Dataverse and digital preservation: an overview of different institutions

(Doc. CO23/16) (4 RDM\Repositori de dades\Preservacio\SEIDOR\2306 DataverseDigitalPreservation_anOverview.docx, 16.06.23)

Report prepared by SEIDOR

for Open Science Area
Consorci de Serveis Universitaris de Catalunya (CSUC)

June 2023

Unió Europea
Fons Europeu de Desenvolupament Regional
Finançat com a resposta a la pandèmia de COVID-19

Next Generation Catalunya

Generalitat de Catalunya
Departament de Recerca i Universitats

Acknowledgements

This document has been prepared by SEIDOR jointly with representatives of the Consorci de Serveis Universitaris de Catalunya (CSUC).

We would like to express our sincere appreciation to all experts cited in this document for the time they devoted to informing and sharing their expertise.

Written by Albert Baucells and David Pérez (SEIDOR)

Coordinated by Mireia Alcalá and Lluís Anglada (CSUC - Open Science Area)

Digital version: http://hdl.handle.net/2072/536532

This project is part of the Next Generation Recovery Plan, as a response by the Union to the COVID-19 pandemic under the Operational Programme FEDER of Catalonia 2014-2020 with a grant of €12,73 million.

Unió Europea
Fons Europeu
de Desenvolupament Regional
Finançat com a resposta a la pandèmia de COVID-19

Next Generation
Catalunya

Generalitat de Catalunya
Departament de Recerca
i Universitats

Version control

Data i versió	Autors
16-05-23 DRAFT	• Albert Baucells and David Pérez
26-05-23 DRAFT-V1	• Mireia Alcalá and Lluís Anglada
31-05-23 DRAFT-V3	• Albert Baucells and David Pérez
13-06-23 DRAFT-V3	• Mireia Alcalá and Lluís Anglada
14-06-23 V1	• David Pérez

Summary

Executive summary

This document was delivered to CSUC by a SEIDOR IT team with expertise and seniority, with the aim of providing support in the acquisition of useful information for a decision-making process to be carried out in relation to the improvement of a Digital Preservation System. The document presents, in a structured way, the result of a process of data collection from a selection of five internationally recognized institutions. The selection was a choice of six institutions from different countries which are providing digital preservation services mainly to preserve open research data.

Guided by FAIR principles, as well as by main technology trends, the most valuable information required to fulfil the goals of CSUC were sought. A comparison grid was created by accessing, collecting, processing, and documenting open documentation published by target institutions. Open documentation related with the Implementation of the CoreTrustSeal certification and Core Trustworthy Data Repositories Requirements was found to be very valuable and was largely collected. The comparison grid is organized by the functional areas defined by NSDA to determine Levels of Digital Preservation: Storage, Integrity, Control, Metadata and Content. Comparison grid was approved as a helpful tool to CSUC for the decision-making process.

A more detailed analysis of CSUC's overall needs to improve its preservation system dictates that a set of questions be asked in person to team members from each institution whose preservation system is of interest to CSUC. The final part of the project is the collection of detailed information that was not being perceived through the open documentation reviews.

1. Introduction

CORA.RDR (https://dataverse.csuc.cat) is a federated and multidisciplinary repository for the publication of research datasets in FAIR mode (Findable, Accessible, Interoperable and Reusable) following the European Open Science Cloud (EOSC) guidelines.

These guidelines request datasets to be reusable for at least 10 years after publication. Leaving aside other aspects, this requires the digital preservation of datasets. This is based on different processes including having a remote copy system. This problem is common to the different repositories of research data and, specifically, to those who use Dataverse software (the same as CORA.RDR).

To be able to establish a system of remote copies of the datasets of the CORA.RDR and in order to do it in the most efficient way possible, we want to know the characteristics of the systems that perform this function in repositories, not only similar in objectives and function but, in addition, in software, and that is why CSUC want to have knowledge of the digital preservation characteristics of these repositories:

- Borealis (Canada)
- dataRepositoriUM (Portugal)
- DataverseNL (The Netherlands)
- DataverseNO (Norway)
- FAIRData (Finland)
- Odum's Research Data Information Systems (RDIS) (EUA)

From the repositories mentioned above, we want to know the technical operating characteristics, both software and hardware, suitable for replicating the infrastructure in the Catalan case. In an indicative way, we want to know: the technological architecture used (local, in the cloud, etc.), the mechanisms of the ingest process (centralized/decentralized), the relationship between the preservation function and the open repository, etc.

