Descripción técnica de e-cienciaDatos

Grupo de trabajo de Ciencia Abierta

1. Descripción general

e-cienciaDatos es el repositorio de datos multidisciplinar para los resultados de investigación del personal investigador de las 6 universidades miembro del 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) y Universidad Rey Juan Carlos (URJC)

e-cienciaDatos está abierto a cualquier tipo de formato de datos aunque se recomienda usar formatos abiertos cuando estén disponibles o, en su defecto, formatos ampliamente aceptados en la disciplina de cada dataset. Cada una de las universidades tiene una comunidad y se encarga de gestionar los datasets de la misma, incluyendo la posibilidad de crear nuevas comunidades si lo considera necesario. e-cienciaDatos admite datasets que ocupen hasta 100Gb. Se pueden admitir datasets más grandes si se considera que puede ser interesante conservarlos, en este caso, lo tendría que aprobar el GT de Ciencia Abierta del Consorcio Madroño.

Los datasets deben ser los datos finales de investigación, y aunque se admiten periodos de embargo, el objetivo final es que los datos estén en abierto. La comunidad a la que se dirige e-cienciaDatos es en primer lugar la comunidad investigadora de las universidades públicas de la Comunidad de Madrid, y por extensión la comunidad científica nacional. Ver las políticas de e-cienciaDatos para más información.

Puede agregar contenido al repositorio el personal investigador de las universidades del Consorcio Madroño. Algunas de las razones para hacerlo pueden ser:

- Necesitan tener los datos de sus investigaciones en acceso abierto por el mandato o recomendación de revistas o agencias de financiación.
- Quieren dar visibilidad a un proyecto de investigación.
- Consideran que los datos de su investigación tienen que estar en abierto por razones éticas o de cualquier otra índole.

e-cienciaDatos está basado en el software libre Dataverse (https://dataverse.org/), un potente y fiable software para la creación de repositorios de datos de investigación que ofrece todas las funcionalidades esperables (https://dataverse.org/software-features) entre las que se encuentran:

- Ofrece DOIs como identificadores persistentes.
- Permite indicar publicaciones y otros datasets relacionados.
- Potente control de versiones.
- Ofrece una forma recomendada para citar los datasets.
- Ofrece búsquedas simples y avanzadas por campo o dentro de una comunidad dataverse. También se pueden hacer búsquedas usando el API-REST o navegar por facetas.
- Además:
    - Cuenta con la certificación CoreTrustSeal, que garantiza que e-cienciaDatos es un repositorio confiable.
    - Los datasets pueden recolectarse mediante el protocolo OAI-PMH en formatos Datacite, oai_dc, oai_ddi, oai_datacite y dataverse_json.
    - e-cienciaDatos es recolectado por el Dataverse de Harvard, Datacite Commons, OpenAIRE EXPLORE y Google Dataset Search.

Para asegurarse la completitud de los datos y metadatos, los administradores del repositorio hacen una curación previa y solicitan al personal investigador una plantilla Readme.txt con instrucciones y preguntas sobre los datos y metadatos del dataset. Solo se admiten datasets del personal investigador de las universidades miembro del Consorcio Madroño. Un dataset de e-cienciaDatos puede actualizarse por las siguientes razones:

- Se ha encontrado algún error en el fichero o en sus metadatos.
- Alguno de los ficheros del dataset puede actualizarse con otro más conveniente.
- Por razones de preservación.

2. Misión / Alcance

Como se indica en el Plan Estratégico 2021-2025, la misión del Consorcio Madroño es “Proporcionar una infraestructura de información que impulse la excelencia de la investigación en las instituciones miembro, contribuyendo al desarrollo de la Ciencia Abierta, a la transformación digital y al desarrollo sostenible, y mejorando la experiencia de sus usuarios a través de los servicios que presta el Consorcio.”

La misión de e-cienciaDatos, como infraestructura proporcionada por el Consorcio Madroño, es garantizar, en el contexto de la ciencia abierta, el depósito, la preservación y la difusión de los datos de investigación generados por las instituciones miembro de acuerdo con los principios FAIR (Findable, Accessible, Interoperable and Reusable), fomentando así una investigación abierta a la comunidad académica y, en un sentido más amplio, a los ciudadanos en general.

La misión de e-cienciaDatos ha sido aprobada por el Consejo de Gobierno del Consorcio Madroño el 22 de febrero de 2022.

