e-cienciaDatos Technical Description

Open Science Working Group

1. General Description

e-cienciaDatos is the multidisciplinary data repository for the data resulting from the research of the researchers from the 6 member universities of the Consorcio Madroño: Universidad de Alcalá (UAH), Universidad Autónoma de Madrid (UAM), Universidad Carlos III (UC3M), Universidad Nacional de Educación a Distancia (UNED), Universidad Politécnica de Madrid (UPM) and Universidad Rey Juan Carlos (URJC).

e-cienciaDatos is open to any type of data format, although it is recommended to use open formats when available or, if unsuccessful, widely accepted formats in the discipline of each dataset. Each of the universities has a community and is responsible for managing its datasets, including the possibility of creating new communities if necessary. e-cienciaDatos accepts datasets that occupy up to 100Gb. Larger datasets can be accepted if the Open Science WG of the Consorcio Madroño considers it interesting to keep them.

The datasets must be the final research data and, even though embargo periods are allowed, the ultimate goal is to make the data open. The target community for e-cienciaDatos is primarily the research community of the public universities of the Community of Madrid and UNED, and by extension the national scientific community. For more information, please visit e-cienciaDatos policies.

Researchers from the universities of the Consorcio Madroño can add content to the repository. Some of the reasons for doing so may be:

- 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.

e-cienciaDatos is based on the open source software Dataverse, a powerful and reliable software for the creation of research data repositories that offers all the expected functionalities, including:

- Provides DOIs as persistent identifiers
- Allowing to indicate publications and other related datasets.
- Powerful version control system.
- Provides a recommended way to cite datasets.
- It offers simple and advanced searches by field or within a dataverse community. The API-REST and the facets could be used to perform searches.
- In addition:
    - e-cienciaDatos is granted as a trustworthy repository with the CoreTrustSeal certification.
    - Datasets can be collected using the OAI-PMH protocol in Datacite, oai_dc, oai_ddi, oai_datacite and dataverse_json formats.
    - e-cienciaDatos is collected by the Harvard Dataverse, Datacite Search, OpenAIRE EXPLORE and Google Dataset Search.

To ensure the completeness of the data and metadata, the repository administrators perform a prior curation and the researchers have to fill a Readme.txt template with instructions and questions about the data and metadata of the dataset. Only datasets created by researchers from member universities of the Consorcio Madroño are accepted. The reasons for updating a dataset are:

- An error was found in the file or its metadata.
- Some of the files in the dataset can be updated with a more convenient one.
- For preservation reasons

2. Mission/Scope

As stated in the Strategic Plan 2021-2025 (Spanish link), the mission of Consorcio Madroño is "To provide an information infrastructure that promotes research excellence in member institutions, contributing to the development of Open Science, digital transformation and sustainable development, and improving the experience of its users through the services provided by the Consortium."

The mission of e-cienciaDatos, as a infrastructure provided by Consorcio Madroño, is to ensure, in the context of Open Science, the deposit, preservation and dissemination of the research data generated by its partner institutions according to the FAIR (Findable, Accessible, Interoperable and Reusable) principles, thus fostering open research to the academic community and, in a broader sense, to the citizens.

e-cienciaDatos' mission (Spanish link) has been approved by Consorcio Madroño Governing Board on 22th February, 2022.

3. Organisational infrastructure

The Consorcio Madroño receives funding from the 6 member universities and the Community of Madrid. Specifically, the funding for e-cienciaDatos comes through the e-ciencia project. The Consorcio Madroño has existed for more than 25 years and the e-ciencia project has been receiving funding for almost 20 years to promote open access and later open science. The purchase of the e-cienciaDatos server was financed by the Community of Madrid with an extra allocation for the project.

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 25% of their time to e-cienciaDatos.