2. Objective

The report aims to explore how different institutions are utilizing Dataverse and implementing digital preservation strategies. For this reason, is to outline the state of the art of the reference technological models for the digital preservation of digital objects or "Datasets" (set of files and metadata) of the research system.

3. Methodology

This section provides information on the methodology followed in this report. First, the institution selection was provided by CSUC and included institutions from Europe (dataRepositoriUM from Portugal, DataverseNL from the Netherlands, DataverseNO from Norwegian and DigitalPreservationFI from Finland) and America (Borealis from Canada and Odum's Research Data Information Systems (RDIS) from EUA) institutions.

Based on these systems, information is collected from the different public websites (see 6. Bibliography and references) and a series of questions are drawn up that should be answered. After a discussion by CSUC, some of these questions are discarded and the nine final questions (see appendix 1) asked information for: number of copies, localisation, media type, integrity and checking.

With the information collected, it is checked that there are still gaps and each institution is requested to have a short one-hour interview (see appendix 2) to go deeper with some of the topics. In general terms, these interviews are done with the technical staff of the repository and allow for fill-in the gaps.

Then, a grid (see appendix 3) is developed with different indicators to easily compare the different systems. And this allows us to make the synthesis and final conclusions (see 5. Conclusions).

4. Analysis of preservation systems

In this section each system is introduced and a short summary of each of them is made.

4.1. dataRepositoriUM

"DataRepositóriUM is an institutional Data Repository to share, publish and manage research data generated and collected by the activity of researchers and in the research units of the University of Minho."

The UMinho considers that digital preservation is important, but they have not yet been able to develop policies and workflows in the repository. For this reason, no information is collected from this system.

4.2. Borealis

"Borealis, the Canadian Dataverse Repository, is a bilingual, multidisciplinary, secure, Canadian research data repository, supported by academic libraries and research institutions across Canada. Borealis supports open discovery, management, sharing, and preservation of Canadian research data."[1]

The preservation system achieves three copies to five different nodes via OLRC (using a software feature of OpenStack Swift) without controlling it. The production data in OpenStack swift is automatically stored in 5 nodes, one copy in the NFS storage system synced nightly, one copy on-site on tape synced nightly, and one copy of tape off-site at Iron Mountain storage facility in Hamilton (shipped every other day or twice a week). A geographic radius of 450 kilometres away from each other copy is reached. A monthly integrity check of all the files in the Dataverse is performed. The integrity check is based on the MD5 hash formula taken from the Dataverse ingest flow. Integrity-calculated information is stored separately in a database.

4.3. DataverseNL

"DataverseNL is a shared service provided by participating institutes and DANS. DataverseNL uses the Dataverse software developed by Harvard University, which is used worldwide. DataverseNL is jointly offered by participating institutes and DANS. Since 2014, DANS has been managing the technical infrastructure; the participating institutes are responsible for managing the deposited data in their dataverses within DataverseNL. Every institute participates in the Dataverse Advisory Board, which determines the policies of the service. The local dataverse managers exchange their experiences regularly, together they develop "good practices" in the field of research data management."[2]

The preservation system consists of two backup replicas, which are stored in geographically distributed locations within a 20km range. The storage type for the copies is a snapshot. UNF checksum and normalization function and SHA256 integrity function are used. One copy resides in the datacenter of the contracted vendor, located in Almere, another at a local CPD in The Hague. Currently, these copies are run offline (tape) and on disk. Once a month, a bit rot check is done. A great number of random files at a time are checked to ensure data integrity is conserved, this contributes to having an insight into the health of the data. No measurements are made of the total amounts of computing resources spent. The main resource spent is Disk IO. Virus checking is done by using https://www.clamav.net/. Integrity information is stored on disk as a file. It is physically stored on the same medium, in a different folder.

