New infrastructure for Ingest and Preservation of Digital Records

Preliminary description of scope.

1. Introduction

The Danish National Archives (DNA) seeks to replace its current infrastructure for digital archiving within a few years with, preferably, an off-the-shelf solution using open standards and open source and providing automation, independence and easy maintenance. This document is a brief introduction to the draft needs and requirements such a solution is expected to include. The actual System Requirement Specification is to be developed by the consultant in close cooperation with the project team at the Danish National Archives. Thus, this document is only preliminary and cannot be used as the basis for an actual market research.

2. Background

The DNA ingests and stores data submitted from different public agencies on both state, regional and municipal level in Denmark, as well as research institutions and private individuals and companies. The DNA also digitizes and preserves significant amounts of paper records. The current infrastructure and its components is, due to historical reasons, mostly developed in-house and consist of both software and hardware.

The new infrastructure for ingest and preservation of digital records, both born-digital and digitized, should be OAIS compliant and designed to ensure that we, at all times, have the right capacity (and scalability) to ingest, process and preserve the relevant amounts of data as efficiently as possible and as securely as necessary. It must be built on the principles of distributed digital preservation and ensure that data can be retrieved for reuse and access as needed.

3. Draft general requirements

3.1 Capacity

Currently, the DNA approves of about 60 TB born digital data per year, and the rate has been increasing quite steadily. However, the amount of processed data is significantly higher, as data usually is submitted more than once due to errors in the first submission. In addition, data goes through several processes from ingest to preservation. The current collection of born digital records is approximately 300 TB per set for submitted data – 900 TB in total.

The largest SIP so far has been around 15 TB but we are expecting an increase in the number of very large SIPs of up to maybe 30TB or even more each. It must be possible for the infrastructure to ingest and validate SIPs, and produce preservation copies within a reasonable timeframe.

For digitized records, we expect the around 600 TB of digitized data, increasing with 75 TB annually, to be managed by the future digital archiving infrastructure.

Since we have no means of knowing exactly what the requirements for capacity will be in the coming years, scalability is a core requirement.

3.2 Efficiency

The new infrastructure should support automated workflows and if possible reduce the number of manual processes. It must be possible automatically to collect relevant metadata about the processes of ingest and preservation and it should be possible to share those metadata with relevant applications outside the infrastructure itself.

3.3 Security
Data processed and preserved by the DNA have various needs for protection in terms of information security. The infrastructure must provide the necessary technical measures to ensure a level of security appropriate to the risks of the confidentiality, integrity, availability and authenticity of the digital records.

The infrastructure must help the DNA avoid dependency on individuals.

4. Draft requirements related to OAIS functional entities

4.1 Ingest
Currently born-digital records are submitted as SIP's according to the Executive Order on Information Packages: https://www.sa.dk/wp-content/uploads/2020/05/Executive-Order-on-Information-Packages128-1.pdf

In a few years, the DNA expects to comply with the E-ARK IP format: https://earkcsip.dilcis.eu/

However, there is currently no fixed IP structure for the digitized records, and it is generally extremely important that the infrastructure is flexible enough to handle various IP structures.

The ingest process includes validation of the submitted data and it must be possible to integrate relevant components in the infrastructure, e.g. the validation tool for SIPs that has been developed by the DNA. Preferably, the new infrastructure includes a variety of validation components.

4.2 Preservation
Preservation actions include, but are not limited to, production of preservation copies on relevant storage media, integrity check and migration.
The principles of distributed digital preservation means that data should be stored in independent copies in various locations using various media technologies.

4.3 Access
The infrastructure must make it as easy as possible to retrieve archived data for reuse and it must have the necessary interfaces to integrate with relevant external solutions such as the infrastructure for access.

The actual access functionalities are not included in the scope of the infrastructure. However, if market research shows that there are relevant off-the-shelf solutions that can include access as well as ingest and preservation, it is possible to widen the scope of the infrastructure.