Each library has at least one librarian with experience in open science in charge of managing the dataverse community of their university and uploading the datasets after explaining the characteristics and rules of e-cienciaDatos to the researchers and performing a previous management and curation of the datasets. Both the IT staff and at least one librarian per institution have been trained in research data management. They have also trained other colleagues from their universities and external institutions and continue to be trained by attending conferences and congresses.

e-cienciaDatos does not have internal partners (insource partners). The only outsource partner is the Global Dataverse Community Consortium, which is responsible for serving DOIs to e-cienciaDatos.

e-cienciaDatos is part of the Global Dataverse Community Consortium and the Ibero-American Dataverse Development Group, which includes repositories in Brazil, Chile, Colombia, Peru, Portugal and Spain. It is also harvested by the Harvard Dataverse, DataCite Search, OpenAIRE EXPLORE and Google Dataset Search among other harvesters.

4. Board

The medium-term objectives, policies and funding of e-cienciaDatos are approved by the Technical Commission of the Consorcio Madroño, which includes the directors of each of the member libraries and the Technical Director of the Consorcio Madroño. The Technical Commission entrusts the Open

Science Working Group with the work to implement these objectives and in turn studies the suggestions of this group of experts.

The Open Science Working Group of the Consorcio Madroño is formed by the technical director of Consorcio Madroño, a library director, librarians with expertise in open science (at least one per university) and two computer scientists with experience in repository management. The librarians are responsible for the management of their universities' data in liaison with the researchers of the member universities, and are supported by their respective legal advisors in case of questions. Both the staff of the libraries and the Consorcio Madroño participate in training events related to research data management and are responsible for the InvestigaM portal, which brings together the open science-related actions of the Consorcio Madroño. University librarians advise and assist individual researchers in the creation of their datasets, and in some cases, to increase their visibility by collaborating in projects.

In addition, from 2019, e-cienciaDatos activates surveys twice a year for its end users in which questions related to the usage of the repository are asked and suggestions are solicited.

5. Archive and Workflows

In e-cienciaDatos, data management and curation is 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 to ensure the homogeneity and the quality of the metadata (as understood in the FAIR principles). e-cienciaDatos also keeps the right to update the datasets, both data and metadata, to ensure their preservation and accessibility in the medium/long term[1]. e-cienciaDatos traces all changes made and saves both data and metadata of all versions when any dataset is modified.

The procedure for depositing data in e-cienciaDatos depends on each university and can be consulted here:

e-cienciaDatos administrators support librarians if they have questions or need to perform work that cannot be done through the e-cienciaDatos administration website. They are also in charge of creating geolocation points for the datasets when requested by librarians.

As described in the e-cienciaDatos policies, e-cienciaDatos administrators can alter the dataset files or metadata for preservation, data curation or metadata curation. If an obsolete file is detected or the dataset should be enriched, both the dataset files and metadata can be altered, but the history changes, previous versions and version differences are always available and accessible.

6. Service and Data Quality

Consorcio Madroño is ISO9001 certified in activities covering software development, including e-cienciaDatos.

e-cienciaDatos is certified with the CoreTrustSeal seal, which implies that it is a trusted repository that performs preservation, curation and is committed to ensuring that both data and metadata follow the FAIR (findable, accessible, interoperable and reusable) principles.

e-cienciaDatos supports DDI-based metadata but does not use a specific metadata schema internally. However, metadata can then be exported in various schemas such as DDI, oai_dc, DataCite, oai_ore... It supports specific metadata for social sciences, astronomy and astrophysics, life sciences, geospatial and new ones could be added if the Open Science WG deems it necessary.

e-cienciaDatos ensures that data will be safeguarded for at least 20 years, although there are no plans to delete datasets due to time expiry. If it is decided to remove a dataset, an attempt will be

[1] http://www.consorciomadrono.es/en/investigam/politicas/

made to connect with the DOI owner researcher, and he will be given the option to point the DOI to the new address. In any case, access to datasets can be removed if it is found that they infringe any of the e-cienciaDatos rules, and then recovered if these problems are solved.

e-cienciaDatos is a repository that has been running since December 2016. The two computer scientists in charge of its administration have been working with repositories, OAI harvesters and metadata transformations for more than 15 years. All libraries have at least one expert librarian in charge of managing their university's community.

