D1.2 Data Management Plan

Lead Partner:	UNIVIE – University of Vienna
Version:	**1.2**
Status:	**Update of DMP**
Dissemination Level:	**PU**
Document Link:	doi:10.5281/zenodo.3786325

Deliverable Abstract

This document describes the Data Management Plan (DMP) of the H2020 project EOSC-Pillar.

The EOSC-Pillar project is co-funded by the European Union Horizon 2020 programme under grant number 857650. Version 1.2 of the DMP is attributed to Lisa Hönegger, Anita Bodlos, Fulvio Galeazzi, Rob Carrillo, Federica Tanlongo, Sara di Giorgio, Philipp von Hartrott, Alessandro Rizzo, Vincent Breton, Leonardo Candela.

Version	Effective Date	Description of Document/ Changes
1.0	28/11/2019	First version of DMP – M5 of project
1.1	28/07/2021	Second version of the DMP – Update at M25 of project
1.2	24/01/2023	Third version of the DMP – Update at the end of the project. Added link to Zenodo URL.

Contents

Administrative Data

Version of DMP, date	1.2, 24th January 2023
Project coordinator	Claudia Battista, GARR, direttore@garr.it
Work package leaders	WP1 Fulvio Galeazzi, GARR, fulvio.galeazzi@garr.it WP2 Rob Carrillo, Trust-IT SRL, r.carrillo@trust-itservices.com WP3 Lisa Hönegger, UNIVIE, lisa.hoenegger@univie.ac.at WP4 Federica Tanlongo, GARR, federica.tanlongo@garr.it WP5 Philipp von Hartrott, Fraunhofer, philipp.von.hartrott@iwm.fraunhofer.de WP6 Frédéric Huynh, IRD, frederic.huynh@ird.fr WP7 Vincent Breton, CNRS, vincent.breton@clermont.in2p3.fr
Authors of DMP V1.2	Lisa Hönegger, Fulvio Galeazzi, Rob Carrillo, Federica Tanlongo, Sara di Giorgio, Philipp von Hartrott, Alessandro Rizzo, Vincent Breton, Leonardo Candela
Authors of DMP V1.1	Lisa Hönegger, Anita Bodlos, Fulvio Galeazzi, Rob Carrillo, Federica Tanlongo, Sara di Giorgio, Philipp von Hartrott, Alessandro Rizzo, Vincent Breton

Authors of DMP V1	Paolo Budroni, University of Vienna, paolo.budroni@univie.ac.at;
	Lisa Hönegger, University of Vienna, lisa.hoenegger@univie.ac.at;
	Anita Bodlos, University of Vienna, anita.bodlos@univie.ac.at
	
	Acknowledgement: We are grateful to the input provided to this DMP by project partners (work package leaders and reviewers, Lisana Berberi and Lars Kaczmirek for Version 1, work package and task leads for Version 1.1).

Data officer and responsible for DMP	Lisa Hönegger, University of Vienna, lisa.hoenegger@univie.ac.at
Project title (Acronym)	EOSC-Pillar; Coordination and Harmonisation of National & Thematic Initiatives to support EOSC
Start and end date of project	Start on 1st July 2019; End on 31st December 2022
Grant number	857650 - H2020

Executive summary

This DMP outlines all activities related to the gathering or managing of research data. For instance, the DMP contains descriptions of the content of data, metadata and software applications generated during several project activities. In addition, this DMP describes the comprehensive lifecycle of data, from collection to storage, preservation, distribution and re-use scenarios.

List of Abbreviations

DMP	Data management plan
NI	National Initiatives
RDM	Research data management
T	Task
WP	Work Package

Introduction

This Data Management Plan (DMP) describes the comprehensive lifecycle of research data in the project, from gathering to storage, preservation, distribution and re-use scenarios.

The purpose of this DMP is twofold: First, investing time and resources in the conceptualization of data management at the beginning of the project allows for finding the best possible solution in the individual steps of the data lifecycle. At a later stage, the updates of the DMP fulfil the same purpose of planning all data related activities. The aim of this planning process is to pave the way for efficient data gathering and a FAIR publication of data. Second, the DMP can offer guidance to project members who can rely on this document regarding many data-related questions. Due to the size of the project and the contribution by many different partners, the formalization of all data related aspects in the form of a DMP is essential to guarantee efficient working processes. Following the convention, this DMP is a living document and therefore versioned by definition.

In more detail, this DMP provides a description of what (kind of) data will be collected along the entire lifecycle of the project. Furthermore, we describe how the data will be processed both during the project and after its completion. This description includes statements about the provenance of data, contextual statements, infrastructures used to store and manage data, as well as information regarding the publication, long-term accessibility and, if necessary, deletion of data during or after the research lifecycle. If personal data is processed, we refer to documents handling legal and ethical aspects, including statements on data protection, terms of use, copyright attribution and exploitation rights for further reuse, and licensing.

Throughout this document, we aim at compliance of data with the FAIR principles in order to improve the findability, accessibility, interoperability, and reusability of the data collected in this project. Therefore, data, related metadata, and contextual tools like software applications and source codes are stored and made available for reuse in suitable repository or archiving systems. Whenever sensible, data are assigned globally unique and persistent identifiers to allow their findability. Furthermore, we share data whenever possible under standardized licenses to facilitate reuse. Of course, we take care to comply with intellectual property rights, data protection regulations or other applicable laws and regulations.

As this DMP is a versioned document, any version can vary in length and detail following the developments of the research focus in the EOSC-Pillar project as well as experiences based on working with different kind of data and different lifecycle-stages. Thus, not all questions addressed in this document may be included in subsequent versions of the DMP.

Rather, readers can regard the DMP as a dynamic document, which can and will be updated along the whole lifecycle of the project. In order to keep track of different versions, the version number of each DMP is always included in the administrative section above.

The guiding principle of data gathering in EOSC-Pillar is to gain insights into the environment of the European Open Science Cloud (EOSC) and to support its implementation. Consequently, data collected by EOSC-Pillar may be of special interest to various stakeholders. In the past, EOSC-Pillar cooperated with the EOSC Governance Board, the EOSC Executive Board and the Working Groups of the Executive Board. In addition, EOSC-Pillar has cooperated with other EOSC supporting projects (EOSC-Synergy; EOSC-Nordic; NI4OS-Europe, ExPaNDs) from the beginning. Further stakeholders who may be interested in the data are the European Commission, e-IRG, ESFRIs, Science Europe, and the so-called “Cluster Projects” (e.g. SSHOC)

This document proceeds as follows. In the beginning, we outline RDM aspects that are relevant for all work packages (WPs). The following chapters are dedicated to the individual Work Packages and their data management activities. As a large number of partners work on this project and on very different aspects, we structure this DMP along the individual work packages (WPs) and discuss for each WP the corner stones of the data lifecycles. Information on data processing within different work packages stems from the WP leaders who are also responsible for updates regarding information on data processing (collection, archiving, accessibility, etc.) related to activities within their WPs.

Data management aspects relevant to all WPs

The project EOSC-Pillar produces various kinds of data in different stages of the project. In this chapter, we describe general procedures that are relevant for all work packages.

Repositories used by EOSC-Pillar and their metadata standards

At the time of writing Version 1.2, we still plan to use the services of three repositories for publication and storage: AUSSDA (hosted at the University of Vienna), Zenodo (hosted at CERN in Geneva) and GitHub. In order to reach compliance with the FAIR principles, all WPs will carefully draft metadata that accompany all publications. These repositories and their metadata standards are described in detail below:

(1) **AUSSDA** (The Austrian Social Science Data Archive) is based at the University of Vienna and available under https://data.aussda.at/. AUSSDA currently supports the export formats DDI, Dublin Core and JSON. Bibliographic entries are available in Endnote XML, RIS, BibTeX formats.

(2) **ZENODO** is located at CERN in Geneva, Switzerland and is available under https://zenodo.org/. Metadata supported by ZENODO are: BibTeX, CSL, DataCite, Dublin Core, DCAT, JSON, JSON-LD, GeoJSON, MARCXML and Mendeley.

