ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Deliverable Title: 1.1 - Data Management Plan

Partner Responsible: CERN

Work Package: WP1

Submission Due Date: M6

Actual Submission Date: 30/06/2019

Distribution: Public

Nature: Report

Abstract: This document describes the initial Data Management Plan (DMP) for the ARCHIVER project. It addresses the project data collected as part of the execution and management of the Pre-Commercial Procurement (PCP) process within the project as well as data used as part of the scientific deployments on the resulting ARCHIVER services. An updated version of this document is foreseen at M17 and a final version at M26.

Document Information Summary	
Deliverable number:	D1.1
Deliverable title: Editor:	Data Management Plan João Fernandes (CERN)
Contributing Authors:	Marion Devouassoux (CERN), Bob Jones (CERN), David Foster (CERN), Miguel Coelho dos Santos (CERN), Jamie Shiers (CERN),
Reviewer(s):	Marion Devouassoux (CERN), João Fernandes (CERN), Bob Jones (CERN)
Work Package no.:	WP1
Work Package Title:	Consortium Management
Work Package Leader:	CERN
Work Package Participants:	CERN, EMBL-EBI, DESY, PIC and TRUST-IT Services
Distribution:	Public
Nature:	Report
Version/Revision:	V 1.0
Draft/Final:	Final
Keywords:	DMPs, Data Preservation, Personal Data

Disclaimer

The ARCHIVER project with Grant Agreement number 824516 is a Pre-Commercial Procurement Action funded by the EU Framework Programme for Research and Innovation Horizon 2020.
This document contains information on the ARCHIVER core activities, findings, and outcomes, and it may also contain contributions from distinguished experts who contribute to ARCHIVER. Any reference to content in this document should clearly indicate the authors, source, organisation, and publication date. This document has been produced with co-funding from the European Commission. The content of this publication is the sole responsibility of the ARCHIVER consortium and cannot be considered to reflect the views of the European Commission.

Grant Agreement Number: 824516
Start Date: 01 January 2019
Duration: 36 Months

Document History

Issue	Date	Description	Author/Partner
V0.1	19/06/2019	First draft version available for partners' input and specific contributions	João Fernandes, Marion Devouassoux (CERN)
V0.2	20/06/2019	Edits and comments	
V0.3	27/06/2019	Final edits and comments	ARCHIVER PMT

Document Approval

Issue	Date	Name
V0.1	19/06/2019	First draft for circulation within the collaboration
V0.2	20/06/2019	Updated version revised by the ARCHIVER project office
V1.0	28/06/2019	Final version for review by the Collaboration Board
V1.0	8/07/2019	Final version to be submitted to the EC

Executive Summary

This document describes the initial Data Management Plan (DMP) for the ARCHIVER project. It addresses the project data collected as part of the execution and management of the Pre-Commercial Procurement (PCP) process within the project as well as data used as part of the scientific deployments on the resulting ARCHIVER services.
The DMP for the scientific use cases are using the "Practical Guide to International Alignment of Research Data Management"[1]:

"*The aim of the initiative was to develop a set of core requirements for data management plans (DMPs), as well as a list of criteria for the selection of trustworthy repositories where researchers can store their data for sharing. In light of the development of the EOSC and an increasing tendency towards data sharing, these requirements and criteria should help to harmonise rules on data management throughout Europe.*"

As such, this DMP focuses on data sharing and re-use and in particular on the following issues:

- What types of data will the project generate/collect?
- How is personal data managed in the context of the ARCHIVER project?
- What standards will be used?
- How will data be exploited and/or shared/made accessible for verification and re-use?

Regular updates to this data management plan will be made according to the following tentative schedule:

Date	Milestone	Issue(s) to be addressed in revision
October 2019	Tender publication	Were all the use-cases retained?
June 2020	Start of Prototype Phase	Could all the use-cases be satisfied?
March 2021	Start of Pilot Phase	Could all the use-cases be deployed?
November 2021	End of Project	Are the plans still valid beyond the end of the project?

[1] https://www.scienceeurope.org/wp-content/uploads/2018/12/SE_RDM_Practical_Guide_Final.pdf

Table of Contents

1. Introduction

This Data Management Plan (DMP) addresses two distinct sets of data to be managed by the ARCHIVER project:

- ARCHIVER Project Data: data collected as part of the execution and management of the Pre-Commercial Procurement (PCP) process within the project.
- Scientific use cases deployments: data managed by some of the use cases that will be supported and deployed during the pilot phase by the services to be developed by the PCP process.

These two sets of data are treated separately by the project and in this document.

This document is the initial version of the data management plan. It explains the provisions for project data and the general approach for scientific use cases. The data management plan will be reviewed and updated at major milestones in the project, once the final set of use case deployments has been refined, with the following tentative schedule:

Date	Milestone	Issue(s) to be addressed in revision
October 2019	Tender publication	Were all the use-cases retained?
June 2020	Start of Prototype Phase	Could all the use-cases be satisfied?
March 2021	Start of Pilot Phase	Could all the use-cases be deployed?
November 2021	End of Project	Are the plans still valid beyond the end of the project?

Table 1 - Tentative Data Management Plan Update Schedule

2. ARCHIVER project data

This section describes the plan for the data to be managed as part of the execution and management of the Pre-Commercial Procurement (PCP) process within the project.

The strategy to manage the data of the ARCHIVER project is based on the following requirements:

- Provide a common toolset for use by the ARCHIVER consortium and contractors during the lifetime of the project
- Allow actions and roles for participants that are distinct from their other/existing roles in the organisations that are not linked to the project
- Create a collaborative environment that adheres to European legislation (e. g. GDPR)
- Take a 'buy not build' approach, using sub-contracting funds, since there are many commodity collaboration services available on the market of high quality and competitive cost
- Reduce the cost of operation during the limited project lifetime
- Be able to export all project data in an archive to another location for backup and/or restore.

2.1. Applicable Legislation

As stated in section 13 of the ARCHIVER Consortium Agreement, "All personal data processed by the Parties for the purpose of the Project shall be processed in accordance with their respective legal frameworks". The consortium partners of the ARCHIVER PCP project are research institutes and companies residing in EU countries and personal data of people in the EU will be processed for the execution and management of the PCP.
Therefore, all personal data that is collected during the ARCHIVER project will be processed in accordance to the European General Data Protection Regulation (GDPR) or internal regulations as applicable, in order to ensure the adoption of best practices for the processing operations of personal data[2]. In section 2.3, the roles and processing operations are detailed.

2.2. Collaborative Tool

The ARCHIVER project needs a collaborative tool to provide an environment to be used across its consortium partners and companies via the tender process for file and document sharing, calendars and meetings workflow.
ARCHIVER is using the Google G-Suite cloud based collaborative tool under a business subscription plan for this purpose.

[2] https://cds.cern.ch/record/2651311?ln=en

The vendor commits to follow GDPR regulation[3] and provides access for customers to independent third-party certifications such as ISO 27001, ISO 27017, ISO 27018, and SOC II/III audit reports[4] in order to allow customers to verify compliance and conduct risk assessments in order to determine whether appropriate technical and organisational measures are in place.

