TESTS.....	18
TABLE 9 ICOS USE CASE STATUS AND TIMING OF THE INTEGRATION TASKS.....	18



Terms and abbreviations

ASTRON	Astron
BSC	Barcelona Supercomputing Center - Centro Nacional de Supercomputación
CESNET	CESNET, z. s. p. o.
CINECA	Cineca
CoE	Centre of Excellence
CSC	CSC – Tieteen Tietotekniikan Keskus Oy
Cyl	The Cyprus Institute
Datacite	DataCite
DKRZ	Deutsches Klimarechenzentrum GmbH
DoA	Description of Action
EC	European Commission
EOSC	European Open Science Cloud
ETHZ	Eidgenössische Technische Hochschule Zürich
EU	European Union
FZJ	Forschungszentrum Juelich GmbH
GA	Grant Agreement to the project
GRNET	National Infrastructures for research and technology
GWDC	Gesellschaft für Wissenschaftliche Datenverarbeitung mbh Göttingen
INFN	Istituto Nazionale di Fisica Nucleare
IT4I	Vysoka Skola Banská - Technická Univerzita Ostrava
KIT	Karlsruhe Institut für Technologie
KNAW-DANS	Koninklijke Nederlandse Akademie van Wetenschappen
KPI	Key Performance Indicator
MPG	Max Planck Gesellschaft zur Förderung der Wissenschaften e.V.
PID	Persistent Identifier
SIGMA	SIGMA2
SNIC	Uppsala Universitet
SURF	SURFsara BV
TRUST	Trust-IT services
UCL	University College London
ULUND	University of Lund
VA	Virtual Access
WP	Work Package



Executive Summary

Effectiveness of the data services integration provided within DICE is a key factor to achieve the objectives of the project. The three use cases serve both as demonstration and direct channel to get feedback to improve DICE services integration with three mature communities' platforms in different scientific domains: biomedicine, radio astronomy, and environmental sciences.

This deliverable contains the work plan to carry out for the three use cases included as demonstrators in the DICE project.

The use cases integration plans have been concretized into defined operational tasks, where the services offered by the project are integrated with existing tools and platforms to serve the corresponding scientific communities.

This deliverable is the first of two that WP5 will produce. The first period of the project, called pilot phase, consists of the planning and first testing of the services to be fully integrated during the next phases of the work, until reaching the production and exploitation phase. This first deliverable contains the plan and pilot design, with the list of the tests to be executed and planned to prove the effectiveness of the solution adopted and promptly have feedback of the work done. For each of the three use cases, this document presents the goal, scope of the use cases inside the community, the DICE services to be integrated, and the execution plan. Finally, the validation tests and their actual status are presented.

The underway work of WP5 – Integration of DICE services with communities' platforms – is carried out in collaboration with WP4 – Integration with other Services and Platforms – such to constitute a feedback for improvement of the DICE services interoperability and technical integration. The collaboration with WP2 - Outreach, stakeholder engagement and service uptake – consists in constant communication, giving content, and disseminate the success stories of the WP5 use cases to the scientific community.

A second deliverable (D5.2) will be presented at the end of the project (M30), with the final integration description, the feedback from communities, and the evaluation of the impact of the work done, in collaboration with the Community Advisory Board, for the extension of the lesson learned extended to more scientific communities.



1 Introduction




The Data Infrastructure Capacities for EOSC (DICE) consortium brings together a network of computing and data centres, and research infrastructures, for the purpose to enable a European storage and data management infrastructure for EOSC, providing generic services and building blocks to store, find, access and process data in a consistent and persistent way. Specifically, DICE partners will offer 14 state-of-the-art data management services together with more than 50 PB of storage capacity. The service and resource provisioning will be accompanied by enhancing the current service offering in order to fill the gaps still present to the support of the entire research data lifecycle; solutions will be provided for increasing the quality of data and their re-usability, supporting long term preservation, managing sensitive data, and bridging between data and computing resources.

1.1 About this deliverable






The purpose of this document is to describe the work plan for the three DICE use cases, to be fully developed over the 30-months duration of the project. This essential initial planning will be followed and updated during the evolution of the project.

