La Gestión de Datos de Investigación

Marisa Pérez Aliende
Universidad Autónoma de Madrid
mp.aliende@uam.es
Sevilla, 13 de Junio 2017

UNIVERSIDAD D SEVILLA
Biblioteca

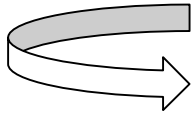
Los datos de investigación

- Definición y tipos
- Ciclo de vida
- Actores en la gestión de los datos
- La gestión a lo largo del ciclo de vida

¿Qué es el *Open Data*?

"El conocimiento abierto es cualquier contenido, información o dato que puede ser libremente utilizado, reutilizado y redistribuido sin restricciones legales, tecnológicas o sociales. Es en lo que se convierten los datos abiertos cuando son útiles, usables y utilizados" (Open Knowledge Foundation).

"Los datos abiertos son datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, cuando más, al requerimiento de atribución y de compartirse de la misma manera en que aparecen" (http://opendatahandbook.org)



INTEROPERABILIDAD

"Denota la habilidad de diversos sistemas y organizaciones para trabajar juntos (interoperar). En este caso, es la habilidad para interoperar o integrar diferentes conjuntos de datos."

Es necesario que estén abiertos:

- Técnicamente: estar disponibles en un formato legible por máquina.
- Legalmente: aplicar una licencia alineada con la "definición de abierto" (http://opendefinition.org/licenses/)

¿Qué son los datos de investigación?

¿Qué son los datos de investigación?

“Material factual registrado comúnmente aceptado por la comunidad científica como necesario para validar los resultados de investigación. Además de los datos , incluye metadatos (ej.: protocolos experimentales, códigos escritos para los análisis estadísticos” .

(National Science Foundation)

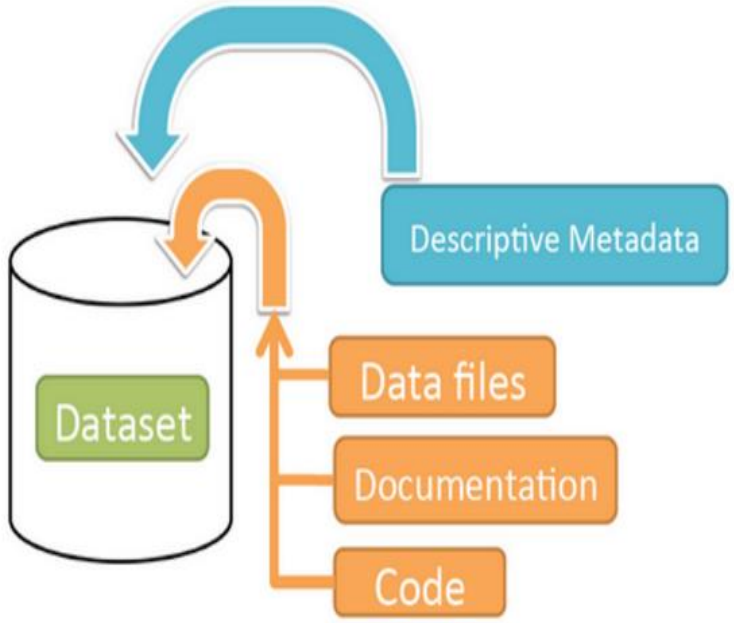
“Los datos de la investigación son hechos, observaciones o experiencias en que se basa el argumento, la teoría o la prueba. Los datos pueden ser numéricos, descriptivos o visuales. Los datos pueden ser en estado bruto o analizado, pueden ser experimentales u observacionales. Los datos incluyen: cuadernos de laboratorio, cuadernos de campo, datos de investigación primaria (incluidos los datos en papel o en soporte informático), cuestionarios, cintas de audio, videos, desarrollo de modelos, fotografías, películas, y las comprobaciones y las respuestas de la prueba. Las colecciones datos para la investigación pueden incluir diapositivas; diseños y muestras. En la información sobre la procedencia de los datos también se podría incluir: el cómo, cuándo, donde se recogió y con que (por ejemplo, instrumentos). El código de software utilizado para generar, comentar o analizar los datos también pueden ser considerados datos” .

(Universidad de Melbourne, en el informe Recolecta)

“Los datos de investigación son registros factuales, que pueden tomar la forma de números, símbolos, texto, imágenes o sonidos, utilizados como fuente primaria para la investigación, comúnmente aceptados por la comunidad científica como necesarios para validar los hallazgos de la investigación”.

(Griffith University)

Los datos de investigación: datasets



Dataverse 4.0 Schematic Diagram of a Dataset

¿Qué es un dataset?:

- Colección de datos reunidos durante la ejecución de un proyecto de investigación.
- Los datasets son objetos digitales compuestos y heterogéneos.
- Pueden comprender diferentes elementos o tipos de datos: documentos de texto, hojas de cálculo, ficheros de operaciones matemáticas, gráficos, imágenes, etc.
- Constituye la base de una investigación y va asociado a una publicación científica como resultado de dicha investigación.
- Adquiere valor añadido si se integra con la publicación relacionada (‘linking data’: cita y enlace), independientemente de su ubicación.

