Data Management Plan (D1.2)

Project	HORIZON-ZEN - EU research programme beneficiary depositing solution in Zenodo.
Work package	WP1
Deliverable no.	D1.2 (Public)
Authors	Nielsen, Lars Holm (CERN)
Reviewers	Gonzalez Lopez, Jose Benito (CERN)
Version	1.0
Date	2023-10-31
DOI	https://doi.org/10.5281/zenodo.10059542

Funded by the European Union

Funded by the European Union under Horizon Europe grant agreement number 101122956

Table of contents

Data Summary **3**
FAIR data **3**
Making data findable, including provisions for metadata 3
Making data accessible 4
Making data interoperable 4
Increase data re-use 5
Data security 5
Ethics 7
Software **7**

Funded by
the European Union

Data Summary

The primary objective of the HORIZON-ZEN project is to enhance Zenodo with FAIR-enabling capabilities to support beneficiaries of the EU programme when depositing their research outputs. This enhancement aims to leverage Zenodo's straightforward and user-friendly interface.

The HORIZON-ZEN project proposal does not include the creation of datasets as part of its scope. Therefore, the data management plan outlines the management of other research outputs within the project and describes Zenodo's data management. If the situation changes, and there is a possibility of creating datasets, the plan will be updated accordingly in advance of the creation.

HORIZON-ZEN primarily produces software in the form of additions to the existing Zenodo repository service. The software is made available for reuse as a technical platform in InvenioRDM.

In addition, Zenodo, as a trusted repository service, hosts research data and outputs created by other Horizon Europe projects. The data management plan describes how Zenodo manages FAIR data, as well as how the software is managed.

FAIR data

The following section describes Zenodo’s data management practices.

Making data findable, including provisions for metadata

Persistent identifiers
All records in Zenodo are identified by registering a Digital Object Identifier (DOI) via DataCite.

Rich metadata
Metadata in Zenodo is stored internally according to the InvenioRDM metadata format which is a superset of DataCite Metadata Schema v4. Zenodo exports all records in the the following metadata formats: JSON, JSON-LD (Schema.org), Citation Style Language, DataCite JSON v4, DataCite XML v4, Dublin Core (according to OpenAIRE Guidelines for Data and Literature Repositories), MARCXML, BibTeX, GeoJSON, W3C DCAT.

Discipline-specific metadata
Zenodo supports DarwinCore Metadata standards for biodiversity (searchable). In addition Zenodo allows users to upload additional metadata files to provide discipline-specific metadata (not searchable). Users of Zenodo are responsible for providing the metadata.

Making data accessible

Zenodo is a trusted repository and part of the EOSC federation.

Data access
All metadata is licensed under Creative Commons Zero. Files, on the other hand, are licensed according to the user's specification, with the default being CC-BY. Zenodo accommodates both public and restricted content, including embargos, allowing users to determine the appropriate license for their materials. HORIZON-ZEN will introduce additional features to ensure that outputs deposited within the Horizon Europe (HE) community align with HE contractual requirements (refer to https://doi.org/10.5281/zenodo.8419426 for details).

Zenodo provides access to data via HTTPS protocol and openly available APIs and metadata standards (see above).

APIs
Zenodo provides an OAI-PMH interface to harvest all or specific subset of metadata. In addition a rich REST API provides access to all content. Record landing page's embeds JSON-LD (Schema.org) metadata according to the *A data citation roadmap for scholarly data repositories*[1] for interoperability with Citation Managers and e.g. indexing in Google Dataset Search. Zenodo further register's all uploads with DataCite Commons.

Retention
Deposited records are retained for the lifetime of the repository. This is currently the lifetime of the host laboratory CERN, which currently has an experimental programme defined for the next 20 years at least. All records of Zenodo can be exported in the Oxford Common Filesystem Layout.

Making data interoperable

Creators/contributors can be identified using ORCID, RORs, GNDs, and ISNIs. Licenses are persistently identified using SPDX. Funders are identified using RORs, grants using the OpenAIRE Graph. Related works can be identified using DOI, Handle, ARK, PURL, ISSN, ISBN, PubMed ID, PubMed Central ID, ADS Bibliographic Code, arXiv, Life Science Identifiers (LSID), EAN-13, ISTC, URNs, and URLs. All resource types are mapped to DataCite, BibTeX, Schema.org and Citation Style Language vocabularies.

Cross-disciplinary metadata is provided via DataCite Metadata Schema and further export formats. Discipline-specific metadata is being implemented as part of the project, in addition to DarwinCore.

[1] Fenner, M., Crosas, M., Grethe, J.S. et al. A data citation roadmap for scholarly data repositories. Sci Data 6, 28 (2019). https://doi.org/10.1038/s41597-019-0031

Increase data re-use

Zenodo offers extensive documentation on the utilization of our APIs and the features provided by Zenodo. You can find more information at https://help.zenodo.org/. Additionally, Zenodo supports users in understanding the potential uses of records through standard license vocabulary. It's important to note that metadata is consistently licensed under CC0.

Data security

Zenodo is powered by CERN Data Centre and InvenioRDM repository platform and is fully run on open source products all the way through.

Physically, Zenodo's entire technical infrastructure is located on CERN's premises which is subject to CERN's legal status. CERN is an intergovernmental organization and has legal personality in the metropolitan territories of all CERN Member States (CERN Convention, Article IX) and enjoys the corresponding legal capacity under public international law.

