ARCHIVER Project

Technical Summary

CERN DIGITAL MEMORY

Problem Definition:
We need to archive the CERN Digital Memory which consists of the digital production of the institution for the 21st century (including new types like web sites, social media, emails, etc) as well as the analog documents produced by the institution in the 20th century, composed of digitized papers (physical archive) and various multimedia: audio (e.g. recordings of meetings), still images and moving images.

The digitized institutional content is loaded and maintained into CERN Live information systems. These information systems use various underlying storage solutions (e.g. systems like DFS, EOS etc.) but none of them is OAIS compliant. The goal is to connect the active services with a dark archive where Archival Information Packages will keep comprehensive information for each 'document'. The goal is to have at our disposal at the end a standard trustworthy ISO16363-compliant digital archive where live systems can deposit content selected for long term preservation.

By ensuring that data are kept in the same archiving solution, we can introduce specific standards and formats for long term preservation and, as a consequence, minimize fragility which is one of the highest risks for long term preservation in the organization. We want to ensure that we keep a globally accepted preservation processes and standard formats for the same type of content (e.g. video).

Aligning the CERN digital archive to best practices (OAIS) for the sake of long term digital preservation is one of the most important benefits for this use case. In addition to this, the existence of successful disaster recovery solution institution-wide could impact all individuals, as everyone produces the same types of content.

Lifecycle - Workflow Characteristics:

- The authentication needs for the basic use case are minimal, as only the Service Manager will need to access the Archiving Service and not the end user.

- One important aspect of the use case is the ability to have partial access to the data, i.e. to recall just one file or even a chunk of a file out thousands of files in a specific dataset. The system manager should be able not only to submit and download data from the service but also to update and re-ingest specific parts.
- We want the files to be accessed reasonably quickly but not live and we expect the files to be accessed relatively rarely. Consequently, the data ingestion needs for this scenario are not particularly high.
- The existence of personal data is also important for this scenario. For example, we need to be able to deal with cases of users asking their photos to be accessed or removed from the archiving service, in compliance with CERN Internal Regulation (Operation Circular No11) or GDPR.

Authentication and Management Functions:

- The authentication needs for the basic use case are minimal, as only the Service Manager will need to access the Archiving Service and not the end user.
- The user access control requirements for this scenario are particularly high as the ability to store files that are readable only by specific users is essential. For example, the Service Manager should not be able to access to the audio of a CERN Council meeting, even if he will still be the one submitting the files to the service.. As a consequence, files need to be encrypted and this drastically increase the R&D challenges in terms of long term preservation.

Data and Metadata Characteristics:

The overall size of the CERN Digital Memory is estimated to be ~700TB of digitized material and approximately ~700TB of digitally born files. For the vast majority of the data we have well defined metadata. However, as there are all different types of content, including documents, images, videos, sound recording etc. the existing metadata also vary drastically in terms of schema, existing fields and volume. There is no universal standards and formats for the contents of the CERN Digital Memory

After the backlog of data is ingested, the daily ingestion rate will require some more evaluation as it will depend on which information systems will be plugged into the archiving service. Ingestion of Invenio-based services like videos.cern.ch or zenodo.org should be sized to be able to deal with an average speed of 1.5MB/s.

Interface Characteristics:

There are many types of interfaces needed. These are not exhaustive.

- API from active system to the dark archive to enable:
    - Automation of the transfer of a SIP (or Baggit) from the information systems into an archiving pipeline to create the AIP
    - Automation of the inclusion of the metadata part into the indexes of the Archive system

    - Automation of the definition of the processing steps (configuration options) to overwrite the default configuration for a given SIP
    - Automation of the fetching of ‘external’ files at the time of archiving
    - Automation of the re-ingestion of a SIP with modified metadata or modified files
    - Automation of the creation of an AIC (Compound object composed of multiple AIPs)
- API from dark archive to the active information systems to enable:
    - Return the status of an archiving process for a given SIP (id)
    - Provide detailed information in case of failures, individually or by batch (e.g.: all the failures on Week 23)
    - Provide direct access to the created AIPs
    - Provide direct access to converted files (depending on the formats), and corresponding checksum files
    - Provide the log of the operations run for the preservation of a given SIP (id)
    - Provide a risk assessment for a given AIP (id)
- Web Interface for active system managers
    - A Dashboard interface with Browsing and Searching capabilities to search/discover the preservation processes. Each Info System manager should only access the dashboards relative to the SIPs/Baggit coming from his system.
    - An Audit log interface where the details of all actions can be analysed.
- Web Interface for the Archive Manager
    - Administrative and Access interface to manage the dashboards and the storage features of the Archive system,

Reliability Requirements:
The possibility to deploy new pipelines to cover the preservation of new content types (e.g. emails, websites and others) and scale up the archive whenever needed is key to the reliability of the service. In addition, the possibility to run the process directly on the CERN Data Cloud would be a plus - so that the Data will not be hosted on external non-controlled storage. This is particularly true for confidential data.

The possibility to support encryption of Data would provide more reliability for the handling of confidential content. Encryption could be at the info system level (bit preservation, provenance, authenticity only) or at the AIP level, after processing the submitted SIP in clear.
To provide more trust in the Archive, some preservation information like the fixity and the provenance could also be published to a public Blockchain (distributed and read-only). The key would be auditable, unchangeable, and open and it would serve as a single source of truth which all users could trust. At any point this information could then be used to audit and validate the information stored in the Archiving system.

Compliance and Verification:
Ideally, the system would be able to support a data management plan. For examples:

- Interval to run fixity checks on the assets should be configurable and managed in the Archiving system admin interface
- Verification on the end of embargoed files and changes of access rights accordingly (closed -> public or closed -> restricted etc) should be supported, and reported.
- Batch migration of formats depending on planned obsolescence, or new emerging standards.
- Additional Documentation Information required by OAIS (e.g. Designated Community, etc) could be integrated into the Archiving System to help its maintenance and to support regular reviews.

Cost Requirements:

Maintenance costs should not jeopardize the preservation funding. Some costs can be well defined, and others are more difficult to estimate. For this deployment scenario, we are particularly interested to receive cost estimations through the Curation Costs Exchange Interface.

Initial Data Management Plan:

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.