Grupo de Trabajo de “Depósito y Gestión de datos en Acceso Abierto” del proyecto RECOLECTA. La conservación y reutilización de los datos científicos en España. Informe del grupo de trabajo de buenas prácticas [en línea] Madrid: Fundación Española para la Ciencia y la Tecnología, FECYT (2012)

- Se almacenan y gestionan en repositorios desarrollados conforme estándares internacionales.

La *Long Tail of Research Data*

BIG DATA

Son los datos de la *big science* los que se comparten entre los grandes equipos. Los científicos suelen contar con sofisticadas infraestructuras que les ayudan en la gestión de esos volúmenes de datos.
Bien organizada, utilizada frecuentemente y citada.

SMALL SCIENCE / LITTLE SCIENCE

The long tail of research data. Procede de equipos pequeños, recogen datos de proyectos específicos, altamente heterogéneos, gestionados por lo general de manera local y en el ámbito del investigador, no preservándose en repositorios.

"Pequeños, diversos datasets gestionados por múltiples actores en diferentes contextos: universidades, bibliotecas,etc." (RDA).

- Hacer que los datos de la *long tail* puedan ser localizables y reusables se convierte en uno de los grandes desafíos.
- Relacionado con la financiación científica.

(Horstmann, W. "Long Tail data Access and the Research Data Alliance". 2014)
(Proll, Meixner y Rauber "Precise data identification services for Long Tail Research Data". 2017)

Los datos de investigación: Clasificación

Los datos de invetigación pueden generarse para diferentes propósitos y mediante diferentes procesos. En función de esto la Research Information Network los clasifica como:

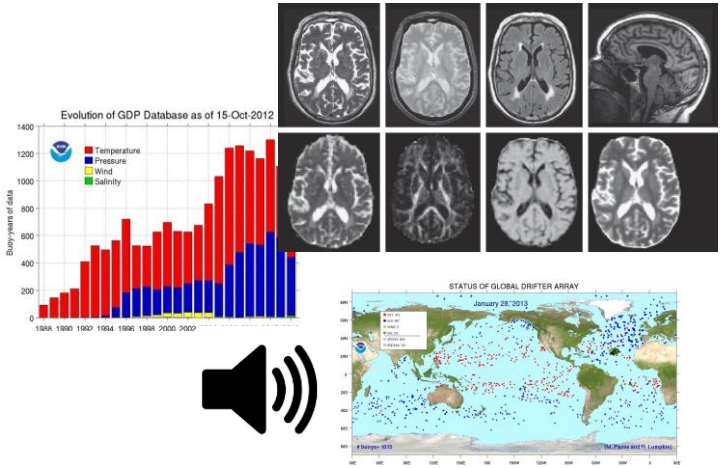
- **Observacional:** capturado en tiempo real, única e irremplazable e.g. *neuroimages, survey data.*
- **Experimental:** datos de experimentos, e.g. equipos de laboratorio, normalmente reproducible, pero con un alto coste e.g. *chromatograms, gene sequences*.
- **Simulación:** datos generados a partir de modelos de prueba, donde el modelo y los metadatos pueden tener más importancia que los datos de salida del modelo, por ejemplo, modelos económicos o climáticos.
- **Derivados o compilados**: resultado de combinar o procesar 'raw' data, a menudo pueden ser reproducibles pero con un alto coste e.g. compiled databases, text mining, aggregate census data.
- **Referencial:** conjunto o colección de pequeños datasets, (peer reviewed) datasets, e.g. gene sequence databanks, chemical structures.

La Universidad de Minessota los clasifica como:

- **Datos brutos o datos primarios:** información grabada como notas, imágenes, vídeos, encuestas, ficheros, etc.
- **Datos procesados:** análisis, descripciones y conclusiones con forma de informes o documentos.
- **Datos publicados:** información distribuida a la gente más allá de aquellos involucrados en la adquisición de los datos y la administración.

Los datos de investigación

Los datos de investigación pueden nacer en formato digital o convertirse a formato digital, pueden ser creados por el propio investigador o utilizar otros ya existentes. Pueden ser:



- Documents (text, Word), spreadsheets
- Laboratory notebooks, field notebooks, diaries
- Questionnaires, transcripts, codebooks
- Audiotapes, videotapes
- Photographs, films
- Test responses
- Slides, artefacts, specimens, samples
- Collection of digital objects acquired and generated during the process of research
- Data files
- Database contents (video, audio, text, images)
- Models, algorithms, scripts
- Contents of an application (input, output, logfiles for analysis software, simulation software, schemas)
- Methodologies and workflows
- Standard operating procedures and protocols

(Edinburgh University Data Library Research Data Management Handbook)

Los datos de investigación

Los datos de investigación pueden estar múltiples formatos:

- Text - flat text files, Word, Portable Document Format (PDF), Rich Text Format (RTF), Extensible Markup Languague (XML).
- Numerical - Statistical Package for the Social Sciences (SPSS), Stata, Excel.
- Multimedia - jpeg, tiff, dicom, mpeg, quicktime.
- Models - 3D, statistical. • Software - Java, C.
- Discipline specific - Flexible Image Transport System (FITS) in astronomy, Crystallographic Information File (CIF) in chemistry.
- Instrument specific - Olympus Confocal Microscope Data Format, Carl Zeiss Digital Microscopic Image Format (ZVI).