(3) **GitHub** provides hosting for software development and is available under https://github.com. Metadata can be provided as json and DCAT.

Publications

The focus of this DMP is the management of ***research*** data. Therefore, we do not discuss the publication of project results as reports or papers in detail. For the sake of completeness, we will however mention some general guidelines in this section.

In the project, we use several types of software for writing publications including proprietary software (Microsoft Office) as well as open-source software (Only Office, Libre Office, LaTeX).

Whenever sensible, we make project output available using the repository Zenodo and CC-BY licenses that allow for easy reuse. Zenodo automatically assigns DOIs to publications. Furthermore, all publications are openly accessible.Data on organizational processes

In all WPs and Tasks, organizational activities like meetings, budget administration or project management result in the various documents (e.g. meeting minutes, recording of meetings, work plans etc.). As this data is conventionally not considered as research data, we do not discuss this material further.

WP1 - Management

The main task of work package 1 lies in the organization, management and administration of all project activities. For this reason, WP 1 does not foresee any gathering or handling of research data.

WP2 – The human factor of the EOSC: Dissemination, Outreach and Community building.

The main aim of WP2 is the dissemination of the project's results, communication activities and the engagement of EOSC stakeholders. The dissemination tasks require that WP2 draws on the results of research data and material gathered by other WPs. In these instances, WP2 cooperates closely with project partners of other WPs and relies on the RDM activities of other WPs. For instance, WP2 focused on disseminating the results of WP3 in the first and second project period. However, WP2 does not collect research data on its own.

Other data relevant to WP2 activities are user data, e.g. event registration and newsletter subscription data. This data will remain confidential and accessible only by the data controllers and data processors as mentioned in https://www.eosc-pillar.eu/privacy-policy. However, a detailed discussion of user data is beyond the scope of this DMP, which focuses on research data.

WP3 – the “National Initiatives” survey

The goal of the first year of WP3 was to collect information on “National Initiatives” (NI) who are stakeholders of EOSC. Since then, the focus of WP3 has shifted to collecting information on researchers. For both activities, WP3 has gathered and published various kind of research data.

Description of data

Data collected by WP3 can be distinguished roughly in two groups: (1) quantitative data from surveys and (2) qualitative data from face-to-face interviews.

Quantitative data:

In the first year of the project, we worked on the “National Initiatives” Survey. In the second period of the project we conducted a second survey on “Researchers (re)use of data and services”. Many RDM aspects are similar for both activities. Both quantitative data sets are managed in the same way and will be archived and published in the repository AUSSDA.

Software and Format:

For the quantitative survey activities, we used the open-source software “LimeSurvey”. The code for the structure of the survey can be exported as .lss files which was used for sharing the code of the NI survey.

Besides, we used Microsoft Excel (or equivalent software) for collecting the contact details of targets who fit the survey frame. Since this information contains personal data, we did not publish the full version (however, we published a list of organisations we invited to participate which is the result of a mapping exercise).

The survey data was exported from LimeSurvey in a .csv format and curated in a .dta format using the statistical software Stata. We conducted the analysis using Stata and R. In order to allow for a broad reuse, the data curators working at AUSSDA (Austrian Social Science Archive) converted the data also to .zsav, a format that can be read by SPSS, a third common software for statistical analyses. In addition, AUSSDA published the collected data as .tab, a non-proprietary format that allows for easy reuse by a broad range of software applications.

All non-pseudonymised versions of the data are stored in encrypted folders using the free software VeraCrypt and only accessible by a minimum number of project members who regularly participate in data protection training.

Accessibility:

The pseudonymized quantitative survey data has been published at AUSSDA under a customized license tailored to the needs of the scientific community (called “Scientific Use”). This license allows for a reuse for scientific purposes. Users need to identify themselves via their institutional account or by creating an account before the download of the files. The documentation material was published under a CC-BY license, which allows for a wide range of reuse.

All data sets and their documentation materials are available in the AUSSDA Dataverse:
https://data.aussda.at/dataset.xhtml?persistentId=doi:10.11587/VOSVGK
https://data.aussda.at/dataset.xhtml?persistentId=doi:10.11587/BDQGOR
https://data.aussda.at/dataset.xhtml?persistentId=doi:10.11587/D0UMOH

Time frame for reuse:

AUSSDA is an established and sustainable repository with Core Trust Seal certification; hence, we count on the long-term availability of the published data.

Qualitative data:

Qualitative data from WP3 activities stems from qualitative interviews conducted among researchers.

Software and Format:

We collect the qualitative data using the video-conference tool Microsoft Teams. This software allows for recording the video material in the format .mp4. We then convert the material to a .mp3 format using the free software VLC player. For the transcription of the interviews, we use MAXQDA, easytranscript or simply a text editor (e.g. Microsoft Word) and an audio player. We code the interviews using MAXQDA. Since the interviews may contain personal data before the pseudonymization process, all material is stored in an encrypted way using the freely available software VeraCrypt.

The pseudonymized and transcribed version of the interviews will be stored in a text format (.txt or .rtf or .doc or .pdf/A)

Accessibility:

The video and audio files will be deleted after transcribing according to the consent form used and signed by participants. The transcripts were pseudonymized and published for reuse under a license for scientific use in the AUSSDA repository. Since qualitative data sets are usually more sensitive, we used a more restrictive access condition for this data set, as is the standard procedure with social science data archives and according to the AUSSDA Access Policy.

Time frame for reuse:

The transcripts were made available for scientific reuse at the AUSSDA repository so the same conditions as for the quantitative data apply and data is available long-term.

Quality control

WP3 undertakes several measures of quality control. Before the collection of any data, WP3 runs pretests of the quantitative and the qualitative, semi-structured questionnaire. During these pretests, team members ask volunteers to respond to the questions and comment on any ambiguities or difficulties.

For the quantitative surveys, we conduct a series of sanity checks to test the responses for plausibility. For instance, we used the curated data for the frequency analyses and compared the results to the raw data that was directly imported from LimeSurvey to test for any errors during the curation process. For the qualitative interviews, we double check all transcripts and regularly discussed uncertainties of the coding process.

For the quantitative data, WP3 additionally makes use of the following services offered by AUSSDA: Data quality checks (e.g. checks of consistency of labels, logical errors in the data), data curation, and version control. Additionally, the repository will perform all necessary transformation of data in relevant formats to ensure the FAIRness and long-term availability of the data and documents.

We do not intend to delete or block access to data after publication. In case of any updates or changes, these are traceable by means of version control. According to the repositories' policies, it is not foreseen to delete data or metadata that has been published.

Documentation

Both surveys have been published in the repository AUSSDA along with several files for documentation. In addition to the data itself, we also published the questionnaires, the methods reports, tabulation reports and a variable identifier file.

- A method report in the form of a list of targets who were invited to participate in the survey. This document contains the list of organisations that were invited to the survey. For reasons of data protection, we did not publish the name or contact details for representatives who we invited to participate in the survey. [1] We will not publish a similar document for the survey on researcher for reasons of data protection.

- A "variable identifier file" provided by AUSSDA, i.e. a codebook containing all variable names, variable labels, values and value labels. This codebook is machine-readable and part of the long-term preservation strategy of AUSSDA.

For the qualitative interviews, the related documentation material include a methods report that contains the description of the research design, the semi-structured questionnaire and a data curation protocol.

At AUSSDA, all documents related to the same dataset are part of a landing page with the same DOI. The terms of use of the repository and specifically for the scientific reuse of the data are available at the following URL:

https://aussda.at/fileadmin/user_upload/p_aussda/Documents/AUSSDA_Terms_of_Service_en.pdf

Data storage strategy

During survey data collection, data is stored with LimeSurvey in Germany, according to a GDPR compliant data processing contract between the University of Vienna and LimeSurvey GmbH. After the collection process is finished and during the curation process, data storage, data security and recovery strategies are ensured by the central storage systems provided by the Computer Centre of the University of Vienna.

