DICE

Data Infrastructure Capacity for EOSC

Deliverable D5.2

Community platforms integration feedbacks and exploitation impact

Responsible Partner:	BSC
Status-Version:	Final – v1.1
Date:	13.07.2023
Distribution level (CO, PU):	PU

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

Title of Deliverable:	Community platforms integration feedbacks and exploitation impact
Due Date of Delivery to the EC	30.06.2023
Actual Date of Delivery to the EC	13.07.2023

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

Abstract:	This deliverable describes the use cases integration of the platforms selected, and validation of the implementation work, with a consideration of their impact so far.
Keyword List:	use case, service, integration, impact
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

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

Document Description

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v0.1	05.05.2023	Template creation	CINECA
v0.2	02.06.2023	First draft version	BSC, SURF, UCL, ICOS, ASTRON
v0.3	16.06.2023	Comments and suggestions received by internal reviewers	CSC, CINECA
V1.0	29.06.2023	Final draft version	BSC
V1.1	13.07.2023	Updated with comments from T5.2	BSC

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

Table of Contents

Table of Contents 4
List of Figures 4
Terms and abbreviations 5
Executive Summary 6
1 Introduction 7
1.1 About this deliverable 7
1.2 Document structure 7
2 CompBioMed platform 8
2.1 Plan 8
2.2 Workflow 8
2.2.1 Description 9
2.2.2 Validation 10
2.3 Impact for users 11
3 LOFAR - Radioastronomy Science Data Repository 12
3.1 Plan 12
3.2 Workflow 12
3.2.1 Description 13
3.2.2 Validation 14
3.3 Impact for users 14
4 ICOS - Carbon Observations benchmarking tool 16
4.1 Plan 16
4.2 Workflow 16
4.2.1 Description 17
4.2.2 Validation 18
4.3 Impact for users 18
5 Communities Advisory Board feedback 19
5.1 Role and responsibilities 19
6 Conclusions 20
APPENDIX A: CAB recommendations track 21

List of Figures

Figure 1 - CompBioMed workflow implemented in DICE 9
Figure 2 - LOFAR workflow implemented in DICE 13
Figure 3 - ICOS workflow implemented in DICE 17

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

Terms and abbreviations

API	Application Programming Interface
ASTRON	Astron
BSC	Barcelona Supercomputing Center - Centro Nacional de Supercomputacion
CESNET	CESNET, z. s. p. o.
CINECA	Cineca
CSC	CSC – Tieteen Tietotekniikan Keskus Oy
CyI	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
EUDAT ltd	EUDAT ltd
FZJ	Forschungszentrum Juelich Gmbh
GA	Grant Agreement to the project
GRNET	National Infrastructures for research and technology
GWDG	Gesellschaft für Wissenschaftliche Datenverarbeitung mbh Göttingen
INFN	Istituto Nazionale di Fisica Nucleare
IT4I	Vysoka Skola Banska - Technicka Univerzita Ostrava
KIT	Karlsruhe Institut für Technologie
KNAW-DANS	Koninklijke Nederlandse Akademie van Wetenschappen
KPI	Key Performance Indicator
LTA	Long term archiving
MPG	Max Planck Gesellschaft zur Foerderung der Wissenschaften e.V.
PID	Persistent Identifier
SDR	SURF Data Repository
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

One of the main objectives of the project is to demonstrate, and improve, the effective integration of the data services provided within DICE. The preliminary integration plan for the three use cases selected for this project have been described in D5.1. In this document we report about the final implementation work, motivating the deviations of the original plan occurred.

For each of the three use cases, we also present the immediate impact for the users collected during the first months of production phase, and last months of the project duration.

The services implementation has been carried out in collaboration with WP4. This close collaboration permitted to get direct feedback during the services implementation process, and improve the material and users support of generic services for other scientific platforms in the future.

An important work of dissemination of the use cases has been coordinated by WP2 (see D2.3), taking advantage of the EOSC Future activity of dissemination, EOSC success stories and researchers' engagement, beyond DICE users pool.

The future work and the sustainability of the integrated services is described in D1.5. The level of overall satisfaction of the three use cases is very good. The will of keeping the services integration is one of the indicators of satisfaction of the scientific communities.

The feedback from other communities, and the evaluation of the impact of the work done, has been also collected in collaboration with the Community Advisory Board, for the extension of the lesson learned to a broader range of scientific communities.

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

1 Introduction

The Data Infrastructure Capacities for EOSC (DICE) consortium brings together a network of computing and data centers, 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, and access data in a consistent and persistent way. Specifically, DICE partners have been offering 14 state-of-the-art data management services together with more than 50 PB of storage capacity.

The service and resource provisioning has been accompanied by enhancing the current service offering to fill the gaps still present to the support of the entire research data lifecycle; solutions have been provided to the use cases selected, 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 result of the integration work executed for the three DICE use cases, with a reference to the initial planning (D5.1 “Pilots description and validation”[1]) updated during the evolution of the project. Part of the results consist of the impact for the research platforms communities, collected and here reported.

1.2 Document structure

The next three sections are dedicated to the three communities involved as demonstrators:

- CompBioMed
- LOFAR
- ICOS

Each section contains:

- A short description of the community involved, relevant to the use case.
- The services integrated and the providers involved.
- The validation of the integration work.
- The feedback received about the impact on the community.