(Edinburgh University Data Library Research Data Management Handbook)

Los datos de investigación

También es importante gestionar durante y después de terminado el proyecto:

- Correspondence (electronic mail and paper-based correspondence)
- Project files
- Grant applications
- Ethics applications
- Technical reports
- Research reports
- Master lists
- Signed consent forms

(Edinburgh University Data Library Research Data Management Handbook)

Los datos de investigación: La Gestión

Research Data Management (RDM) se refiere a las mejores prácticas en la planificación, recopilación, almacenamiento, uso, distribución y conservación de los datos generados en cualquier proyecto de investigación (Universidad de Edimburgo).

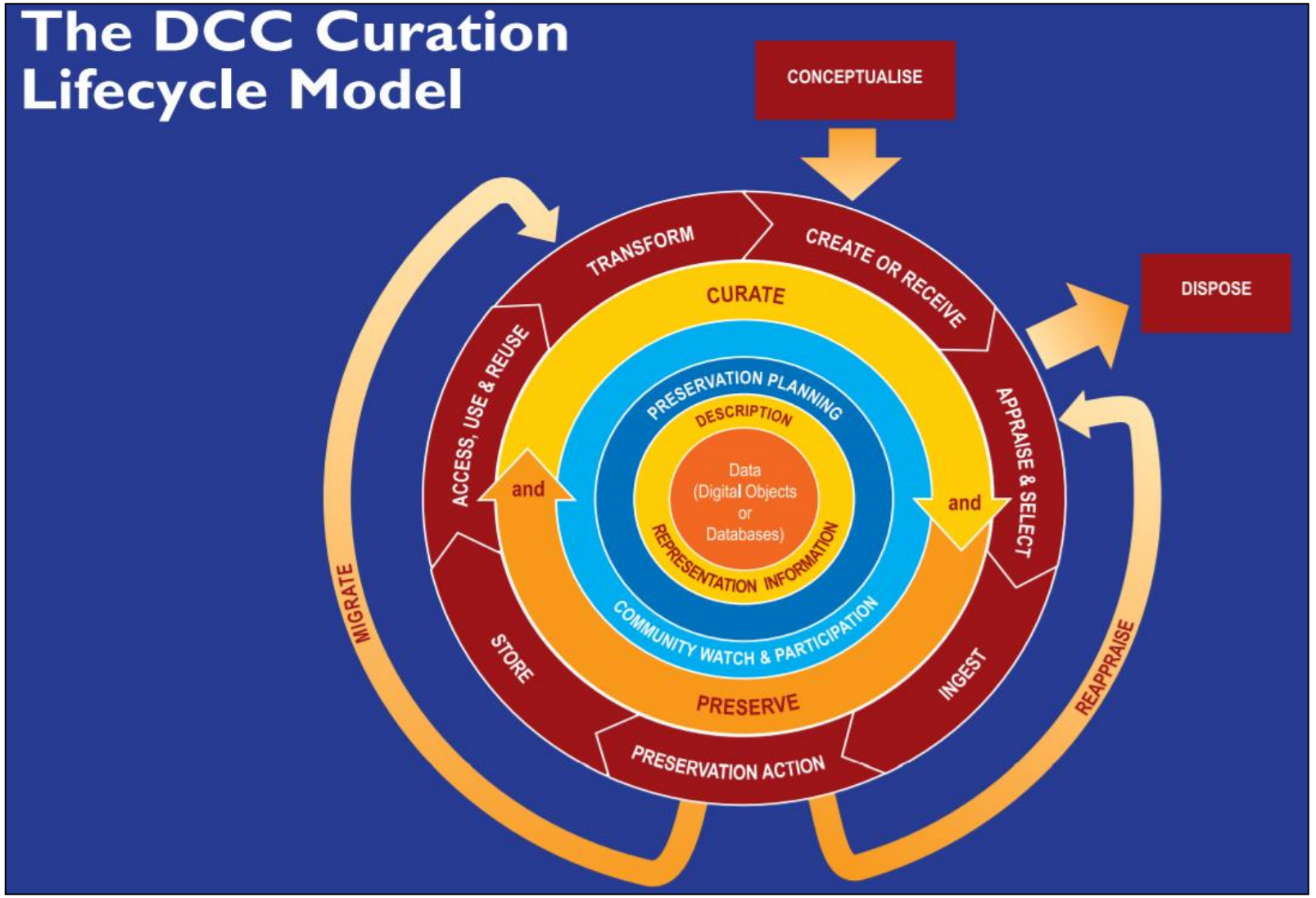
Ayuda a los investigadores a realizar una mejor investigación.

