e-cienciaDatos Preservation Plan

Open Science Working Group

1 March 2022

Scope and purpose

e-cienciaDatos is the multidisciplinary data repository of the six-member universities of the Consorcio Madroño. It is open to any data format, although it is recommended to use open formats when available or otherwise widely accepted formats in the discipline of each dataset. Each of the universities has a dataverse community and is responsible for managing its datasets, including the possibility of creating new collections if necessary.

Researchers from the universities of the Consorcio Madroño can add content to the repository. They may decide to publish their datasets for any of the following reasons:

- Making their research data open access by the mandate or recommendation either from journals or funding agencies.
- Enhancing the visibility of a research project.
- Making their research data available for ethical or other reasons.

The datasets must be the final research data and, even though embargo periods are allowed, the goal is to deposit them in open access. As e-cienciaDatos is a multidisciplinary repository, it is not possible to consider it targeted to a specific community of users but to the scientific community as a whole.

In e-cienciaDatos, data management and curation are carried out beforehand. Each university has one or more librarians in charge of managing the institution's datasets, guiding the researchers in the process of creating their datasets and creating/reviewing the metadata and their files together with the researcher. e-cienciaDatos has no self-archiving enabled to ensure the quality of the metadata (as understood in the FAIR principles) and their homogeneity. e-cienciaDatos also keeps its right to update the datasets, both data and metadata, to ensure their preservation and accessibility in the medium/long term. e-cienciaDatos allows tracing all changes made and saves both data and metadata changes in all versions when any dataset is modified.

This preservation plan is created to ensure medium and long-term access to the datasets contained in e-cienciaDatos and to guarantee access for researchers who believe in this repository to store their datasets and to show the rest of the scientific community that it is a trusted repository. In addition to guaranteeing access, the aim is also to ensure that these datasets can be reused and validated by the scientific community in the medium and long term.

This preservation plan distinguishes between standard files, for which migration can be guaranteed if the file format becomes obsolete, and specific formats for which migration cannot be guaranteed.

Objectives

The objective of this preservation plan is to guarantee medium and long-term access to digital objects accessible from e-cienciaDatos, as well as their possibility of reuse and validation by the scientific community. Given the file formats heterogeneity, we can only guarantee the migration of the most popular formats, but as far as possible, we will also try to monitor the potential obsolescence of all files and plan their migration when possible. Looking at the list of files as of 4 February 2022, we can see that more than 95% of the files have known and common extensions, from which it can be easy to migrate: pdf, zip, txt, tar.gz, wav, csv, mp4, ogg, webm, xlsx, tab, csv, ods, rar, docx...

Collections and users

e-cienciaDatos has a dataverse community per university and, within each university, collections have been defined by projects when deemed necessary by the university administrators. The datasets uploaded to e-cienciaDatos must be in open access for possible reuse and validation by any researcher. However, they may be in draft status for a while, accessible only through a private URL while the dataset is validated by authors, journals, collaborators...

The users are:

- Society in general and researchers in particular who access and download e-cienciaDatos datasets.
- Librarians, administrators of their university's dataverse community in charge of creating datasets and managing their dataverse community and collections.
- The IT administrators of the repository.

Roles and responsibilities

Each dataverse community is managed by one or more university librarians, who are responsible for the following tasks:

- Create datasets at the request of their researchers.
- Verify that datasets are complete to allow validation and re-use.
- Request the necessary data from researchers to ensure the above point.
- Generate suitable dataverse communities.
- Create guestbooks if requested by a researcher to analyse the use of a dataset.

There are also two IT staffs in charge of managing the repository whose responsibilities are:

- Maintain the repository software and hardware secure and well-maintained.
- Update the repository software when the Open Science working group deems it necessary, usually when there is some interesting new functionality in the software used to manage the repository.
- Upload large files when, due to their size, they cannot be uploaded via the e-cienciaDatos web interface.
- Report statistics and possible incidents in the repository to the Open Science group.
- Provide technical support to the dataverse community librarian administrators on the operation of the repository.
- Geolocate the datasets requested by the repository administrators of the dataverse communities.
- Verify the possible obsolescence of the files.
- Migrate files to other formats when they become obsolete.