As outlined above, we use free encryption software (VeraCrypt) to store all non-pseudonymized data in an encrypted way. Access to these folders is secured by random passwords with a minimum length of 20 characters. These passwords are again stored in an encrypted way in a free password manager (KeePass). The master password to access the encrypted data is only available to selected project staff who participates in data protection training at least once per year.

Data preservation strategies:

WP3 uses the preservation services offered by AUSSDA. The repository has standards in place to allow for the long-term preservation of stored data. The repository does not delete data after the assignment of persistent identifiers (DOIs). Access to data is secured (according to the set access conditions). In certain exceptions, it is possible that the repository blocks access to data (e.g. if a data protection issue has been identified after a data set has been published), however, metadata will remain accessible.

Data interoperability strategy

All data published in WP3 will be made accessible in various file formats and thus be readable by various software applications. As described above, Stata (.dta) and SPSS (.zsav) are two common formats used for statistical analyses. Besides, AUSSDA also publishes data in non-proprietary formats (.tab or .csv or .txt) to allow for reuse by a wide range of software applications including non-proprietary software applications like R or Python.

Legal and Ethical Aspects

WP3 has ensured compliance with legal regulations, as well as scientific standards and ethical guidelines during all stages of the “National Initiatives” survey, the survey on “Researchers (re)use behaviour” and of the qualitative interviews with researchers.

Quantitative data:

To begin with, data collection is based on an information sheet/consent form for respondents including all relevant information on data processing. Only respondents who agreed to the informed consent were allowed to proceed with the survey. Also, the identity of participants to the survey is protected and no negative effects of the participation in the survey is foreseen. All non-pseudonymized data is stored in encrypted folders and only selected project staff with training in data protection has access to this raw data.

Before publication, the data was pseudonymized to protect respondents’ privacy according to the GDPR and Austrian national law. In addition, WP3 chose access conditions offered by AUSSDA for sensitive data: The data set is available for scientific use, following AUSSDA’s principle to set access conditions to data “as open as possible, but as closed as necessary”. Data are published under licenses appropriate for the type of publications and potential necessary restrictions (e.g. tailored, CC and further free licences of use).

All legal aspects related to the storage of data including virtual and physical security procedures are under the authority of AUSSDA and the Computer Center of the University of Vienna. The party responsible for processing data of the “National Initiatives” survey is the University of Vienna.

We aligned all legal considerations of the survey on “Researchers (re)use of data and services” to the legal considerations described for the “National Initiatives” survey above. In general, we had fewer challenges concerning data protection because the second survey was anonymous to begin with: We chose technical settings for the survey that do not allow us to connect survey responses to the names of respondents. Of course, we will nevertheless take all precautions to minimize the risk of re-identification for respondents.

Qualitative data:

The qualitative data collected in T3.4 is also based on an information and consent sheet including all relevant information on data processing. Only respondents who signed the informed consent were allowed to proceed with the interview. Also, the identity of participants is protected and no negative effects of the participation in the interviews is foreseen. As long as the pseudonymization process was not completed, all data was stored in encrypted folders.

WP4 – From National Initiatives to Transnational Services

Description of data

Data collected in WP4 consists of data from interviews/consultations with national initiatives (please see section “Transversal Task force” for this information at the end of the document) and the metadata (Provider and Resource instances) and metadata and software contributing to the National Service Registry Prototype.

Software and Format:

The National Service Registry Prototype is based on the gCube open source software. In particular, the Registry is an instance of the gCube Catalogue service configured to make it possible to register Provider and Resource items according to the metadata format promoted by the overall EOSC Catalogue. This software in mainly written in Java and consists of several artefacts in multiple versions.

Accessibility:

The source code of gCube Catalogue service is made available by the Gitea repository https://code-repo.d4science.org/gCubeCI/gCubeReleases The Software is released with the EUPL Licence.

Access to the Service Registry instance is made available by the EOSC-Pillar gateway at https://eosc-pillar.d4science.org/web/eoscpillaritserviceregistry/catalogue. Besides the GUI, the catalogue offers a REST API for accessing its content (the API is documented at https://wiki.gcube-system.org/index.php?title=GCat_Service). The catalogue content is made available using the W3C DCAT standard https://ckan-eoscpillar.d4science.org/catalog.rdf.

Time frame for reuse:

The Service Registry content will be maintained for two months after the project end.

Quality control

Service Registry content is curated by a dedicated team via the accompanying Virtual Research Environment. In practice, whenever a new item is published the members of the VRE are informed (by a post) and can review the metadata characterising the item according to completeness and correctness. Registry items are described according to defined profiles defining the typology of information to be collected and controlled vocabularies (if any) to be used to compile every metadata field. Moreover, the mandatory nature of certain information is forced by the system preventing users from registering incomplete items.

Documentation

The Virtual Research Environment accompanying the Service Registry is provided with a dedicated Wiki describing the collected information.

The Service Registry REST API is described by:

https://wiki.gcube-system.org/index.php?title=GCat_Service

Data storage strategy

Registry contents (metadata) are stored in a Relational DB as well as indexed by SOLR.

Data preservation strategy

Incremental copies of registry contents (metadata) are automatically created on a daily basis as backup.

Data interoperability strategy

Registry contents is exposed by DCAT standard https://ckan-eoscpillar.d4science.org/catalog.rdf. Moreover, an OAI-PMH endpoint could be activated as well as the REST API is available for clients to collect registry contents,

The catalogue is equipped with several harvesters (DCAT, CKAN, CSW, OAI-PMH) enacting the systematic collection of contents from data sources.

Legal and Ethical Aspects

Every Service Registry entry is annotated with a licence governing its usage.

T4.5 Business models

T4.5 aims to foster the sustainable operation of open science services. These services support the research community and offer the potential to accelerate research efforts to solve societal challenges. To fully exploit this potential, the services need to have reliable business models that allows their continuous, long-term operation. Therefore, this task evaluated business models for services related to open science in order to initiate a discussion in management and policy on how to sustain open science.

Description of data

For the systematic capture of business models, business model workshops and interviews were conducted based on Osterwalder et al. (2005) Business Model Canvas. Overall, ten Open Science Services from five different European countries (Austria, Belgium, France, Germany, and Italy) were selected for the analysis. Based on an adapted version of Osterwalder et al. ‘s (2005) Business Model Canvas. Between two and six people from management and implementation participated in the workshops and interviews conducted by trained members of T4.5.

Documentation

The data was collected in a table (Business Model Canvas) in Microsoft Word (.docx) and only accessible to members of T4.5. Before the data collection and analysis started, all members of the team agreed to treat the collected data confidential and present the results anonymized.

Data storage strategy

The data is confidential and will be deleted after the analysis (October 2022).

WP5 - The Data layer: establishing FAIR data services at the national and transnational level

T5.1

Description of data :

T5.1 implements the metadata repository for F2DS (Federated FAIR Data Space), which stores descriptive metadata in DCAT format for datasets from external repositories.

Software and Format:

The code is written in Java and JavaScript. In particular, the Spring Boot framework is used for the backend, and the Angular framework on the front end, and is stored on a Git-lab instance managed by GARR. The application is deployed on a Kubernetes cluster through Docker images which are also published on Git-lab. The metadata harvested from the connected repositories are mapped against the DCAT format and stored on Blazegraph database.

Accessibility:

The source code is openly available on Git-lab. The type of licence for re-use has yet to be determined.

The DCAT metadata can be harvested by the D4Science to be made available in the catalogue. It is also possible to query the Blazegraph triplestore through the FDP (FAIR Data Point) from the following URL: https://f2ds.eosc-pillar.eu/app/

Timeframe for reuse:

Metadata is accessible on the FDP as soon as it has been harvested from the registered repositories, and is made available in the data catalogue thereafter (see T5.2). The repository content will be maintained for two months after the project end.

Quality control

The quality of the metadata stored in F2DS is highly dependant on the quality of the DCAT mapping done by repository managers when registering on F2DS.

