Early Adopter: Stockholm University Library

Early Adopter's Name:
Stockholm University Library (**SUL**)

Organisation type:
research institutions and universities

Organisation size:
Stockholm University Library **- medium-sized enterprise:** with 50-249 persons employed
Stockholm University (**SU**) - **large enterprise:** with 250 or more persons employed

Organisation Research Field(s):

- Social Sciences
- Natural Sciences
- Humanities

Organisation Profile:
SU is a research and higher education organisation, with currently more than 27,000 students, 1,400 doctoral students, and 5,700 members of staff, offering 300 programmes and 1,700 courses, including 75 master's programmes taught in English within the wide research areas above. The university has a total revenue of SEK 5.3 billion. [https://www.su.se/english/about]
Our prime stakeholder group for the Archiver project are the researchers at all levels – from PhD students, post-docs to senior researchers and professors. Other important stakeholder groups are funding organisations, requiring good quality documentation of funded research results, and Research Data Management staff (analysts, archivists, counsellors, curators, IT-staff) at SU. Another stakeholder is defined by the *Swedish Freedom of the Press Act*, in which *the principle of public access to official documents* has been enshrined. This means that "[i]n principle, all Swedish citizens and aliens are entitled to read the documents held by public authorities." [https://www.government.se/information-material/2009/09/public-access-to-information-and-secrecy-act/]

Organisation website URL:
https://www.su.se/english/

Suggested Use case title:
Multi-Repository Research Data Harvester and Transformer for Swedish Archival Standard

Problem definition:
We want to be able to harvest and transform datasets from different research data repositories, enriching them as needed with metadata from other sources, and at the same time comply with Swedish law and National Archive regulations as well as **GDPR**. We currently use and *curate* collections from these repositories: *dataverse.harvard.edu*; *su.figshare.com*; *zenodo.org* and *snd.gu.se*. SU also hosts the Bolin Centre Database for climate and earth system data, which is curated by domain specialists at SU. Individual researchers at SU may also use other repositories, e.g. *Datadryad, Pangaea.de etc*. While most repositories in use are cloud-based, we now consider having also our own repository, on a SU server, possibly a local Dataverse instance.

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

We also prefer a local storage for our long-term preservation digital archive. As a partner of the Swedish National Data Service consortium (https://snd.gu.se/en), users should be able to retrieve DIPs from our digital archive through the SND metadata catalogue. Further, the Archiver solution must handle **version control** when transforming SIPs to AIPs and DIPs, so that different versions of the same dataset SIPs, harvested from repositories at different occasions, are recognized and "bundled" together in the AIPs and DIPs. **DIPs** should be derived from AIPs, and be able to be **searched, requested and delivered by the same method**, **independently of original source** repository. At **all stages**, from ingest to DIP delivery, **authority control, access rights** and **potential confidentiality management** must be possible.
Here follows a set of further links to rules and regulations that the Archiver solution should comply with:

- SU Research data policy and Rules for research data management
- https://dilcis.eu/specifications/ : Common Specification for Information Packages (CSIP), E-ARK (SIP, AIP, DIP)
- **OAIS-model** (Magenta-book, ISO 16363:2012), **FAIR** principles: https://doi.org/10.1038/sdata.2016.18
  TRUST principles: https://doi.org/10.1038/s41597-020-0486-7, **CTS – Core Trust Seal**

Archive size: 500 TB, expansion possible.
Lifecycle: storage min. 10 years, most files requiring preservation indefinitely.

Is this use case new for your organisation?

If so, what is the envisaged timeline for implementation?

The bulk of this use case is new (for what is already implemented, see below), involving a local repository, an OAIS and GDPR compliant real digital archive (with prospective *Core Trust Seal* certification), producing SIPs, AIPs and DIPs according to selected metadata standards, with the enrichment of preservation metadata (PREMIS, PROV) and conversion / migration to sustainable file formats for those that are subject to obsolescence in the near future.
The envisaged timeline for full implementation is estimated to 2-3 years, with control stations on the way.