Finally, we summarize the results of the activity realized at WP5 to gather feedback by the Communities Advisory Board in relation with all the project WPs.

[1] https://b2share.eudat.eu/records/febb345de1174ca287bc3792fcd21a01

The DICE project has received funding from the European Union’s Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

2 CompBioMed platform

The CompBioMed Center of Excellence (CoE)[2] renders predictive models of health and diseases by providing to clinical practice a personalized aspect to treatment.

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 somewhere else, or kept closer to the HPC services for post processing. This use case addresses the need for safe data replication and large data transfer between HPC centres, supporting a FAIR data cycle.

2.1 Plan

The integration work plan was prepared during the first stage of the project and is described in D5.1. Internally, it has been published and continuously updated in an internal wiki page accessible to the whole project consortium.

2.2 Workflow

The objectives of the integration of generic data services for CompBioMed were:

- The development of a data management workflow to facilitate simulations using large datasets and for the preparation of Exascale simulations.
- The promotion of the access to the workflow to different HPC centres (e.g.: BSC, SURF, LRZ) as well as academic partners (UCL) and potentially to the research and medical centres in the future.
- The share and preservation of synthetic or simulated data, with a mechanism that support a possible extension to sensitive data in the future.

The B2SAFE[3] federation between SURF, BSC and UCL did provide a network that facilitates the transfer and replication of large datasets between these partners: two are HPC centres (BSC and SURF) and one is a research/academic partner (UCL). Through the collaboration with the LEXIS project[4] (an EU project involving EUDAT partners, with similar needs for B2SAFE federation), the B2SAFE endpoint at SURF is also federated with LRZ in Germany. Figure 1 shows the B2SAFE federated network of partners and sites involved in CompBioMed. To accomplish this work, we had monthly task meetings and get support from the B2SAFE core development team of EUDAT.

The B2SAFE instance at SURF is connected to a tape archive storage in the backend and ensures that the data lands on tape for long term preservation. For data publication, UCL is planning to deploy an own instance of B2SHARE[5] as repository. SURF has a dedicated instance of B2SHARE which is called SURF data repository[6] that can be used for publishing large data.

[2] https://www.compbiomed.eu/

[3] https://eudat.eu/catalogue/b2safe

[4] https://lexis-project.eu/

[5] https://b2share.eudat.eu/

[6] https://repository.surfsara.nl/

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

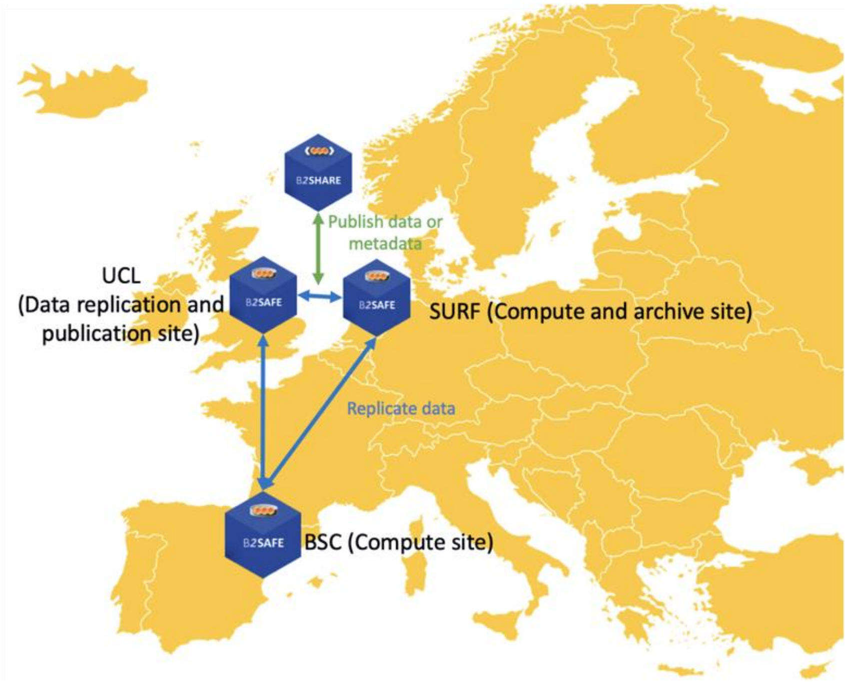

Figure 1 - CompBioMed workflow implemented in DICE

2.2.1 Description

The CompBioMed workflow is schematically shown in **Error! Reference source not found.** above. It includes the deployment and configuration of the services for data replication:

- **B2SAFE and B2HANDLE** at SURF
- **B2SAFE** at BSC
- **B2SAFE** at UCL

BSC and SURF both run the B2SAFE service in production. We have setup the federation of B2SAFE between these two centers. As a B2HANDLE service provider, SURF provided the handle prefix to UCL as a prerequisite for setting up B2SAFE instance and has federated its B2SAFE instance with UCL.

UCL has established a local instance of B2SAFE at the University to enable the federated storage and transfer of data for analysis. In addition to the project goals, UCL tested the process for establishing the EUDAT services at the institutional level, to become a part of the university's suite of centrally provided services for all researchers across all academic disciplines. Part of this work involved testing the federation of the service with the partner institutions involved in CompBioMed.