1.2 Document structure

Three sections are dedicated to the three communities involved as demonstrators:

-  CompBioMed,
-  LOFAR,
-  and ICOS.

Each section will contain:

-  Description of the community involved, relevant to the use case focus of the work in DICE,
-  The use case design (services to be integrated and providers involved),
-  Description of the validation tests which will prove the effectiveness of the solution,
-  Timing: status and evolution plan,
-  Expected impact of the foreseen results on the community.

The final conclusions will summarize the activity we are carrying out in WP5 with respect to WP4, which will receive the feedback for improvement of the services integration, and with WP2, in charge of the dissemination and outreach activities, in which the use cases activities have an important role.



2 CompBioMed pilot

Within the CompBioMed community, the use case considered within DICE is to put in place two workflows for data replication and data publication. In the data replication workflow, we plan to make a federation of the High-Performance Computing (HPC) centres involved in the pilot (BSC, SURF and potentially UCL) in order to enable share and exchange of large data among institutions that are using those HPC facilities. In the data publication workflow, the aim is to provide the possibility to publish the results of simulations or final data in an open data repository to be findable and accessible by the wider community in long-term. Another objective is to promote access to the workflow to other HPC centres (e.g. LRZ, EPCC), research and medical centres in the community.

2.1 The CompBioMed community

The CompBioMed Center of Excellence (CoE) seeks to exploit the third pillar of science in order to render predictive models of health and disease more relevant to clinical practice by providing a personalized aspect to treatment. One of the clear trends in the biomedical community is its ever-increasing demand for storing more data as well as the transfer, management and longer-term preservation of this data. Frequently, large data sets need to be moved closer to High Performance Computing (HPC) services prior to performing computational work. Once the computational work is done, the resulting data is then moved to somewhere else or kept closer to the HPC services for post processing work. This use case addresses the need for safe data replication and large data transfer within a system that can support a FAIR data cycle, an important data requirement within this international community.

2.2 Integration use case design

This use case work to setup the workflows, focuses on the development of a data management solutions to facilitate Alya¹ simulations using large datasets and to explore the capabilities and challenges towards the use of Alya for Exascale simulations. This work will be also supported by parallel activities within the CompBioMed consortium, where the results of this work will be extended, and it will promote the access to the workflow to different HPC (e.g.: LRZ, EPCC), research and medical centres in the future.

The services in this use case will treat research synthetic and simulation data, no sensitive data will be treated. A possible extension to sensitive data treatment will be explored in collaboration with DICE WP4 during the evolution of the project.

For the data replication workflow, the B2SAFE² and B2HANDLE³ services will be used, both part of the DICE offering. SURF already has B2SAFE as a running service. The plan includes the deployment of an instance of B2SAFE in BSC (and potentially in UCL) and the federation of the nodes, which facilitates replicating data.

For the Data Publication workflow, an own instance of B2SHARE⁴ will be deployed in UCL. The reason for this is that the data being generated in CompBioMed is large and does not fit to be published in the community instance of B2SHARE, which is a paid facility for big data such as the

¹ <https://www.bsc.es/research-development/research-areas/engineering-simulations/alya-high-performance-computational>

² <https://www.eudat.eu/services/b2safe>

³ <https://www.eudat.eu/services/b2handle>

⁴ <https://www.eudat.eu/services/b2share>



ones produced by CompBioMed⁵, the proposal which got accepted for the second phase of the Computational Biomedicine Centre of Excellence (CoE) (> TB).

Table 1: DICE services for CompBioMed use case

Service	Description	Resources Needed	Provider
B2SHARE	Data Repository for data publication. Metadata schema can be implemented in this repository. Integration with B2FIND for harvesting data and facilitating findability of the data.	50 TB	UCL
B2HANDLE	Tool required to make persistent identifiers (PIDs) for the data to facilitate findability of the data. The PIDs will potentially be used in B2SAFE and B2SHARE.	1 prefix 10000 PIDs	SURF
B2SAFE	Data staging and safe replication of research data between HPC centres in CompBioMed. The archival storage on tape facilitates long-term preservation of the data.	50 TB 50 TB	SURF BSC

The workflow to implement and the interconnection between the services is schematically shown in Figure 1.