[1] https://borealisdata.ca/about/

[2] https://dataverse.nl/

4.4. DataverseNO

"DataverseNO (https://dataverse.no/) is a national, generic repository for open research data, owned and operated by UiT The Arctic University of Norway. DataverseNO is aligned with the FAIR Guiding Principles for scientific data management and stewardship. The technical infrastructure of the repository is based on the open-source application Dataverse, which is developed by an international developer and user community led by Harvard University. DataverseNO is CoreTrustSeal certified"[3].

The preservation system of DataverseNO has three copies of the backup replicas. There are two location copies stored in two separate datacenters within the same building in Oslo, Norway. The two datacenters are separated with a fire resistant wall. In addition, there is one copy stored in a Microsoft Azure Datacenter in The Netherlands. Currently implementing storage of another additional copy on immutable storage.

MD5 integrity functions are used.

4.5. FAIRData

"Fairdata services are part of the digital preservation services of the Ministry of Education and Culture, Finland ("Minedu"). The Fairdata services support the research process and management of digital data with some of the components described in the Framework for Open Science and Research. The Fairdata services consist of the following service components: IDA – Research Data Storage, Etsin – Research Data Finder, Qvain – Research Dataset Metadata Tool, auxiliary services, such as Metax, identity management, download component for published data and Digital preservation service for Research Data (including management and packaging)."[4]

The preservation system consists of four total backup replicas: three online copies and one offline. The copies are geospatially distributed into three separate locations (data centers in Finland): the first one is local in disk storage, second and third on tape storage and the fourth and fifth on the dark archive. Related to integrity, the system signed SMIME manifests.

4.6. ODUM

"The Odum Institute Data Archive is a leader in research data stewardship, with over 50 years of experience beginning with the acquisition of the Louis Harris Data Center in 1965. Our longstanding commitment to data access and research transparency has been a driving force behind ongoing efforts to enhance our infrastructure, workflows, and policies to ensure that the

[3] https://info.dataverse.no/

[4] https://www.fairdata.fi/en/about-fairdata/fairdata-services/

data assets in our care remain FAIR—findable, accessible, interoperable, and reusable–now and into the future.

We are home to one of the largest catalogues of social science research data in the U.S. which includes the Harris Polls, North Carolina Vital Statistics, and the most complete collection of 1970s U.S. Census data. In addition, we manage and provide access to the UNC Dataverse, a web-based data repository, that enables scientists, research teams, scholarly journals, and other members of the UNC research community to archive and share their own datasets."[5]

The preservation system consists of four total amount backup replicas. The copies are housed in geographically distributed storage locations (offsite local server, cloud server in Northern Virginia and Northern California and a copy in the private cloud). MD5 integrity functions are used.

5. Conclusions

In conclusion, several key findings can be drawn:

Storage:

- European institutions rely on two to five copies for preservation systems, while Canadian and American institutions use a higher **number of copies**, ranging from five to twelve.
- Using cloud storage allows for a larger number of copies compared to on-premises infrastructure.
- **Key strategies for safeguarding copies** include storing them in different locations, utilizing services from multiple providers, and using different storage media types. Some institutions heavily rely on cloud providers and their certifications, considering the type of media used for storage less important.
- There is a trend to **shift from file/block storage to object storage**, especially in cloud-based digital preservation procedures. The percentage of object storage increases with the adoption of cloud-based preservation. However, this transformation requires software changes and poses some difficulties.

Integrity:

- Many interviewees believe that integrity checks should be activated in more situations. Dataverse ensure **integrity calculation only at ingestion time**.
- Integrity information, added to content for long-term preservation, is usually **calculated from the entire content** of the ingested file rather than from split parts. Computed integrity information is commonly stored separately from the preserved content. Some

[5] https://odum.unc.edu/archive/#archive1

systems comply with this requirement, and those relying on cloud providers trust their certifications.

- The **MD5 cryptographic** hash function, previously widely used for integrity validation, is considered too weak. Teams are replacing it with stronger alternatives such as SHA256 checksum verification.
- Teams relying on cloud provider infrastructure are less concerned about knowing when integrity checking is performed compared to those relying on on-premise solutions.

Control:

- **Virus checking** is not considered a top priority by most digital preservation IT teams. Open-source tools for virus checking are challenging to keep up to date. Institutions working with large internet service providers include virus checking in their ingestion processes without significant workload. Virus checking is considered unimportant for executable files rarely included in datasets.