2.2.1. Google G-Suite Business Plan Subscription for ARCHIVER

A business subscription for ARCHIVER based on the Google G-Suite product has been purchased with reserved project funds. The license is under a flexible prepaid plan of 10.40 EUR per user per month.
The objective is to maintain the license active for a period of three years covering the lifespan of the project.

A subset of integrated tools is being used for ARCHIVER activities:

- Google Docs, Sheets and Slides
- Groups for Business: groups of users depending on their role and contributions (Buyers Group members, Consortium members, Contractors, etc.)
- Google Drive spaces: repository for documents, shared across the E-groups of users, with specific access restrictions based on E-group user policies
- Calendar scheduling

Among others, the business subscription plan includes the following features[5] by default:

- Unlimited cloud storage
- 24/7 support by phone, mail and online
- Security and administration controls for archiving and retention policies for mail and chats, including alerts via an alert center
- Audit reports to track user activity

[3] https://cloud.google.com/security/gdpr/
[4] https://gsuite.google.com/learn-more/security/security-whitepaper/page-5.html
[5] https://gsuite.google.com/pricing.html

The ARCHIVER project members are also benefiting from enhanced interoperability between Google Docs and Microsoft Office files since it is possible to edit, comment, and collaborate on Microsoft Office files using Google Docs, Sheets, and Slides without converting file types.[6] Additional tools included in the business subscription will be used as needed (e.g. Google Hangouts video conferencing tools).

The current public ARCHIVER website is not under the G-Suite subscription, being directly managed by an ARCHIVER partner (TRUST-IT) on its own infrastructure. This approach ensures the public ARCHIVER website will continue to be available after the end of the project and the end of the G-Suite subscription.

The G-Suite subscription has been made operational in six main steps (a total of 1 Person-Month (PM) of effort):

1. Purchase of a domain for the ARCHIVER organisation (archiver-project.eu).
2. Start a G-Suite 14 day trial at: https://gsuite.google.com/signup/basic/welcome
3. Verification of the domain ownership according to a set of provided instructions[7] in order to associate the G-Suite Business subscription to the project domain.
4. Setup the G-Suite MX records to the project domain host and Domain Name Server (DNS).
5. Add credit via a temporary credit card to the subscription based on the number of users to support and the time needed for the subscription (project lifespan of 3 years).
6. Add user accounts (via bulk import, manually or both ways combined).

[6] https://gsuiteupdates.googleblog.com/2019/04/office-editing.html
[7] https://support.google.com/a/topic/1409901

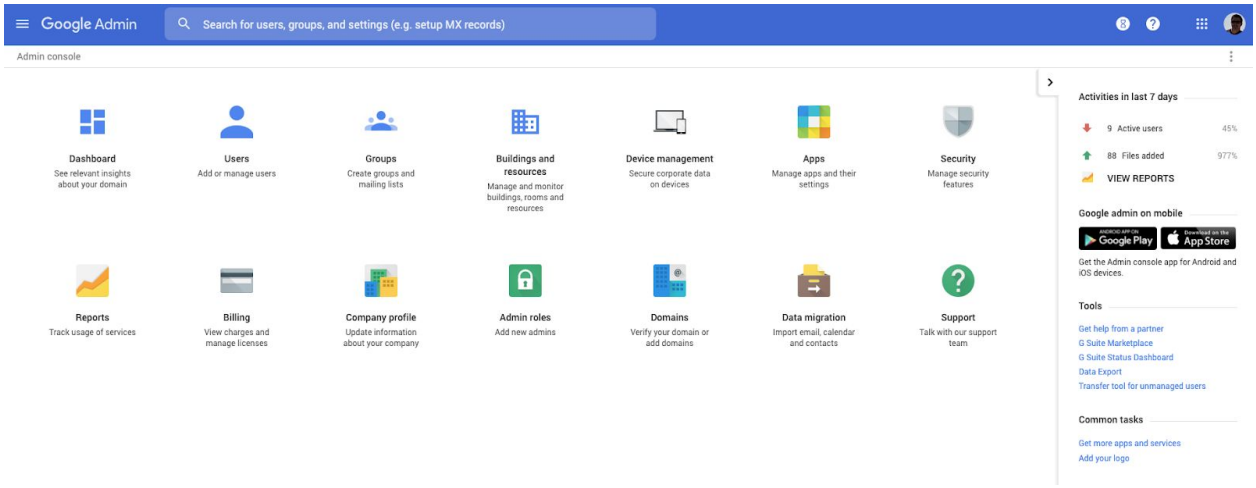

Figure 1: G-Suite Business Subscription admin dashboard

2.2.2. Data Location Policy

The G-Suite business subscription allows the ARCHIVER project to choose to store project data in specific geographic locations (the United States or Europe) by using a data region policy. The data region policy for the ARCHIVER project domain in G-Suite is set to Europe only.
The table below illustrates the G-Suite services covered by the data region policies for data-at-rest, including backups.

Google Calendar	Event titles, descriptions, dates, times, frequencies, invitees, locations
Google Drive	Original file content uploaded to Drive
Google Forms	Text, embedded images, responses
Gmail	Subjects, bodies, attachments, senders, message recipients
Google Docs, Sheets and Slides	File body text, embedded images, embedded drawings, associated end user-generated comments
Hangouts Chat	Messages, attachments
New Sites	Text, embedded images, embedded site information,

	embedded HTML/CSS/Javascript
Google Vault	Exports

Table 1: Data-at-rest (primary copy and backup) per service under the chosen data location policy.[8]

At the time this document is produced, data center locations in the Europe data region include Ireland, The Netherlands, Denmark, Finland and Belgium[9]. Data region policies can't yet be applied to all customer data types such as logs or cached content.