Cuando hablamos de gestión de datos de investigación hemos de tener presente cuál es el **ciclo de vida** de los datos y cuál el del proyecto de investigación asociado. Cada etapa requiere consideraciones, actividades y prácticas distintas con objeto de preservar, permitir el acceso y uso a los datos una vez completado el proyecto. Se trata de optimizar los datos que estarán accesibles para su reutilización.

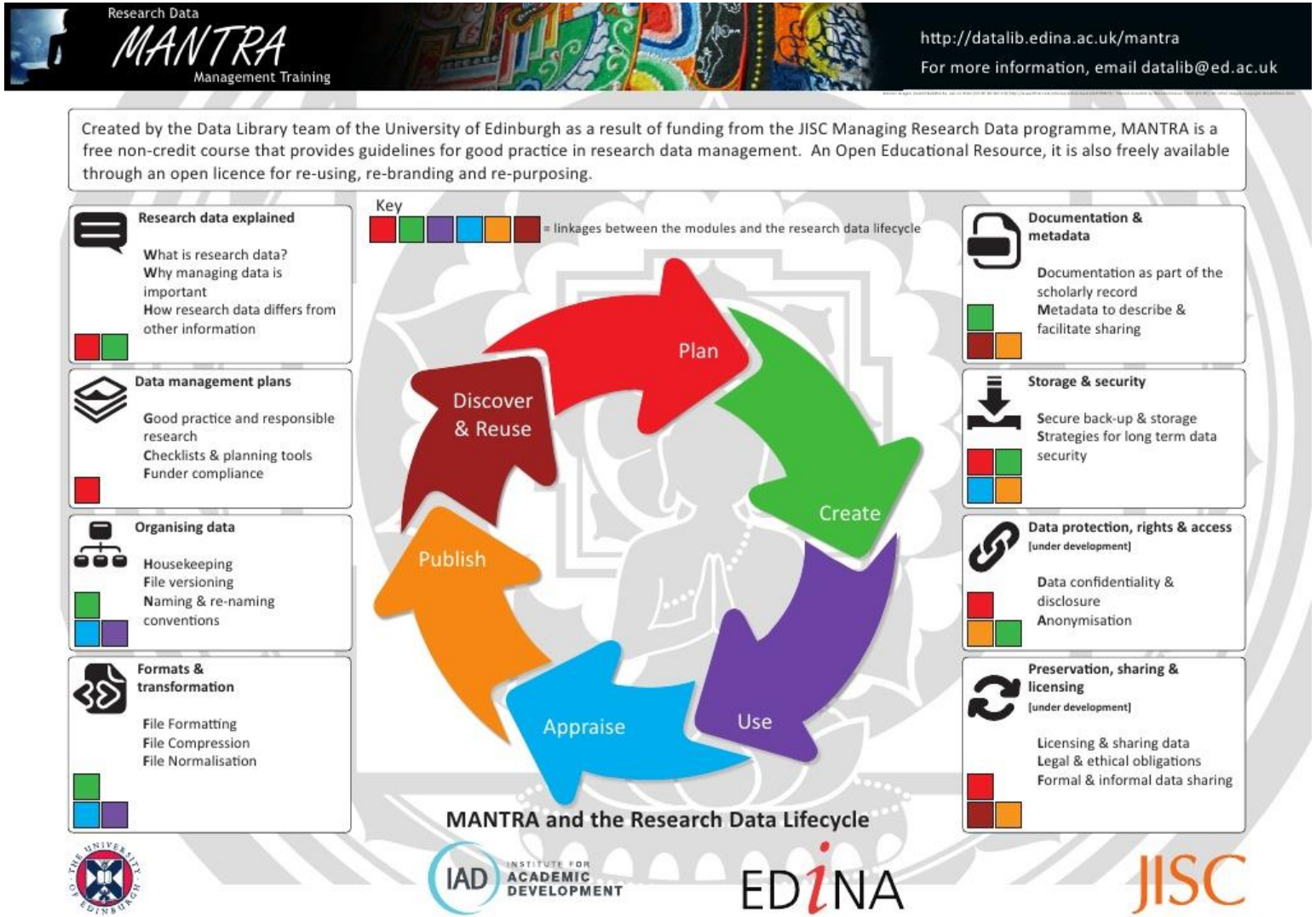
Las principales actividades de la gestión de datos son:

- Planificación
- Documentación (documentar los datos)
- Dar formato
- Almacenamiento
- Anonimización
- Control de acceso