Other:

- The analysis should have included a broader range of Digital Preservation System components, not just limited to Dataverse.

6. Bibliography and references

6.1. CoreTrustSeal certifications documentation

DataverseNL	https://www.coretrustseal.org/wp-content/uploads/2018/04/Tilburg-University Dataverse.pdf
DataverseNO	https://www.coretrustseal.org/wp-content/uploads/2020/03/DataverseNO.pdf
ODUM	https://www.coretrustseal.org/wp-content/uploads/2020/10/Odum-Institute-Data Archive.pdf

6.2. General information

Borealis	https://spotdocs.scholarsportal.info/plugins/servlet/mobile?contentId=228196817# content/view/228196817
	https://spotdocs.scholarsportal.info/display/DAT/B.+Successful+CTS+Certification +Applications
DataverseNL	https://dans.knaw.nl/en/preservationplan/1
DataverseNO	https://site.uit.no/dataverseno/about/policy-framework/preservation-policy/
	https://site.uit.no/dataverseno/about/policy-framework/preservation policy/preservation-plan/
	https://site.uit.no/dataverseno/
FAIRDATA - CSC IT Center for Science (Finland)	Digital Preservation with added resilience in Finland
	https://www.fairdata.fi/en/
	https://digitalpreservation.fi/mets-validator
ODUM	https://www.coretrustseal.org/wp-content/uploads/2020/10/Odum-Institute-Data Archive.pdf

FAIRDATA - CSC IT Center for Science (Finland)	Digital Preservation with added resilience in Finland
	https://www.fairdata.fi/en/
	https://digitalpreservation.fi/mets-validator
UMinho	https://www.um.edu.mt/library/oar/bitstream/123456789/92709/1/University%20of %20Minho%20Data%20Repository.pdf

6.3. Tools

Libnova	https://www.libnova.com/
TwoRavens Dataverse Project	https://guides.dataverse.org/en/5.9/installation/r-rapache-tworavens.html
Digital Preservation Coalition	https://www.dpconline.org/handbook/technical-solutions-and-tools/tools
Apps, Dataverse Project	https://guides.dataverse.org/en/latest/api/apps.html
Archivematica	https://www.archivematica.org/es/

6.4. Others

Dataverse Project	https://guides.dataverse.org/en/latest/
Wikipedia	https://es.wikipedia.org/wiki/Open_Archival_Information_System
NDSA	https://ndsa.org/publications/levels-of-digital-preservation/
CoreTrustSeal	https://www.coretrustseal.org/

Appendix

1. Final questions

1. How many complete copies the preservation system achieves?
2. In how many different geographic locations are the copies stored?
    a. Interest in knowing whether they are local copies, private cloud, or public cloud b. Also, interest in knowing if copies access is online or offline.
3. How many of the copies have different disaster threat?
4. Is there, at least, one copy on a different storage media type?
    a. Interest in knowing the type of media: optical media, hard disk, tape.
5. How often is object integrity checked? Is integrity verification performed for the full content? Does backup system allow for checking integrity of a part of the content?
6. What integrity calculation cryptographic functions are used?
7. What amount of computing resources do integrity checking consume?
8. What amount of computing resources do virus checking consume?
9. Is integrity information managed as special metadata? Is integrity information stored in a different destination from standard metadata?

2. Institutions interviews

Name	**Borealis**
Country	**Canada**
Software	**Dataverse**

Analysis:

1. How many complete copies the preservation system achieves?
The system is achieving at least three copies. And it can be up to five *Anytime anything is deposited into Dataverse, it is automatically copied into backup and also into the preservation system. Then that copy in preservation is replicated at least three times across the Ontario Library research cloud, which is cloud-based. The infrastructure has nodes at five different universities in Ontario,* *Daily backup of all files to tape using IBM Tivoli Storage Manager (TSM)* *For active files:* *Seven versions of a file are available for restore for 30 days* *If a file has not been modified for over 30 days, the most recent version of the file is retained permanently in backup* *The six previous versions of a file are discarded after 30 days* *For deleted files:* *The most recent version of a deleted file is available for restore for 60 days* *Two copies of the tape backup are retained onsite and one copy is retained offsite*