Progress throughout the project was hindered by a combination of factors, such as the processes imposed by UCL's Information Services Division (ISD) for setting up new, secure, and conformant services at the university level, the fact that the documentation for B2SAFE was not fully maintained, and some technical incompatibilities. As an example, at the enterprise level, UCL insists upon using Red Hat Enterprise Linus version 8, but the version of the iRODS[7] policy engine that forms a core part of B2SAFE was not compatible with this. This illustrates the problems of developing researcher-led open-source software that can be straightforwardly deployed into the sort of centrally architected enterprise-service hosting environments that characterize large

[7] https://irods.org/

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

multidisciplinary universities. UCL ISD's need to provide consistent, up-to-date, and secure platform infrastructure meant that there were challenges, both technical and bureaucratic, at various stages of the project. This is an issue that EUDAT will need to confront if they are to be able to benefit from centrally supported take-up at the largest research universities in Europe, and beyond. Despite the slower than planned progress towards implementing B2SAFE at UCL, this work was completed, and data transfers between UCL and the partner institutions conducted.

Although the delays in implementing B2SAFE at UCL have delayed the implementation of B2SHARE, the Research Data team at UCL will nevertheless be implementing this service and evaluating both its appropriateness as part of the data publication workflow for the current CompBioMed use case and, longer term, whether it can form part of their future institutional suite of research data management services. The team consider B2SHARE a strong candidate for this, given that it enables far more customizable community metadata schemas than their current institutional data repository, which is based on Figshare[8]. At present the central B2SHARE instance provided by EUDAT CDI is being used by the CompBioMed consortium.

2.2.2 Validation

The validation tests planned have been already described in D5.1 (see Table2). We have adapted them to the final workflow configuration as described below.

B2SAFE federation was tested by replicating test data. The data transfer speed from SURF to BSC for 1.2TB files was on average 104.214 MB/s. During the tests we did face issues in transferring big files from BSC to SURF. The issue was narrowed down to the fact that iRODS failed to transfer large file between zones if the receiving zone has SSL enabled and the sending zone has no SSL. For this, a GitHub issue has been open at iRODS consortium[9].

UCL also successfully performed tests to measure the transfer speed between UCL and SURF federation in different directions. The following tests were conducted:

Transfer Direction	File Size	Transfer Speed	Attempts & Time
SURF -> UCL	1 GB	36.8 ± 4.1 GB/s	Based on 5 attempts at around midnight
SURF -> UCL	10 GB	30 ± 14 MB/s	Based on 5 attempts, mid-afternoon on a Friday
UCL -> SURF	10GB	221 ± 23 MB/s	Based on 5 attempts, mid-afternoon on a Friday
UCL -> SURF	100 GB	253 ± 4.5 MB/s	Based on 5 attempts at around midnight

In the above UCL tests, 4 threads were used in all transfers. During our testing, we achieved approximately 1/5 of the maximum theoretical bandwidth when moving data from UCL to SURF, which is a positive result. Input speeds from UCL to SURF were significantly faster, which we expect to be caused by the firewalls that the data goes through on the way into UCL. Time of

[8] https://figshare.com/

[9] https://github.com/irods/irods/issues/6537

day is also affecting transfer speed and speed variability, which is probably caused by contention with other traffic.

For data publication, with support of DICE project we have provided a data publication hackathon to the CompBioMed community on data publication and metadata. The researchers within CompBioMed have used the central B2SHARE instance[10] to publish CompBioMed related data. A dedicated community has been created for CompBioMed and the data are published under this community. The published data are assigned a DOI and PID for findability. Examples are:

- Dataset of compounds as potential inhibitors for ROS1 kinase https://b2share.eudat.eu/records/89da477f828048dd833272b51360d13a
- Dataset of Isoxazole Amides as SMYD3 Inhibitors https://b2share.eudat.eu/records/039b83d15eff48cebb34968801761696

2.3 Impact for users

The objectives of the integration of generic data services for CompBioMed (ref. section 2.2) have been achieved, with the deployment of a data management workflow to facilitate simulations using large datasets. The integration of B2SAFE between UCL, SURF, and BSC has fulfilled expectations and will be retained for the purposes of completing the use cases required by CompBioMed. Once the CompBioMed2 project is completed (end of Sept 2023), it will be reviewed to assess likely future demand and a decision will be made as to whether it should form part of an ongoing long-term suite of services.

The services deployed at UCL and the federation network of research and HPC centres are expected to help the researchers in the CompBioMed community with large data sharing, transfer, and FAIR data management. The publication of CompBioMed data in a generic platform might facilitate collaborations inside and outside the scientific community. Multidisciplinary studies are facilitated by publishing data with interoperable and standardized information.

In the initial use case, we anticipated the need for transfer of around 50TB of non-identifiable data. This was deprioritized during the project due to the changing requirements of the researchers we were working with, although towards the end of the project funding period a new use-case was identified which involved the transfer of sensitive data from UCL's Data Safe Haven to BSC.

UCL is currently working with the relevant researchers as well as the team who maintain the UCL Data Safe Haven to implement the data pipeline between UCL, BSC, and SURF, and will continue to do so beyond the formal end of the DICE project. The sensitive nature of the data to be transferred has required discussion with UCL's research data governance staff, which is delaying the conclusion of this. UCL is reviewing the additional adoption of B2SHARE, providing a more flexible data publication service to the existing Figshare-based institutional repository, since B2SHARE offers more customizable metadata schemas than Figshare, which might benefit research groups wishing to set up institutionally supported community data repositories.