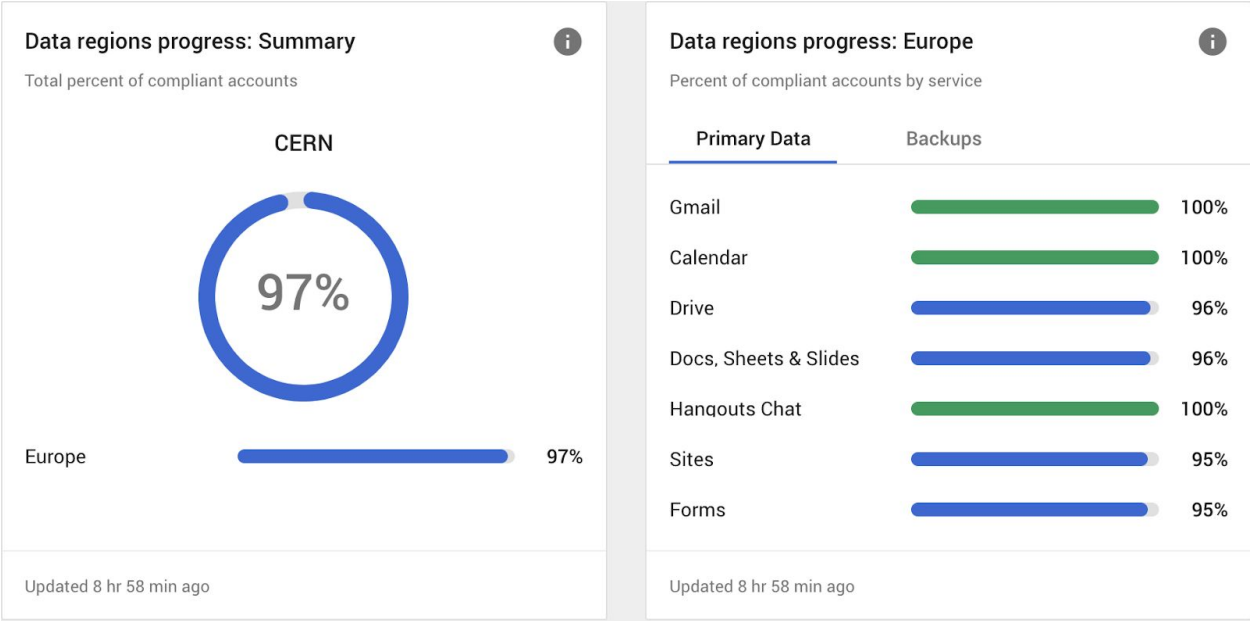

Figure 2: Breakdown services data location by region (data in Europe only)

2.2.3. Additional Security Features

G-Suite includes Google Drive, an ISO 27001 compliant service[10] for file synchronization and sharing, supporting both encryption for data-at-rest and in transit[11]. Essentially, when users log in to their account and try to access a file, it is decrypted and presented for viewing or editing. In transit, links or attachments are sent to a receiver. If the receiver has access permission

[8] https://support.google.com/a/answer/9223653?visit_id=636965194905928067-2781997998&rd=1
[9] https://www.google.com/about/datacenters/inside/locations/index.html
[10] https://cloud.google.com/security/compliance/iso-27001/
[11] https://support.google.com/googlecloud/answer/6056693?hl=en

authorized, the file is unencrypted and presented for viewing or editing. However, access by the vendor is possible at any time as end users don't hold the private keys. In order to mitigate the risk, the ARCHIVER project uses an application from the G-Suite marketplace[12], where third-party software support including open source is provided.

The app used, called Secure File Encryption[13], provides an additional level of security, allowing bank-grade AES256 encryption to protect sensitive files to be stored in Google Drive.

This is particularly important for confidential documentation such as the PCP Contract Notice tender documents.

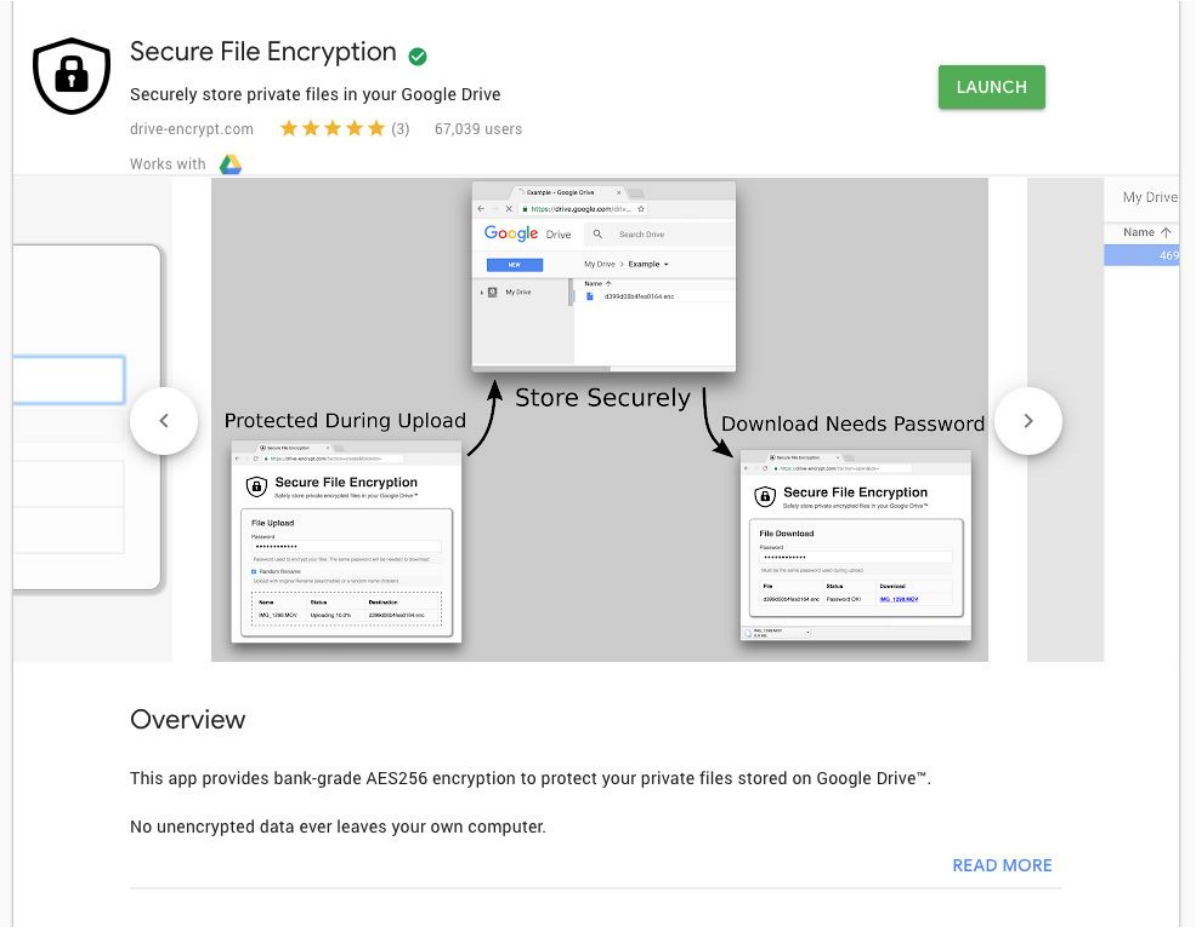

Figure 3: Third-party app example allowing AES256 encryption locally before storing files in Google Drive.

[12] https://gsuite.google.com/marketplace/

[13] https://gsuite.google.com/marketplace/app/secure_file_encryption/464708669615

2.2.4. Data Backup and Repatriation

A typical issue in using off-prem cloud based services from a broad range of cloud service providers is the risk of data and vendor lock-in. In order to mitigate the risk, the ARCHIVER project backs up and exports all data created under the G-Suite business subscription out of Google infrastructure.