e-cienciaDatos offers the possibility to contact the dataset manager, who can be the researcher or the librarian who uploaded the data, to ask for more information about the dataset. Users can also contact the repository administrators. In addition, from 2019, e-cienciaDatos activates surveys twice a year for its end users in which questions related to the use of the repository are asked, and suggestions are solicited.

e-cienciaDatos is a multidisciplinary repository and therefore cannot have staff with knowledge of all possible disciplines of science from which datasets can be created. However, when a researcher wants to publish a dataset in e-cienciaDatos, has to contact the librarian of his university and fill in a Readme.txt form explaining the data and metadata of the dataset, as well as how it can be used. The librarian performs a pre-curation task and requests more information from the researchers if the dataset is not complete.

As explained below, there is a plan to recover from system crashes, hardware failure, accidental data loss or an attack on Consorcio Madroño's servers. e-cienciaDatos services and storage devices are physically located in Consorcio Madroño devices, making e-cienciaDatos more resilient to funding cuts than if it were on a cloud system. As the hardware can continue to operate on a shoestring budget.

Consorcio Madroño's funding is set out in a framework agreement with the Community of Madrid, which is renewed on an annual basis. The e-cienciaDatos repository licence includes a commitment to maintain the datasets for a minimum of 20 years, but the aim of the commitment to maintain the datasets for a minimum of 20 years, but the vocation of the repository is to maintain permanent access. To guarantee access to each individual dataset, the Consorcio Madroño has contracted a DOIs service with DataCite through the GDCC. In this way, access to the dataset is guaranteed through a persistent URL that can be maintained even if the e-cienciaDatos software is changed or if there are changes in the membership of the Consorcio Madroño.

e-cienciaDatos has defined a preservation plan. Preservation focuses on metadata and the most common file types. As e-cienciaDatos is a multidisciplinary repository, file format migration cannot be guaranteed for all possible files, but a periodic check is made to ensure as far as possible that there are no files with obsolete formats in e-cienciaDatos.

7. Data reuse

Researchers who want to publish datasets in e-cienciaDatos have to send a Readme.txt template with the data, where they are asked for the completeness of the data to validate the research and where they have to include information to fill the metadata. The university library staff evaluates the Readme.txt, the datasets and their metadata before publication. Descriptive metadata about the software needed to use the files and a description of how to validate the dataset are requested, but subject-specific metadata in astronomy, life sciences, social sciences or geospatial sciences are also welcome.

Researchers have to accept the repository policy, which includes the possibility of migrating files and metadata due to obsolescence or to enrich the dataset.

Researchers are offered a set of licences to publish their data and have to choose one of them. If a researcher needs a different license, that would have to be assessed by the Open Science Working Group.

8. Technical infrastructure

e-cienciaDatos, being based on Dataverse, covers several information dissemination standards: OAI-PMH for content collection, API SWORD for data deposit, JSON, OAI-ORE bags…

The server supporting e-cienciaDatos is a SuperMicro machine supported by a Synology disk server. Both brands are well known and reliable.

The disk server uses a Linux distribution created by Synology itself, specialised in this type of server, and the main server uses Ubuntu, one of the distributions used by the community, named in the Dataverse prerequisites website.

e-cienciaDatos servers are secured by a firewall based on the Linux iptables standard to close access to all non-essential ports, which is also behind the UNED firewall, the university where the servers are physically installed.

The application software installed is the one recommended in the Dataverse prerequisites page and the applications that complement it in the generation of statistics and previews:

Specifically, as of 15 July 2024, the software offering service on the Ubuntu Focal Fossa distribution is:

- Dataverse 5.10.1
- Dataverse previewers 1.1.1
- counter processor 0.0.1
- Matomo 5.0.3

The above software in turn, relies on:

- Payara 5.2021.6
- java 11.0.22
- Mysql 8.0.36
- postgresql 14
- python 3.10.6
- apache 2.4.52
- solr 8.11.1
- postfix 3.6.4.1
- jq 1.6
- php 8.1

All supporting software is included in the Ubuntu distribution except Payara. Both Payara and the software that provides the services are obtained from official sources.