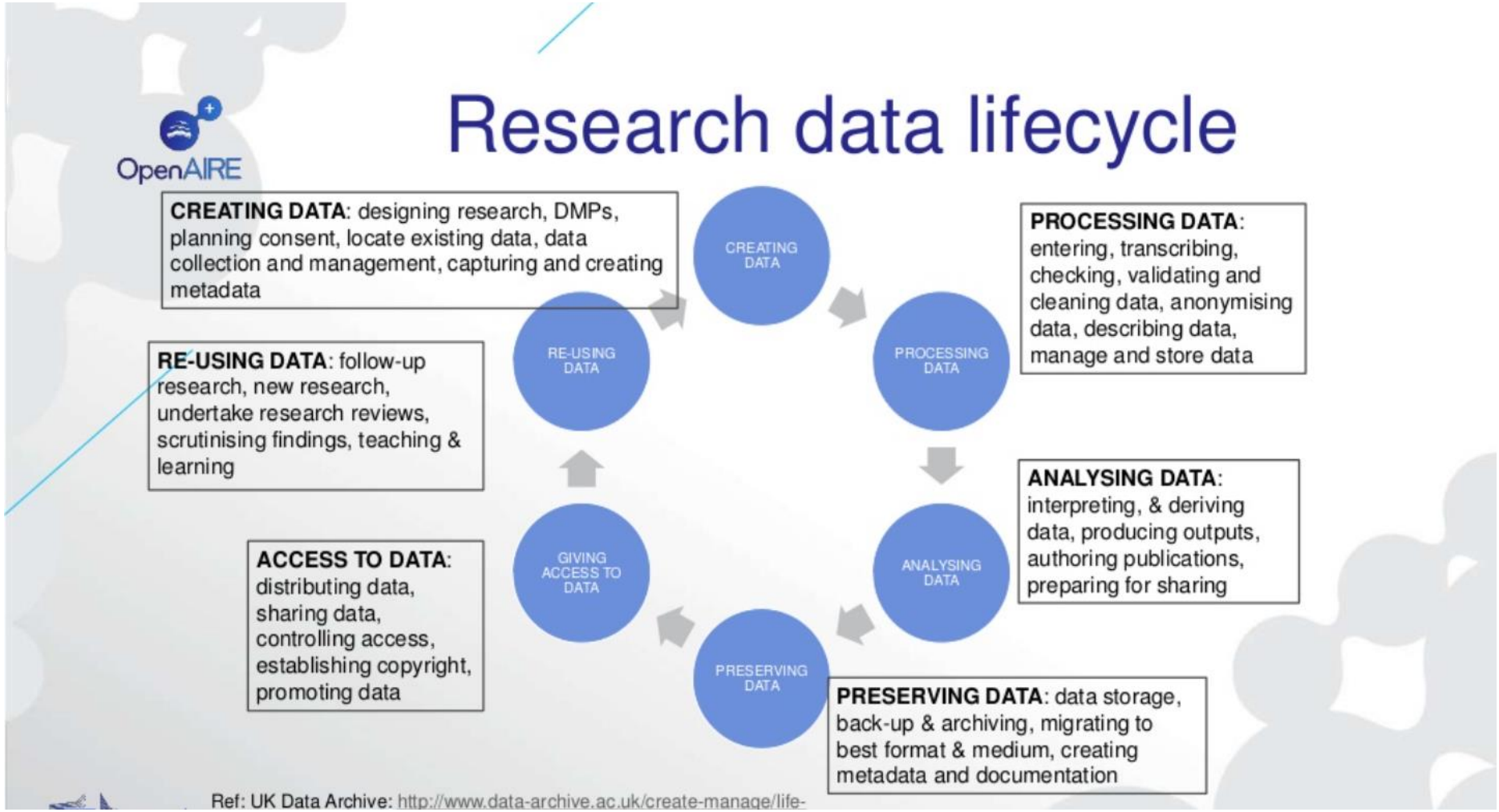
Los datos de investigación: Ciclo de Vida