Documentation

Documentation of the F2DS metadata repository back-end is available here: https://f2ds.eosc-pillar.eu/smart-harvester/swagger-ui/index.html?configUrl=/smart-docs/swagger-config

Data storage strategy

Information is recorded/stored in a relational database (MongoDB) as well as in a triplestore (Blazegraph). Backups are performed on a daily basis.

Data preservation strategy

The F2DS has been developed as a proof of concept. There is no plan for preservation in the long-term.

Data interoperability strategy

The metadata repository content is exposed by DCAT standard. In addition, a REST API is available for clients to collect repository contents (e.g. F2DS data catalogue, see T5.2),

Legal and Ethical Aspects

Not applicable as there are no personal or sensitive data collected as part of this proof of concept.

T5.2

T5.2 is responsible for the development of data catalogues. The contents of the catalogue stems from the metadata repository discussed in T5.1. Moreover, new items can be explicitly published into the catalogue.

One instance of the Data Catalogue, actually of the VRE dedicated to showcase the services stemming from WP5, is available at https://eosc-pillar.d4science.org/web/eoscpillarresdatactlg. The content of this catalogue is systematically collected from the F2DS metadata repository. This content is stored both in a Relational DB and indexed by SOLR. The software technology is the same discussed above for T4.4 (the Service Registry).

T5.3

T5.3 manages the metadata records that are made available through the EOSC Pillar RDM Training and Support catalogue. Each metadata record describes a training or support resource or set of resources, which are already available online.

Software and Format: the EOSC Pillar RDM Training and Support catalogue has been built as an instance of the D4Science services and is a component of the EOSC-Pillar Federated FAIR Data Space (F2DS), which relies on CKAN technology. Besides the default metadata fields, a specific metadata profile has been created to cater for the need to describe training resources.

Accessibility: Metadata of each record is openly accessible for any user. The (URL of the) resources associated to each record are accessible for the catalogue members under registration.

Time frame for reuse: Metadata records in the catalogue are available for consultation and reuse as soon as they are published.

T5.4

Description of data

T5.4 designs and creates custom training courses delivered in different formats to cover the variety of topics and to meet the specific needs of the different types of audience and their level.

Software and Format

Types of training formats defined are:

- Workshop- usually intended as a hands-on, collaborative and interactive activity;
- Course/class- classes can be considered short courses, usually focused on a specific topic, and delivered in a few hours;
- Lecture/seminar- presenting a specific theme, usually with a discussion on innovative or controversial aspects;
- Self-training material: short video tutorials on specific topics and available tools

Since the pandemic outbreak hit Europe, all training events were turned into online remote events. The EOSC-Pillar training team adopted a set of approaches to maximise the attendees participation and interaction:

- *Use of specific tools* such as Mentimeter (www.mentimeter.com) to allow for easier interaction with the public; All questions designed are shared with the attendees.
- *Use of D4Science Virtual Research Environments* (VREs) to allow for a virtual class experience; the VRE allows for sharing of material and includes a social networking tool where attendees can discuss even after the course ends; it also allows the interaction with EOSC-Pillar training team for specific support such as to ask questions between modules or even after the course ends; the VREs is also used to acquire information and messages after the course, such as new events or information that may be relevant to the specific audience.
- *Long courses* were necessarily divided into shorter modules; the slides and recordings are made available right after each module through the VRE to allow attendees to attend off-line modules, giving more flexibility to plan the course attendance.
- The *mandatory test to download a certificate of attendance* was used also to be sure participants attended the course either live or via self-learning view of recording/slides

or via a mixed approach; this allows users to attend the course also off-line, and it is supported by the VRE social networking where attendees following one or more modules via recording can pose questions and interact with trainers for clarifications. The tests are designed to fix some important take away messages from the courses. In some of the courses a questionnaire is designed to ask for participant's feedback in order to evaluate the course based on different aspects such as: the overall course itself, the presentation, materials, the application used for the webinar, audio, use of the VRE as a communication hub and social networking and use of the Mentimeter for discussion, and a scoring formula.

All these data were collected and processed internally to get an overview on the effectiveness of the training course through KPIs which were published in the latest Training Plan Update, Deliverable 5.8.

- Self-learning materials

  During the last half of the project, we will also focus on creating e-learning content dedicated to self-learning that can be used as a support for the training provided, or as tools in their own for the themes identified above.

Furthermore, these training material resources (in ppt/pdf/ or .mp4[video]) will be deposited at the EOSC Pillar RDM Training and Support catalogue deployed in T5.3 (please see section on T5.3. for more information).

Accessibility: free with a login option

Time frame for reuse: Training materials are made available as soon as they are presented to the audience; no time constraints to the re-usability are foreseen**.**

T5.5

T5.5 has collected information about semantic artefacts used in the projects use cases (WP6). This collection comprises references to semantic artefacts, references to repositories for such artefacts and possibly email contacts of experts in the field (internal and external to the project). All this was collected in a spreadsheet in the project repository. The primary data is intended for project internal use only due to privacy restrictions.

Derived from this primary data T5.5 has created a data collection that systematically analyzes and classifies the semantic artefacts. This collection is currently maintained as a spreadsheet in the project repository. It is intended to be published as a project result following the usual project publication procedure and represents a FAIR version of the primary data (please see the introduction section “Data management aspects relevant to all WPs” for more information on the dealings with project results publications).

Future data may comprise copies of the above mentioned semantic artefacts for further processing. This will involve the operation of a software system (likely OntoPortal) as a repository for the artefacts. For the data interoperability strategy for this data please see the section for WP6.

WP6 EOSC in Action: use cases and community-driven pilots

WP6 is the largest work package of EOSC-Pillar and consists of 10 tasks. For this reason, we will discuss in this chapter only RDM aspects that are relevant for all Use Cases. Information on the individual Use Cases can be found in the following chapters.

Quality control

To assure data quality, a process (i.e., data flow and the involved communities/groups) is planned to be described. A data quality (DA) dimension will be addressed which consists of accuracy, completeness and consistency or other measures/indicators. Some of the data quality issues can be “missing data”, “incorrect data”, “irrelevant data” etc. Each use-case community will identify their relevant and specific quantifiable data quality criteria.

Data interoperability strategy for WP5 and WP6

After data is stored in the EOSC-Pillar repository and available for access, it should also be interoperable with other platforms among different researchers, institutions, countries etc. To allow inter-disciplinary interoperability a specific ontology will be developed and maintained as proposed in WP5.

Datasets collected/generated from use-cases in WP6 will be in compliance with the FAIR principles. Here, a short list of some common tools and mechanisms for FAIR research data providers and consumers will be defined from WP5 deliverables.

Different data formats and schema mapping tools (such as X3ML) will be defined and used to make data use feasible not only by state, but also transnationally.

WP6 - Use case T 6.1

Description of data

Software and format:

Jupyter Notebook (open-source web application) and PANGEO software suite, ENES Climate Analytics Service, Ophidia analytics framework, Python scripts and libraries (e.g. PROV library), workflow documents (in JSON format).

Input data: gridded climate data from the ESGF/CMIP data archive in NetCDF format.; log files (text documents).

Generated output: documents in XML, JSON, RDF and graphical formats compliant to the W3C PROV standard (https://www.w3.org/TR/prov-overview/).

Accessibility:

The source code and the notebook scripts are publicly available on GitHub:

- https://github.com/OphidiaBigData/ophidia-provenance
- https://github.com/OphidiaBigData/ophidia-container/tree/main/single-node-backend-hub
- https://gitlab.dkrz.de/data-infrastructure-services/climate_data_provenance
- https://github.com/t0m-R/STM_images

Time frame for reuse:

The content of the Github repositories is expected to stay available as long as possible (long term availability is planned).

Quality control:

The EOSC-Pillar project does not manage UC6.1 community data as such, the Research Infrastructures involved in UC6.1 do it for the data repositories they handle and therefore ensure the data quality according to their own DMP.