[10] https://b2share.eudat.eu/

3 LOFAR - Radioastronomy Science Data Repository

The LOFAR[11] Observatory operates LOFAR, a unique radio astronomical instrument with stations distributed over Europe.

ASTRON is working for offering further data services associated with the instrument data archives.

The DICE use case is focused 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 users and ASTRON managed processing services connected to the LOFAR Long Term Archive.

3.1 Plan

The workplan was set up during the first stage of the project and is described in D 5.1.

Internally, it has been published and continuously updated in an internal wiki page which is accessible to all the project members.

3.2 Workflow

The objectives for the implementation of generic data services connected to the LOFAR data platform are:

- Create an advanced data product repository, supporting initial operations for safe storage, discoverability, and distribution of science-level data connected to the LOFAR Long Term Archiving (LTA).
- Integrate those services with automated data processing services running on compute clusters co-located with the LTA data storage infrastructure.

The first objective has been fully realized with the publication of a large LOFAR science data release.

The second objective has been also achieved: we now have a process in place to generate data releases from products generated by automated data processing services as described. The chosen approach is to not automatically publish output as it is produced, but only following a careful selection and curation phase. Scripts have been created to generate deposits for curated data collections.

[11] https://www.astron.nl/telescopes/lofar/

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

3.2.1 Description

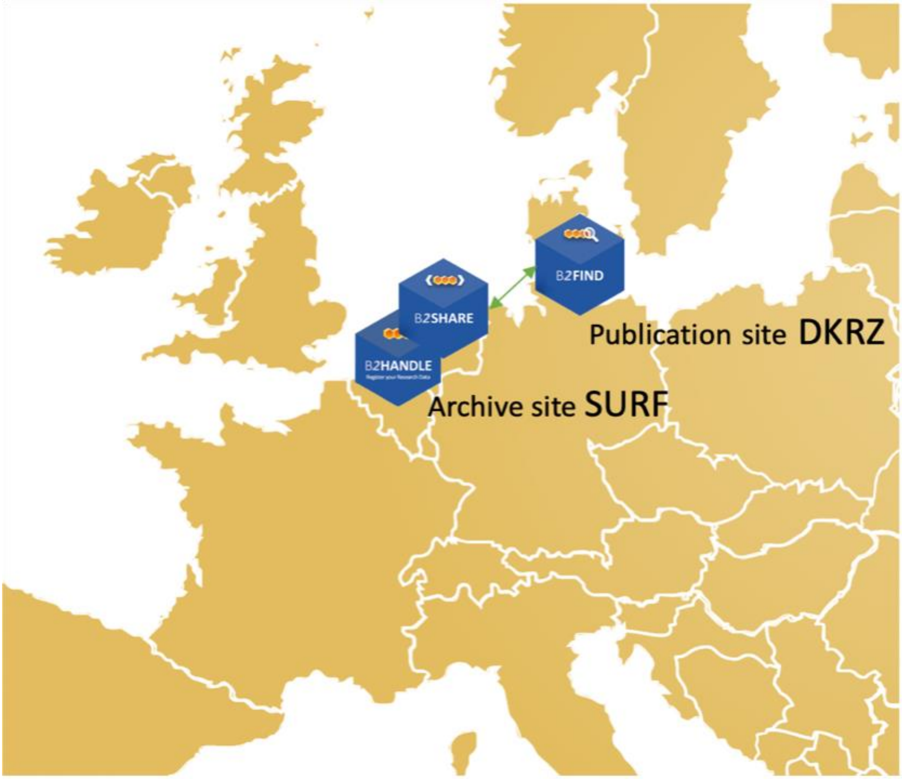

Figure 2 - LOFAR workflow implemented in DICE

The LOFAR workflow is schematically shown in the Figure above.

The advanced data product repository at SURF, provided by DICE, is supporting the initial operations for safe storage, discoverability, and distribution of science-level data connected to the LOFAR Long Term Archiving.

It is integrated with automated data processing services, running on compute clusters co-located with the LTA data storage infrastructure.

The SURF Data Repository[12] (SDR) integration enables:

- the automatic ingestion of active data (e.g. images/cubes);
- the direct access to the ingested data;
- the user access requests to less active data (e.g. visibilities).

The publication of the data is done at DKRZ B2FIND public instance[13]. The B2FIND integration permits:

- the registration of LOFAR scoped PIDs for data products.
- the harvesting of metadata by Virtual Observatory (VO) and B2FIND.

The integration of the SURF Data Repository & B2FIND has been supported by DICE technical experts at SURF, while the integration of the central VO registry & B2FIND has been provided by DKRZ.

[12] https://repository.surfsara.nl/

[13] https://b2find.eudat.eu/

The collaboration between ASTRON and SURF permitted the integration of the ASTRON VO repository with SURF Data Repository and the central VO registry.

The documentation that allowed the integration with the SURF repository is public and accessible at https://sdrdocs.readthedocs.io/en/latest/

3.2.2 Validation

The list of tests planned for the validation of the integration work is available in D5.1 (see Table 5). They consisted of:

- Ingest of a LOFAR data collection in the SURF Data Repository.
- Publishing a LOFAR data collection in B2FIND
- Automation of data registration.