3. Infraestructura de la organización

El Consorcio Madroño recibe financiación de las 6 universidades miembro y de la Comunidad de Madrid. En concreto, la financiación para e-cienciaDatos llega a través del proyecto e-ciencia. El Consorcio Madroño existe desde hace más de 25 años y el proyecto e-ciencia lleva casi 20 recibiendo financiación para promocionar en un principio el acceso abierto y más adelante la ciencia abierta. La compra del servidor e-cienciaDatos fue financiada por la Comunidad de Madrid con una partida extra para el proyecto.

Dos informáticos con más de 20 años de experiencia y más de 15 trabajando con repositorios se encargan del mantenimiento del hardware y software de e-cienciaDatos, dedicando aproximadamente un 25 % de su tiempo a e-cienciaDatos.

Cada biblioteca tiene al menos responsable de su biblioteca con experiencia en ciencia abierta encargado de administrar la comunidad dataverse de su universidad y publicar los datasets tras explicar al personal investigador las características y normas de e-cienciaDatos y realizar una gestión y cuidado (curation) previa de los datasets. Tanto los informáticos como al menos un responsable de la biblioteca por institución han recibido formación sobre gestión de datos de investigación. Además han impartido formación a otros colegas de sus universidades y de instituciones externas y continúan formándose mediante la asistencia a jornadas y congresos.

e-cienciaDatos no cuenta con colaboradores internos (insource partners). El único colaborador externo (outsurce partner) es eScire que se encarga de servir DOIs a e-cienciaDatos.

e-cienciaDatos forma parte del Global Dataverse Community Consortium y el Grupo Iberoamericano de desarrollo de Dataverse, del que forman parte repositorios de Argentina, Brasil, Chile, Colombia, Perú, Portugal y España.

4. Dirección del proyecto

Los objetivos a medio plazo, políticas y financiación de e-cienciaDatos se aprueban por la Comisión Técnica del Consorcio Madroño, de la que forman parte los directores de cada una de las bibliotecas miembro y el Director Técnico del Consorcio Madroño. La Comisión Técnica encomienda al Grupo de Trabajo de Ciencia Abierta los trabajos para implementar estos objetivos y a su vez estudia las sugerencias de este grupo de expertos.

El Grupo de Trabajo de Ciencia Abierta del Consorcio Madroño está formado por el director técnico del Consorcio Madroño, un director de biblioteca, personal bibliotecario experto en ciencia abierta (por lo menos uno por universidad) y dos informáticos con experiencia en gestión de repositorios. Los responsables de cada biblioteca son los responsables de la gestión de los datos de sus universidades en conexión con el personal investigador de las universidades miembro y, cuentan con el apoyo de sus respectivas asesorías jurídicas en caso de dudas. Tanto el personal de las bibliotecas como el Consorcio Madroño participan en eventos de formación relacionados con la gestión de datos de investigación y son los responsables del portal InvestigaM, que agrupa las actuaciones relacionadas con la ciencia abierta del Consorcio Madroño. El personal bibliotecario responsable de las universidades asesora y ayuda al personal investigador de forma individual a la creación de sus datasets, y en algunos casos, a aumentar su visibilidad mediante la colaboración en proyectos.

Además, desde el año 2019, e-cienciaDatos activa dos veces al año encuestas para sus usuarios finales en las que se preguntan cuestiones relacionadas con el uso del repositorio y se solicitan sugerencias.

5. Archivo y flujo de datos

En e-cienciaDatos se realiza una gestión y cuidado (curation) previa de datos. Cada universidad tiene uno o varios responsables de su biblioteca encargados de administrar los datasets de su institución, guiar al personal investigador en el proceso de creación de sus dataset y crear/revisar junto al investigador tanto los metadatos como sus ficheros. De esta forma se garantiza la calidad de los metadatos (tal y como se entiende en los principios FAIR), y la homogeneidad de los mismos. e-cienciaDatos también se asegura el derecho a actualizar los datasets, tanto datos como metadatos, para asegurar su preservación y accesibilidad a medio/largo plazo[1]. e-cienciaDatos traza todos los cambios realizados y guarda tanto los datos como los metadatos de todas las versiones cuando se modifica cualquier dataset.