Documentation:

A guidelines document about the implemented demonstrator is published into the EOSC Pillar Training and Support Catalogue as well as the EOSC-Pillar Zenodo community.

WP6 – Use case T 6.2

Description of data

Software and Format:

The partners of the use-case use Jupyter Lab (https://jupyter.org/), Jupyter Notebook (open-source web application) and PANGEO software suite: PANGEO community libraries (https://pangeo.io/), Conda (https://docs.conda.io/en/latest/), XArray data structure (http://xarray.pydata.org/en/latest/), Dask parallel computing (https://docs.dask.org/en/latest/), Pandas, Python ecosystem and its libraries (NumPy, Cartopy, Matplotlib), Intake catalogue (https://intake.readthedocs.io/en/latest/), Intake-esm catalogue (https://intake-esm.readthedocs.io/en/latest/), the Ophidia analytics framework (https://ophidia.cmcc.it).

The partners of the use-case use different datasets to run their Notebooks. These datasets are described below:

	DATA SOURCES from UC 2 partners			
Ifremer/ CNRS for DATA TERRA RI	**Argo** dataset is a collection of 3000 in-situ data floats for temperature and salinity in the first 2000 m of the water column. Access on IFREMER HPC data storage or from Sextant catalogue	**CORA** CMEMS dataset built from a various collection of in-situ data, whether profiles or time series. This gridded product provides values for temperature and salinity between 1990 and 2018. Access on IFREMER HPC data storage or from Sextant catalogue	**SMOS** gridded data issued by the eponym satellite on ocean surface salinity from 2010 to 2019. CATDS is responsible for data processing and provides access on IFREMER HPC storage or from Sextant catalogue	**GSLH** dataset is a collection of gridded multimission sea surface heights computed with respect to a twenty-year mean. Several variables are proposed: Sea Level Anomaly, Absolute Dynamic Topography, Geostrophic Velocities. Access on CNES HPC data storage (xarray-zarr format) or from Copernicus catalogue
DKRZ	**CMIP6** climate model data (~3 PByte), Access from local intake cat (*Optional data sources :ERA 5 data*	**CMIP5** climate model data (~1.5 PByte, online as well as offline tape data) Access from Harvesting, catalog etc.	**CORDEX** regional model data (~ 400 TByte) Access from (https://cordex.org/)	

CMCC	*- > 1PByte), Grand Ensemble data set, ..)*			
	CMCC models contribution to CMIP5 Access from (https://esgf-node.cmcc.it/thredds) and **CMIP6:** Access from (https://esgf-node2.cmcc.it/thredds)	**ESGF** specific data collections Access from https://esgf-data.dkrz.de/search/cmip5-dkrz And https://esgf-data.dkrz.de/search/cmip6-dkrz		

Ifremer and CNRS re-used a Jupyter Notebook developed by CNES/CLS to compute sea level anomalies interpolation on ARGO float position and GSHL data. This Notebook is accessible here: http://gallery.pangeo.io/repos/pangeo-gallery/physical-oceanography/01_sea-surface-height.html#

Accessibility:

The Notebooks developed by UC2 partners are available on D4Science VRE (/Workspace/VREFolders/EOSCPillar4EarthScience/notebooks) with authentication and *at https://gitlab.ifremer.fr/eoscpillarforearthscience/data-science-notebooks*

Time frame for reuse: no time constraints to the re-usability are foreseen.

Quality control:

The EOSC-Pillar project does not manage UC6.2 community data as such, the Research Infrastructures involved in UC6.2 do it for the data repositories they handle and therefore ensure the data quality according to their own DMP.

Documentation: Guidelines about the use-case demonstrator EOSCPILLAR4EarthScience VRE are available on EOSC-PILLAR Training and Support catalog and on the EOSC-Pillar Zenodo community under a CC-BY 4.0 license.

WP6 – Use case T6.3

Description of data

Software and Format:

The 6.3 use case used the following software:

- The Dataverse project data repository software (https://dataverse.org) distributed under the Apache License, Version 2.0. The source code is accessible at https://github.com/IQSS/dataverse. The Dataverse instance used by the use case is Recherche Data Gouv (formerly Data INRAE), accessible at https://entrepot.recherche.data.gouv.fr.
- The D4Science Virtual Research Environment data infrastructure (https://www.d4science.org/) distributed under the European Union Public Licence (EUPL v.1.1) license. The VRE used by our use-case namely “EOSCPillar4AgriFood” is accessible at https://eosc-pillar.d4science.org/group/eosc-pillar-gateway/workspace
- JupyterHub (https://jupyter.org) computing platform software distributed under the under the terms of the Modified BSD license. The service/source code is accessible at https://github.com/jupyterhub/jupyterhub.
- Renku (https://renku.readthedocs.io/en/stable/index.html) computing platform software, distributed under the Apache License, Version 2.0. The service/source code is accessible at https://github.com/SwissDataScienceCenter/renku.
- Kubernetes (https://kubernetes.io) container orchestrator to deploy computing platforms, distributed under the Apache License, Version 2.0. The source code is accessible at https://github.com/kubernetes/kubernetes.
- VITAM (https://www.programmevitam.fr/) for data archiving, distributed under the CeCILL license. The source code is accessible at https://github.com/ProgrammeVitam/vitam/.
- Fraunhofer Materials Marketplace (https://www.the-marketplace-project.eu/en/AbouttheMarketplaceProject.html) distributed under the MIT License. The source code is accessible at https://github.com/materials-marketplace.
- OpenSilex (http://www.opensilex.org/) ontology-driven Information System as a data source, distributed under the GNU AFFERO GENERAL PUBLIC LICENSE. The source code is accessible at https://github.com/OpenSILEX.
- B2SAFE (https://www.eudat.eu/b2safe) for data archiving, distributed under Copyright (c) 2018, EUDAT CDI - www.eudat.eu. The source code is accessible at https://gitlab.eudat.eu/b2safe/B2SAFE-core.

The use case produced the following outputs:

- A connector between Dataverse based repositories and Jupyterhub distributed under the Apache License, Version 2.0. The source code is accessible at https://forgemia.inra.fr/dipso/eosc-pillar/dataverse-jupyterhub-connector.
- A connector between Dataverse based repositories and D4Science VRE distributed under the Apache License, Version 2.0. The source code is accessible at https://forgemia.inra.fr/dipso/eosc-pillar/dataverse-d4science-connector.
- A batch program to read datasets from Dataverse based data repositories and archive them in Vitam and B2SAFE distributed under the Apache License. The source code is accessible at https://forgemia.inra.fr/dipso/eosc-pillar/dataverse-vitam-archiving.
- A batch program to populate Dataverse based repositories with datasets from OpenSilex ontology system. The source is integrated int OpenSilex source code distributed under the GNU AFFERO GENERAL PUBLIC LICENSE. The source code is accessible at https://github.com/OpenSILEX/opensilex/.
- A connector between Fraunhofer Materials Marketplace and Dataverse based data repositories distributed under the MIT License. The source code is accessible at https://github.com/materials-marketplace/dataverse-app.
- An improvement to D4Science to use external computing platforms. The source code has been integrated into D4Science's own source code at https://code-repo.d4science.org/gCubeSystem.

Accessibility:

All the source codes for the software used and produced above are accessible as mentioned in each of their dedicated paragraph.

Quality control

The EOSC-Pillar project does not manage UC6.3 community data as such, the Research Infrastructures involved in UC6.3 do it for the data repositories and archiving platforms they handle and therefore ensure the data quality according to their own DMP.

Documentation

An overall documentation of the use case demonstrators can be found in the deliverable 6.2 of the project[2], and specific documentation for each of the services used or produced under the repositories or website linked above.

[2] Rizzo, Alessandro, Pierkot, Christelle, Vernet, Marine, Sudre, Joël, von Hartrott, Philipp, & Berberi, Lisana. (2022). EOSC-Pillar D6.2 - Demonstrators and success stories from the use cases. Zenodo. https://zenodo.org/record/6541565

Data preservation strategy

As mentioned in Quality Control, the EOSC-Pillar project does not manage UC6.3 community data and therefore its preservation.