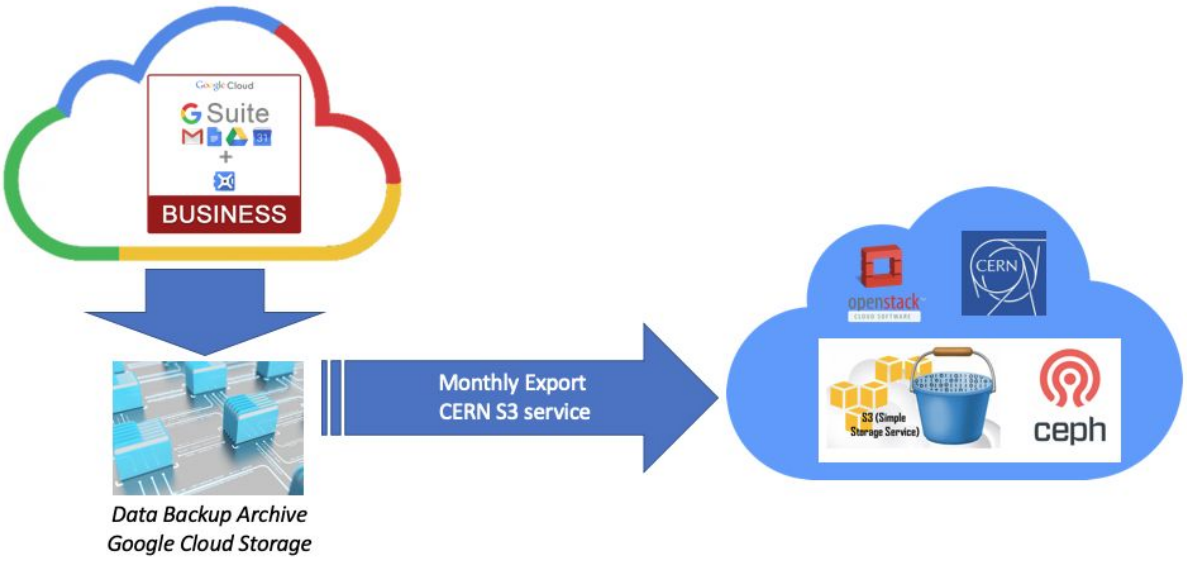

Figure 4: Implementation of the ARCHIVER project data archive & backup strategy.

Data from ARCHIVER G-Suite core services (for example Google Drive documents, Calendars, and Groups for Business) can be backed up and exported for all users of the ARCHIVER project. The backup and export can only be initiated by super admins of the ARCHIVER G-Suite domain. Once the export is complete, a confirmation email is sent with a link to the archived data in Google Cloud Storage.

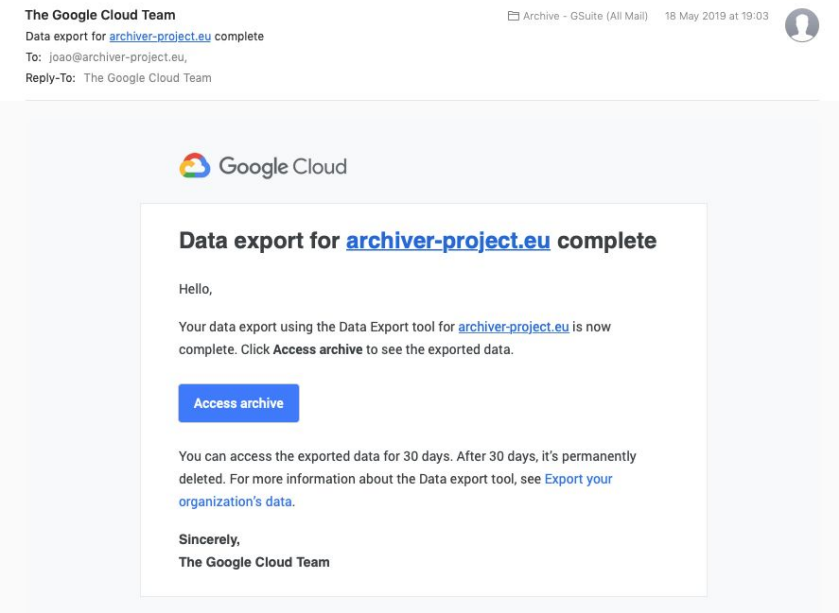

Figure 4: Confirmation the data export is complete.

For security purposes, the archived data is only available to super admins of the domain. From there, one is able to download data in several formats. Standard formats are supported for the backed up data such as zip, plain text, pdf, docx, xlsx, pptx and html for mail, documents and web pages (see Figure 7), allowing easy data restore.

In addition, the data export feature enforces additional security controls:

- It can only be initiated from super admins accounts created more than 30 days prior to the request.
- Admins must be authenticated using 2-step verification.
- When a data export is initiated, other admins of the domain will be notified immediately.
- Export events are logged in the admin audit logs.

Super admins can find the link to the archive at any time in the Data Export tool by clicking "Access archive." The data will be available in Google Cloud Storage for 30 days before it's permanently deleted. Admins can also enable end users to download their own individual data via the existing "*Download your data*" tool.

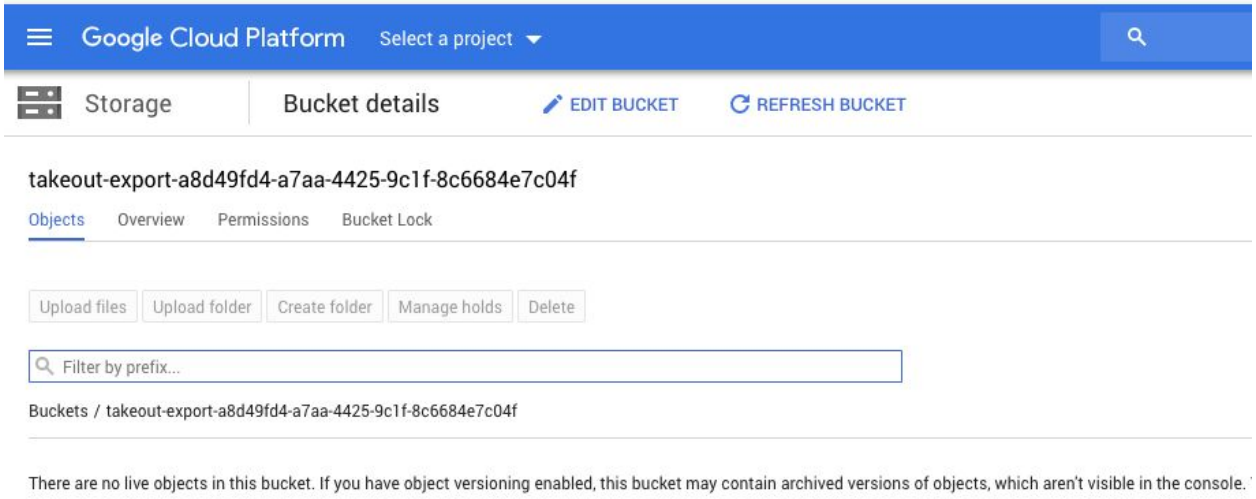

Figure 5: Archive snapshot of the full ARCHIVER domain stored in Google Cloud.