If not, how is it currently implemented?

It is partly implemented currently through a locally developed software package for harvest and transform of research data from su.figshare.com (described here: 10.5281/zenodo.1203726) and now also from zenodo.org.
The harvested and transformed research data and metadata are then deposited (as SIPs conforming to the Swedish National Archive METS standard *FGS-CSPackage*, which is essentially the same as the *dilcis.eu* CSIP referred to above) in a temporary file storage archive, *MADI* on a SU server, currently holding some 200 GB in total, of which over 80% are harvested and transformed research data (a proportion that may change over time). This is while we are awaiting the implementation of a full-fledged digital archive (OAIS model), in which a further transformation to AIPs and DIPs can be made.

Data and metadata Characteristics:

Currently, research data from SU researchers within all three research areas (Natural Sciences, Social Sciences and Humanities) comprise a wide variety of file formats and file sizes. For data files, we encourage researchers to deposit in *non-proprietary*, commonly used and *sustainable* file

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

formats (e.g. from the Library of Congress list), but we cannot force anyone to deposit only recommended file formats. This means the Archiver solution should allow for file format conversion when needed, also as part of preservation measures according to a migration plan, requiring monitoring of obsolescence risks.
Dataset sizes, roughly corresponding to sizes of resulting SIPs (added metadata xml-files being, range from < 1MB to > 10GB. Individual file sizes may also vary considerably, almost within the same range as entire datasets, which are sometimes deposited as compressed .zip-files.
Repositories currently used by SU for research data deposit have different limits on file sizes and storage limits ranging from 500MB (SND) to 5 GB (Figshare) for individual data files for self-deposit web-upload. (Here is an overview of some properties of the four repositories curated by SU).
As for metadata standards, a selection of preference would be: DDI, DataCite, DublinCore, **OAI-PMH** and as an essential "wrapper-format" METS (required by the *dilcis.eu*) – all handled today in XML (preferred over JSON). Further, for the creation of AIPs metadata records must be able to be enriched with PREMIS and possibly also PROV preservation metadata.

Cost requirements:
The estimated cost requirements will be specified on demand in direct negotiations with offering vendor consortia, considering also our local investment costs for storage, servers, staff and maintenance, which will naturally limit the means at disposal for the Archiver software solution.

Benefits and expected impact:

A harvest- & transform mechanism to archival format (SIP) for our use case should be platform- and metadata standard agnostic to the extent that users (content creators/depositors/researchers) should be able to use several different repositories (a selection meeting certain criteria, notably the FAIR principles) for upload and deposit. To ease the administrative burden of the content creators, the researchers, we want to use various metadata sources for enriching harvested metadata with funding information, e.g. from swecris.se, local user identification (*orcid.org* and SU staff directory, *sukat.su.se*), ethical vetting documents etc. The main benefit of the Archiver solution would be that of helping develop a workflow for Research Data Management, that eases the administrative burden on the researchers and to the extent possible automates the process of digital archiving and preservation. The SU-RDM staff (curators, analysts, archivists) would also benefit from this automation, by making it easier to meet an expected future increase of data deposits and demands for RDM support from researchers.
It would contribute to secure sustained long-term preservation of research data information packages by transformation to AIPs holding also preservation metadata (PREMIS, PROV) and supporting file format monitoring, conversion and migration, all in compliance with National Swedish Archive regulation, Swedish law, and GDPR. The production of DIPs within the archive system would finally ensure the availability of research data files, even if these files are no longer available in the repositories from whence they were harvested. Preferably, the solution should also contribute to an increased trust in our RDM system, eventually allowing SU to acquire the Core Trust Seal.

Contact person & details:
Philipson, Joakim
https://orcid.org/0000-0001-5699-994X
https://www.linkedin.com/in/joakimphilipson/
@JoakimPhilipson

e-mail: opendata@su.se

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.