Data interoperability strategy

All connectors have been developed to interface with standard built-in features of the Dataverse project, allowing their reuse for all Dataverse-based repositories, of any scientific community and their data.

For other tools, the use of their APIs has been favoured to try to make the demonstrators as interoperable as possible.

WP6 – Use case T6.4

Description of data

Software and Format:

This Use Case produces and provides two main outputs:

1. the content of the Software Heritage Archive and
2. the software used to run the archive.

The content of the Software Heritage Archive is all the available source code and its development history as captured by modern distributed version control systems.

Accessibility:

1. All the gathered source code present in the Software Heritage Archive is available at https://archive.softwareheritage.org via the web interface and APIs.
2. The source code of the Software Heritage project itself is Free and Open Source Software and is available at https://gitlab.softwareheritage.org via standard mechanisms.

Time frame for reuse:

1. The content of the Software Heritage Archive is expected to stay available as long as possible (long term availability is planned).
2. The source code of the Software Heritage Archive infrastructure is expected to be available as long as useful (and is self-archived on (1) as well).

Quality control

For the general case, nothing specific is done: we archive source code as is. We do not check the content of archived source code artefacts. But we do verify the integrity of archived source code artefacts by the mean of cryptographic checksums (the data model is a Merkle Directed Acyclic Graph).

For the case of the deposited source code (using the Deposit service), we guarantee that the deposited content is the one uploaded by the authenticated user (partner, e.g. the HAL repository) and that the CodeMeta metadata linked to the deposited software is available. But it remains the responsibility of the partner of the Software Heritage project that have access to the Deposit service to ensure / enforce the quality of the deposited software.

Documentation

All the documentation for the Software Heritage project is available on https://docs.softwareheritage.org and on the main we site https://www.softwareheritage.org/

Data storage strategy

The Software Heritage Archive uses as data model based on a Merkle DAG which is stored in a Postgresql relational database, with a copy stored in a Cassandra database. The source code file content (blobs) are stored in a content-addressable object storage using cryptographic hash of the content as key.

Data preservation strategy

The Software Heritage Project being a preservation project, we keep several copies of the archive:

- the main copy is hosted on a private infrastructure in Inria's datacenter, with 2 copies of the Postgresql database, plus one copy in Cassandra database and one copy as a complete Kafka jourrnal,
- there is one complete copy of the Archive (both the database and the object storage) hosted on the Azure cloud provider (eu-west),
- a network of complete mirrors of the Archive is currently work in progress,
- a full copy of the Archive is expected to be stored in CINES' Vitam service as part of this UC of the EOSC-Pillar project; this is still work in progress.

Data interoperability strategy

All the content of the Software Heritage Archive is available via public APIs. For now, these APIs are custom REST-like APIs (not described using OpenAPI specifications), and a GraphQL API is a work in progress.

For the identification of software source code artefacts, the SWHID specification has been published and its adoption is increasing. Its normalization by ISO is a work in progress.

Legal and Ethical Aspects

1. For the content of the Software Heritage Archive, please see legal statements published on https://www.softwareheritage.org/legal/ and note that:
   - each source code artefacts is available under its own license,
   - both API and bulk use are submitted to ToS to avoid DoS and privacy issues.
2. All the source code written and used to operate the Software Heritage Project is licensed under FOSS licences (mostly GPLv3).

WP6 – Use case T 6.5

Description of data

This Use Case aimed at creating a Proof Of Concept (POC) of creating a link between Social Sciences and Humanities (SSH) data stored in Nakala national repository and publications deposited on the French Archive portal HAL.

HAL is a multidisciplinary open access archive with a specific part for SSH resources the POC is focussing on. HAL contains published and unpublished scientific literature. For the HAL open archive the CCSD has been working since its creation so that the publications and the metadata that describe them fully comply with the guiding principles of open science https://www.ccsd.cnrs.fr/en/fair-guidelines/

Nakala is a research data repository for the SSH. The data repository in Nakala offers services on several stages of the life cycle of research data in SSH: on their preservation, their publication and their reuse

Software and Format:

Nakala accepts all types of digital research data coming from scientific projects in the Humanities and Social Sciences, whether these data are associated with a publication or not. It can therefore be text, image, video, sound files, etc. In the same way, Nakala accepts all file formats. However, we recommend the use of open formats for a long-term readability and interoperability of the data. For more information on this subject, please consult the paragraph "Preparing the data" in our online documentation: https://documentation.huma-num.fr/nakala-preparer-ses-donnees/#choisir-les-formats.

Several types of scientific documents are accepted in HAL; The Main goal of HAL is to provide access to the full-text of journal articles, communications, reports, thesis and so on…The document may or may not have been published, so you can find journal articles as well as preprints in the archive. A lot of formats are accepted. For source files in lateX format, a PDF is generated after compilation.

Accessibility:

For HAL, all the documents are available at https://hal.science/ via the web interface and APIs (https://api.archives-ouvertes.fr/docs). The source code of HAL is available at https://github.com/CCSDForge/FinHal

The term "data" in Nakala designates the association of one or several files and a set of information describing them (the metadata). The deposit of a data item therefore necessarily implies loading one or several files in Nakala and describing them with metadata.

To validate the registration, one must choose between the following two options: "Create" or "Publish".

These two modes of recording the data imply a notable difference in the status of the data which impacts its accessibility.

If the data is just "created", it keeps its status of private data. Thus, creating data makes it visible only to the submitter and to users who have been granted rights by the submitter of the data. This is a transitory state before final publication, for example for an intermediate stage of review or validation.

For data whose status is private, Nakala limits the storage space. Thus, each user can deposit up to 10,000 private data (or the equivalent of 100 GB maximum).

A private data can be deleted at any time by any user who has modification rights on this data.

On the other hand, "publishing" a data in Nakala makes it publicly accessible and reusable.

By default, the files as well as the description record (metadata) of a public data can be consulted by any visitor of Nakala. However, it is possible to define an access limit (embargo) on the files of your choice. They will then only be visible at the end of the time limit. However, these files remain accessible to people who have read rights on the data. The description record will always be visible and freely accessible. There is no storage limit for published data.

The publication of the data is definitive. Indeed, a perennial identifier of type DOI is automatically assigned to a published data. This DOI is assigned instantly, as soon as the data is published. Each DOI is declared to DataCite.

The function of a persistent identifier is to identify a resource in a stable and long-term manner. This is why once a data item is published on Nakala, the depositor can neither delete nor unpublish it.

Time frame for reuse:

Except for the data under embargo, the delay of reuse is immediate for the data published in Nakala.

Quality control

HAL team checks the descriptive metadata (title, authors, journal…) of deposited files before being put on-line, the technical quality of the file (readable file), and the compliance with publisher’s policy for published documents (https://doc.archives-ouvertes.fr/en/check-of-the-deposits/ and https://doc.archives-ouvertes.fr/en/legal-aspects/

For the moment, each user depositing in Nakala is responsible for the quality of the deposit made and for respecting the recommendations for FAIR data. Specifications and rules are also put in place to guide the use of metadata and their encoding (see : https://documentation.huma-num.fr/nakala-guide-de-description/). For example, mandatory and strongly recommended metadata are specified to validate the deposit.

The implementation of a curation process coordinated by Huma-Num agents is planned for 2024.

Documentation

All the documentation for HAL is available on https://doc.archives-ouvertes.fr/en/homepage/

All the documentation on Nakala is available here : https://documentation.huma-num.fr/nakala/

The documentation of the POC itself is documented in the following deliverables: D6.3, D6.2 (Demonstrators and success stories from the use cases), MS26 (Communications Proof of Concept between NAKALA and HAL). See also the video presentation and the dedicated poster.

Data storage strategy

Nakala and HAL store their data in a secure environment (Computing Center of IN2P3) and accessible via open protocols.