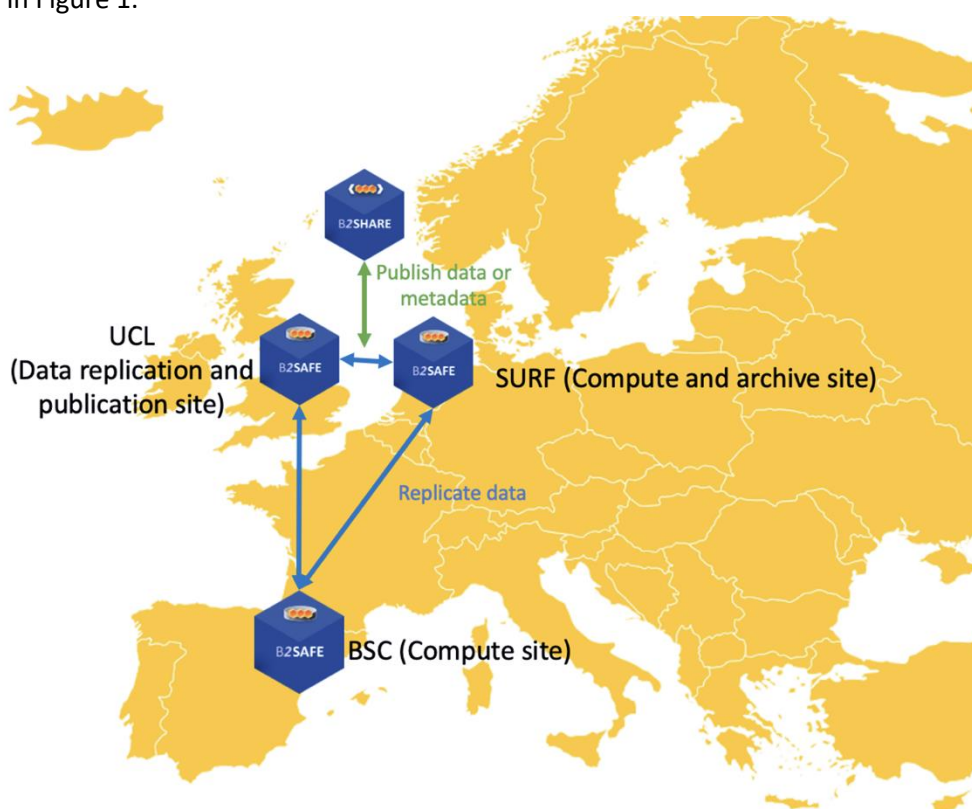


Figure 1 CompBioMed Workflow

⁵ <https://cordis.europa.eu/project/id/823712>



2.3 Validation tests

Table 2: CompBioMed use case validation tests

Requirement	Validation	Expected result	Comments
Simulated data produced at CompBioMed HPC centres BSC or SURF must be accessed to CompBioMed analysis centres	First data replication between B2SAFE sites: transfer with reduced data set (30Gb, total 1Tb)	B2SAFE is deployed at BSC and UCL. Services are properly configured. Resources are allocated. The data replica test will be moderated by SURF and performances compared with reference values.	The configuration will be automated to deploy B2SAFE potentially also to other sites (LRZ and EPCC).
CompBioMed data are stored in long term archive facility for long-term preservation	Data replicated to B2SAFE at SURF will land on a Data Archive facility based on tape to ensure long-term preservation of the data.	Connection between B2SAFE and Data Archive at SURF is in place	Costs for the long-term storage to be reimburse via Virtual Access for the project duration. Sustainability after the project to be discussed among service providers and CompBioMed Consortium, within “Long-term sustainability” task in CompBioMed2 ⁶ .
CompBioMed data will be published in an open data repository	Final data processed at UCL or BSC will be published in a B2SHARE instance at UCL or at SURF Data Repository (SDR)	B2SHARE service deployed at UCL Published data with PIDs assigned in B2SHARE at UCL or SDR at SURF	Metadata schema definition for CompBioMed is to be done in CompBioMed project WP3, Task T3.4.
Scientists are able to execute simulations, and services are accessible to the users’ community.	Transfer real data and automatic execution of workflows: Automatic execution of the workflow for Alya simulations on distributed HPC systems.	The services are running in a production environment and made accessible to the users’ community.	Extension of the workflow to other use cases will be explored.