The LoTSS DR2 data collection has been successfully deposited in SDR and was automatically picked up by B2FIND from both the Virtual Observatory and SDR publications. The content has been inspected thoroughly and found to be correct:

- DR2 data collection in SURF repository
  https://repository.surfsara.nl/collection/lotss-dr2
- ASTRON VO publication, harvested from central VO registry
  https://b2find.eudat.eu/dataset/bbe9cb0b-9d73-5347-8fb9-b926bc4c6cf9
- Example dataset from the deposit, harvested from SDR
  https://b2find.eudat.eu/dataset/810c16db-68f5-50b5-9a9c-a83cd65cd7c4

The process of service integration has been well supported but ultimately was executed by delivering input for the SDR deposit by ASTRON to SURF and SURF ingesting it into SDR. A documented API for the SDR could be used to further automate generating a deposit without requiring execution support by SURF.

One other aspect of integration that has been identified and is not currently available is monitoring. This will be further investigated by ASTRON & SURF. It is therefore not possible yet for ASTRON to assess usage of the publications in SDR & B2FIND.

Finally, the cost model for continued use of SDR beyond the DICE project as advertised by SURF is prohibitive for ASTRON given the data volumes that the deposits will have. This has resulted in caution being applied with creating more data deposits in the SURF Data Repository, since the impact of using the service is difficult to measure. The cost model is subject of discussion between ASTRON and SURF.

3.3 Impact for users

The initial expected impact for the Community was the enhancement of the science output from LOFAR by lowering the threshold to science-level data, improving traceability, findability, and attribution for data. In general, the integration work is fulfilling the expectations.

As far as it is currently known, no new collaborations have been formed from users accessing the DICE services. However, as it has been noted in the Validation section, we have no direct access to monitor interest in the DICE services integrated in the LOFAR workflow.

Feedback has been received from the science group that generated the data collection and at several fora where the activity has been presented. Overall, the response has been positive but also cautious with respect to direct impact and the additional value the services provide for the (radio) astronomical community, as compared to existing (Virtual Observatory) community practices.

The paper associated with the data deposit (also harvested by B2FIND from the central VO registry) has been cited 11 times since publication in February 2022, 5 of which by papers that did not include the authors of the original paper[14].

Other features that the new services are providing and positively evaluated by the users are the complete coverage of published data products by associated PID's, and the integrated (open & public) data staging capabilities for data that is stored on near-line medium (tape).

[14] https://www.aanda.org/component/citedby/?task=crossref&doi=10.1051/0004-6361/202142484

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

4 ICOS - Carbon Observations benchmarking tool

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.

This work extends 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.

4.1 Plan

The workplan was set up during the first stage of the project and is described in D 5.1. Internally, it has been published and continuously updated in an internal wiki page and available to the whole DICE consortium.

4.2 Workflow

The objectives of the services integration are:

- Extending the ICOS community platform into a benchmarking environment for the so called atmospheric inverse transport models.
- Storing the data from ICOS and staging it for analysis at computing platforms.
- Providing an alternative service for publishing results in addition to the ICOS Carbon Portal.

The main work on the extension is still ongoing, originally planned to be finished in May 2023. All the data have been gathered and prepared for staging in the Carbon Portal. An important example of the data to be used in the tool are the monthly updated NRT fluxes data at https://doi.org/10.18160/20Z1-AYJ2 (350 files, 2 TB), Global Vegetation Model simulations with LPJ (178 files, 9 GB) and modelled atmospheric footprints at all atmospheric stations (228 files, 5 GB).

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

4.2.1 Description

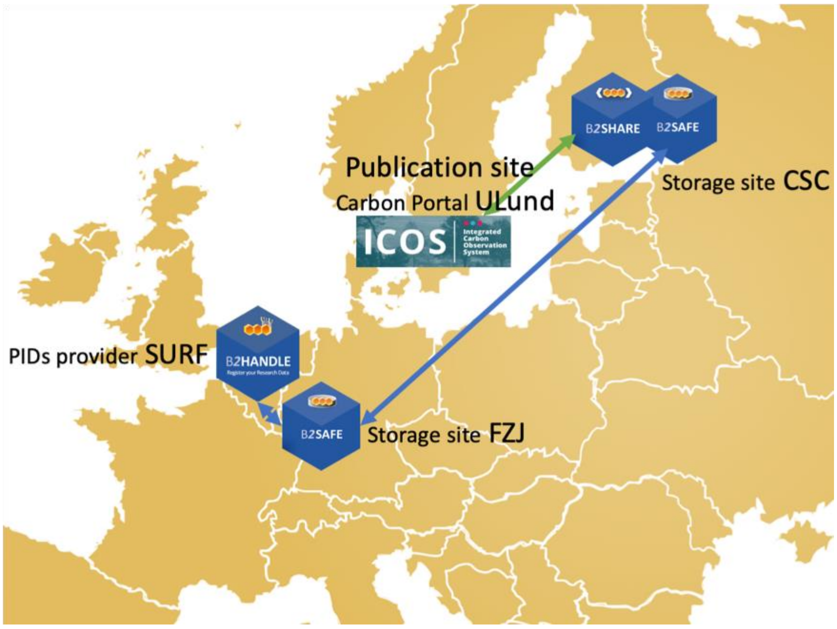

Figure 3 - ICOS workflow implemented in DICE

The services that have been integrated and development activities carried out during the project are:

- B2SAFE - Safe storage for the research data, with staging to computing platforms at CSC and potentially at FZJ;
- B2SHARE- Publication of datasets resulting from analysing the research data CSC;
- Port of inversion code to production environment;
- Develop Jupyter interface to inversion code including publication of result at B2SAFE and B2SHARE;
- Publish Jupyter VM on ICOS Portal.