It's important to note that when the Data Backup and Export tool is used, all supported data is exported for all active users in the ARCHIVER domain. It's not possible to backup and export only some types of data or data for only some users. Backing up and exporting data for deleted users, or any user created in the past 24 hours is not possible and some newly-created data might not be included in the archive.

As a second step of the process, data from the ARCHIVER project is data then copied monthly to a S3 bucket in the CERN infrastructure, provided by the S3 service at CERN[14].

[14] https://cern.service-now.com/service-portal/function.do?name=S3ObjectStorage

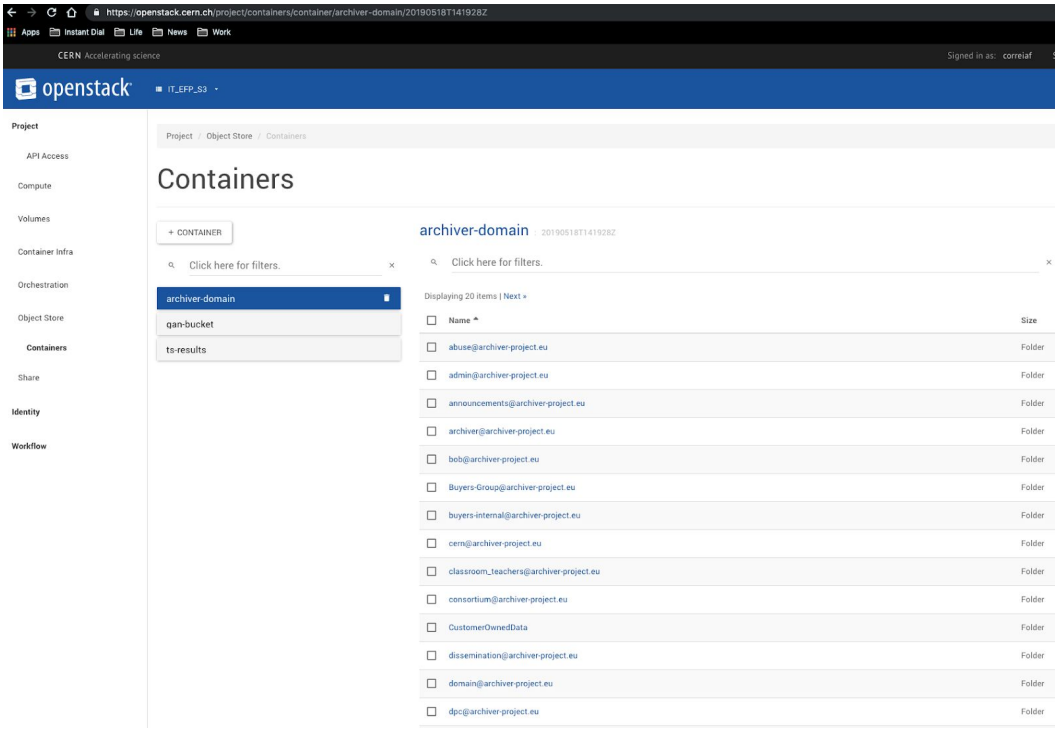

Figure 6: ARCHIVER project data archive backup stored in CERN S3 service.

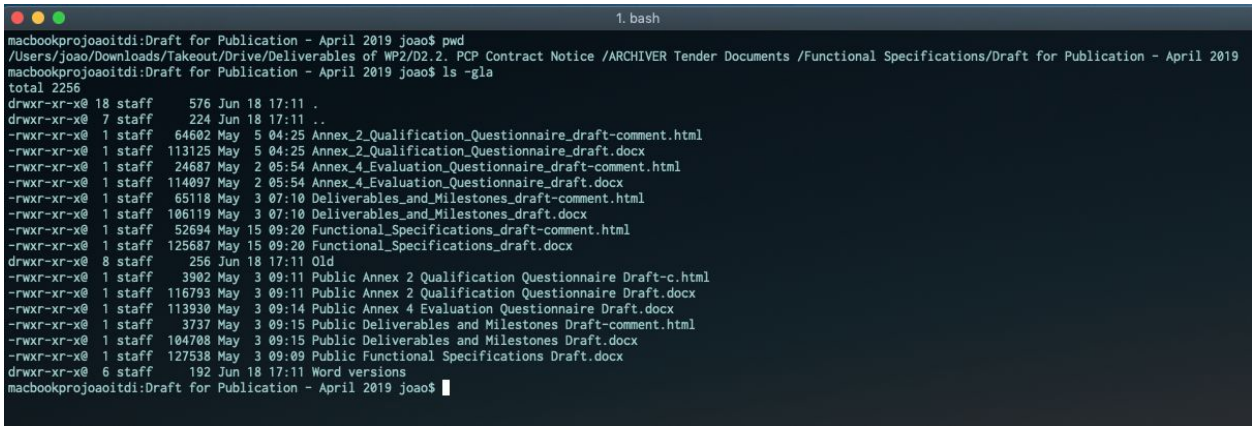

Figure 7: Partially extracted archive downloaded from CERN S3 service backup.

The ARCHIVER project team is working to fully automate a monthly backup copy of the data export ARCHIVER project archive to the CERN S3 service with the option of encrypting the export. This feature will be treated in the next revision of this document.

The table below summarizes the risks of using Google G-Suite as a collaborative platform for the activities of the project. These risks will be added to the project risk register.

Risk Identified	Mitigation actions
ARCHIVER public website not available after the end of the project.	Public website managed by project partner in order to ensure it will be available after the project is finished.
Data location outside of Europe.	Setup of data policy to Europe only. 97% of data resides in Europe. Cached content and logs are not covered by the policy yet.
Data and Vendor lock-in.	Data exports of data on a monthly basis to the CERN storage infrastructure.

Table 2: Risks to be added to the ARCHIVER Risk Register.

2.3. Personal Data Management

This section details the processing of personal data in the context of the ARCHIVER project.

2.3.1. Data Subject Rights

Data Subjects, identifiable individuals to whom particular personal data relates to, have rights they can exercise under particular conditions. Fulfilling GDPR requirements means, among others, enabling the exercise of these rights. Applicable legislation such as CERN internal regulation, ensures as well alignment with the data subject rights[15].

The eight fundamental rights are the following:

- The right to know whether data concerning him or her are being processed.

[15] https://home.cern/data-privacy-protection-policy

- Right to rectification: when personal data are inaccurate, then controllers need to correct them.
- The right to be forgotten with additional stipulations, among others if personal data has been made public.
- The data subject right to restriction of processing i. e., to limit the processing of his/her personal data.
- The right to be informed about how personal data is processed and who got these data.
- The right to data portability.
- The right to object.
- The data subject right not not to be subject to a decision based solely on automated processing, including profiling.

2.3.2. Data Controllers

i) CERN

The data collected as part of the execution and management of the PCP process within the project is centrally processed by the Consortium Management work package (WP1) led by CERN.