⁶ See also the deliverable D4.1 on CompBioMed2 compute and data services strategic plan: https://www.compbiomed.eu/wp-content/uploads/2020/11/D4.1_CompBioMed2_Compute_and_Data_Services_Strategic_Plan_v1.0.pdf



2.4 Status and timing of integration tasks

Requirements and validation tests are here separated into tasks, with their planification. The timing will be followed during the execution of the project and updated whenever needed until the end of the project. Other tasks, like the dissemination done in collaboration with WP2, are not listed here, since we are considered just the services integration activities.

Table 3 CompBioMed use case status and timing of the integration tasks

Task	Status	Planned Due time
Elaboration of the integration plan	Done	July 2021
Deployment and configuration of the services	In progress	November 2021
Data replication test	Planned	January 2022
Connection with long-term storage	Planned	January 2022
Connection with publication site	Planned	March 2022
Execution of the automated workflow, accessible to the users' community	Planned	Mid 2022

2.5 Expected impact for the community

In this collaboration, we are setting up workflows for data replication and publication within the CompBioMed community, using the existing services of DICE and EUDAT. One major impact for the community is alignment with international standards and approaches for FAIR data management. The platform we are building, promotes FAIR data by enabling sharing and reuse of the data within the community. Generated data will be stored and archived for long-term preservation and published in open data repositories to ensure findability and reusability of the data.

Moreover, the CompBioMed CoE is moving towards exascale in compute. Processing large data will require an infrastructure that can handle storing and transferring large amounts of data as well. The data infrastructure we build, will facilitate large data exchange between multiple international sites.



3 LOFAR pilot

The work of this use case will develop a data service that supports manual and automated ingestion of processed data, in particular integrating automated data processing services running on compute clusters co-located with the LOFAR Long Term Archive data storage infrastructure. The proposed case addresses all aspects from Findability, Accessibility, Interoperability, and Reproducibility principles with a focus on the first two (FA).

3.1 The LOFAR community

The LOFAR Observatory operates LOFAR, a unique radio astronomical instrument with stations distributed over Europe, interconnected through a 10 Gbps wide area network and a central processing facility hosted by the University of Groningen. Following initial processing on the central processing facility, with the objective to assess quality and reduce data volume, observation data is stored in the Long-Term Archive (LTA). All data is registered in a central catalogue of LOFAR data (<https://lta.lofar.org>) and stored in one of three associated data centres hosted by SURF in the Netherlands, FZJ in Germany, and PSNC in Poland. The total volume of archived data is approximately 50 Petabyte (early 2020).

The revolutionary multi-beaming capabilities of the LOFAR telescope allow astronomers to engage in multiple lines of research at once: they can look back billions of years to a time before the first stars and galaxies were formed (the so-called 'Dark Ages'), they can survey vast areas of the low-frequency radio sky, and they can be constantly on the lookout for radio transients originating from some of the most energetic explosions in the universe.

ASTRON has started preparing for offering further data services associated with the instrument data archives. Initially in the EOSCpilot project, followed up in the EOSC-Hub project where LOFAR is at the core of the Radio Astronomy Competence Center, and now also as a partner in the ESCAPE project, various aspects of integration with the EOSC infrastructure are evaluated and prototyped. In 2020, ASTRON has formed the Science Data Center program to develop and offer new and enhanced data services for unlocking the science potential of the data archives it operates. As a first large-scale activity, the LOFAR Data Valorization project has started with the generation of value-added data products from existing data in the archive with a focus on providing homogenized curated data that forms the basis for further processing and generation of science level data products. The developed framework is offered as a user requested data processing service, supported by the EGI-ACE project, and will deposit advanced data products in a science data repository, targeted to be built on an integration of the EUDAT B2SHARE service and dCache through the DICE project. Service access will be offered via AARC compliant federated AAI solution where LOFAR community management is implemented to handle community and science project membership for use by the integrated services.

