Plan de Preservación de e-cienciaDatos

Grupo de trabajo de Ciencia Abierta

1 de Marzo de 2022

Alcance y propósito

e-cienciaDatos es el repositorio de datos multidisciplinar de las 6 universidades miembro del Consorcio Madroño. 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 sus dataset, incluyendo la posibilidad de crear nuevas colecciones si lo considera necesario.

Pueden agregar contenido al repositorio los investigadores de las universidades del Consorcio Madroño. Los investigadores pueden decidir publicar sus datasets por alguna de las siguientes razones:

- Necesitan tener los datos de sus investigaciones en acceso abierto por 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.

Los datasets deben ser los datos finales de la investigación y, aunque se admiten períodos de embargo, su objetivo final debe de ser estar en acceso abierto. La comunidad designada (usuarios a los que se dirige el repositorio) es la comunidad científica de forma global. Al ser e-cienciaDatos un repositorio multidisciplinar no se puede concretar más.

En e-cienciaDatos se realiza una gestión y cuidado (curation) previa de datos. Cada universidad tiene uno o varios bibliotecarios encargados de administrar los datasets de su institución, guiar a los investigadores en el proceso de creación de sus dataset y crear/revisar junto al investigador tanto los metadatos como sus ficheros. e-cienciaDatos carece de autoarchivo para asegurar 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. e-cienciaDatos permite trazar todos los cambios realizados y guarda tanto los datos como los metadatos de todas las versiones cuando se modifica cualquier dataset.

Este plan de preservación se crea para asegurar el acceso a medio y largo plazo de los datasets contenidos en e-cienciaDatos, de forma que se garantice el acceso a los mismos tanto a los investigadores que han confiado en este repositorio para guardar sus datasets como para mostrar al resto de la comunidad científica que se trata de un repositorio confiable. Además de garantizar el acceso, también se tratará de garantizar que a medio y largo plazo que dichos datasets puedan ser reutilizados y validados por la comunidad científica.

En la preservación se distinguirá entre ficheros comunes, de los que se puede garantizar una migración si llegado el caso el formato del fichero queda obsoleto, y formatos específicos sobre los cuales no se puede garantizar dicha migración.

Objetivos

El objetivo de este plan de preservación es garantizar el acceso a medio y largo plazo a los objetos digitales accesibles desde e-cienciaDatos, así como su posibilidad de reutilización y validación por la comunidad científica. Dada la heterogeneidad de formatos de ficheros, solo se puede garantizar la migración de los formatos más conocidos, pero en la medida de lo posible

también se intentará vigilar la posible obsolescencia de todos los ficheros y planificar su migración cuando sea posible. Mirando la lista de ficheros a 4 de febrero de 2022, podemos observar que más del 95% de los ficheros tienen extensiones conocidas y comunes, de las que puede ser sencillo realizar una migración: pdf, zip, txt, tar.gz, wav, csv, mp4, ogg, webm, xlsx, tab, csv, ods, rar, docx ...

Colecciones y usuarios

e-cienciaDatos tiene una comunidad dataverse por universidad y, dentro de cada universidad, se han definido colecciones por proyectos cuando los administradores de cada universidad lo ha considerado necesario. Los datasets subidos a e-cienciaDatos tienen que estar en acceso abierto para su posible reutilización y validación por cualquier investigador. No obstante, pueden estar un tiempo en estado de borrador, accesibles solo mediante una URL privada mientras el dataset es validado por los autores, revistas, colaboradores…

Los usuarios son:

- La sociedad en general y los investigadores en particular que acceden y descargan los datasets de e-cienciaDatos.
- Los bibliotecarios, administradores de la comunidad dataverse de su universidad encargados de crear los datasets y administrar su comunidad dataverse y sus colecciones.
- Los informáticos administradores del repositorio.

Funciones y responsabilidades

Cada comunidad dataverse es administrada por uno o varios bibliotecarios de una universidad, que son los responsables de los siguientes tareas:

- Crear los datasets a petición de sus investigadores.
- Verificar que los datasets estén completos para permitir su validación y reutilización.
- Solicitar los datos necesarios a los investigadores para garantizar el punto anterior.
- Generar las colecciones dataverse que crean convenientes.
- Crear libros de visitas si lo solicita un investigador para analizar el uso de un dataset.