The integration of B2SAFE was performed in cooperation with CSC and KTH, for the combination of minting of Handle PIDs and the streamed upload of data objects from uploader directly to both Carbon Portal and B2SAFE using the new REST API.

Technical support was provided by CSC experts. In the whole project period, we had three occasions where, next to announced service windows, the B2SAFE service had problems. The B2SAFE documentation is available at https://eudat.eu/services/userdoc/b2safe. The feedback given during the execution of the work contributed to improve the documentation of the service.

We employed the well-established regional inversion model LUMIA (*Monteil et al*, 2021, *Monteil & Scholze* 2020) for our studies. It requires data on meteorology, estimates of anthropogenic emissions, biosphere-atmosphere interactions, sea surface-atmosphere interactions, background concentrations of CO2 and CO2 fluxes at the borders of the modelled region and in the context of which any measured carbon concentrations are evaluated. We relied on ERA5 (https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5) for the meteorological data, the biosphere-atmosphere exchange fluxes for CO2 were taken from simulations with the Vegetation Photosynthesis and Respiration Model VPRM (Mahadevan et al., 2008, https://doi.org/10.1029/2006GB002735), the ocean fluxes and fossil fuel emissions are taken from Mikaloff-Fletcher et al. (2007) https://doi.org/10.1029/2006GB002751 and EDGAR_v4.3 https://doi.org/10.18160/2M77-62E6, respectively. The CO_2 background fields were created with the global transport model TM5 and CO_2 footprints with the atmospheric transport model FLEXPART 10.4, both as described in Monteil (2020, 2021). Output is written in netcdf data format. More information can be found at https://meta.icos-cp.eu/objects/sNkFBomuWN94yAqEXXSYAW54.

Thus far, LUMIA has been a specialist research tool, where all input data had been specifically prepared in local files for any particular study. Our first steps have been to move the code onto the ICOS Carbon Portal itself, where the bulk of the data is located and where we have fast access to meteorological data. We added the capability to auto-discover any existing CO_2 dry mole fraction concentration observations available through the Carbon Portal including near real-time data (NRT) as well as the ability to calculate any missing footprints seamlessly. Anthropogenic, air-sea and air-biosphere net exchange files are also discovered and read automatically without specification of file locations through the harvested PID handles. All user choices, like the geographic coverage or a filtering of stations to be excluded e.g. based on altitude or inlet height, can be tuned in a single yaml configuration file.

A simple python notebook has been created on the Carbon Portal to simplify interaction with the complex modelling tool in order to make the inversion tool more accessible to a wider scientific audience. This notebook exposes only some of the most common user choices, limits itself to a spatial resolution of 0.25x0.25 degrees, but provides scientifically sound and reproducible results. An advanced user may still achieve more specific configurations by accessing the yaml configuration file directly.

4.2.2 Validation

Data is automatically transferred at real time to the new storage and publication services.

The Jupyter VM published in the ICOS Portal is connected to the new generic data services offered by DICE. The data transfer is fully aligned with the data transfer between Carbon Portal and B2SAFE, so is daily used by all automated data updates, now at least 500 times per day.

4.3 Impact for users

The satisfactory integration of the service has permitted to create a unique staging environment for Greenhouse Gas emission validation systems, that will be applied in the framework of the Paris climate change mitigation agreement.

We will organize webinars and summer school(s) to introduce the new services to the scientific community. ICOS organises annual Summer Schools to introduce students to the climate system, the carbon cycle and use of modelling tools like inverse models for emission verification in the framework of mitigation of climate change.

Concerning the direct users feedback, we have not been able to collect it, as the service has not been exposed to users yet. The tool will be made available through the open Jupyter environment at Carbon Portal in September 2023. Impact will be followed by tracking the use of the notebook, its data usage and the amount of data produced, including tracking of citations of these data.

The new service will be an important product to showcase the value of ICOS data and the value of well-documented PID'ed reproducible workflows.

5 Communities Advisory Board feedback

5.1 Role and responsibilities

The Communities Advisory Board (CAB) is composed by 5 experts from users' communities (in different scientific domains) and not already partners of the DICE project. While the use cases part of the project have been important to pilot some of developed technical solutions, the CAB aims to strengthen the connection and engagement of the scientific communities in the whole project activities and vision.

The CAB, with its participation in the project meetings, has given feedback to the services and the service provision from their respective communities. The CAB organized its proper internal structure, meeting, and communication channels, both internally and externally, in connection with their own community. The connection between the activities of the project and the CAB, the collection of their reports and the logistic support has been carried out as part of Task T5.1 (BSC). The CAB activity received communications and had the possibility to interact to the project, as it was offered the possibility to give presentations to the general meetings and receive the reports of the Project Management Board.

