Digital Preservation Plan – Aleia

September, 2023

This plan was developed to preserve the digital resources of the Aleia Repository for a long period of time. Due to the importance of digital resources, it is imperative that a framework is established to ensure the authenticity, reliability and long-term availability of these materials.

Digital preservation ensures that the scientific community and society in general can access the repository's digital resources in the long term. Thus, this plan will provide strategies and action steps that encompass digital preservation in accordance with established best practices, and will articulate a common understanding of such activities in the Aleia repository.

For the long-term preservation and archiving of research data stored in the Aleia repository, the OAIS model and Archivematica software are used. This open source platform makes it possible to automate the ingestion, preservation, management and access of digital objects.

Background (ALEIA REPOSITORY)

The Aleia repository, managed by the Brazilian Institute of Information in Science and Technology (Ibict), makes available research data published by the institutional scientific community.

The Aleia repository organizes and makes research data available, in order to protect and preserve the rights of authors and facilitate sharing. The repository stores the following data:

I - sets of research data contained in reports from the Institutional Training Program (PCI);

II- sets of research data that appear in research grant reports coordinated by Ibict;

III- sets of research data contained in pro-doc consultancy products;

IV- sets of research data that were produced during the execution of research highlighted in the biannual TCG;

V- sets of data from research carried out within the scope of the Postgraduate Program in Information Science (PPGCI), more specifically from theses and dissertations defended.

The Aleia repository uses open software from the Harvard Dataverse Project, which meets most of the requirements of the FAIR metadata principles.

The historical context of the repository can be accessed at ALEIA - Ibict research data repository

The data

Preserving research data is critical to ensuring long-term integrity, accessibility, and usability. A comprehensive digital preservation plan involves several steps, from collection to ongoing data management. Here is a basic digital preservation plan for the Aleia Repository:

The types of data that will be deposited comprise digital files, preferably in open formats such as source codes, documents, images, videos, spreadsheets, forms, statistical data, among others.

The stored data is based on the premise of its good representation, with a description of all metadata of the project, the research arm, collection data and the description of the variables of the collected data, composing a set of detailed metadata.

Backup Protocol

In order to guarantee the preservation of the structure and data existing in the repository, it is necessary to back up the content with a certain frequency. This must guarantee the easy return of the platform in case of problems with the machine where it was stored. To this end, the main components will have their data preserved, as will be described below.

1. Backup Schedule

To better guarantee the preservation of the state of stored data, the backup will be carried out on a scheduled and frequent basis. Based on the frequency of deposits and similar repository schedules, the minimum weekly execution scheduled in the system is defined, storing the contents in defined external storage. A time of low use to carry out the processes must be defined. A possible weekly schedule can be held every Tuesday at 3:00.

2. Backup Process

As mentioned, different applications and repository configurations that are located in different locations must be preserved. The following must be preserved: the database, Payara, data storage and customization files.

The backup process can be executed with the following commands. They will be displayed according to each part, the first being the preparation of the restoration and the following ones referring to each of the contents involved.

```
#alterar o diretório para o caminho de armazenamento externo

cd /diretorio/backup/dataverse
```

2.1. Database

Commands for backing up the PostreSQL database used by Dataverse

```
cd /diretorio/backup/dataverse
su postgres #crie o diretorio apos estar com o usuario postgres
mkdir postgres
pg_dump dvndb > dvndb.sql
```

2.2. Payara

Commands for backing up the Payara web-server used by Dataverse

```
cd /diretorio/backup/dataverse
service payara stop

/usr/local/payara5/glassfish/bin/asadmin backup-domain --backupdir
/diretorio/backup/dataverse/payara/
```

2.3. Data storage

Copy of content stored in Dataverse in the configured /filesDataverse directory

```
cd /diretorio/backup/dataverse
mkdir /files
cp -a /filesDataverse/. /diretório/backup/dataverse/files #caso erro de
permissão, usar sudo
```

2.4. Customization files

Copy of customization contents of pages in the configured /http directory

```
cd /diretorio/backup/dataverse
mkdir /www
cp -a /var/www/. /diretório/backup/dataverse/www #caso erro de
permissão, usar sudo
```

3. Restoration Process

In a similar way to the backup, the restoration will act on the contents relating to each part of the content: the database, Payara, the data storage and the customization files.