También hay dos informáticos encargados de administrar el repositorio cuyas responsabilidades son:

- Mantener el software y el hardware que alberga el repositorio seguro y en buen estado.
- Actualizar el software del repositorio cuando el grupo de trabajo de Ciencia Abierta lo considere necesario, normalmente cuando hay alguna nueva funcionalidad interesante en el software usado para gestionar el repositorio.
- Subir ficheros grandes cuando, por su tamaño, no pueden subirse mediante la interfaz web de e-cienciaDatos.
- Informar de estadísticas y posibles incidencias del repositorio al grupo de Ciencia Abierta.
- Dar soporte técnico a los bibliotecarios administradores de las comunidades dataverse sobre el funcionamiento del repositorio.
- Geolocalizar los datasets que soliciten los administradores de las comunidades dataverse de los repositorios.
- Verificar la posible obsolescencia de los ficheros.
- Migrar los ficheros a otros formatos cuando queden obsoletos.

A nivel organizativo, el grupo de trabajo de Ciencia Abierta del Consorcio Madroño es el grupo de expertos encargado de solicitar cambios y mejoras en el repositorio y tomar las decisiones. Algunas de las decisiones tomadas por el grupo de trabajo, bien por su coste económico o por su tiempo de desarrollo, tienen que ser aprobadas por la Comisión Técnica del Consorcio Madroño,

uno de sus órganos de gobierno. La Comisión Técnica también puede solicitar desarrollos, informes o funcionalidades al grupo de trabajo.

Compromisos y políticas institucionales

El Consejo de Gobierno del Consorcio Madroño apoya la distribución y preservación de la documentación científica. En 2013 los rectores de las universidades que forman el Consorcio firmaron la Declaración de apoyo al acceso abierto que fue renovada en 2017 con la Declaración de apoyo a la Ciencia Abierta. En ambas declaraciones se recomienda a las universidades que forman dicho consorcio "*Adoptar políticas que aseguren el archivo, preservación y difusión en abierto de la producción académica y científica de sus instituciones*".

La Comisión Técnica ha firmado el 22 de febrero de 2022 la misión del repositorio e-cienciaDatos, donde la preservación de los datos de investigación es uno de sus puntos fundamentales.

Acciones de preservación y control de calidad

La preservación sigue el modelo OAIS del siguiente modo:

Componentes del modelo de referencia OAIS

A continuación indicamos la relación entre los componentes del modelo de referencia OAIS y e-cienciaDatos.

Productores

Individuos o entidades que transfieren información al sistema OAIS. El sistema OAIS tiene que negociar con ellos tanto los formatos, la información relativa al objeto a preservar, así como los derechos de transformación y de difusión. Los productores de información son los investigadores del Consorcio Madroño aunque no son ellos los que crean los datasets en e-cienciaDatos, sino los bibliotecarios administradores de las universidades en su nombre. Los bibliotecarios se aseguran, entre otras cosas, de que los ficheros tienen un formato adecuado para su preservación y que los metadatos contenidos son suficientes.

Consumidores y comunidad designada

Consumidores: Individuos o entidades que usan la información que provee el sistema OAIS. Los consumidores pueden acceder, solicitar acceso o solicitar información al sistema OAIS.

La comunidad designada son los consumidores para los que se ha diseñado el sistema OAIS y

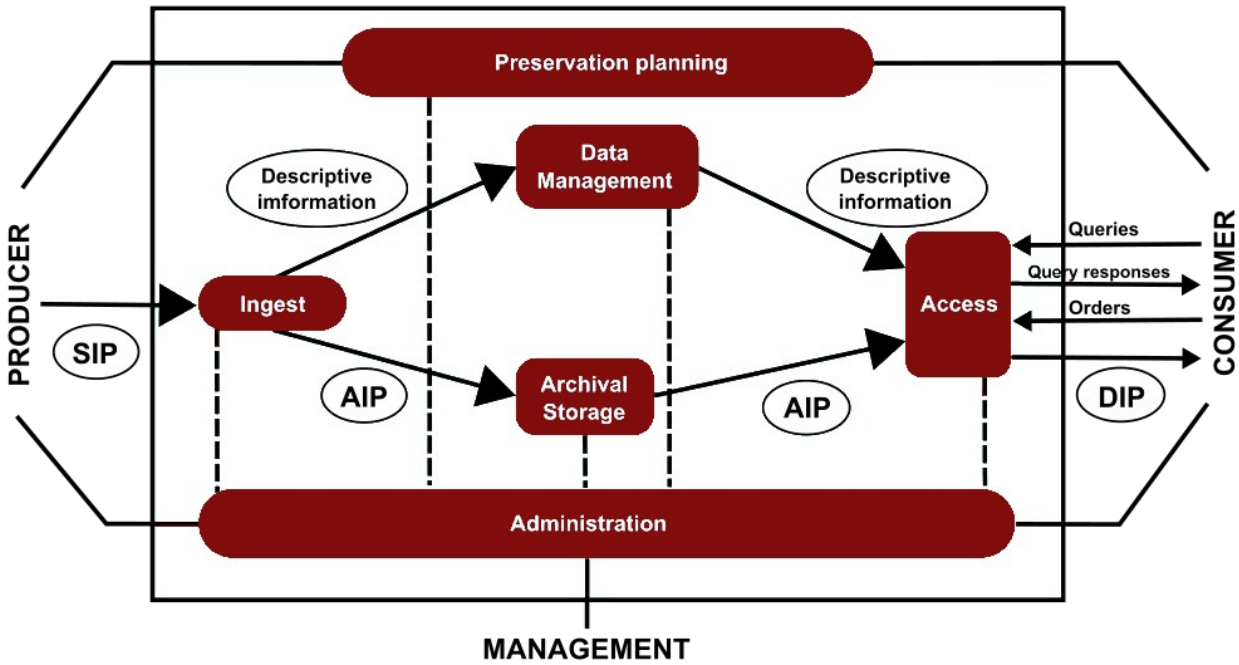
que deben ser capaces de entender la información almacenada. En primer lugar, consiste en los investigadores de las universidades del Consorcio Madroño como depositantes de los datasets y, en segundo lugar, el resto de investigadores como usuarios de los mismos. e-cienciaDatos tiene sus datasets en acceso abierto, por lo que cualquier persona con acceso a internet puede descargarse los mismos.