At the organisational level, the Open Science working group of the Consorcio Madroño is the group of experts in charge of requesting changes and improvements to the repository and making decisions. Some of the decisions taken by the working group, either because of their economic cost or their development time, have to be approved by the Technical Commission of the Consorcio Madroño, one of its governing bodies, which can also request developments, reports or functionalities from the working group.

Institutional commitments and policies

The Governing Council of the Consorcio Madroño supports the distribution and preservation of scientific documentation. In 2013 the rectors of the universities that formed the Consortium signed

the Declaration on Support for Open Access, and it was renewed in 2017 with the Declaration on Support for Open Science. Both declarations recommend to the universities that make up the consortium to "Adopt policies that ensure the open archiving, preservation and dissemination of the scholarly and scientific output of their institutions".

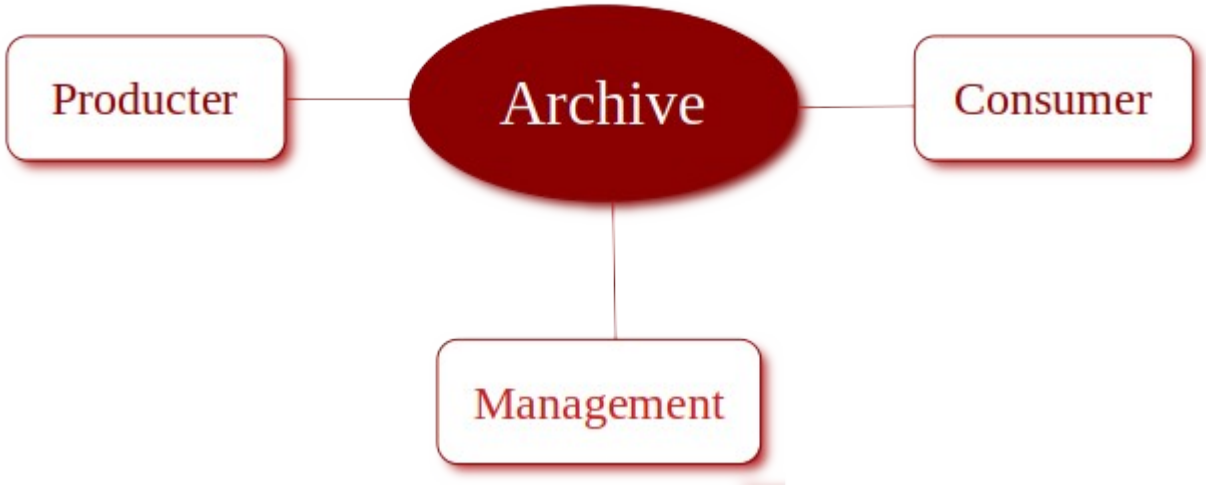
The Technical Commission has signed the mission of the e-cienciaDatos repository on 22 February 2022, where the preservation of research data is one of its fundamental points.

Preservation and quality control actions

Preservation follows the OAIS model as follows:

Components of the OAIS reference model

Below we indicate the relationship between the components of the OAIS reference model and e-cienciaDatos.



Producers

Individuals or entities that transfer information to the OAIS system. The OAIS system has to discuss with them the formats, the information related to the object to be preserved, as well as the transformation and dissemination rights. The information producers are the researchers of the Consorcio Madroño, although they are not the ones in charge of creating the datasets in e-cienciaDatos, but the university librarian administrators on their behalf. The librarians ensure, among other things, that the files are correctly formatted for preservation and that their metadata are sufficient.

Consumers and designed community

Consumers: Individuals or entities that use the information provided by the OAIS system. Consumers can access, request access to or request information from the OAIS system.

