ARCHIVER Project

Technical Summary

CERN OPEN DATA

Problem Definition:
The CERN Open Data portal disseminates more than one petabytes of primary and derived datasets from particle physics as they were released by LHC collaborations. The portal offers datasets together with accompanying software examples, virtual machine images, condition databases, configuration files and necessary associated documentation to enable non-specialists to use the data.

The CERN Open Data service managers seeks an easy-to-use, easy-to-achieve independent archiving and backup service for its holdings. The archiving service will be based on SIPs [Submission Information Packages] encompassing research data. The Archive should offer simple retrieval mechanisms that will be used mostly for disaster recovery purposes.

As a more advanced use case, we want to explore the possibility of offering reproducibility services on top of the Archive infrastructure. This would consist of opening the Archive also to independent researchers (not only service managers) which would imply to support community protocols (such as XRootD) to access the files. The most complex use case would consist of the Archive offering complementary “live” services such as instantiating and running software repositories on the cloud so that “research reproducibility” can be achieved completely independent of the CERN infrastructure.

Lifecycle - Workflow Characteristics:
The CERN Open Data portal coordinates data releases with LHC experiments. The end users are not allowed to self-deposit the data. The data is curated by the CERN Open Data portal team together with physicists and is being released together after checks.

There are about three or four release batch campaigns happening per year. We therefore do not need Archive to be constantly ingesting data throughout the year but we can restrict data ingestion to a few well defined “ingestion campaign” periods per year. The ingestion therefore does not have to be particularly agile and can run on the order of ~500TB per month.

Due to the open nature of the data assets, it is possible to think of both a “push” scenario (where the CERN Open Data portal would initiate deposits into the Archive) and a “pull” scenario (where Archive would periodically crawl CERN Open Data portal for new content). The ingestion can happen over any protocol such as HTTP.

The recall of the data from the archive is rare. The recall may happen on demand by the Service Manager in order to replace a defect file in the portal. The end users won’t have access to the Archive. In case of data recall by the Service Manager, the full dataset unit will be downloaded from the archive, although it would be preferable to be able to access one particular file from the dataset. Note that one dataset can be several Terabytes big and can consist of several thousands of files of the Gigabyte size. In the case of a data loss of one file on the CERN Open Data portal, the size of which would be typically ~4 GB, we would like to fetch from the Archive the given file only, and not the full dataset of ~4 TB size.

Layer 4 Features based on Remote Object Storage and Software Reproducibility

For the advanced use case we are looking at offering some “reproducible” services on top of the Archive. This notably concerns running additional services on top of the data storage in order to be able to run example analyses using non-CERN infrastructure.

The CERN Open Data portal disseminates Virtual Machine images that use CernVM File System to serve the software and condition database during analysis runtime.

The CernVM File System provides a scalable, reliable and low-maintenance software distribution service. It was developed to assist High Energy Physics collaborations to deploy software on the worldwide-distributed computing infrastructure used to run data processing applications. CernVM-FS is implemented as a POSIX read-only file system in user space. Files and directories are hosted on standard web servers and mounted in the universal namespace /cvmfs.

As CernVM-FS can use S3 protocol for storage, we want to explore three possibilities:

(a) use CernVM-FS servers at CERN but with its S3 backend storage run via external infrastructure
(b) run CernVM-FS servers completely in an external service (such as cvmfs.example.com)
(c) allow end users to run examples of open data analyses via VMs or containers residing in external infrastructure

This would ensure the reusability of open data as all of the data, code, compute environments, and example analyses will run independently of the original CERN cloud platform.

Authentication and Management Functions:
The authentication needs of the CERN Open Data portal use case are minimal, because all the data are fully public and do not require any special access control. The only person allowed to deposit data to the Archive is the Service Manager. However, Federated IAM becomes essential for the case of research reproducibility on public infrastructure.

Due to the particular nature of research data material, the Archive does not have to process the deposited material in any special way beyond the usual conformance tests and fixity checks. It would be useful to have periodic reports of any data management processes and activities happening in the Archive.

Data and Metadata Characteristics:
There are about 10'000 bibliographic records consisting of 600'000 files of the total size 2 PB stored in the portal. The formats vary from ROOT files for primary datasets through H5 files for data science applications up to CSV files for simplified datasets.

The data is packaged in the form of datasets. A typical dataset size is of the order of 3 TB. One dataset typically contains about 3000 files. Each individual file is of the order 3 GB size. The individual files can be smaller or larger but do not usually exceed 5 GB.

The software is provided in various formats, usually packaged in zip files. The software can be Python or C++ files. The software can be preserved in a tarball manner.

The documentation is provided usually in PDF and Markdown formats.

The metadata is described by means of custom JSON Schema which is inspired by WC3 DCAT standard. However, the rich information about particle physics nature of the data is stored in custom JSON fields as there is currently no specific ontologies or vocabularies describing research data in particle physics. The JSON Schema describing the data comes with a set of mandatory fields (such as "creation date") and field types (such as "string") that can be used for schema validation.

The CERN Open Data portal provides serialisation of its custom JSON format into JSON-LD using schema.org.

One particularity of the CERN Open Data portal is the wide-spread usage of "fetch.txt" technique in the BagIt packages generated by the portal. Due to large number of associated files in the dataset, and due to the fact that these files may exist in a safe copies elsewhere, certain BagIt packages are small in size but contain a large number of files listed in "fetch.txt". It would be beneficial if the Archive triggers the preservation action for the files specified in the "fetch.txt" upon ingestion.

Interface Characteristics:
For the basic use case, there are no particular requirements on interfaces. The Service Manager should have a well-defined way how to deposit data for ingestion by the Archive and how to retrieve data from the Archive. For both ingestion and recall, the HTTP protocol with well-defined REST API services are preferred.

For the advanced use case, we would need the Archive to support XRootD access to files so that the Archive could be used by the physicists directly without passing by the CERN Open Data portal.

For the service manager dashboard, a web access should be offered for interactive peruse. A REST API access to report logs for individual assets or for certain time periods would be a plus.

Reliability Requirements:
For the basic use case, the Archive should be able to guarantee file recall within a few hours.
For the advanced use case, the Archive should be able to guarantee “immediate” file recall.
In both cases, the Archive is expected to store sufficient number of copies of data on disks or tapes in order to guarantee bit preservation.

Compliance and Verification:
The Archive should validate the integrity of deposited material through JSON Schema validation upon ingestion.
The Archive should run periodic fixity checks to verify the checksums of the deposited files.

Cost Requirements:
For the basic preservation and archiving part, we are looking at the most effective cost solution that would guarantee minimally reasonable ingestion and recall speeds (estimated around 10 Gbps). However, we expect that for the “research reproducibility” use case, the speeds should be at least one order of magnitude higher.
The focus is on providing excellent data integrity in the archiving service in order to avoid data loss. The data recalls from the Archive are expected to be very rare. The service cost should include data preservation management tasks to guarantee the service over a long period (more than 5 years).
For the advanced for the “research reproducibility” use case, the cost will have to be studied in detail with respect to offering (i) XRootD protocol for recall; (ii) S3 storage for CVMFS service at CERN; (iii) running CVMFS server on external infrastructure; (iv) sufficient bandwidth to reply to N connected physicists.

Initial Data Management Plan:

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 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.