Administración

Se encarga de formular, revisar y asegurar el cumplimiento de las políticas del sistema OAIS. Incluye:

- Ámbito de la colección a preservar. La colección a preservar son todos los datasets contenidos en e-cienciaDatos, aunque solo se garantiza la migración de formatos obsoletos para los tipos de ficheros más comunes que abarcan más del 95 % de los ficheros contenidos en los datasets de e-cienciaDatos.
- Garantía de preservación para que el archivo sea confiable. La preservación se garantiza por un mínimo de 20 años, aunque la idea es que se pueda albergar la información y preservar más allá de este tiempo. Con este fin se ha desarrollado el plan de análisis de riesgos que se comenta más adelante. También se detallará la viabilidad tanto técnica como económica del repositorio.

Servicios funcionales OAIS



* Imagen procedente de la Wikipedia con licencia Creative Commons Attribution-Share Alike 4.0 International.

A continuación se indica cómo se mapean los servicios funcionales OAIS en e-cienciaDatos:

- Depósito: Procesos para aceptar la información enviada por los productores. Es realizada por los bibliotecarios administradores de la universidad en colaboración con el investigador y de los administradores de los repositorios cuando es necesario. Consiste en:
  - Recepción de información en el sistema enviada por el investigador (productor).
  - Validación para comprobar que es completa y sin errores por parte del bibliotecario.
  - Transformación en un formato admitido por el sistema. Si los formatos o nombres de archivos no son considerados adecuados por el bibliotecario, solicita un cambio al investigador.
  - Creación de metadatos descriptivos que permitan la búsqueda y subida al sistema del