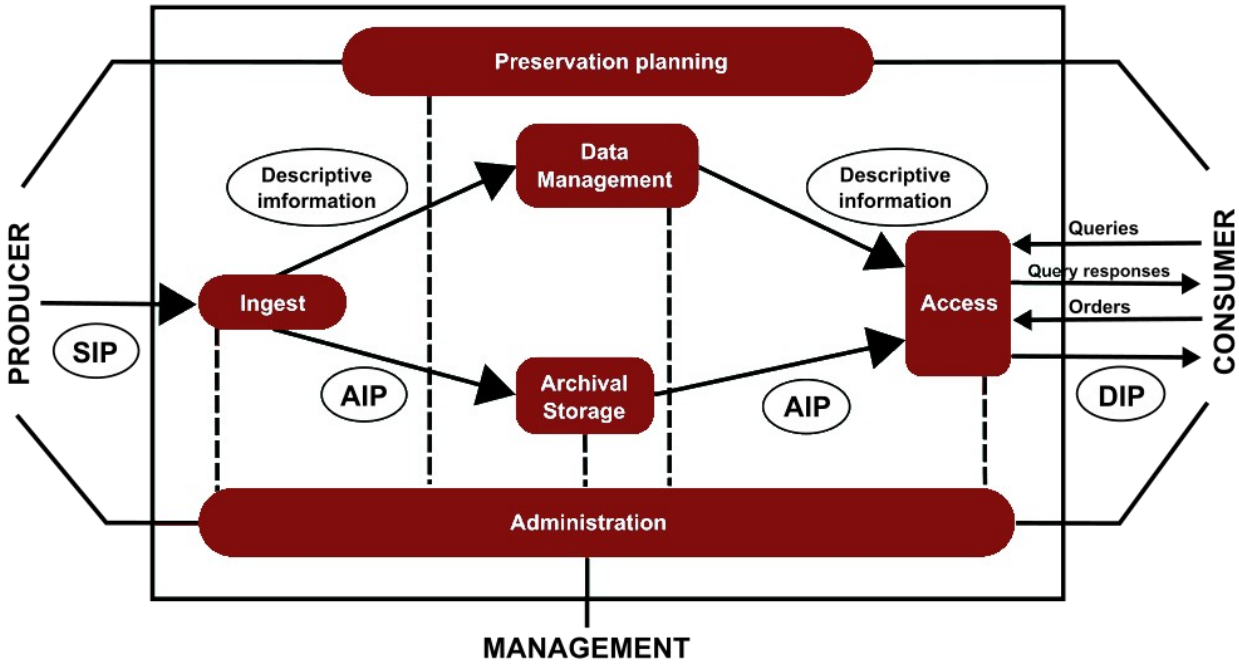
The Designed Community are the consumers for whom the OAIS system has been designed and who must be able to understand the information stored. Firstly, they are the researchers of the universities of the Consorcio Madroño as depositors of the datasets and, secondly, the rest of the researchers as users of the datasets. e-cienciaDatos has its datasets in open access so that anyone with internet access can download them.

Administration

Responsible for formulating, reviewing and ensuring compliance with OAIS policies. It includes:

- The scope of the collection to be preserved. The collection to be preserved is all the datasets contained in e-cienciaDatos, although the migration of obsolete formats is only guaranteed for the most common file types, including more than 95% of the files contained in the e-cienciaDatos datasets.
- The preservation guarantee for the reliability of the repository: Preservation is guaranteed for a minimum of 20 years, although the idea is that the information can be housed and preserved beyond this time. To this end, a risk analysis plan has been developed, which is discussed below, and the technical and economic feasibility of the repository will also be detailed.

OAIS functional services



* Imagen from Wikipedia with Creative Commons Attribution-Share Alike 4.0 International licence.

The following points show how the OAIS functional services are translated to e-cienciaDatos:

- Ingest: Processes for accepting information submitted by producers. It is performed by university librarian administrators in collaboration with the researcher and repository administrators when necessary. It consists of:
  - Receipt of information in the system sent by the researcher (producer).
  - Validation for completeness and accuracy by the librarian.
  - Transformation into a format supported by the system. If the formats or file names are not considered suitable by the librarian, he/she requests a change from the researcher.
  - Creation of descriptive metadata to enable the item to be preserved and searched, and uploaded to the system. The researcher carries out this task with the help of the librarian.
  - Transfer of the information to the storage location. This task is performed by the librarian, with assistance from repository managers if necessary.
- Archival storage. It manages the long-term storage, maintenance and security of materials within the OAIS system:
  - It ensures that information is stored in the appropriate form and is accessible in the long term. e-cienciaDatos uses the Dataverse software, with extensive community support