3.2 LOFAR Integration use case design

The proposed case will follow-up in particular on the deployment and further development of the advanced data product repository, supporting initial operations for safe storage, discoverability, and distribution of science-level data that has been generated by user- and ASTRON managed processing services connected to the LOFAR LTA. Since the processed data will typically be orders of magnitude smaller than the source (instrument) data, it is considered that the advanced data-product repository can in its initial phase be centrally hosted (i.e. hosted by a single partner). It is considered to offer data analysis services with a low demand on compute resources in conjunction with the repository using an EOSC cloud compute service, e.g. a cut-out service to generate specific cut out slices from data stored in the repository.



Table 4 DICE services required by LOFAR use case

Service	Description	Resources Needs	Provider
B2SHARE/ SURF Data Repository (SDR)	Differentiate between active data (e.g. images/cubes) for direct access and less active data (e.g. visibilities) for user requested access. Support data access as a linked service by VO software. Scalability requirement is to (technically) allow growth to a petabyte by 2023.	500 TB	SURF
B2HANDLE	Registration of LOFAR scoped PIDs for data products.	1 prefix ~15000 PIDs	SURF
B2FIND	DKRZ has already developed an integration with the astronomy standards based Virtual Observatory (VO) by harvesting published VO datasets. The service is to be enabled to support harvesting of metadata by VO and B2FIND.	Some tens of datasets, encompassing approximately 10,000 data objects.	DKRZ

3.3 LOFAR use case validation tests

Table 5 LOFAR use case validation tests

Requirement	Description of the validation test	Expected result	Impact, comments
LOFAR data collections are stored, preserved and shared in a generic data service	POC ingest of a LOFAR data collection (B2SHARE/SDR)	Fully described data collection registered, findable, and retrievable in accordance with FAIR principles. PID's associated with the collection and individual data products.	Gain experience for final integration and preparation of publication of a LOFAR data collection.
LOFAR Data are findable by the scientific community in a multidisciplinary discovery data service	Publish a LOFAR data collection into B2FIND	Initial LOFAR metadata schema defined & implemented in science data repository First LOFAR data release ingested	Increased visibility and interest for LOFAR science level data. Improved traceability, findability and attribution for the data collection as compared to current practices
LOFAR Data are automatically registered and published to be discovered and	Automation of data registration	LOFAR data ingest service as part of the LOFAR workflow framework implemented	The integrated services are running in a production environment and



accessed by all the scientific community			made accessible to the user community. (Transition of pre-production to production not expected to incur additional work)
--	--	--	--

The interconnection between the services is schematically shown in Figure 2.