Los datos de investigación: La Gestión



Los datos de investigación: Ciclo de Vida



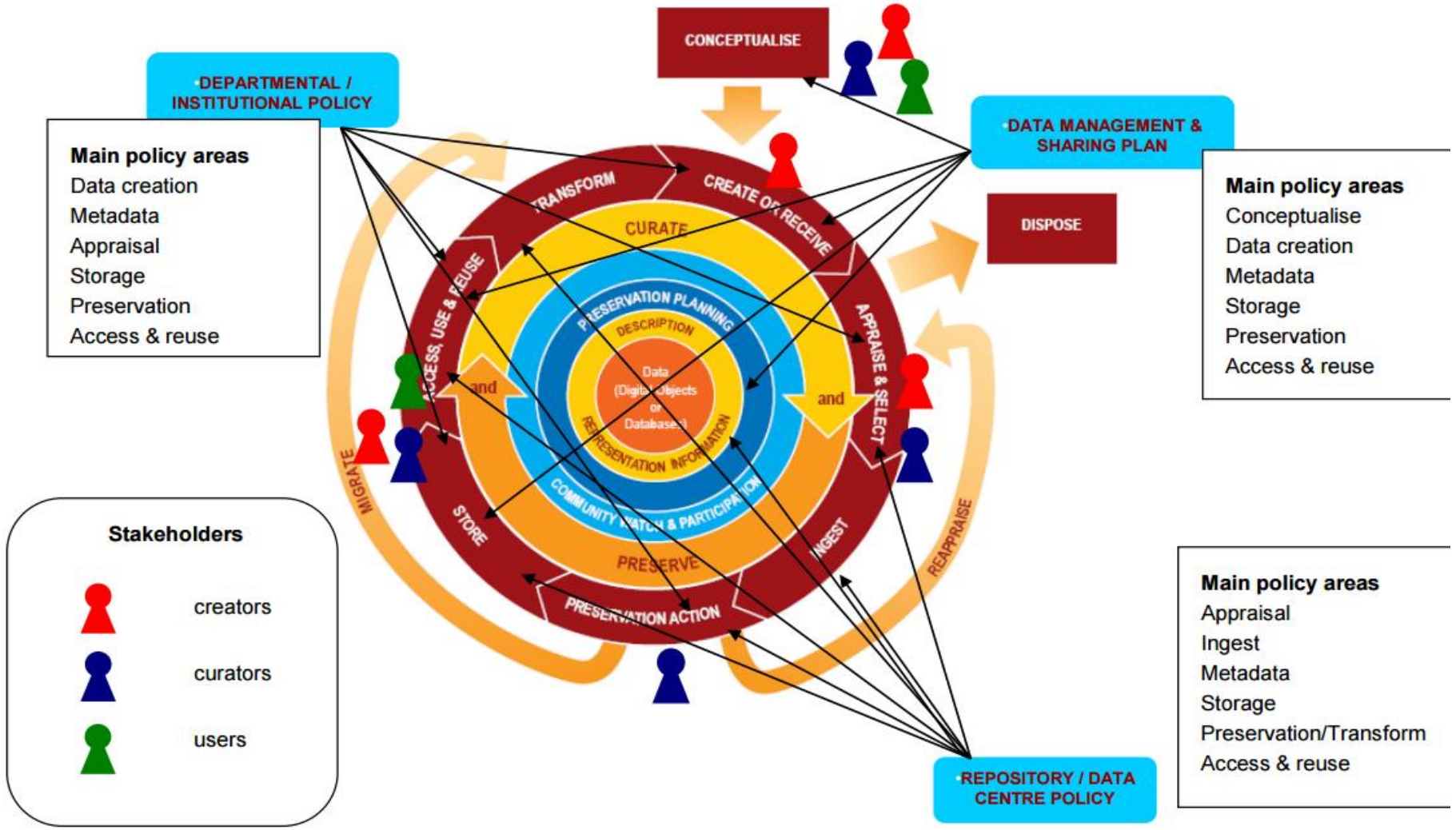
GROOTVELD, M. y ROSS-HELLAWER, T. “Open Research Data in H2020”. 2017.
https://www.slideshare.net/OpenAIRE_eu/20170530open-research-data-in-horizon-2020

Los datos de investigación: Actores/Stakeholders

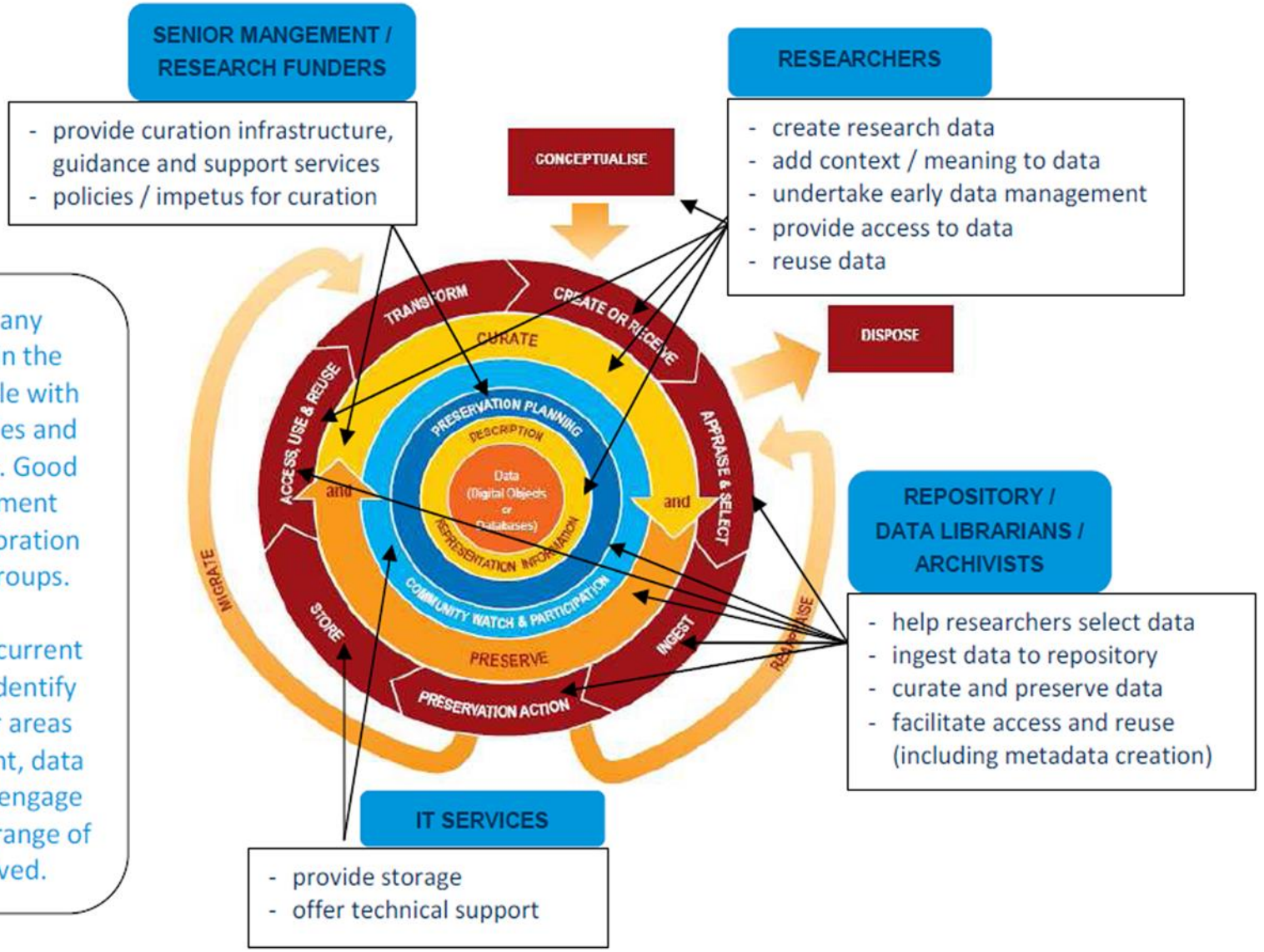
Por lo general se puede hablar de:

- INVESTIGADORES (INVESTIGADOR PRINCIPAL):
  - Crea y usa los datos
  - Diseña el estudio
  - Especifican los datos que se van a recoger
  - Determina como analizar los datos
  - Define conclusiones a partir de los análisis
- INSTITUCIÓN:
  - Define su política de gestión de datos
  - Proporciona recursos a los investigadores:
    - Formación en data management
    - Apoyo en la creación de planes de gestión de datos
    - Servicios de archivo
- REPOSITORIOS DE DATOS:
  - Tareas de curación de datos
  - Preservación a lo largo del tiempo
  - Proporciona acceso a los datos
  - Se debe trabajar en colaboración con los creadores/recolectores de datos para obtener información sobre embargos, copyright, acceso, restricciones, políticas de la agencia de financiación y seguridad
- USUARIOS SECUNDARIOS:
  - Verifican los resultados publicados
  - Realizan análisis secundarios
  - Formar
- AGENCIAS DE FINANCIACION:
  - Proporcionan los recursos
  - La financiación
  - Requiere a los investigadores gestionar los datos activa y adecuadamente
  - Requiere realizar planes de gestión de datos:
- EDITORES Y REVISTAS:
  - Difunden los descubrimientos
  - Mantienen la integridad de los registros científicos
  - Animan a los investigadores a citar los datos
  - Publican políticas de compartir datos ej.: PLoS (datos junto con los artículos)

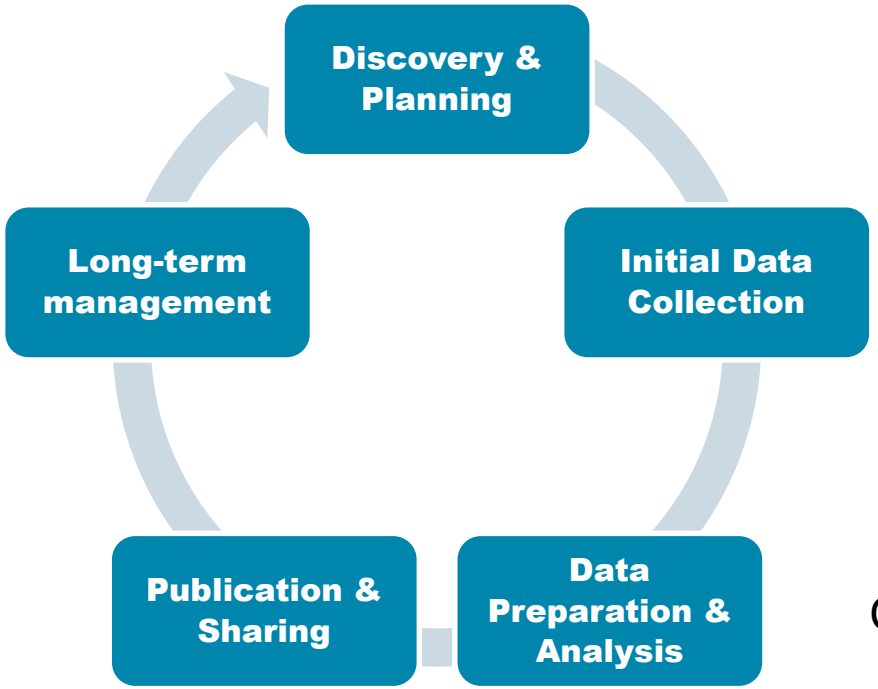
Los datos de investigación: Ciclo de Vida



Los datos de investigación: Ciclo de Vida



Los datos de investigación: La Gestión



Basado en Data Documentation Initiative – DDI (2008)

¿Por qué se gestionan los datos de investigación?