El procedimiento depende de cada universidad y puede consultarse aquí:

Los administradores de e-cienciaDatos apoyan a los responsables de la biblioteca si tienen dudas o necesitan realizar alguna labor que no se puede realizar mediante la web de administración de e-cienciaDatos. También se encargan de crear puntos de geolocalización para los datasets.

Como se describe en las políticas de e-cienciaDatos, los administradores de e-cienciaDatos pueden alterar los ficheros o metadatos de un dataset por razones de preservación o gestión y cuidado (curation) de los datos o metadatos. Si se detecta algún fichero obsoleto o se decide enriquecer los datasets de algún modo, tanto los ficheros como los metadatos de un dataset pueden actualizarse, quedando siempre disponible la versión anterior y el histórico de los cambios de versión con sus diferencias.

6. Calidad de los datos y del servicio

e-cienciaDatos está certificado con el sello CoreTrustSeal, lo que implica que es un repositorio confiable que realiza tareas de preservación, cuidado de datos (curation) y que está comprometido

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

con que tanto los datos como los metadatos siguen los principios FAIR (acrónimo en inglés de encontrable, accesible, interoperable y reutilizable).

El Consorcio Madroño, está certificado según la norma ISO9001 en actividades que cubren el desarrollo de software, incluyendo e-cienciaDatos.

e-cienciaDatos admite metadatos basados en DDI, pero no usa internamente un esquema concreto de metadatos. Sin embargo los metadatos pueden luego exportarse en varios esquemas como DDI, oai_dc, DataCite, oai_ore, …. . Admite metadatos específicos para ciencias sociales, astronomía y astrofísica, ciencias de la vida, geoespaciales y se podrían añadir nuevos si el GT de Ciencia Abierta lo considera necesario.

e-cienciaDatos garantiza la salvaguarda de los datos al menos durante 20 años, aunque no hay previsto eliminar datasets por caducidad de tiempo. Si en algún momento se decide eliminar un dataset, se intentará conectar con el investigador propietario y, se le dará la opción de que el DOI apunte a la nueva dirección. En cualquier caso, se puede eliminar el acceso a datasets si se comprueba que infringen alguna de las normas de e-cienciaDatos, recuperándolo después si se solventan dichos problemas.

e-cienciaDatos es un repositorio que lleva en funcionamiento desde diciembre de 2016. Los dos informáticos encargados de su administración llevan más de 15 años trabajando con repositorios, recolectores OAI y haciendo transformaciones de metadatos. Todas las bibliotecas tienen al menos un experto bibliotecario encargado de administrar la comunidad de su universidad.

e-cienciaDatos ofrece la posibilidad de ponerse en contacto con el responsable del dataset, que puede ser el personal investigador o el responsable de la biblioteca que subió los datos, para pedir más información sobre el dataset. También se puede poner en contacto con los administradores del repositorio. Además, desde el año 2019, e-cienciaDatos activa dos veces al año encuestas para sus usuarios finales en las que se preguntan cuestiones relacionadas con el uso del repositorio y se solicitan sugerencias.

e-cienciaDatos es un repositorio multidisciplinar y, por tanto no puede tener personal con conocimientos sobre todas las posibles disciplinas de la ciencia de las que se pueden crear datasets.

Cómo se explica más adelante, existe un plan para recuperarse ante caídas de sistema, rotura de hardware, pérdida accidental de datos o ante un ataque a los servidores del Consorcio Madroño. Los servicios y dispositivos de almacenamientos de e-cienciaDatos se encuentran físicamente en dispositivos del Consorcio Madroño, por lo que e-cienciaDatos es más resistente a recortes en la financiación que si estuviera en algún sistema en la nube. Ya que el hardware puede continuar funcionando con muy poco presupuesto.

La financiación del Consorcio Madroño se recoge en un acuerdo marco con la Comunidad de Madrid que se renueva de forma anual. En la licencia de depósito de e-cienciaDatos se recoge el compromiso de mantener los datasets un mínimo de 20 años, pero la vocación del repositorio es mantener el acceso de forma permanente. Para garantizar el acceso a cada dataset individual, el Consorcio Madroño tiene contratado un servicio de DOIs con DataCite a través de eScire. De esta forma, se garantiza el acceso al dataset a través de una URL persistente que se puede mantener aunque se cambiara el software de e-cienciaDatos o haya cambios en la membresía del Consorcio Madroño.

e-cienciaDatos tiene definido un plan de preservación. La preservación se centra en los metadatos y en los tipos de ficheros más comunes. Al ser e-cienciaDatos un repositorio multidisciplinar no se puede garantizar la migración de formatos de ficheros para todos los ficheros posibles, pero se hace una comprobación periódica para asegurar en la medida de lo posible que no hay ficheros con formatos obsoletos en e-cienciaDatos.