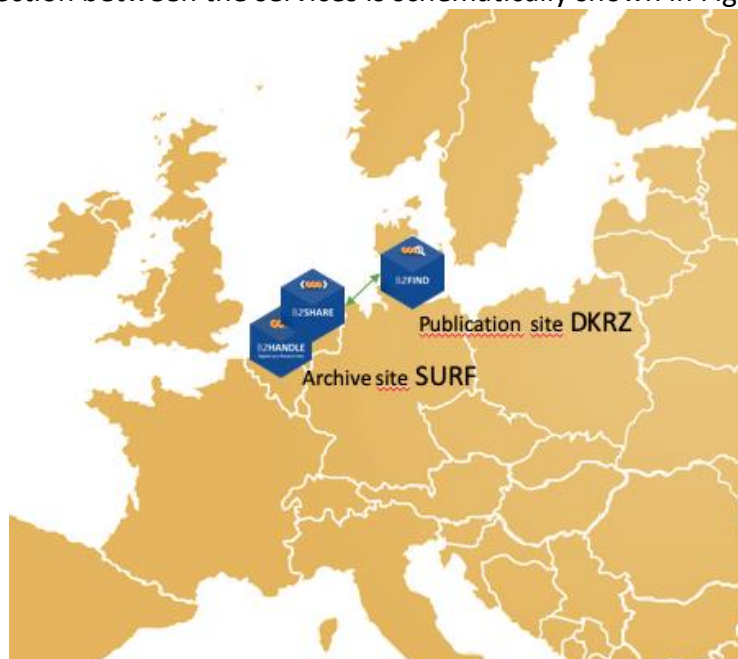


Figure 2 LOFAR instrument data is captured in the Central Processing facility hosted by the University of Groningen and archived at one of the LOFAR Long Term Archive data centres. Advanced data products are generated, primarily on compute infrastructure co-located with archive storage, and deposited in the SURF hosted science data repository in Amsterdam. Published data collections are registered in the Virtual Observatory. Published data collections are harvested by the B2FIND service hosted by DKRZ.

3.4 Status and timing

Requirements, listed in Table 5, and validation tests are here described as actions and tasks, with their status and planification. The timing will be followed during the execution of the project and updated whenever needed until the end of the project. Other tasks, like the dissemination done in collaboration with WP2, are not listed here, since we are considered just the services integration activities.

Table 6 LOFAR use case integration status and timing

Task	Status	Planned Due time
Elaboration of the integration plan	Done	August 2021
POC ingest of a LOFAR data collection	Planned	October 2021
Publishing a LOFAR data collection	Planned	December 2021



Automation of data registration	Planned	March 2022
Operation and exploitation	Planned	December 2021 – June 2023

3.5 Expected impact for the community

By generating science-ready data LOFAR will attract a much wider community to exploit its data for scientific aims including cosmology, extragalactic, and galactic astrophysics. The impact will be further enhanced by depositing the data in a common science data repository that follows FAIR principles which also has the potential of attracting the interest of a wider public outside of the astronomical community. The actions undertaken within the DICE project are therefore expected to support the generation of a dramatic increase of the science output from the instrument.

Offering advanced data products will result in a more attractive LOFAR data archive as it will bring data to a higher level of pre-processing, attracting science users to data in the LOFAR LTA that do not necessarily have a high level of expertise with respect to the LOFAR instrument, but also because the size of this type of dataset will be less demanding than the lower level instrument data products currently offered in the LOFAR Long Term Archive, making them easier for scientists to access and handle for further analysis.

Integration with DICE supported infrastructure is expected to result in increased visibility and interest for LOFAR science level data. The integrated support for persistent identifiers and metadata standards enables automatic harvesting by data discovery services such as B2FIND and will improve traceability, findability and attribution for data.



4 ICOS pilot

This work will extend the ICOS community platform from a Jupyter based data analysis and cooperation tool where the model results are shared, compared and analysed, into a benchmarking environment that controls the whole processing chain from selection and preparing the prior information datasets, running the transport models to be tested, consecutively running the Bayesian inversion method on the model results, followed by the benchmarking and analysis of the results.

4.1 The ICOS community

ICOS is a pan-European research infrastructure for quantifying and understanding Europe's greenhouse gas balance. Its mission is to collect high-quality observational data and to promote its use, e.g. to model greenhouse gas fluxes or to support verification of emission data. ICOS brings together more than 130 measurement stations across the atmosphere, ecosystem and marine domains.

This task will be focused on the methane emission analysis. More complex cases are planned that would demonstrate the modularity of the system by application of the approach to the assessment of natural CO₂ fluxes over Europe using two models' setups from an earlier comparison in the framework of the EUROCOM project. Three inversion model setups (Carbon-tracker, LUMIA and STILT) are already in operational use at the institutes that work together in Carbon Portal (ULUND and Wageningen University).

4.2 Integration use case design

This task will utilize the EUDAT data service B2SAFE for storing the data from ICOS and staging it for analysis at computing platforms. The suitability of B2SHARE service for publishing results in addition to the ICOS Carbon Portal is investigated. ICOS data in B2SAFE will be given persistent identifiers via the B2HANDLE service in addition to the main Handle PIDs directly minted by ICOS that are based on the data object checksum. The case addresses all aspects from Findability, Accessibility, Interoperability, and Reproducibility principles, with a focus on the reproducibility.