elemento a preservar. Tarea realizada por el investigador, con la ayuda del bibliotecario.
    - Transferencia de la información al lugar de almacenamiento. Tarea realizada por el bibliotecario, con ayuda de los administradores del repositorio en caso necesario.
- Almacenamiento de archivos. Administra el almacenamiento a largo plazo, mantenimiento y seguridad de los materiales dentro del sistema OAIS:
    - Asegura que la información se guarda en la forma apropiada y que además es accesible a largo plazo. e-cienciaDatos usa el software Dataverse, con amplio soporte de la comunidad e instalaciones por todo el mundo. Dispone de mecanismos como chequeo de checksums y control de versiones. Permite acceder a versiones anteriores de los datasets y hacer comparaciones entre las mismas.
    - Migración de los medios de almacenamiento y formatos. Esta tarea la realizan los administradores del repositorio cuando un formato se ha quedado obsoleto, hay fallos en medios de almacenamiento o se considere conveniente por cualquier razón cambiar de medio de almacenamiento.
    - Realiza funciones de seguridad como chequeo de errores y procedimiento de recuperación ante desastres. Tareas realizadas por los administradores del repositorio con la ayuda del software Dataverse.
    - Se comunica con el módulo de acceso cuando hay solicitudes de los consumidores. Esta tarea es realizada por el software de Dataverse.
- Gestión de datos: Mantiene los metadatos descriptivos que permiten identificar y servir de soporte a las herramienta de búsquedas, monitores de rendimiento o sistemas de estadísticas. Estas tareas las realiza el software de Dataverse e incluyen:
    - Mantenimiento de las BD.
    - Interrogar a las BDs ante peticiones de otras entidades funcionales del OAIS.
    - Actualizar las BDs cuando se incorpora nueva información, se elimina o se actualiza.
    - Dar soporte a la búsqueda, recuperación y administración de los datos del sistema OAIS.
- Plan de preservación. Detallado en este documento. Mapea la estrategia de preservación y recomienda revisiones a esta estrategia con la evolución del sistema OAIS:
    - Monitoriza el entorno externo buscando cambios y peligros: nuevas tecnologías para almacenamiento o acceso, cambios en la comunidad designada o en sus expectativas...
    - Crea recomendaciones para actualizar las políticas y procedimientos para adaptarse a los cambios.
- Acceso: Controla los procesos y servicios ofrecidos a los consumidores. Todas estas tareas son realizadas por el software de Dataverse:
    - Localiza, solicita y recibe los documentos e información de la entidad funcional de almacenamiento.
    - Presenta los resultados de las búsquedas al consumidor.
    - Envía los ítems solicitados, transformándolos si es necesario.
    - Implementa mecanismos de seguridad en el acceso. El acceso como consumidor está abierto a todo el mundo, aunque dirigido a la comunidad científica. Tienen acceso para subir los datasets los bibliotecarios administradores de cada universidad y los informáticos administradores del repositorio.
- Administración: Realiza las tareas administrativas del día a día y coordina las actividades de las otras entidades funcionales. Estas dos tareas también son realizadas por los administradores de e-cienciaDatos.
    - Monitorización del rendimiento.
    - Actualizaciones del sistema.

Servicios comunes OAIS

Complementan a las entidades funcionales para garantizar el acceso a largo plazo al material preservado. Estas funciones son supervisadas por los administradores de e-cienciaDatos e incluyen:

- Computación básica.
- Recursos de red.
- Servicios del sistema operativo.
- Servicios de seguridad.

Versiones del Information Package (IP)

- Submission Information Package (SIP). Producto transferido para su archivo con los metadatos iniciales. El SIP es subido por el bibliotecario administrador de la universidad con los ficheros y metadatos facilitados por el investigador, tras comprobar que los metadatos son completos para asegurar su preservación y permitir búsquedas consistentes en el sistema. Pueden colaborar los administradores de e-cienciaDatos si es necesario en tareas como subir ficheros grandes, responder dudas técnicas, crear puntos de geolocalización...
- Archival Information Package (AIP). Incluye el "Content data object" (objeto a preservar) y su información de representación asociada (metadatos). Este archivo se distribuye entre la base de datos de e-cienciaDatos y el sistema de ficheros y lo crea el software Dataverse a partir del SIP. A estos dos elementos se les conoce como Content Information y a ellos se les añade metadatos de preservación (Preservation Description Information), que incluyen a su vez cinco componentes:
  - Información de referencia: Identificador único dentro del sistema.
  - Información de contexto: Relación con otros objetos.
  - Información de procedencia: Fecha de creación, fechas de modificaciones, traza de modificaciones.
  - Información de estabilidad (Fixity information): Checksums.
  - Información de derechos de acceso: Licencia, permisos de acceso.

  e-cienciaDatos almacena el AIP como un AIC (Archival Information Collection), dado que se almacenan de forma independiente cada fichero a preservar y por otro lado sus metadatos y cada fichero puede tener sus metadatos propios asociados e incluso una estrategia de preservación diferente.
- Dissemination Information Package (DIP): Información y ficheros que se entregan al consumidor. La distribución de estos ficheros y metadatos la realiza el software Dataverse, permitiendo visualizar dichos metadatos en distintos formatos.

Sostenibilidad financiera

La financiación del Consorcio Madroño se recoge en un acuerdo marco con la Comunidad de Madrid que se renueva de forma anual. El Consorcio Madroño recibe financiación de las 6 universidades miembro y de la Comunidad de Madrid. En concreto, la financiación para el repositorio e-cienciaDatos llega a través del proyecto e-ciencia. El Consorcio Madroño existe desde hace más de 20 años y el proyecto e-ciencia lleva más de 15 recibiendo financiación para promocionar en un principio el acceso abierto y más adelante la ciencia abierta, incluyendo en ambos casos la preservación de la documentación y datos científicos y académicos. La compra del servidor e-cienciaDatos fue financiada por la Comunidad de Madrid con una partida extra para el proyecto.

Sostenibilidad técnica

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 20 % de su tiempo a e-cienciaDatos.