An upgrade to Dataverse 6.3 is planned in summer 2024. In that case, Java and Payara versions will be updated.

The computer scientists of Consorcio Madroño adapt and customise Dataverse, according to the agreements reached with the Open Science Working Group of Consorcio Madroño. These

agreements are in the minutes of the meetings and, among other issues, indicate when a software upgrade or new development should be done.

Dataverse, Dataverse previewers and counter processor are open source tools maintained by communities with which Consorcio Madroño has collaborated from time to time, while Matomo Analytics, although also open source, is maintained by the company Matomo.

The modifications made to the main Dataverse branch are publicly available on Consorcio Madroño's GitHub.

The e-cienciaDatos servers are located on the UNED campus. They communicate with the outside world through a network maintained by redimadrid (network infrastructure for high-speed data exchange between research, educational and innovation institutions in the Community of Madrid, which allows connection with other national and international research networks through RedIRIS) that offers 10 Gbps for the entire university. In the years that e-cienciaDatos has been operating, there have been no network saturation problems and, if the client's connection is good enough, datasets of many GB can be downloaded in a few minutes.

9. Security, Data integrity and authenticity

Both the software and hardware, including storage and backup, is located at one of the universities of the Consorcio Madroño and is managed at the IT level by the staff of its technical office.

e-cienciaDatos is based on the open source software Dataverse, a powerful and reliable software for the creation of research data repositories the next functionalities:

- Checksums to prevent file corruption.
- Powerful version control system that allows seeing the differences between two versions of the same dataset and the author of the changes. It allows to differentiate between major and minor changes, being major changes whenever a file is changed.
- All versions of the dataset with their files and metadata are publicly accessible.
- The identity of the dataset creator and each person making the version change is stored.
- Dataset changes should be performed by their administrators.

e-cienciaDatos is located in the computing centre of the UNED campus, behind the UNED firewall that prevents access from the outside from most of the ports, and is protected by a firewall based on iptables of e-cienciaDatos itself to allow access only to the essential ports.

Physically, the computing centre is locked, and the UNED keeps a record of all persons entering and leaving the centre.

At the application level, each university has users in e-cienciaDatos with permission to administer their dataverse community. Their passwords are stored encrypted using the "salted hash" algorithm in a database so that even if the database were accessed, the passwords could not be found out. The Dataverse software, on which e-cienciaDatos runs, validates that created passwords have sufficient length and contains at least numbers and letters.

At the operating system level, only the two IT staff of the Consorcio Madroño have access to e-cienciaDatos servers, and these servers can only be accessed physically or via a secure remote connection using SSH. Unbuntu, the operating system on which the server runs, warns if the user passwords created are not sufficiently secure and requests that they should be changed. Consorcio Madroño's IT staff are responsible for keeping the system secure, up-to-date and accessible.

The monthly full backup copies of e-cienciaDatos includes all AIPs and are stored on an isolated external hard disk in a different room from the e-cienciaDatos server in Madrid. Another backup copy is stores in the CSUC server in Cataluña. This copy would take two working days to retrieve due to the

amount of data it contains. Backup copies are checked to ensure that they can be restored twice a year.

The e-cienciaDatos servers have redundancy in the power and disk supplies employing a raid 5 system. Each power supply is connected to a different power line and protected against power cuts by a UPS. In any case, if there is a system failure, the recovery process is started as indicated in the preservation plan:

- If a computer were rendered unusable by an attack, a clean operating system would be installed, and the system would be 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 disk server:
  - At present, and as long as there is space on other servers of the Consorcio Madroño, the datasets will 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%) will be left out while a long-term solution is sought, and the researchers who own the dataset and the university concerned would be notified of the problem.
- Once a month, the integrity of the files is tested by checking the file checksum. If a file is damaged or corrupt, an attempt is made to find out the reason, and, in any case, it will be restored from its backup copy.

e-cienciaDatos does not store sensitive information in the datasets without the explicit permission of those affected. On the other hand, the Consorcio Madroño computer scientists review the logs of access to the system every month to investigate attempted attacks and possible intrusions.