As a result, CERN is the data controller for the several processing operations that include the collection of personal data (name, email address, affiliation, role within the company, company name, and company postal address provided by individuals representing companies and organisations interested or participating in the PCP process. In addition, it includes storing and backing up personal data on G-Suite platform for the duration of the project (3 years), storing data in a CERN Operational Circular no 11 (OC11) compliant platform during and/or at the end of the project, create mailing lists and delete data at the end of the retention period.

CERN might also request Trust-IT, the project partner in charge of the communication and outreach of the project, to run Google Analytics on the project website[16] and share results with CERN. The objective is to use these results as the basis for statistics summarizing the level and scope of engagement in the procurement process and will be reported (anonymously) in the mandatory deliverable reports. In this case, Trust-IT is the data processor.

CERN as a data controller, provides access to subsets of personal data to the respective data processors as detailed in section 2.3.2.

[16] http://archiver-project.eu

The legal basis for collecting and processing this data is consent er processing operation from individuals using the ARCHIVER ROPO (Record of Processing Operations)[17].

ii) TRUST-IT

Trust-IT is the data controller of the data collected on the project website using cookies. The legal basis for collecting and processing this data is consent from the individual using the following privacy policy[18].

2.3.3. Data Processors

The data processor is a person or organization who deals with personal data as instructed by a controller for specific purposes and services offered to the controller that involve personal data processing. In the ARCHIVER project, the data processors are the following:

- The ARCHIVER consortium partners bound by the ARCHIVER Grant Agreement (GA), Consortium Agreement (CA) and Joint Procurement Agreement (JPA).
- External organisations supporting locally the organisation of ARCHIVER events, in an in-kind contribution basis, such as the EEN[19] and its local branches such as ACCIÓ[20].
- Google G-Suite collaboration tool, under the terms and conditions of the G-Suite business subscription plan.
- CERN services under CERN internal legislation describing rights and obligations at CERN (Operational Circular No 11).

2.3.4. Retention Periods

The project has a lifespan of 3 years from January 2019 to December 2021. In addition, being a project funded under the "Horizon 2020 EU Framework Programme for Research and Innovation", the coordinator has the obligation of archiving the project data for a period of at least 5 years, as stipulated under article 18 of the "AGA - Annotated Model Grant Agreement"[21].

[17] https://cern.service-now.com/service-portal/privacy-policy.do?se=external-funded-projects&notice=ARCHIVER
[18] https://www.archiver-project.eu/privacypolicy
[19] https://een.ec.europa.eu/
[20] http://www.accio.gencat.cat/ca/inici
[21] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf

The retention period of the data collected as part of the execution and management of the PCP is then 8 years (2019-2026).

2.3.5. Data Transfers

CERN, as data controller, is responsible for data transfers to other members of the consortium. Data transfers foreseen as part of the execution and management of the PCP process in the project that include giving access to certain data sets to the members of the consortium (e.g. list of participants to events scheduled as part of the process of the PCP), grant rights to consortium members to send emails to mailing list created by CERN. CERN may also receive the results of Google Analytics from partner TRUST-IT.
When possible, CERN will give access to anonymised data set to the consortium members.
The ROPO (Record of Processing Operations) for the ARCHIVER project, contains a data transfer section as depicted below.

Personal Data shared with entities or individuals outside ARCHIVER		
Personal Data	**3rd Parties**	**Purpose**
Name, email address, Affiliation, and E-group membership, photographs and audio recordings	Project members of the consortium partners of the ARCHIVER PCP project	Organisation of project events and all communication in the context of the ARCHIVER project.
Name, email address, Affiliation, and E-group membership, photographs and audio recordings	Google G-Suite cloud based collaborative tool under a business subscription plan	The ARCHIVER project needs a collaborative tool to provide an environment to be used across its consortium partners and companies via the tender process for file and document sharing, calendars and meetings workflow. The collaborating partners are research institutes and companies residing in EU countries. The collaborative tool needs to ensure that personal data processed in the context of

		the ARCHIVER project is compliant with the GDPR (https://ec.europa.eu/info/law/law-topic/data-protection_en). This is because such collaborating organisations are required to assure themselves that appropriate safeguards are in place when sharing personal data as these organisations are themselves subject to the GDPR.
Name, email address, Affiliation, E-group membership, country of origin, photographs and audio recordings	External organisations supporting the local organisation of ARCHIVER events	Organisation of project events and all communication in the context of the ARCHIVER project.

Table 2: Personal Data that may be transferred to entities or individuals outside the project[22].

3. DMPs for the ARCHIVER deployment use cases

This section describes the initial plan for the data to be managed as part of the scientific use cases that will be deployed during the pilot phase of the pre-commercial procurement process.
As a general rule, a second copy of data will be used for the R&D activities during ARCHIVER. The second copy derives from the current data in the institutional and/or discipline-specific repositories for archiving and long-term preservation, curation and sharing.
Many of these repositories, e. g. CERN Digital Repository for its holdings (Scientific Data, Associated Software, Documentation, Metadata, etc.) are currently in the process of self-certification according to ISO 16363:

ISO 16363:2012 defines a recommended practice for assessing the trustworthiness of digital repositories. It is applicable to the entire range of digital repositories. ISO 16363:2012 can be used as a basis for certification.

This initial version of the Data Management Plan includes initial examples from the scientific use cases to be deployed, others will maybe be added during subsequent document revisions.

[22] https://cern.service-now.com/service-portal/privacy-policy.do?se=external-funded-projects&notice=ARCHIVER

The DMPs will be part of the technical summaries of each use case deployment[23] as a core component for the resulting services of the ARCHIVER project.

The plans are using the "Practical Guide to International Alignment of Research Data Management"[24]. In the context of the EOSC and with the increasing trend of sharing data and make it reusable, experts of Science Europe member organisations developed a set of core requirements for data management plans (DMPs) as well as a list of criteria for the selection of trustworthy repositories where researchers can store their data for sharing. The objective is to harmonise data management across Europe and across different funders and organisations.

The core requirements are summarized in the table below.

DMP Topic	Guideline in what needs to be addressed
Data description and collection or re-use of existing data	a. How will new data be collected or produced and/or how will existing data be reused? b. What data (for example the kinds, formats, and volumes) will be collected or produced?
Documentation and data quality	a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data? b. What data quality control measures will be used?
Storage and backup during the research process	a. How will data and metadata be stored and backed up during the research process? b. How will data security and protection of sensitive data be taken care of during the research?
Legal and ethical requirements, codes of	a. If personal data are processed, how

[23] https://www.archiver-project.eu/deployment-scenarios-technical-summaries
[24] https://www.scienceeurope.org/wp-content/uploads/2018/12/SE_RDM_Practical_Guide_Final.pdf

conduct	will compliance with legislation on personal data and on data security be ensured? b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable? c. How will possible ethical issues be taken into account, and codes of conduct followed?
Data sharing and long-term preservation	a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons? b. How will data for preservation be selected, and where will data be preserved long-term (for example a data repository or archive)? c. What methods or software tools will be needed to access and use the data? d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?
Data management responsibilities and resources	a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)? b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be **FAIR (Findable, Accessible, Interoperable, Re-usable)**?