Cada biblioteca del Consorcio Madroño tiene al menos un bibliotecario experimentado encargado de administrar la comunidad dataverse de su universidad y subir los datasets tras explicar a los investigadores las características y normas de e-cienciaDatos y realizar una gestión y cuidado (curation) previa de los datasets para asegurarse de que cumplen los requisitos de preservación de e-cienciaDatos. Tanto los informáticos como al menos un bibliotecario por institución han recibido formación sobre gestión de datos de investigación, que incluía preservación de objetos digitales y han impartido formación tanto a compañeros de su institución como a personal de otras instituciones.

Plan de contingencia y análisis de riesgos

Plan de contingencia ante fallos tecnológicos

Tanto el servidor e-cienciaDatos como su servidor de discos tienen redundancia en las fuentes de alimentación y de discos mediante un sistema raid 5, por lo que si falla un disco, el sistema sigue funcionando. Cada fuente de alimentación se encuentra conectada a una línea de potencia diferente, protegidas contra cortes de luz mediante un SAI. Los servidores están en un centro de cálculo al que solo tiene acceso el personal autorizado y donde se registran todas las entradas y salidas.

Los accesos lógicos a los servidores están protegidos por un firewall de dos capas, a nivel institución y a nivel del propio servidor, para prevenir accesos no autorizados.
En caso de fallo de algún disco, el servidor crea una alarma y se planifica la actuación necesaria para arreglar la situación en el menor tiempo posible. Dado el gran volumen de la copia de seguridad, recuperar e-cienciaDatos de un fallo hardware llevaría dos días de trabajo.
Las copias de seguridad completas 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. En cualquier caso:

- 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 NAS o en el que alberga el software de e-cienciaDatos:
  - 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 (probablemente menos del 1%) mientras se busca una solución a largo plazo y se avisaría del problema a los investigadores propietarios del dataset y a la universidad correspondiente.
- Una vez al mes se comprueba 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.

Las copias de seguridad completas, debido a su tamaño, se realizan una vez al mes. También se realizan copias de seguridad incrementales de forma semanal que se guardan en otro servidor propiedad del Consorcio Madroño.

En cuanto a la vigilancia ante posibles fallos:

- Se hace una revisión mensual de los discos de e-cienciaDatos. Si uno falla, se inicia el procedimiento de sustitución por uno nuevo.
- Anualmente se hace un estudio de los formatos de los ficheros que forman los datasets de e-cienciaDatos. Los resultados se muestran en un informe que se envía al Grupo de Trabajo de Ciencia Abierta del Consorcio Madroño y se estudian los ficheros que hay que migrar.
- Desde un ordenador externo se monitoriza cada hora que el servidor de e-cienciaDatos está accesible de forma automática.
- Se realizan copias de seguridad completas mensuales y semanales incrementales.
- Se activan dos veces al año durante un mes encuestas para los usuarios de e-cienciaDatos, en las que se les pregunta, entre otras cuestiones, sobre posibles mejoras del sistema.

Plan de contingencia financiera

Como se comentó anteriormente, el Consorcio Madroño está financiado por las 6 universidades miembro y la Comunidad de Madrid, con lo que cuenta con una sólida base para continuar funcionando después de sus más de 20 años de existencia. No obstante, en el caso altamente improbable de que alguna crisis cortara la financiación para el proyecto, este podría seguir funcionando con el hardware actual al estar todos los servidores, incluyendo el almacenamiento, en hardware del Consorcio Madroño. Si tras el corte de financiación, los discos fueran fallando y no hubiera espacio para todos los datasets, se irían eliminando del repositorio los datasets de más tamaño y se devolverían los mismos a sus investigadores o instituciones para que pudieran seguir haciendo uso del mismo a través el DOI asignado.

Formación

Tanto los informáticos como al menos un bibliotecario por institución del Consorcio Madroño que trabajan en e-cienciaDatos han recibido formación sobre gestión de datos de investigación, en concreto, el curso *Research Data Management and Sharing* de Coursera que incluye nociones de preservación digital. A su vez, el personal que ha recibido formación también ha participado en formaciones tanto a bibliotecarios de otras instituciones como de instituciones externas y sigue participando en seminarios y jornadas sobre gestión de datos y preservación digital

Evaluación, seguimiento y revisión del propio plan

Este plan de preservación ha sido creado por el Grupo de Trabajo de Ciencia Abierta del Consorcio Madroño. Está vigente desde el 1 de marzo de 2022 y se revisará de forma bianual por los mismos grupos de forma que se adapte a las necesidades futuras del repositorio.