2. In how many different geographic locations are the copies stored? **a) Interest in knowing whether they are local copies, private cloud or public cloud** **b) Also, interest in knowing if copies access is online or offline**
Copies are in different nodes. One of them is the University of Toronto. The other ones are, in Queen's University, which is based in Kingston Ontario, Guelph University, which is in Guelph Ontario, the University of Ottawa which is based in Ottawa. The other one is in McMaster based in Hamilton Ontario. Copies are replicated. Across those nodes. *A geographic radius of like 450 kilometers away from each other*

3. How many of the copies have different disaster threat?
This can be deduced from the two topics above.

4. Is there, at least, one copy on a different storage media type? **a) Interest in knowing the type of media: optical media, hard disk, tape.**
Backup uses tape *One in a private cloud it is not known*

5. How often is object integrity checked? Is integrity verification performed for the full content? Does the backup system allow for checking integrity of a part of the content?
A monthly check of all the files in the dataverse is performed *They check against all the files in the cloud storage* *They have been running the file integrity for two years end they have found 124 corrupted files, and they found that most of them were uploaded with zero* *Integrity checks are stored separately from the content*

6. What integrity calculation cryptographic functions are used?
Dataverse uses MD5 and their storage uses MD5

7. What amount of computing resources do integrity checking consume?
There aren't measures. It takes 24 hours to run, six terabytes of data

8. What amount of computing resources do virus checking consume?
When there is an upload, there's virus checking happening, there is no checking on the storage side

9. Is integrity information managed as special metadata?

Is integrity information stored in a different destination from standard metadata?
It is stored separately, the record of each fixity check (both positive and negative) is stored in an internal MySQL database

Name	**DataverseNL**
Country	**Holanda**
Software	**Dataverse**
Website	**https://dataverse.nl**
Related institutions	**4TU.ResearchData**
	NIOO-KNAW
	Trimbos Instituut
	Utrecht University
	Vrije Universiteit Amsterdam
	Hanzehogeschool Groningen University of Applied Sciences
	Leiden University
	Avans University of Applied Sciences
	Tilburg University
	Protestantse Theologische Universiteit
	Maastricht University
	University of Groningen
	University Medical Center Utrecht
	Hogeschool Rotterdam
	Fontys Hogeschool
	Rijksdienst voor het Cultureel Erfgoed

Analysis:

1. How many complete copies the preservation system achieves?
Has two total amount of backups replicas
Quantitative information of interest: ● *Amount of backup replicas*
Two
Additional or reinforcement information of interest: ● *Is there documentation that can be shared with CSUC regarding this point of interest of CSUC, as well as documentation about the decision-making process that ended up with the current amount of backup replicas?*
There is not additional information

2. In how many different geographic locations are the copies stored? **a) Interest in knowing whether they are local copies, private cloud or public cloud** **b) Also, interest in knowing if copies access is online or offline**
Quantitative & qualitative information of interest: ● *Copies are online/offline* ● *Storage location strategy and physical separation of copies (local copies, private cloud, or public cloud)*
Geographically distributed storage locations 20 km
Additional or reinforcement information of interest: ● *Is there documentation that can be shared with CSUC regarding this point of interest of CSUC, as well as documentation about the decision-making process that led to the current geographic location of the copies and the local vs cloud strategy?*
There is not additional information

3. How many of the copies have different disaster threat?

There is not information

Quantitative information of interest:

- *Main threat for each copy*

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about specific threats of the copies? This will allow CSUC whether they have similar threats.*

4. Is there, at least, one copy on a different storage media type?
a) Interest in knowing the type of media: optical media, hard disk, tape.

Copies storage type is snapshot.

Quantitative & qualitative information of interest:

- *List of backup types*

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current list of used backup types?*

Snapshot

5. How often is object integrity checked? Is integrity verification performed for the full content? Does the backup system allow for checking integrity of a part of the content?