- Optimizar el uso de los datos durante la fase activa del proyecto de investigación.
- Colaborar y compartir datos con otros investigadores.
- Asegurar que los datos se preservan para futuros investigadores a la hora de descubrir, interpretar y reusar.
- Mantener el valor de los datos permitiendo a otros verificar y construir a partir de resultados publicados.

Otras razones (del UK Data Archive):

- Incrementar el impacto y visibilidad de la investigación.
- Maximiza la transparencia y rendición de cuentas.
- Nuevas colaboraciones entre usuarios y creadores de datos.
- Reduce los costes por duplicidad en la recolección de datos.

Los datos de investigación: Gestión a lo largo del Ciclo de Vida I

Las mejores prácticas según el ciclo de vida de la Data Documentation Initiative (2008)

- DESCUBRIMIENTO Y PLANIFICACION
  - Determina si el proyecto:
    - Producirá un nuevo dataset (recogida de nuevos datos)
    - Combina datasets existentes
    - Analiza datasets existentes
  - Tipo y formato de los datos
  - Temas éticos, privacidad, confidencialidad
  - Tipo de documentación, y estándares de metadatos
  - Identificación de usuarios potenciales: quién, qué, acceso y restricciones, útiles para análisis secundarios
  - Repositorio de datos adecuado
  - Costes de la gestión:
    - Documentación de los datos
    - Darles formato
    - Almacenamiento
    - Anonimización
    - Archivo

Los datos de investigación: Gestión a lo largo del Ciclo de Vida II

- RECOGIDA INICIAL DE DATOS (en esta fase se determinan los flujos de trabajo):
  - Organización de ficheros:
    - Denominación de ficheros
    - Versionado (evitar duplicar versiones y mejor sincronización)
    - Localización de documentos
    - Menor riesgo de perder contenido
  - Copias de seguridad y estrategias de almacenamiento
  - Protocolos que aseguren la calidad:
    - Documentar los procesos de comprobación y revisión de recogida de datos
    - Control de acceso y la seguridad de los datos
    - Informar al Servicio de Informática de temas de seguridad y necesidades
    - Especial cuidado con datos sensibles
- PREPARACION Y ANALISIS DE LOS DATOS
  - Los investigadores:
    - Procesar los datos en bruto
    - Documentar cambios
    - Crear una versión “master” de los datos a analizar
    - Documentar procedimientos de análisis:
      - Modificaciones adicionales
      - Modelo usado
      - Código empleado
      - Especificaciones de hardware y software

Los datos de investigación: Gestión a lo largo del Ciclo de Vida III

- PUBLICAR Y COMPARTIR

  En este estadio se consulta con el repositorio para determinar:

  - Preparar los ficheros de datos y otros materiales para reutilizarlos
  - Asegurarse que los datos cumplen con los requerimientos de los repositorios:
    - Ficheros con formatos apropiados
    - Cumplir las especificaciones del repositorio
  - Revisar la documentación y aplicar los metadatos para asegurar que se usen por terceras partes

- GESTION A LARGO PLAZO

  Cuando los investigadores comparten sus datos y descubrimientos, a través de las publicaciones y depositan los datos en el repositorio:

  - Los datos se depositan en repositorios de datos que:
    - Aseguran su integridad
    - Protegen ante la pérdida
    - Proporcionan acceso
  - Documentar cambios
  - Crear una versión “master” de los datos a analizar
  - Documentar procedimientos de análisis

Una gestión de datos efectiva tiene lugar a lo largo de todas las fases del ciclo de vida de la investigación, desde planificar el proyecto, la recogida de datos, preparar, analizar, publicar y compartir los datos a través de un repositorio.

Los datos de investigación: Resumen

Si los datos están:

- Bien organizados
- Documentados
- Preservados
- Accesibles
- Verificados en cuanto a exactitud y validez

Los resultados son:

- Datos de calidad alta
- Fáciles de compartir y reutilizar
- Citación y credibilidad al investigador
- Ahorro de costos para la ciencia

(DataONE. Education Modules)

Los datos de investigación: Buenas prácticas

Entre las buenas prácticas que identifica DataONE para gestionar los datos podemos mencionar:

- Asignación de nombres descriptivos a los ficheros
- Tareas de backup (copias de seguridad)
- Compatibilidad en la integración de datos
- Definición de tareas y roles en un equipo de gestión de datos
- Descripción de técnicas de medición de datos
- Tareas de control de calidad
- Identificación de software más adecuados para el proyecto
- Presupuesto de actividades
- Gestión de propiedad de los datos y selección de licencias de uso

(DataONE. Best practices)
https://www.dataone.org/all-best-practices

La Gestión de Datos de Investigación

Marisa Pérez Aliende
Universidad Autónoma de Madrid
mp.aliende@uam.es
Sevilla, 13 de Junio 2017

UNIVERSIDAD D SEVILLA
Biblioteca