The restoration process can be performed with the following commands. They will be displayed according to each part, the first being the backup preparation and the following ones referring to each of the contents involved.

```
#alterar o diretório para o caminho de armazenamento externo
```

```
cd /diretorio/backup/dataverse
```

3.1. Database

Commands for backing up the PostreSQL database used by Dataverse

```
psql dvndb < /diretorio/backup/dataverse/payara/dvndb.sql

psql -d dvndb
GRANT ALL PRIVILEGES ON ALL TABLES IN SCHEMA public to dvnapp;
GRANT ALL PRIVILEGES ON ALL SEQUENCES IN SCHEMA public to dvnapp;
GRANT ALL PRIVILEGES ON ALL FUNCTIONS IN SCHEMA public to dvnapp;
```

3.2. Payara

Commands for backing up the Payara web-server used by Dataverse

```
cd /diretorio/backup/dataverse

asadmin restore-domain --backupdir
/diretorio/backup/dataverse/payara/domain1/arquivodomain1.zip domain1
(opção dezipar o arquivo no domain1)
/usr/local/payara5/glassfish/domains
```

3.3.Data storage

Copy of content stored in Dataverse in the configured /filesDataverse directory

```
cd /diretorio/backup/dataverse/
mkdir /filesDataverse
sudo chown dataverse /filesDataverse #garantir acesso ao app
cp -a /diretório/backup/dataverse/filesDataverse/. /filesDataverse
```

Archivematica

In addition to the backup protocol, Archivematica was implemented to preserve the repository's data in the long term.

Archivematica is an open source system developed for managing digital preservation and creating reliable digital archives. It is designed to assist institutions, archives, libraries and other organizations in preserving their digital heritage over the long term, ensuring that digital materials remain accessible, authentic and readable over time.

The admission (ingestion) of data into Archivematica from Dataverse is carried out semi-automatically, with the incorporation of new data sets. Identification and filing is carried out monthly, every first Monday of the month. The action is performed by the operator who defines which datasets and which versions will be sent to Archivematica.

Periodically, an audit is carried out on the data stored in the Archivematica Storage and in the Dataverse files, checking the incidence of viruses and their checksum.

Digital preservation requires constant monitoring of technological advances, in some cases the need to migrate data formats in order to avoid the obsolescence of old formats. In this way, Archivematica was designed to guarantee the long-term accessibility of digital materials, even as technologies evolve.

The responsibilities of digital preservation

Ibict ensures active management of digital content for long-term preservation and access:
Aligning the Repository and the organization with the OAIS Reference Model;
External data integrity checks to ensure all data remains intact;
Appropriate security measures to ensure that data cannot be accessed or modified by unauthorized persons;
A regular file format review to identify content stored in at-risk file formats and recommend corrective actions such as migration or emulation;
Creation of persistent identifiers (PIDs) in the form of digital object identifiers (DOIs) via DataCite for all objects in the Repository.

Training

All usage processes and reports are recorded, as well as repository anomalies.
Training of the repository management team is carried out continuously and according to needs.
Operation manuals and system configuration documentation are also available.

Review of the Digital Preservation Plan

This document must be reviewed and updated by the Management Committee within a period not exceeding two years from the date of publication of the last version, provided that there are no other factors that require its early review. Examples of occurrences that may require an untimely update of the Plan are:

- Update of institutional policies for operation, access, preservation and others related to the Aleia Repository;
- Structural change in the management and operation of the Aleia Repository;
- Change or update of the Aleia Repository system;
- Inclusion or exclusion of document types and file formats in the system;
- Update of digital preservation standards and best practices;
- Adoption of new routines that impact the digital preservation workflow.

MARTINS JUNIOR, A. A. Open access to research data in Brazil: Archivematica: installation documentation on CentOS. Research Report. Available in: http://hdl.handle.net/20.500.11959/1271