and installations worldwide. It has mechanisms such as checksum checking and version control. It allows access to previous versions of datasets and comparisons between them.
    - It carries out the migration of storage media and formats. This task is performed by repository administrators when a format has become obsolete, there are storage media failures, or it is deemed desirable for any reason to change storage media.
    - It performs security functions such as error checking and disaster recovery procedures. Repository administrators perform these tasks with the help of Dataverse software.
    - It communicates with the access module when there are consumer requests. This task is performed by the Dataverse software.
- Data management: it maintains descriptive metadata to identify and support search tools, performance monitors or statistical systems. These tasks are performed by the Dataverse software and include:
    - Maintaining the Dbs.
    - Interrogating the DBs in response to the requests from other OAIS functional entities.
    - Updating the DBs when new information is added, deleted or updated.
    - Supporting the search, retrieval and management of OAIS data.
- Preservation plan. Detailed in this document. It maps the preservation strategy and recommends revisions to this strategy as the OAIS system evolves:
    - It monitors the external environment for changes and dangers: new technologies for storage or access, changes in the designed community or their expectations, etc.
    - It creates recommendations for updating policies and procedures to adapt to changes.
- Access: Controls the processes and services offered to consumers. All these tasks are performed by the Dataverse software:
    - It locates, requests, and receives documents and information from the archival storage functional entity…
    - It presents search results to the consumer.
    - It sends the requested items, transforming them if necessary.
    - It implements security mechanisms for access. Access as a consumer is open to everyone, although it is aimed at the scientific community. Access to upload datasets is granted to the librarian administrators of each university and the IT administrators of the repository.
- Administration: it performs the day-to-day administrative tasks and coordinates the activities of the other functional entities. e-cienciaDatos administrators also perform these two tasks.
    - Performance monitoring.
    - System updates.

OAIS Common Services

They complement the functional entities to ensure long-term access to preserved material. These functions are overseen by e-cienciaDatos administrators and include:

- Basic computing.
- Network resources.
- Operating system services.
- Security services.

Information Package (IP) versions

- Submission Information Package (SIP). Product transferred for archiving with initial metadata. The university librarian administrator uploads the SIP with the files and metadata provided by the researcher after checking that the metadata is complete to ensure preservation and to allow consistent searches in the system. The e-cienciaDatos

administrators can collaborate if necessary in tasks such as uploading large files, answering technical questions, creating geolocation points...

- Archival Information Package (AIP). It includes the "Content data object" (object to be preserved) and its associated representation information (metadata). This file is distributed between the e-cienciaDatos database and the file system and is created by the Dataverse software from the SIP. These two elements are known as Content Information and Preservation Description Information (PDI) is added to them, which in turn includes five components:
  - Reference information: Unique identifier within the system.
  - Context information: Relationship to other objects.
  - Source information: Date of creation, dates of modifications, trace of modifications.
  - Fixity information: Checksums.
  - Access rights information: Licence, access permissions.

  e-cienciaDatos stores the AIP as an AIC (Archival Information Collection) since each file to be preserved and its metadata are stored separately, and each file can have its own associated metadata and even a different preservation strategy.
- Dissemination Information Package (DIP): Information and files that are delivered to the consumer. The distribution of these files and their metadata is carried out by the Dataverse software, allowing the metadata to be displayed in different formats.

Financial sustainability

The funding of the Consorcio Madroño is set out in a framework agreement with the Community of Madrid, which is renewed on an annual basis. The Consorcio Madroño receives funding from the six-member universities and the Community of Madrid. Specifically, the funding for the e-cienciaDatos repository comes through the e-ciencia project. The Consorcio Madroño has been in existence for more than 20 years, and the e-ciencia project has been receiving funding for more than 15 years to promote initially open access and later open science, including in both cases the preservation of scientific and scholarly documentation and data. The Community of Madrid financed the purchase of the e-cienciaDatos server with an extra allocation for the project.