More precisely, Nakala data is stored on a network storage device. An image of the data (snapshot) is taken at regular intervals. In addition, a backup on tape is performed daily. Nakala's metadata are stored in a SQL database that is backed up daily on Huma-Num's infrastructure.

HAL then sends the published documents to CINES (https://www.cines.fr/en/) to preserve their accessibility and readability in the long term. On the side of Huma-Num, this approach of preservation is also sealed by a partnership, but it is not automatic.

Data preservation strategy

Huma-Num accompanies in a long-term preservation process the projects that request it through a partnership with the CINES, the National Computer Center for Higher Education. An audit of the data to be preserved is then carried out by the "Data and user support" department. Discussions take place to bring the data (and metadata) into compliance with the requirements expected for long-term preservation:

- General data organization;
- Quality of formats used and compliance of data with format specifications;
- Verification of metadata and addition of information needed for long-term preservation (e.g. status, discoverability, etc.).

Once these different points are examined, the choice of the type of long-term preservation is made within a "liaison committee" defined by the collaboration agreement with the CINES, Huma-Num's partner for preservation.

Data interoperability strategy

HAL

Metadata are accessible via open APIs (no prior registration), OAI-PMH and in a triplestore. The contents of the documents are available in open and free access.
Standards and protocols: OAI-PMH, API, RDF Triplestore
Identifiers : DOI, PMID, SWHid, arxivid
Alignment with idRef, ORCID, RNSR
Vocabularies : DC, RDF, FOAF, SKOS, BILBO, Fabio

Nakala

Metadata are accessible via open APIs (no prior registration), OAI-PMH and in a triplestore. The contents of the documents are available in open and free access.

Standards and protocols: OAI-PMH, API, RDF Triplestore

Identifiers : DOI

Alignment with idRef, ORCID,

Vocabularies : DC, RDF, SKOS

Legal and Ethical Aspects

Nakala

Within the framework of the COMMONS project financed for 8 years by the French National Research Agency, it is planned to recruit a legal affairs officer to deal with legal and ethical issues related to the development and implementation of the project which concerns in large part Nakala at Huma-Num. This person will be in charge of the coordination of the documents produced relating to these questions and will This person will be in charge of the coordination of the documents produced relating to these questions and will ensure the respect of the regulations in force.

Huma-Num also relies on CNRS legal affairs department for those topics.

HAL

CCSD is currently working with CNRS legal affairs department and lawyers on these issues.

WP6 – Use case T 6.6

Description of data

Software and Format:

The 6.6 usecase uses the Galaxy software (https://galaxyproject.org/), deployed through a specific service, namely the Laniakea@ReCaS service (https://laniakea-elixir-it.github.io/). This service is based on Laniakea software stack, which provides the possibility to automate the creation of Galaxy-based virtualized environments through an easy setup procedure, providing an on-demand workspace ready to be used by life scientists and bioinformaticians. At the end of the process, the user gains access to a private, production-grade, fully customizable, Galaxy virtual instance. Laniakea features the deployment of a stand-alone or cluster backed Galaxy instances, shared reference data volumes, encrypted data volumes and rapid development of novel Galaxy flavours for specific tasks.

For its testing purposes, usecase 6 is using training data for somatic variant calling. These datasets are open and available from https://doi.org/10.5281/zenodo.2582555

Accessibility:

The Laniakea Software is published through Apache License 2.0 and GNU General Public License v3.0.

Quality control

The EOSC-Pillar project does not manage UC6.6 community data as such, the Research Infrastructures involved in UC6.6 do it for the data repositories they handle and therefore ensure the data quality according to their own DMP.

Documentation

The Laniakea documentation is available online: https://laniakea.readthedocs.io/en/latest/

All documentation produced as part of usecase 6 within EOSC-Pillar is made available and stored using platforms and tools provided at project scale.

WP6 – Use case T 6.7

Description of data

In task 6.7, *dastools* (Quinteros, 2021[3]) was developed as a software package written in Python programming language and consists of a set of tools to work with seismic data generated by Distributed Acoustic Sensing (DAS) systems. These systems specifically generate large volume of a new data type as additional data of N-Large experiments that have been already carried out in seismology and archived at some FDSN data centres[4] using the actual standards. Therefore, a significant challenge that all seismological data centres are facing is how to archive, deliver, and process the data they host safely and accurately. Another need they face is how to support on-the-fly conversion of a big volume of data from proprietary to standard formats and investigate how current standard web services (which are very stable and mature) can manage it. Hence, we conducted a user survey (seismologists as target group) to summarize the landscape in seismology regarding big datasets and what this implies for standard services and data formats in the community. This analysis of the challenges related to the inclusion of big datasets in seismological data centres, regarding data formats and services was published in a high-impact-factor journal (Seismological Research Letters) (Quinteros et al., 2021[5]) as one of the main outcomes of this task.

Software and Format:

- The *dastools* package[6] is hosted in a GitLab repository and provides a library to access and manage TDMS datasets generated by DAS systems. This consists of Python classes which are the core of the tools provided. These classes can be imported by the users in their own code and include in it all the flexibility of *dastools*. It includes:
  - *dasws* as a stand-alone implementation of the FDSN Dataselect web service, which is able to serve miniSEED data extracted from a folder with DAS files.
  - *dasconv* as a tool which lets you convert and manipulate seismic waveforms in TDMS format and export them into the standard miniSEED

Accessibility: the software is free and open source***;***

[3] Quinteros, J. (2021), dastools - Tools to work with data generated by DAS systems. V. 0.5. GFZ Data Services. doi: 10.5880/GFZ.2.4.2021.001

[4] https://www.fdsn.org/datacenters/

[5] Quinteros, J., J. A. Carter, J. Schaeffer, C. Trabant, H. A. Pedersen (2021). Exploring Approaches for Large Data in Seismology: User and Data Repository Perspectives. Seismol. Res. Let., doi: 10.1785/0220200390

[6] https://git.gfz-potsdam.de/javier/dastools

Time frame for reuse: ready to be run as it's in production; possible to redistribute it and/or modify it under the terms of the GNU General Public Licence v3.0[7]

Quality control

As this software is under the copyright of Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, the data quality is verified according to their DMP.

Documentation

Documentation about the software package is available here.

The outcome of this task is reported internally in the project through *MS27 - Code Repository for the Dataselect - WS* and publicly available in "Exploring Approaches for Large Data in Seismology: User and Data Repository Perspectives" (Quinteros et al., 2021)[8].

[7] https://git.gfz-potsdam.de/javier/dastools/-/blob/master/LICENSE

[8] https://doi.org/10.1785/0220200390

WP6 – Use case T 6.8

Description of data

Task 6.8 will provide a production-ready implementation of a lightweight system following the RDA Recommendations from the Research Data Collections WG[9].

This system, named Data Collection System, was being developed and tested in the seismology discipline with the intention to extend current requirements and adopt it by other disciplines. The system provides a formal API to share and link pre-assembled datasets as data collection[10] objects (aggregated data from different data sources) and apply to them machine-actionable CRUD (*Create, Retrieve , Update, Delete*) operations.

Software and Format:

Data Collection System as a web service.

Since the current version the system is capable to include generic collections and can be used either by a partner or by downloading and deploying the container provided. By means of it, any institution could provide their own Data Collections Service if needed.

Our instance is being used internally at GEOFON with thousands of collections.

Accessibility:

free to access

Time frame for reuse:

Ready to be operated.

Quality control

As this software is under the copyright of Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, the data quality is verified according to their DMP.

Documentation

The outcome of this task will be updated internally in the project through *MS28 - Code Repository for improved Data Collection - WS ready* and the public documentation on terms of use for the software being developed.

The full technical specification of the generic Data Collection System is available on the RDA website[11].