3.1. CERN Digital Memory

DMP Topic	What needs to be addressed
Data description and collection or re-use of existing data	Patrimony data, such as historical images, videos, audios, as well as documentation, publications or conference content. Depending on policy, it could also include other types of data, like official emails, web pages, social media content, etc. All newly produced data selected as part of CERN heritage should be loaded to the archive.
Documentation and data quality	The descriptive metadata will come from the live information systems where users directly submit information. For example, photographers enter images into CERN Document Server organized as albums with title, captions, abstract, etc. The entire metadata should always be transferred to the archive, even if only part of it is actually well identified by the archiving system (e.g. the mandatory Dublin Core fields). The completeness of the metadata, the validity of the checksum and the proper identification of the file formats (plus other services like virus checking, fonts inclusion, etc) should be provided by the archiving service.
Storage and backup during the research process	Not relevant. Digital Memory does not focus on Research Data. Before the start of the Archival process, the DM data is maintained within existing active CERN information systems (like CDS, EDMS, Indico, AIS or others) who are in charge of ensuring the storage and availability of the data. These systems are all relying on the CERN Data Center infrastructure, with robust backup and restore procedures.

Legal and ethical requirements, codes of conduct	The communication between live systems and the archive should guarantee the transfer of information relative to personal data and copyrights. The data stored and managed by the archive is only at the disposal of the information system that deposited the SIP. Acting as a 'dark archive' leaves responsibility on the access rules to the system that has initially captured the content. The rule is that all the systems should try to prevent legal issues by making sure GDPR and copyright concerns are dealt with at the submission time. In case of failures in acquiring/transferring such information, the archive should always allocate to the transferred data the most secure access rules. These rules should actually be inspired (or copied) directly from the rules governing the CERN central (paper) Archive, with up to 100 years embargo period.
Data sharing and long-term preservation	Each single record of data should have its own ACL defining who has access, and optionally who should gain access within a given timelapse. This must be part of the definition of an AIP and it should therefore either be transferred inside the submitted SIP or added by the archiving system when ingesting the data. If an acl changes within the initial information system, it must be reflected within the archival system (by the creation of a new AIP or the update of an Archival Information Compound (AIC) object). The selection of the data to be preserved is not the responsibility of the archival system. A CERN wide policy (like OC3) and a governing body (like the Heritage Committee) should decide the content types/collections that must be part of this new long-term

	digital preservation platform. The archival system final output is to provide information systems with access to well formed standard, complete, verified and up to date AIPs. There are no specific software or methods to retrieve these AIPs. DOIs or equivalent should be minted by the Live systems, not by the Archive.
Data management responsibilities and resources	A mandate to run a Digital OAIS Archive governed by an Archive Committee should be given to an expert unit (logically within CERN IT). There should be no change of responsibilities at the level of the existing data producers/maintainers within the live Information systems. It is difficult at this point to precisely weigh the resources needed but the idea of running a 'dark archive' (instead of a user-oriented one) is to face the risks of digital obsolescence at a minimum cost, by avoiding duplication of services and interfaces. The maintenance of 'pipelines' to transfer data in the Archive should not be too resource-consuming; an average of 2 FTEs would sound reasonable for the Digital Memory type of content.

3.2. CERN Open Data

DMP Topic	What needs to be addressed
Data description and collection or re-use of existing data	The CERN Open Data portal manages several Petabytes of open data from LHC particle physics. The data are released by LHC collaborations in periodic batches after a certain embargo period to ensure their correctness. The data consists of raw data

	samples, experimental collision datasets and simulated datasets suitable for physics research use cases, the dedicated samples for designated communities such as Machine Learning and Data Science, up to simplified derived data formats and event display files suitable for education use cases. The data includes detailed provenance information with configuration files, virtual machines images, Docker containers, data production and analysis examples demonstrating how to work with the data. The data formats include physics-specific ROOT format, machine-friendly H5 format, up to simplified CSV and JSON formats for derived datasets.
Documentation and data quality	All released data is managed by a digital repository and is described as bibliographic records in JSON format. The format follows a custom JSON Schema describing the particle physics domain and allows to ensure the conformance of metadata information to described standard. The data is exportable in general formats via schema.org and JSON-LD. The bibliographic records contain information about data selection, validation and reuse. Several analysis examples help to ensure the data quality by allowing to rerun example code against data periodically. The data is released on the CERN Open Data portal is carried out in close collaboration with LHC experiments and follows their centralised DMP plans and QA practices.
Storage and backup during the research process	The data is stored on a CERN EOS distributed storage platform using several disk copies. The critical datasets are to benefit from tape storage for longer term. The ARCHIVER use

	case seeks to establish an independent copy of the open data portal content on cloud premises, looking at increasing safety via independent archive. The data is fully public and can therefore be does not contain any personal or sensitive information. The use case does not require any particular data protection plan.
Legal and ethical requirements, codes of conduct	The data concerns with disseminating open particle physics datasets and software and therefore does not contain any particular personal information. The data is typically released under CC0 waiver (for datasets) and GNU/GPL, ASL, BSD and MIT ASL licenses (for software). The data are fully open and can be accessed by anybody. The data does not contain any personal information and does not require anonymisation.
Data sharing and long-term preservation	The data is managed by an Invenio digital repository instance that offers FAIR services for the general public to discover, access, cite and reuse the data. The access protocols include HTTP and XRootD. The most important data assets such as collision and simulated datasets are minted with a DOI; the accompanying supplementary material such as configuration file snippets are accessible via local PID. The data sharing is governed by corresponding open licenses such as CC0 waiver for data and GNU/GPL for software. The ARCHIVER use case seeks to establish an independent preservation-friendly archive on cloud, for example for disaster recovery purposes (basic

	use case), in which case the access is mostly for Service Managers only. The open nature for the data makes is easy to seek independent data exposure and reuse on non-CERN cloud infrastructure (advanced use case), in which case the general public is encouraged to access and explore the data on independent computing infrastructure.
Data management responsibilities and resources	The data stewardship responsibility is being shared by the CERN Open Data portal repository team and the data preservation experts in LHC experiments. This concerns all the open data lifecycle steps from data ingestion, description and curation, up to data publishing and releasing. The ARCHIVER use case is mostly concerned with independent archiving and reuse services for the data. The data management responsibilities will remain with CERN Open Data portal repository team and the LHC experiment data preservation experts.

3.3. EMBL Cloud Caching

DMP Topic	What needs to be addressed
Data description and collection or re-use of existing data	The data to be used is primarily DNA sequence data from the European Nucleotide Archive (ENA), located in the EMBL-EBI data centres. Data is submitted to ENA by research teams around Europe.
Documentation and data quality	Data quality is measured as part of the DNA

	sequencing process, the quality measurement is included with the raw data. The origin of the data (organism, sample etc) forms part of the accompanying metadata collected during the submission process.
Storage and backup during the research process	Data is stored in the FIRE (FIle REplication) archive, consisting of one copy on a distributed object-store and one on a tape archive, all hosted on EMBL-EBI data centres
Legal and ethical requirements, codes of conduct	Much of the data in ENA is freely available, by anonymous FTP. Some data is tightly controlled, since it contains highly sensitive personal information - e.g. cancer tumour DNA sequences that can be traced to an individual. However, for the ARCHIVER project, we will only use the freely available data
Data sharing and long-term preservation	This use-case is about caching data in the cloud, there is no long-term preservation issue. Similarly, being a cache, the data will be transparently and freely available, given that the source data will be the freely available subset of ENA.
Data management responsibilities and resources	Management of the original copy of the data will remain the responsibility of EMBL-EBI. We make redundant copies of the data on different storage technologies (tape and object store) to minimise risk of loss.