As an intergovernmental organization CERN enjoys certain privileges and immunities, including e.g. immunity from jurisdiction of the national courts to ensure our independence from individual Member States. This does not mean that CERN operate in some kind of legal vacuum as protocols require that CERN settle its disputes by arbitration.

Server management
Zenodo servers are managed via the OpenStack and OpenShift container management platform and have the latest security patches applied. Servers are monitored via CERN's monitoring infrastructure based on Flume, OpenSearch, OpenSearch Dashboards and Hadoop. Application errors are logged and aggregated in a local Sentry instance. Traffic to Zenodo frontend servers is load balanced via a combination of DNS load balancing and HAProxy load balancers.

We are furthermore running two independent systems: one **production** system and one **quality assurance** system. This ensures that all changes, whether at infrastructure level or source code level, can be tested and validated on our quality assurance system prior to being applied to our production system.

Front End servers
Zenodo frontend servers are responsible for running the Invenio repository platform application which is based on Python and the Flask web development framework. The frontend servers are running nginx HTTP server and uwsgi application server in front of the application and nginx is in addition in charge of serving static content.

Data storage
All files uploaded to Zenodo are stored in CERN's EOS service in an 18 petabytes disk cluster. Each file copy has two replicas located on different disk servers.

For each file we store two independent MD5 checksums. One checksum is stored by Invenio, and used to detect changes to files made from outside of Invenio. The other checksum is stored by EOS, and used for automatic detection and recovery of file corruption on disks.

EOS is the primary low latency storage infrastructure for physics data from the Large Hadron Collider (LHC) and CERN currently operates multiple instances totalling 1+ exabyte of data with expected growth rates of 100-500 petabytes per year.

Metadata storage
Metadata and persistent identifiers in Zenodo are stored in a PostgreSQL instance operated on CERN's Database on Demand infrastructure with 24-hourly backup cycle with one backup sent to tape storage once a week. Metadata is in addition indexed in an OpenSearch cluster for fast and powerful searching. Metadata is stored in JSON format in PostgreSQL in a structure described by versioned JSONSchemas. All changes to metadata records on Zenodo are versioned, and happening inside database transactions.

In addition to the metadata and data storage, Zenodo relies on Redis for caching and RabbitMQ and python Celery for distributed background jobs.

Security policies

- CERN Data Centre: Our data centres is located on CERN premises and all physical access is restricted to a limited number of staff with appropriate training and who have been granted access in line with their professional duties (e.g. Zendo staff do not have physical access to the CERN Data Centre) .
- Servers: Our servers are managed according to the CERN Security Baseline for Servers, meaning e.g. remote access to our servers are restricted to Zenodo staff with appropriate training, and the operating system and installed applications are kept updated with latest security patches via our automatic configuration management system Puppet.
- Network: CERN Security Team runs both host and network based intrusion detection systems and monitors the traffic flow, pattern and contents into and out of CERN networks in order to detect attacks. All access to zenodo.org happens over HTTPS, except for static documentation pages which are hosted on GitHub Pages.
- Data: Zenodo stores user passwords using strong cryptographic password hashing algorithms (currently PBKDF2+SHA512). Users' access tokens to GitHub and ORCID are stored encrypted and can only be decrypted with the application's secret key.
- Application: We are employing a suite of techniques to protect your session from being stolen by an attacker when you are logged in and run vulnerability scans against the application.
- Staff: CERN staff with access to user data operate under CERN Operational Circular no. 5, meaning among other things that
    - staff should not exchange among themselves information acquired unless it is expressly required for the execution of their duties.
    - access to user data must always be consistent with the professional duties and only permitted for resolution of problems, detection of security issues, monitoring of resources and similar.
    - staff are liable for damage resulting from any infringement and can have access withdrawn and/or be subject to disciplinary or legal proceedings depending on seriousness of the infringement.

Ethics

Zenodo is an open dissemination research data repository for the preservation and making available of research, educational and informational content. The uploader shall ensure that their content is suitable for open dissemination, and that it complies with these terms and applicable laws, including, but not limited to, privacy, data protection and intellectual property rights.

Zenodo operates under CERN's Personal Data Protection which provides similar protection as EU GDPR.

Software

All software that is part of or delivered by HORIZON-ZEN is distributed under an Open Source Initiative-approved open-source license (by default the MIT license) and maintained publicly on GitHub repositories and installable via package management repositories according to Zenodo's and InvenioRDM's existing development practices. See https://inveniordm.docs.cern.ch/install/

Standards and Metadata
Both Zenodo and InvenioRDM are extensively documented according to best practices and the extensions developed in HORIZON-ZEN will follow the same documentation pattern that involves architecture, developer, administrator, and end-user documentation. See https://inveniordm.docs.cern.ch/develop/

Software archiving and preservation
All software is made accessible through GitHub and will be archived and preserved through Software Heritage.

Software development practices
HORIZON-ZEN relies on the mature development process already employed for the development of Zenodo/InvenioRDM. The development involves a collaborative and open design process taking into consideration user experience, accessibility, security, and performance and employing iterative development, code reviews, testing, quality assurance, internationalization, and documentation and thus ensures a very high-level of software quality.