CAB have been involved in these activities:

- Invitation to participate to project meetings to have an up-to-date overview on the project achievements.
- Provide feedback on the user requirements and on the integration work plans (in connection to project milestones).
- Periodically assess, from a users' perspective, the VA usage based on the statistics and information provided by the project.
- Advise on communication and outreach activities to increase uptake from communities.
- Advise on how to clearly present project offering and results to users (i.e. description of service offers, documentation, other material).
- Consultation on sustainability and approach to users after DICE.

A part of the regular project activities and meetings, we organized three online meetings specifically with the CAB members (one per year of activity of the project), to inform about the approach taken by DICE on specific topics, such as the Virtual Access provisioning, and the DICE users transition plan.

The detailed collection of the clarifications requested, and recommendations received, has been organized in the project wiki, so to make it available at any time to all project members. Appendix A reports the table in its state at M28.

The impact of this activity can be evaluated looking at the last column of the table in Appendix A: the collection and management of the recommendations, allowed the project to track the activities in the project WPs, because of the input received, with a general improvement of several aspects of our work.

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

6 Conclusions

The data services offered by DICE have been proven to fit the needs of the communities that in this project constitute our use cases. They have been successfully integrated the DICE services they considered having the greatest impact to their communities, with the support of the DICE services providers and giving precious feedback regarding the integration phase and the characteristics of the services from their perspective. On the other hand, the integration of a generic service with a discipline platform is not a straight-forward process from the administrative point of view, which in some cases can add a level of complexity and delay that is difficult to evaluate in advance. The mid- and long-term character of the services for data storage, share and publication, make the evaluation of the impact just limited to the short-term period. The results show that integrating EUDAT services to community services clear improvements to workflows can be made, without the need to build everything from the scratch. EUAT services' APIs and expert support allow easy integration, and in fact most problems arise on administrative and financial sides.

The feedback received from CAB experts, belonging to other scientific fields, and with broad experience and expertise, has been very useful to have a more open view of how to improve our services and approach, naturally focused to our use cases. The main consequence of the collecting and sharing the CAB advices among the project was having a global approach, interconnecting the activities done in different work packages of the project, with a clear common objective towards offering the best possible data services to the whole scientific community.

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

APPENDIX A: CAB recommendations track

We report here the full table of questions, recommendations and suggestions received by the CAB members, and the follow-up activities (as of beginning of May 2023).

The table is available, accessible to all the project members, who have the possibility to update and comment.

	Questions / Recommendations	Status	To	Action
Q1	How do you plan to provide long term archiving, considering the limited duration of the project?	Solved	WP1-WP4	The services will continue to be offered in the long term by the providers even if the sustainability model will change and in some cases the services will be offered as pay per use (details are provided in the sustainability section of D1.5) Task 4.3 developed a long-term preservation policy for the central B2SHARE and a template for B2SAFE instances (see D4.2 and D4.3).
Q2	How do you handle the AAI problem in different platforms, especially for HPC?	Solved	WP4	Services are accessible through B2ACCESS, which allows for user authentication with edugain and various social IDPs. The authorization for services could be done by group attributes of B2ACCESS. HPC resources can be accessed through Jupyter. If HPC resources should be accessed by ssh, the HPC site has either to provide another authentication mechanism or use additional components, e.g. WATTS (https://github.com/indigo-dc/tts) Federation with other AAIs, like the Fenix one, is technically possible but the decision needs to be taken at political level.
Q3	How B2*services can be used in the different batch systems?	Solved	WP4	On HPC systems the compute nodes usually have no direct connection to the internet. Thus, if a job that is submitted through a batch system onto the compute nodes needs to access data of B2*Services the data needs to be staged into a file system accessible on the compute nodes. Work on how to better integrate B2*Services with HPC systems has been described in D4.3.
Q4	How B2*services can be used in the different storage systems in different sites?	Solved	WP4	B2SAFE is a policy driven storage. Data can be accessed via GridFTP, Webdav, and iRODS commands. WebDAV can be mounted into file systems.

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

	Questions / Recommendations	Status	To	Action
R5	Needs of data experts: solid system to rely on for data taking, raw data consolidation and early data processing/transformation (HEP)	Solved	WP4	These activities are in the early life-time of data. We think that more than a single system for all these activities are necessary, which are optimized for the different tasks, e.g. fast input/output for data taking, fast random access for data consolidation, fast distributed access for data processing.
R6	Data services need to provide rich data management capabilities (replication rules, access policies) and highly reliable bulletproof data life cycles abilities: secure raw data, quick early data process for quasi-online data quality monitoring, etc. (HEP)	Solved	WP4	B2SAFE is a rule-based service and offers replication and access policies. The set of data management policies can be extended to the users' needs.
R7	Investigators and students (PhD and others) need transparent access to data (easy AAI), powerful data search/browsing capabilities (metadata boosted), and seamless and effective data access (low latency, parallelism and scalability to handle many cores processing for hot files) (HEP)	Solved	WP4	B2ACCESS offers the "easy AAI" and B2FIND the powerful search/browsing capabilities. With B2SAFE hot data could be replicated to various sides to allow for low latency access.
R8	Prioritize AAI rationalization across platforms and services: researchers' token/certificate/identity is easy to acquire, handle and provide seamless access to the required services at all levels: access to storage, computing and file catalogues or metadata services, platforms access. (HEP)	Solved	WP4	This is done with B2ACCESS.
R9	Data access protocol coherence for the users: common way to get data and to upload data, ideally through an abstraction layer, e.g. the API, CLIs are embedded in the notebook they are using, the script they are running or the container application they deploy in a large computing cloud or batch system. Researchers are interested in the results not in the process or tools to get/process/put data. (HEP)	Solved	WP4	With the WebDAV interfaces B2DROP folder could be mounted in Jupyter notebooks, container or HPC file systems. B2SAFE can replicate data from file systems, which are available on computing/HPC systems.
R10	Data provision protocol coherence for sites and resources	Solved	WP4	True!