[9] https://doi.org/10.15497/RDA00022

[10] http://hdl.handle.net/21.T11148/2037de437c80264ccbce

[11] https://www.rd-alliance.org/system/files/rda-collections-recommendation_ref10112017.pdf

WP6 – Use case T 6.9

Description of data

In task 6.9, a cloud-native web service called THESPIAN-NER (T*ools for HEritage Science Processing, Integration, and ANalysis - Named Entity Recognition*) was developed (*Bombini et al., 2021*)[12] and integrated into the Software-as-a-Service (SaaS) suite THESPIAN. The suite THESPIAN offers a service for FAIR data storage, called THESPIAN-Mask, which is tailored on a metadata model based on CRMhs, an extension of the CIDOC CRM ontology CIDOC_CRM, designed for modeling the complex entities typical of heritage science, developed by INFN and VAST-LAB PIN (*Niccolucci and Felicetti, 2018*)[13].

THESPIAN-NER is an AI-based web service to assist users while generating FAIR metadata or composing queries. The system provides both the front end web page user interface, the back end REST API with basic MLOps, and a containerisation system for both front end and back end, which allows for horizontal scaling.

Software and Format:

The software was developed using open source frameworks and languages; it is a full stack web service, comprising front end and back end.

Currently, only Italian written documents can be parsed to perform the NLP operations; to extend it to more languages additional training is required.

The implementation is being used internally at INFN.

Accessibility:

Free to access to registered users via INFN cloud platform.

Time frame for reuse:

Ready to be operated as soon as it is in production state.

Quality control

[12] Bombini, A., Castelli, L., dell'Agnello, L., Felicetti, A., Giacomini, F., Niccolucci, F., Taccetti, F.: CHNet cloud: an EOSC-based cloud for physical technologies applied to cultural heritages. In: GARR (ed.) Conferenza GARR 2021 - Sostenibile/Digitale. Dati e tecnologie per il futuro. vol. Selected Papers. Associazione Consortium GARR (2021), https://doi.org/10.26314/GARR-Conf21-proceedings-09

[13] Niccolucci, F., Felicetti, A.: A CIDOC CRM-based model for the documentation of heritage sciences. In: 2018 3rd Digital Heritage International Congress (DigitalHERITAGE) held jointly with 2018 24th International Conference on Virtual Systems Multimedia (VSMM 2018). pp. 1–6 (2018). https://doi.org/10.1109/DigitalHeritage.2018.8810109

The EOSC-Pillar project does not manage UC6.9 community data as such, the Research Infrastructures involved in UC6.9 do it for the data repositories they handle and therefore ensure the data quality according to their own DMP.

Documentation

All documentation produced as part of use case 6 within EOSC-Pillar is made available and stored using platforms and tools provided at project scale.

WP6 – Task 6.10

At the core of Task 6.10 lies the recruitment of service providers who will then receive support in their endeavour to become a part of the EOSC environment. For this purpose, Task 6.10 organises an open call, an event for the applicants and the support of the successful applicants. Therefore, Task 6.10 foresaw the gathering of plenty organizational data, but no collection or reuse of research data.

WP7

Description of data

Data collected by WP7 are:

- quantitative data from a survey on National Catalogues,
- qualitative data from interviews and calls of interest
- technical data related to services.

Data related to the joined work with WP2 and WP4 is described in the section of the Transversal Task force WP2/WP4/WP7.

Quantitative data from a survey on National Catalogues:

In T7.1 we worked on a National Catalogues survey. The objectives were to identify existing or prospective national catalogues in order to help driving the update of the MS7.2 document: “Procedures to include the national services into the EOSC-hub service catalogue and marketplace released”. The information collected is also important for the T4.4 activities.

According to relevant criteria for this task, 387 contacts were selected out of the WP3 National Initiatives Survey respondents and they were asked to answer a set of questions regarding catalogues.

Qualitative data from interviews and calls:

The survey on National Catalogues was completed for some items by dedicated interviews and during calls with catalogue representatives.

Technical data related to services:

In T7.2 and T7.3 we collected data from the WP6 use cases:

A questionnaire was sent to all of them to identify their requirements about AAI services.

A gap analysis and requirements collection for WP7 services was conducted in collaboration with WP6 to identify the use cases that are interested in or need the ready-to-use-services or the in-kind services provided by WP7.

Tasks T7.2 and T7.3 also collected data from the service providers (ready to use services and in-kind services involved in T7.4).

A first collection is related to a service self-assessment.

A second collection is about the service maturity model assessment tool. We used a tool developed as a checklist in a spreadsheet format from the EOSC-Nordic project (Grant agreement ID: 857652) that was slightly adapted to the EOSC-Pillar context.

In T7.4 we have collected metrics data from the services in order to be able to provide usage statistics.

These data are gathered in the project repository as .docx files according to each template.

Software and Format:

The National Catalogues survey was conducted in collaboration with WP3 team in the same conditions as the National Initiatives Survey.

The interviews and calls were not recorded.

The summary information collected through this survey (still incomplete and intended for internal use only) are stored under .xlsx files in the project repository as .xlsx sheets, presentations (.pptx files) and reports (.docx and .pdf).

Accessibility:

The National Catalogues survey was conducted in collaboration with the WP3 team in the same conditions as the National Initiatives Survey (with the consent of participants), however this data is not planned to be published since this data is process data to serve project and WP processes (see description below, especially legal and ethical aspects).

Time frame for reuse:

Data collected is directly used to work in the framework of the project either to ease discussions with potential users and to adapt the services to the use cases needs. It is not foreseen to publish them as is.

Quality control

Technical data are collected directly from use cases staff and WP7 services owners. The templates used were discussed during meetings and pre-tested by a subset of the target group. When possible we use and adapt templates from other projects that were already used in similar contexts.

Documentation

Information of respondents is received through discussions, tasks meetings and specific emails. Final documentation is part of the WP7 reports and deliverables.

Data storage strategy

Data is stored in the project repository for reuse within the project for project activities.

Data preservation strategy

It is not foreseen to preserve data that is the basis for reports or deliverables in WP7.

Data interoperability strategy

n/a

Legal and Ethical Aspects

WP3 has ensured compliance with legal regulations, as well as scientific standards and ethical guidelines during all stages of the “National Initiatives” survey, WP7 relies on this compliance for the survey on National Catalogues. For this survey we informed our contacts that their responses would have been treated in anonymous and aggregated form in publications, the questionnaire would not be circulated outside of the EOSC-Pillar project and any copy of it would be deleted at the latest after 6 months: internally, we agreed with WP3 colleagues that the list which was used as input for this questionnaire would also be deleted within 6 months.

Transversal Task force WP2/WP4/WP7

Description of data

The EOSC-Pillar Transversal Task force combined tasks from WP2, WP4 and from WP7 and aimed to gather research data relevant to 4 different areas: the status of national initiatives in all EOSC-Pillar countries, the legal and ethical framework, the onboarding of services (from national, thematic services) as well as business models of thematic, national services. We have therefore conducted semi-structured interviews with representatives from national or thematic infrastructures or services in the different EOSC-Pillar countries.

Software and Format:

The qualitative research data are transcripts or written interviews in text format (word.doc) and are saved as long term PDF format as well (PDF/A).

Accessibility:

This data will not be published in the form of the transcripts themselves, as per consent of the participants, we will only publish summaries of the statements without identifying the participants themselves unless specific authorization is given by the interviewee. Since this activity was to gather information used by the different tasks and WPs, the data gathered will be used by all relevant tasks individually and in their reports and publications.

Time frame for reuse:

During the project’s lifetime

Quality control

We have conducted multiple pretests while designing the questionnaire and have adapted it to the specific needs of the individual countries in order to align to the national developments.

Data storage strategy

The data itself is stored with individual participating partners, as only the summaries were shared. Individual partners are data controllers and are responsible for adequate saving and securing of the data.

Data preservation strategy

n/a

Data interoperability strategy

n/a

Legal and Ethical Aspects

Due to legal and ethical aspects, we do not publish the data for open reuse. Participants were asked among else sensible questions on their services and development plans which is why we stated in the consent form that the data (transcripts) will not be shared outside of the project and only summaries of the information provided will be passed on to the relevant tasks for further use.