Technical sustainability

Two computer scientists with more than 20 years of experience and more than 15 years working with repositories are in charge of maintaining the hardware and software of e-cienciaDatos, dedicating approximately 20% of their time to e-cienciaDatos.

Each library in the Consorcio Madroño has at least one experienced librarian responsible for managing their university's dataverse community and uploading the datasets after explaining the e-cienciaDatos features and standards to researchers and performing pre-curation of the datasets to ensure they meet e-cienciaDatos preservation requirements. Both the IT staff and at least one librarian per institution have received training in research data management, including preservation of digital objects, and have provided training both to colleagues at their institution and the staff at other institutions.

Contingency plan and risk analysis

Contingency plan for technological failures

The e-cienciaDatos server and its disk server, have redundancy in the power supplies and disks through a raid 5 system so that if a disk fails, the system continues to function. Each power supply is connected to a different power line, protected against power cuts by a UPS. The servers are in a

computer centre to which only authorised personnel have access and where all inputs and outputs are logged.

A two-layer firewall protects logical access to the servers to prevent unauthorised access at the institution level and at the server level itself.

In a disk failure, the server creates an alarm, and the necessary action is planned to fix the situation in the shortest possible time. Given the large backup volume, recovering e-cienciaDatos from a hardware failure would take two days of work.

Two complete backup copies of e-cienciaDatos including all AIPs are stored. One of them on an isolated external hard disk in a different room from where e-cienciaDatos is located. The other is located in a remote server in Cataluña owned by the CSUC. In any case:

- If a computer were to be rendered unusable by an attack, a clean operating system would be installed and the system restored from its backup.
- If a dataset is accidentally deleted or damaged, it will be recovered from the backup.
- If there is a hardware failure in the main e-cienciaDatos server, the server will be moved to another machine of the Consorcio Madroño from the backup.
- If there is a hardware failure in the NAS disk server or in the server hosting the e-cienciaDatos software:
  - At present, and as long as there is space on other servers of the Consorcio Madroño, the datasets would be moved to other servers of the Consorcio Madroño.
  - If there is not enough space to store all the datasets, the datasets that take up the most space (probably less than 1%) would be left out while a long-term solution is sought, and the researchers who own the dataset and the university concerned would be alerted to the problem.
- Once a month, the integrity of the files is checked by checking the file checksum. If any file is found to be damaged or corrupted, an attempt is made to find out the reason, and, in any case, it is restored from its backup copy.

Due to their size, complete backups are performed once a month. Incremental backups are performed once a week and they are stored in a separate Consorcio Madroño server.

In terms of monitoring for possible failures:

- The e-cienciaDatos disks are checked monthly. If one fails, the procedure to replace it with a new one is initiated.
- An annual study is made of the formats of the files that make up the e-cienciaDatos datasets. The results are shown in a report sent to the Open Science Working Group of the Consorcio Madroño, and the files to be migrated are studied.
- An external computer monitors every hour that the e-cienciaDatos server is accessible.
- Monthly complete and weekly incremental backups are performed.
- Surveys for e-cienciaDatos users are activated twice a year for one month, in which they are asked, among other questions, about possible improvements to the system.

Financial contingency plan

As mentioned above, the Consorcio Madroño is funded by the six-member universities and the Community of Madrid, so it has a solid base to continue operating after more than 20 years of existence. However, in the unlikely event that some crisis cuts funding for the project, it could continue to operate with the current hardware as all servers, including storage, are on Consorcio Madroño hardware. If, after the funding cut, the disks were to fail and there was not enough space for all the datasets, the largest datasets would be removed from the repository and returned to the

researchers or institutions to allow them to continue to use their datasets through the assigned DOI.

Training

The IT staff and at least one librarian per institution of the Consorcio Madroño working in e-cienciaDatos have received training in research data management, specifically the Research Data Management and Sharing course from Coursera that includes notions of digital preservation. At the same time, staff who have received training have also participated in training for librarians from other institutions and external institutions and continue participating in seminars and conferences on data management and digital preservation.

Evaluation, monitoring and revision of the plan itself

The Open Science Working Group of the Consorcio Madroño has created this preservation plan. It is effective from 1 March 2022 and will be reviewed biannually by the same group to adapt to the future needs of the repository.