Table 7 DICE services for ICOS use case

Service	Description	Resources Needs	Provider
B2SHARE	Assessing the possibility to publish datasets results from analysing the research data.	1-5 TB	CSC
B2SAFE	Safe storage for the research data, with staging to computing platforms.	50-500 TB	CSC, FZJ
B2HANDLE	Standard data persistent identifiers, in addition to the present ICOS identifiers in the ICOS Portal.	1 prefix, 100000 PIDs/year	SURF



4.3 Validation tests

Table 8 ICOS use case validation tests

Requirement	Description of the validation test	Expected result	Impact, comments
Data must be securely stored in B2SAFE	All ICOS data is at production automatically streamed to the B2SAFE repository at CSC	Automation of data registration	Storage in the remote repository is in production
Jupyter interface to inversion code and results are published in B2SAFE and B2SHARE	Publication of ICOS data through B2SHARE	Data available at B2SHARE	Data is published

The interconnection between the services is schematically shown in Figure 3.



Figure 3 Safe storage of the research data will be granted by B2SAFE at FJZ and CSC. SURF is providing PIDs for datasets. B2SHARE at CSC will give a generic publication platform for datasets from research data analysis, as an alternative to the ICOS Portal, sited at University of Lund.

4.4 Status and timing

Requirements and validation tests are here separated into tasks, with their planification. The timing will be followed during the execution of the project and updated whenever needed until the end of the project. Other tasks, like the dissemination done in collaboration with WP2, are not listed here, since we are considered just the services integration activities.



Table 9 ICOS use case status and timing of the integration tasks

Task	Status	Planned Due Time
Automate transfer at real time	Done	April 2021
Port of inversion code to production environment	Ongoing	September 2021
Develop Jupyter interface to inversion code including publication of result at B2SAFE and B2SHARE	Ongoing	April 2022
Publish Jupyter VM on ICOS Portal	Planned	May 2022
Organise webinars and introduce at summer school(s)	Planned	September 2022

4.5 Expected impact for the community

The publication of a near-real time flux inversion system based on observations from the ICOS network would be a revolutionary breakthrough in the timely availability of carbon cycle balance information, that right now is only available after delays of at least one year and at relatively low resolution.

An interactive interface with the possibility of generating custom data sets that are published on B2SHARE including a transparent data workflow will also be unique. Some training will be needed for the community to understand the possibilities and limitations of the methodology used in the inversion model.



5 Conclusions

This deliverable describes the initial plan, as established during the first months of project activity, of three use cases. Adaptable, generic, and scalable DICE services are integrated into mature communities, as demonstrators.

The three communities' platforms involved in the use cases have been set up for serving three different scientific areas: Life Sciences (CompBioMed), Radio-astronomy (Lofar), and Environmental Sciences (ICOS). Their requirements can be fulfilled with generic data services, with significant impact to their communities.

Having a positive, tangible impact is one of the main objectives of the project. The use cases experience will demonstrate the advantages of having generic services to share, preserve and bring data closer to computing. For this reason, the use cases have been selected to be rather circumscribed and very specific, to be able to start the exploitation phase during the lifetime of the project. The impact of the use cases goes beyond their communities thanks to the availability of the shared data to everyone, independently of the discipline. Automation makes the sharing a nearly real-time procedure, according to the Open Science principles.

The Community Advisory Board (CAB) of the project will be kept continuously informed about the integration activities, for the CAB members to share their experience and give further feedback from the point of view of their respective scientific communities.

The integration experience of the use cases is shared in and out of the scientific community, among the CAB members, and the rest of the consortium members, in collaboration with WP2, with the organization of webinars and other dissemination activities.

This deliverable is a live document. Its final update represents the work effectively executed for the integration of generic services, offered through the EOSC Portal, into discipline-specific platforms. The final version will collect the lessons learned during this integration work, in strict collaboration with WP4, in charge of the preparation of services and their functionalities, according to the researchers' requirements.

The final version of this implementation plan will be delivered at the end of the project.