There is not information

Quantitative & qualitative information of interest:

- *Full contents / partial content (integrity information)*
- *Integrity check frequency*

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current strategy/technique to manage full/partial integrity information?*

6. What integrity calculation cryptographic functions are used?

There is not information

Quantitative & qualitative information of interest:

- *Cryptographic function (integrity calculation)*

- Universal Numerical Fingerprint (UNF)

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current cryptographic function to calculate and verify the integrity information?*
- *What software tools are involved in the generation/verification of the integrity information?*

7. What amount of computing resources do integrity checking consume?

There is not information

Quantitative & qualitative information of interest:

- *Consumption of computing resources for integrity verification*

Reinforcing or additional information of interest:

- *Is there documentation, such as monitoring data or traces, that can be shared with the CSUC about the consumption of IT resources dedicated to integrity verification?*

8. What amount of computing resources do virus checking consume?

There is not information

Quantitative & qualitative information of interest:

- *Consumption of computing resources for virus checking*

Reinforcing or additional information of interest:
• *Is there documentation, such as monitoring data or traces, that can be shared with the CSUC about the consumption of IT resources dedicated to the virus inspection?*

9. Is integrity information managed as special metadata? Is integrity information stored in a different destination from standard metadata?
There is not information
Quantitative & qualitative information of interest:
• *Integrity information (metadata) storage strategy*
Reinforcing or additional information of interest:
• *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current strategy about the way integrity information is treated in relation to metadata?*

Name	**DataverseNO**
Country	**Norway**
Software	**Dataverse**
Website	**https://dataverse.no/**
Related institutions	**Inland Norway University of Applied Sciences**
	MF Norwegian School of Theology, Religion and Society
	Nord University
	Norwegian University of Life Sciences (NMBU)
	NTNU – Norwegian University of Science and Technology
	UiT The Arctic University of Norway
	Nofima
	University of Agder
	University of Bergen
	University of Oslo
	University of Stavanger
	VID Specialized University
	Western Norway University of Applied Sciences
	Østfold University College

Analysis:

1. How many complete copies the preservation system achieves?
Has two total amount of backups replicas
Quantitative information of interest:
• *Amount of backup replicas*
Three (for data files; two for database/metadata records)
Additional or reinforcement information of interest:
• *Is there documentation that can be shared with CSUC regarding this point of interest of CSUC, as well as documentation about the decision-making process that ended up with the current amount of backup replicas?*
There is not additional information

2. In how many different geographic locations are the copies stored? a) Interest in knowing whether they are local copies, private cloud or public cloud b) Also, interest in knowing if copies access is online or offline

There are two location copies stored in two separate datacenters within the same building in Oslo, Norway. The two datacenters are separated with a fire-resistant wall. In addition, there is one copy stored in a Microsoft Azure Datacenter in The Netherlands. Currently implementing storage of another additional copy on local, immutable storage.
a) Two copies are in private cloud at partner institution in Oslo, Norway. One copy is in private Virtual Data Center in public cloud (Microsoft Azure) in The Netherlands.
b) The copies access is online.

Quantitative & qualitative information of interest:

- *Copies are online/offline*
- *Storage location strategy and physical separation of copies (local copies, private cloud, or public cloud)*

Three copies are stored online. Currently implementing storage of fourth copy on immutable storage.

Additional or reinforcement information of interest:

- *Is there documentation that can be shared with CSUC regarding this point of interest of CSUC, as well as documentation about the decision-making process that led to the current geographic location of the copies and the local vs cloud strategy?*

The choice of storage mode and location is based on the UiT cloud deployment policy.

3. How many of the copies have different disaster threat?

There is not information

Quantitative information of interest:

- *Main threat for each copy*

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about specific threats of the copies? This will allow CSUC whether they have similar threats.*

4. Is there, at least, one copy on a different storage media type?
a) Interest in knowing the type of media: optical media, hard disk, tape.

All copies are on hard disk. Currently implementing storage of another additional copy on local, immutable storage.

Quantitative & qualitative information of interest:

- *List of backup types*

Back up by the storage at partner institution in Oslo, Norway, and in Azure blob storage in virtual datacenter in The Netherlands.

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current list of used backup types?*

Snapshot

5. How often is object integrity checked? Is integrity verification performed for the full content? Does backup system allow for checking integrity of a part of the content?

At file ingest. After that relying on file integrity support from cloud provider.