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

	Questions / Recommendations	Status	To	Action
	providers: a common way to integrate new resources (or existing resources) for data infrastructure scalability. Well defined way to integrate storage into the existing infrastructure relying on well-known and standard protocols is fundamental, e.g. http, S3/Swift, etc. (HEP)			
R11	Advertising or demo-ing new systems or tools is a good start, but usually the researchers are too busy to learn a new system or moving to a new platform or service infrastructure. The way to engage with research communities is to offer them the possibility to walk with you through the new system, starting simple and picking a standard workflow where they might think current things can be done better, then take this workflow and start a Proof-of-Concept together with them (you need someone on their side to talk to), the goal is to demonstrate after the PoC that they can do things faster or cheaper or easier or more efficient in the new system. If the PoC manages to prove a sizable improvement this PoC might become a use case in the experiment and if the community sees an advantage, they will pick-up these new services, tools or platforms. **Suggestion**: Start a dedicated meeting one-to-one with your research communities and identify an item/workflow where they would be interested in improving and from there start a PoC for each one of them, most likely you will see the needs and requirements from the different communities are not so different and probably there is a chance to mix some PoC in a joint cross-experiment activities. (HEP)	Solved	WP2, WP5, WP6	Different engagement strategies have been put in place (see D2.3). Different webinars have been organised and also use cases have been presented to the wider community. Additional effort has been put also in the early engagement with users (with dedicated one to one calls being organised) to discuss in details the specific use case requirements, explain how the services work, and provide suggestion on the best solutions for their needs.
S12	Suggestions on the VA report (text): • Put accent in the data access service, which is the real added value of a long-term safe storage.	Solved	WP7	Text updated in D7.1 and suggestion taken up also in D7.2.

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.

	Questions / Recommendations	Status	To	Action
	• Explain what is a "meaningful service offer", by giving examples • Put a reference to the description of the installations offered for VA.			
R13	VA Balance of the provides: • If data are daily generated, like in forecasts, replication and balance between providers is very useful. Consider this possible use case.	Solved	WP7	Work has been done at order management level to assign users to less requested providers, when possible, also by interacting with the users to understand if there were specific features that only one provider could offer. This has helped in getting a more balanced distribution of users/projects on the different installations. Replication of data was put in place when required by the users.
R14	VA improve the accounting: • The accounting system for VA is manual. Try and get and automated accounting system.	Solved	WP7	Manual data collection has been acknowledged as an issue for all the INFRAEOSC07 projects and reported as lessons learnt in D7.2. However, automatic data collection was a challenge due to the diversity of services, and accounting metrics for each of the installation. For automation of the VA accounting data collection, DICE became a test project for the EOSC accounting service. The availability of the EOSC accounting might help with this issue in future projects.
R15	VA Impact of the services usage: • Users must acknowledge resources providers. The project should provide a standard text for the acknowledgement.	Solved	WP7	Standard acknowledgement defined and disseminated. "We acknowledge access to the <name of the service> resources at [hosting site], which are partially funded from European Union H2020-INFRAEOSC-2018-2020 programme through the DICE project (Grant Agreement no. 101017207)."
S16	Comments about the low services request: • Large communities have difficulties in trusting something they do not control. If services are attached to a project, and not directly to EUDAT CDI, it can be a barrier.	Solved	All	The consortium have acknowledge the trust aspect has been one of the most important one. To address this different actions have been taken: ▪ encourage service providers to contact their user base / network as the long term relationship in place can increase the trust on the offering

	Questions / Recommendations	Status	To	Action
	• Webinars are good, but 90% people you get are already engaged. Bottom up approach: pick some activities as entry point, and then gather the communities around the use successful cases. • Since DICE services are not designed for special users, make emphasis on the data cycle. Provide cases of different access mechanisms and full data cycles, for linking platforms. • Fenix and EOSC Future are collaborating with ESCAPE to talk the same language. HPC and data lake storage are complementary. Use a common language for moving data from one side/platform to the other.			▪ make more evident in presentations/communication the long lasting offering of the services (which were already offered before DICE started and will continue to be offered in the longer term)
S17	Feedback on the transition plan for users and projects: • All DICE approach makes sense. Big users can be served by ERIC national funding, small users try to look for calls that are open to scientific communities. National funding is essential to support the national infrastructures to support the national researchers. • Look for pulling resources from the countries to fund services to continue. • In the proposals, give priorities to small users, that might have no other option.	Solved	WP1-7	Priority has been given to individual researchers/small teams in trying to support the offering of the resources still free at the point of use. The consortium will monitor the work of the EOSC TF Long term data preservation on long term data storage and funding mechanisms, as well as the Sustainability TF. Unfortunately, the EOSC procurement is not covering the Data Projects use cases at mid-long term.

The DICE project has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-2018-2020 under Grant Agreement no. 101017207.