3.4. EMBL on FIRE

DMP Topic	What needs to be addressed
Data description and collection or re-use of	See EMBL Cloud Caching.

existing data	
Documentation and data quality	See EMBL Cloud Caching.
Storage and backup during the research process	The FIRE archive maintains primary responsibility for the safe storage of our data, using a distributed object store and a tape archive for redundant copies. The ARCHIVER project will explore the possibility of replacing the tape store with a cloud-based archive, but any decision to do that will not be taken until after the conclusion of the ARCHIVER project itself.
Legal and ethical requirements, codes of conduct	Data is encrypted before being uploaded into FIRE, if there is any requirement that it be protected. FIRE itself does not manage access keys, the data is opaque bits as far as it's concerned.
Data sharing and long-term preservation	Long term preservation remains the primary responsibility of FIRE, using the redundant copies in different technologies for safety. Data is made available via FTP, HTTP and other protocols.
Data management responsibilities and resources	See EMBL Cloud Caching.

3.5. The BaBar Experiment

DMP Topic	What needs to be addressed
Data description and collection or re-use of existing data	BABAR is a particle physics experiment designed to study some of the most fundamental questions about the universe by exploring its basic constituents - elementary particles. The BABAR Collaboration's research

	topics include the nature of antimatter, the properties and interactions of the particles known as quarks and leptons, and searches for new physics.
Documentation and data quality	See the project website at http://www-public.slac.stanford.edu/babar/default.aspx. BaBar publications can be found at http://www-public.slac.stanford.edu/babar/Publications.aspx.
Storage and backup during the research process	During the research process the primary data was stored in IBM's HPSS system on robotic tape storage in the SLAC computer center. Copies were made to TierA sites, including several in Europe (RAL in the UK, IN2P3 in France, KIT in Germany etc.) Given the data volumes, the data itself was not "backed up" but HPSS was responsible for ensuring data integrity. BaBar (SLAC) was a member of the DPHEP Study Group and details of its data preservation objectives and benefits can be found in http://arxiv.org/pdf/1205.4667.
Legal and ethical requirements, codes of conduct	N/A. Information on the organisation of the BaBar collaboration, responsibilities etc can be found at https://www.slac.stanford.edu/BFROOT/www/Organization/index.html (including the Collaboration Governance, Management Plan and Membership).
Data sharing and long-term preservation	The BaBar experiment does not make "Open Data" releases. Its data may only be used by members of the BaBar Collaboration. However, in principle anyone can become a collaboration member by writing to the

	BaBar Spokesperson. Some of the BaBar data is unique and a double copy is already "preserved" at CERN. Additional copies also exist at sites that are members of the Collaboration and the intent is to store a further copy using ARCHIVER services to study feasibility, cost of entry, cost of ownership etc.
Data management responsibilities and resources	See https://arxiv.org/pdf/1205.4667.pdf.

3.6. Petra III_EuXFEL Data Archiving

DMP Topic	What needs to be addressed
Data description and collection or re-use of existing data	Raw, Calibration and derived data are going to be handled in ARCHIVER. Raw and Calibration data is produced by detectors directly - other (derived) data is generated by computer based data analysis. In most cases all data is in HDF5 format and handed over to ARCHIVER for long term storage (cold data) and to generate a second copy (with respect to the primary storage used while high access rates are expected) as soon as possible.
Documentation and data quality	https://www.xfel.eu/data_privacy_policy/index_eng.html https://www.xfel.eu/users/experiment_support/policies/scientific_data_policy/index_eng.html
Storage and backup during the research process	Today, DESY is using its dCache/OSM based service to archive 'cold data' in a HSM like (disk + automated tape) service. That layer is going to be replaced/extended with ARCHIVER services using an additional copy of the data. (cleanup at the end of the project

	requires no further data movement).
Legal and ethical requirements, codes of conduct	The selected scientific data used in Phase II+III will not contain any personal data. Ownership and IP property is addressed in the scientific data policy, Section 5: https://www.xfel.eu/sites/sites_custom/site_xfel/content/e51499/e51503/e52947/e56273/e56274/xfel_file56275/ScientificDataPolicyapprovedbyCouncilon30June2017_eng.pdf
Data sharing and long-term preservation	Open Data is already covered by the selected data policies (second item) and selected (subset) scientific data stored in ARCHIVER, will be candidates to verify open data handling according to the policies mentioned before. All data will be HDF5 formatted - HDF5 is provided by the HDF collaboration (https://www.hdfgroup.org). The idea to use a dedicated copy of open data residing in a public cloud, will include a seamless method to use public cloud compute resources to process that scientific data.
Data management responsibilities and resources	Responsibilities are addressed in the scientific data policy, Section 4 (main responsibilities): https://www.xfel.eu/sites/sites_custom/site_xfel/content/e51499/e51503/e52947/e56273/e56274/xfel_file56275/ScientificDataPolicyapprovedbyCouncilon30June2017_eng.pdf

3.7. PIC Data Distribution and Large/Mixed File Remote Storage

DMP Topic	What needs to be addressed
Data description and collection or re-use of existing data	Raw, calibration and derived datasets will be handled in the ARCHIVER project. These data will be from scientific projects in which IFAE

	participates and which have explicitly given their consent. Initially, data will be from the MAGIC Telescopes https://magic.mpp.mpg.de/ . Data will be re-used in the context of multi-instrument gamma-ray astrophysics.
Documentation and data quality	Data documentation will be done by adapting the internal Data Management and Preservation Plans (DMPP) of the scientific projects who own the data. MAGIC has already given permission to use their DMPP plan. Data quality is measured continuously by the scientific instruments and data quality tags will be available as metadata.
Storage and backup during the research process	Today, PIC is using its dCache/Enstore based service to archive 'cold data' in a HSM like (disk + automated tape) service. That layer is going to be replaced/extended with ARCHIVER services using an additional copy of the data. (cleanup at the end of the project requires no further data movement).
Legal and ethical requirements, codes of conduct	Citation should be given when data is used by permission.
Data sharing and long-term preservation	Easy to use data sharing, both amongst access-controlled groups as well as Open Data, is a major objective of placing the data in the ARCHIVER-developed platforms. Long-term preservation will be tested by simulating in an accelerated manner the Data Management and Preservation cycle.
Data management responsibilities and resources	High-level data management will remain the responsibility of IFAE, while low-level data management and resources will be provided by the vendors.