Quantitative & qualitative information of interest:

- *Full contents / partial content (integrity information)*
- *Integrity check frequency*

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current strategy/technique to manage full/partial integrity information?*

6. What integrity calculation cryptographic functions are used?

There is not information

Quantitative & qualitative information of interest:

- *Cryptographic function (integrity calculation)*

- Universal Numerical Fingerprint (UNF)
- MD5

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current cryptographic function to calculate and verify the integrity information?*
- *What software tools are involved in the generation/verification of the integrity information?*

7. What amount of computing resources do integrity checking consume?
There is not information
Quantitative & qualitative information of interest:
• *Consumption of computing resources for integrity verification*
Reinforcing or additional information of interest:
• *Is there documentation, such as monitoring data or traces, that can be shared with the CSUC about the consumption of IT resources dedicated to integrity verification?*

8. What amount of computing resources do virus checking consume?
There is not information
Quantitative & qualitative information of interest:
• *Consumption of computing resources for virus checking*
Reinforcing or additional information of interest:
• *Is there documentation, such as monitoring data or traces, that can be shared with the CSUC about the consumption of IT resources dedicated to the virus inspection?*

9. Is integrity information managed as special metadata? **Is integrity information stored in a different destination from standard metadata?**
There is not information
Quantitative & qualitative information of interest:
• *Integrity information (metadata) storage strategy*
Reinforcing or additional information of interest:
• *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current strategy about the way integrity information is treated in relation to metadata?*

Name	**FAIRDATA CSC IT Center for Science**
Country	**Finlandia**
Software	**Fairdata**
Website	**https://www.fairdata.fi/en/dps-for-research-data/**
Related institutions	**CSC – IT Center for Science Ltd.**

Analysis:

1. How many complete copies the preservation system achieves?
Has four total amount of backups replicas
Quantitative information of interest:
• *Amount of backup replicas*
Four
Additional or reinforcement information of interest:
• *Is there documentation that can be shared with CSUC regarding this point of interest of CSUC, as well as documentation about the decision-making process that ended up with the current amount of backup replicas?*

There is not additional information

2. In how many different geographic locations are the copies stored?
a) Interest in knowing whether they are local copies, private cloud or public cloud
b) Also, interest in knowing if copies access is online or offline

Quantitative & qualitative information of interest:

- *Copies are online/offline*
- *Storage location strategy and physical separation of copies (local copies, private cloud, or public cloud)*

- 3 online copies & one offline copy
- EMP-protected deep underground shelter, with airgap
- geospatially distributed into 3 separate locations (data centers in Finland)
- 1st (local disk storage) + 2nd (Tape storage) + 3rd (Tape storage) + 4th/5th (Dark archive)

Additional or reinforcement information of interest:

- *Is there documentation that can be shared with CSUC regarding this point of interest of CSUC, as well as documentation about the decision-making process that led to the current geographic location of the copies and the local vs cloud strategy?*

There is not additional information

3. How many of the copies have different disaster threat?

There is not information

Quantitative information of interest:

- *Main threat for each copy*

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about specific threats of the copies? This will allow CSUC whether they have similar threats.*

4. Is there, at least, one copy on a different storage media type?
a) Interest in knowing the type of media: optical media, hard disk, tape.

Quantitative & qualitative information of interest:

- *List of backup types*

Snapshot (virtual machine)

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current list of used backup types?*

5. How often is object integrity checked? Is integrity verification performed for the full content? Does the backup system allow for checking integrity of a part of the content?

There is not information

Quantitative & qualitative information of interest:

- *Full contents / partial content (integrity information)*
- *Integrity check frequency*

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current strategy/technique to manage full/partial integrity information?*

6. What integrity calculation cryptographic functions are used?

Quantitative & qualitative information of interest:

- *Cryptographic function (integrity calculation)*

Signed SMIME manifests

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current cryptographic function to calculate and verify the integrity information?*
- *What software tools are involved in the generation/verification of the integrity information?*

7. What amount of computing resources do integrity checking consume?

There is not information

Quantitative & qualitative information of interest:

- *Consumption of computing resources for integrity verification*

Reinforcing or additional information of interest:

- *Is there documentation, such as monitoring data or traces, that can be shared with the CSUC about the consumption of IT resources dedicated to integrity verification?*

8. What amount of computing resources do virus checking consume?

There is not information

Quantitative & qualitative information of interest:

- *Consumption of computing resources for virus checking*

Reinforcing or additional information of interest:

- *Is there documentation, such as monitoring data or traces, that can be shared with the CSUC about the consumption of IT resources dedicated to the virus inspection?*

9. Is integrity information managed as special metadata?
Is integrity information stored in a different destination from standard metadata?