7. Reutilización de datos

El personal investigador que quiere publicar datasets en e-cienciaDatos tiene que rellenar una platilla Readme.txt, indicando la completitud de dichos datos para poder validar la investigación y en el que tienen que incluir información que permita rellenar los metadatos. El Readme.txt, los datasets y sus metadatos son evaluados por el personal de la biblioteca de su universidad antes de su publicación. Se solicitan metadatos descriptivos sobre el software necesario para ejecutar los archivos y una descripción sobre cómo usar los datos para validar el dataset, pero también se admiten metadatos específicos temáticos de astronomía, ciencias de la vida, ciencias sociales o geoespaciales.

El personal investigador tiene que aceptar la política del repositorio que incluye la posibilidad de migrar ficheros y metadatos por obsolescencia o para enriquecer el dataset.

Al personal investigador se le ofrece un conjunto de licencias para publicar sus datos y tienen que escoger una de las ofrecidas salvo causas excepcionales que tendrían que evaluar el responsable de su biblioteca.

8. Infraestructura técnica

e-cienciaDatos, al estar basado en Dataverse, cubre varios estándares de diseminación de información: OAI-PMH para recolección de contenidos, API SWORD para depositar datos, JSON, OAI-ORE bags, ….

El servidor que sostiene e-cienciaDatos es una máquina SuperMicro apoyada en un servidor de disco Synology. Ambas marcas son conocidas y fiables.

El servidor de discos utiliza una distribución de Linux creada por la propia Synology especializada en este tipo de servidores y el servidor principal utiliza Ubuntu, una de las distribuciones usadas por la comunidad, nombradas en la web de prerrequisitos de Dataverse.

A parte de esto, los servidores están asegurados por un firewall basado en el estándar de Linux iptables para cerrar el acceso a todos los puertos que no son indispensables, que además se encuentra tras el firewall de la UNED, la universidad dónde se encuentran instalados físicamente los servidores.

El software de aplicación instalado es el recomendado en la página de prerrequisitos para Dataverse y las aplicaciones que le complementan en la generación de estadísticas y previsualizaciones:

En concreto, a 15 de julio de 2024 el software que ofrece servicio sobre la distribución de Ubuntu Jammy Jellyfish es:

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

El software anterior a su vez se apoya en:

- 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

Todo el software de apoyo está incluido en la distribución de Ubuntu menos Payara. Tanto Payara como el software que ofrece los servicios se obtienen de fuentes oficiales.

Está prevista una actualización a Dataverse 6.3 durante el verano 2024. En ese caso, se instalarán java 17 y una nueva versión de Payara.

Los informáticos del Consorcio Madroño adaptan y personalizan Dataverse, según los acuerdos alcanzados con el Grupo de Trabajo de Ciencia Abierta del Consorcio Madroño. Estos acuerdos están en las actas de las reuniones y, entre otros temas, se indica cuando hay que hacer una actualización de software o un nuevo desarrollo.

Dataverse, Dataverse previewers y counter processor son herramientas de software libre mantenidas por comunidades con las que ha colaborado el Consorcio Madroño de forma puntual, mientras que Matomo Analytics, aunque también es software libre, es mantenido por la empresa Matomo.

Las modificaciones realizadas sobre la rama principal de Dataverse están disponibles públicamente en el GitHub del Consorcio Madroño.

Los servidores de e-cienciaDatos se encuentran dentro del campus de la UNED y se comunican con el exterior mediante una red mantenida por redimadrid (infraestructura de red para el intercambio de datos a alta velocidad entre instituciones de investigación, educativas y de innovación de la Comunidad de Madrid, que permite la conexión con otras redes de investigación nacionales e internacionales a través de RedIRIS) que ofrece 10 Gbps para toda la universidad. En los años que lleva funcionando e-cienciaDatos no ha habido ningún problema de saturación de red y permiten descargar, si la conexión del cliente es suficientemente buena, datasets de muchos Gb en unos minutos.