There is not information

Quantitative & qualitative information of interest:

- *Integrity information (metadata) storage strategy*

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current strategy about the way integrity information is treated in relation to metadata?*

Name	**ODUM**
Country	**Estats Units**
Software	**Dataverse**
Website	**https://odum.unc.edu**
Related institutions	**Odum institute for research in social science**

Analysis:

1. How many complete copies the preservation system achieves?

Has four total amount of backups replicas

Quantitative information of interest:

- *Amount of backup replicas*

Four

Additional or reinforcement information of interest:

- *Is there documentation that can be shared with CSUC regarding this point of interest of CSUC, as well as documentation about the decision-making process that ended up with the current amount of backup replicas?*

There is not additional information

2. In how many different geographic locations are the copies stored?
a) Interest in knowing whether they are local copies, private cloud or public cloud
b) Also, interest in knowing if copies access is online or offline

- Copies housed in geographically distributed storage locations
- Offsite local server
- Amazon S3 storage in Northern Virginia
- Amazon S3 storage in Northern California
- A copy in the Unitrends private cloud

Quantitative & qualitative information of interest:

- *Copies are online/offline*
- *Storage location strategy and physical separation of copies (local copies, private cloud, or public cloud)*

Additional or reinforcement information of interest:

- *Is there documentation that can be shared with CSUC regarding this point of interest of CSUC, as well as documentation about the decision-making process that led to the current geographic location of the copies and the local vs cloud strategy?*

There is not additional information

3. How many of the copies have different disaster threat?

There is not information

Quantitative information of interest:

- *Main threat for each copy*

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about specific threats of the copies? This will allow CSUC whether they have similar threats.*

4. Is there, at least, one copy on a different storage media type?
a) Interest in knowing the type of media: optical media, hard disk, tape.

Quantitative & qualitative information of interest:

- *List of backup types*

Snapshot (virtual machine)

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current list of used backup types?*

5. How often is object integrity checked? Is integrity verification performed for the full content? Does the backup system allow for checking integrity of a part of the content?

There is not information

Quantitative & qualitative information of interest:

- *Full contents / partial content (integrity information)*
- *Integrity check frequency*

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current strategy/technique to manage full/partial integrity information?*

6. What integrity calculation cryptographic functions are used?

Quantitative & qualitative information of interest:

- *Cryptographic function (integrity calculation)*

MD5 (data and metadata)

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current cryptographic function to calculate and verify the integrity information?*
- *What software tools are involved in the generation/verification of the integrity information?*

7. What amount of computing resources do integrity checking consume?

There is not information

Quantitative & qualitative information of interest:

- *Consumption of computing resources for integrity verification*

Reinforcing or additional information of interest:

- *Is there documentation, such as monitoring data or traces, that can be shared with the CSUC about the consumption of IT resources dedicated to integrity verification?*

8. What amount of computing resources do virus checking consume?

There is not information

Quantitative & qualitative information of interest:

- *Consumption of computing resources for virus checking*

Reinforcing or additional information of interest:

- *Is there documentation, such as monitoring data or traces, that can be shared with the CSUC about the consumption of IT resources dedicated to the virus inspection?*

9. Is integrity information managed as special metadata?
Is integrity information stored in a different destination from standard metadata?

There is not information

Quantitative & qualitative information of interest:

- *Integrity information (metadata) storage strategy*

Reinforcing or additional information of interest:

- *Is there documentation that can be shared with CSUC about the data preservation decision-making process that led to the current strategy about the way integrity information is treated in relation to metadata?*

3. Grid

https://docs.google.com/spreadsheets/d/e/2PACX-1vSiArVX1M67lDwIyhPuccn_-RgTFUeA8mF2pGzNu_xEm0HjNsMskFoUPO10cOcTjA/pubhtml#

4. Recording and transcription of the interviews

The recordings and transcriptions were sent to CSUC; however, they are not included in this report.