9. Seguridad, Integridad y autenticidad de los datos

Tanto el software como el hardware, incluyendo el de almacenamiento y backup se encuentra en una de las universidades del Consorcio Madroño y se administra a nivel informático por el personal de su oficina técnica.

e-cienciaDatos está basado en el software libre Dataverse, que ofrece las siguientes funcionalidades:

- Sumas de verificación para evitar la corrupción de ficheros.
- Potente gestor de versiones que permite ver las diferencias entre dos versiones del mismo dataset y el autor del cambio de versión. Permite diferenciar entre cambios de mayores y menores, siendo las mayores siempre que se cambie algún fichero.
- Todas las versiones del dataset con sus ficheros y metadatos están accesibles de forma pública.
- Se guarda la identidad del creador del dataset y de cada persona que hace el cambio de versión.
- Solo los administradores de un dataset o del repositorio pueden hacer cambios en el mismo.

e-cienciaDatos está situado en el centro de cálculo del campus de la UNED, tras el firewall de la UNED que impide el acceso desde el exterior desde la mayoría de los puertos y protegido por un

firewall basado en iptables del propio e-cienciaDatos para permitir acceso solo a los puertos indispensables.

Físicamente, el centro de cálculo está cerrado con llave y la UNED mantiene un registro de todas las personas que entran y salen en el mismo.

A nivel de aplicación, cada universidad tiene usuarios en e-cienciaDatos con permiso de administración de su comunidad dataverse. Sus contraseñas se guardan codificadas mediante el algoritmo “salted hash” en una base de datos de forma que, aunque se tuviera acceso a la base de datos, no se podrían averiguar las contraseñas. El software Dataverse, sobre el que se ejecuta e-cienciaDatos, valida que la contraseña tenga una longitud suficiente y que las contraseñas creadas tengan por lo menos números y letras.

A nivel de sistema operativo solo tienen acceso al servidor los dos informáticos del Consorcio Madroño, y solo se puede acceder a los servidores físicamente o mediante conexión remota segura usando SSH. Ubuntu, el sistema operativo sobre el que se ejecuta el servidor, avisa si la contraseña creada no es suficientemente segura y solicita que se cambie. Los informáticos del Consorcio Madroño son los encargados de mantener el sistema seguro, actualizado y accesible.

Las copias de seguridad completas mensuales que se realizan de e-cienciaDatos abarcan todos los AIPs y se guardan tanto en un disco duro externo aislado fuera del centro de cálculo donde encuentra e-cienciaDatos en Madrid, como en un servidor remoto en Cataluña propiedad del CSUC. Dado el gran tamaño de los datos, esta copia de seguridad tarda varios días en realizarse. Se comprueba que se pueden restaurarse dos veces al año.

El servidor e-cienciaDatos tiene redundancia en las fuentes de alimentación y de discos mediante un sistema raid 5. Cada fuente de alimentación se encuentra conectada a una toma de corriente distinta y protegida contra cortes de luz mediante un SAI. Ante un fallo del sistema, se seguiría el procedimiento indicado en el plan de preservación:

- Si algún equipo quedara inutilizable por un ataque, se instalaría un sistema operativo limpio y se restauraría el sistema a partir de su copia de seguridad.
- Si se borra o daña accidentalmente un dataset, se recuperaría de la copia de seguridad.
- Si hubiera un fallo hardware en el servidor principal de e-cienciaDatos, se movería el servidor a otra máquina del Consorcio Madroño a partir de la copia de seguridad.
- Si hubiera un fallo hardware en el servidor de discos:
    - En el momento actual y mientras haya espacio en otros servidores del Consorcio Madroño, se moverían los datasets a otros servidores del Consorcio Madroño.
    - Si no hubiera espacio para almacenar todos los datasets, se dejarían fuera los datasets que más espacio ocupan (menos del 1%) mientras se busca una solución a largo plazo y se avisaría del problema a los investigadores dueños del dataset y a la universidad correspondiente.
- Una vez al mes se comprueban la integridad de los ficheros mediante su checksum de los ficheros. Si se encuentra algún fichero dañado o corrupto, se intentaría averiguar la razón y, en cualquier caso se restauraría desde su copia de seguridad.

e-cienciaDatos no almacena información sensible en los datasets salvo permiso explícito por parte de los afectados. Por otro lado, los informáticos revisan mensualmente los logs de accesos al sistema para investigar los intentos de ataque y posibles intrusiones.