PLAN DE DATOS DE LA INVESTIGACIÓN DEL CNRS

Noviembre de 2020

DFFGTRHGFDB BGBSDGF
3DFBFDB FDGDSFBGDFB
AF DFFGTRHGFDB BGBSDGF
DFGDFBFDB FDGDSFBGDFB
DFBVCXCVWERFDFBMG NM VC XCVCX
+20%
-15% +10%
19%
12%
WERGDBFNBMBN
MQWEW TW
RTXCV EWSFSDG VERG
TBVCB
5157170
875957
3458234
32458
342098
1098
342098
3451
25298
412-8079
1-362-570-6859

ÍNDICE DE CONTENIDOS

**1	Introducción**	4
**2	Principios generales**	6
**3	Objetivos**	7
**4	Política de datos de la investigación**	8
**5	Gobernanza de los datos de la investigación**	9
**6	Plan de acción**	10

1 | INTRODUCCIÓN

Este plan se inscribe en un contexto en el que la necesidad de acelerar el desarrollo hacia la Ciencia Abierta es patente. Por lo tanto, el conjunto de los datos resultantes de la investigación, en un sentido amplio tal y como están definidos en el Plan Nacional de la Ciencia Abierta[1] (PNSO), se incluyen en este plan.

No todos los datos de la investigación tienen por qué ser de acceso abierto o divulgados. Existen excepciones evidentes como los datos específicos de carácter confidencial, ya sea debido a su índole personal, debido a motivos de competencia industrial o para salvaguardar los intereses fundamentales o reglamentarios de los Estados. Abrir el acceso a los datos se entiende según la expresión de la comunidad europea «un (acceso) tan abierto como sea posible, cerrado en la medida de lo necesario». La decisión de la apertura o de protección de los datos de la investigación debe adoptarse con los servicios competentes del CNRS: los Servicios de asociaciones (partnerships) y valorización de la propiedad intelectual, la Delegación de la protección de datos para los datos de índole personal y la Dirección de la seguridad para las cuestiones relativas a la soberanía. De esta manera, este plan, así como las acciones que se proponen en el mismo, se refieren a los datos de acceso abierto, ya sean estos datos brutos o reprocesados en todos sus formatos: textos y documentos, softwares, algoritmos y «workflows». Este plan incluye el ecosistema de los datos, es decir, los aspectos relacionados con las infraestructuras digitales así como el conjunto de los servicios, en particular en el contexto de la aplicación de los principios FAIR[2] (Fáciles de encontrar, accesibles, interoperables y reutilizables).

La diversidad de los temas en torno a los datos de la investigación queda ilustrada por su ciclo de vida, constituido de 6 etapas:
- creación o recolección
- procesamiento o tratamiento
- análisis
- preservación
- puesta en común
- reutilización

Es importante entender cada una de estas etapas para poner en pie una gestión adecuada de los datos de la investigación a lo largo de todo el ciclo. Este ciclo incluye varios niveles de intervención (por ejemplo, la producción de los datos, incluidos los cómputos, el almacenamiento de los datos, la incorporación de los principios FAIR a los datos que implican su procesamiento y conservación).

El almacenamiento (transitorio o a largo plazo) es diferente de la noción de archivo. El almacenamiento se refiere a la persistencia de los datos, su identificación, indexación y la optimización de su acceso con miras a procesamientos frecuentes e intensivos. El archivo, en cambio, se refiere a la conservación de los datos por motivos legales o históricos. Las tecnologías de almacenamiento y de archivo son, en general, diferentes. Incluso si el material puede ser el mismo, el rigor y el carácter sistemático de los procedimientos no son iguales.

Existe una gran cantidad de maneras de caracterizar los datos de investigación, en función de su tipo, de su modo de producción, de las diferentes etapas de su transformación y de los diferentes modos de presentación al usuario. Entre los tipos de datos, se distinguen los datos digitales o simbólicos, los textos, las imágenes, los gráficos, los sonidos, etc. Los datos son el resultado de observaciones, de cómputos o de transformaciones diversas. Los datos pueden encontrarse en un estado bruto, curado, integrado o agregado. Esto incluye a los datos procesados y a los productos derivados de los datos, los datos incluidos en las publicaciones o que constituyen publicaciones específicas, denominados «data papers» en inglés.

Es necesario hacer que los datos de la investigación sean reutilizables e incluso compartibles para que otras personas, que no hayan originado su creación, puedan volverlos a utilizar.

Cabe notar una diferencia de definición según las diferentes comunidades en torno a los términos de «almacenamiento de los datos de la investigación» y de «repositorio de los datos de la investigación». Para algunos, depositar sus datos, por ejemplo, en el disco duro del laboratorio, no significa necesariamente que esos datos se hayan procesado o que se hayan creado los metadatos

asociados, lo que permitiría a otras personas reutilizarlos. Para otras comunidades (en particular la de la ciencia abierta), el repositorio de datos puede definirse como un servicio en línea destinado a gestionar la descripción del conjunto de los datos en vistas a su preservación y reutilización. Según esta definición, el depósito de los datos de la investigación no incluye el espacio físico o la infraestructura material que contiene los datos. En este documento, vamos a optar por esta definición.

El archivo y la curación de los datos y más ampliamente la gestión y puesta en común de los datos brindan diversas ventajas que no se recordarán aquí de manera exhaustiva. La puesta a disposición de los datos vinculados a una publicación científica es un elemento constitutivo de la comprensión y de la validación de resultados científicos, así como una base para la reproducibilidad de los procesos que han conducido a la obtención de dichos resultados. Por lo demás, volver a utilizar los datos existentes en vez de volver a hacer los cómputos o los experimentos que los han producido, permite un mejor uso de las inversiones públicas. Finalmente, nuevos conocimientos pueden surgir del cruce de datos provenientes de comunidades muy diferentes y pueden conducir a temáticas de investigación originales, a condición de ser compartidos en un contexto de investigación, lo cual implica un cierto nivel de calidad, de contextualización e incluso una validación por los pares.

Este plan se sustenta en las reflexiones que condujeron a la redacción, en enero de 2018, de un libro blanco sobre los datos en el CNRS[3]. Mediante un análisis detallado, instituto por instituto, se llegó a la conclusión de que era urgente promover una verdadera «cultura de los datos», dotar al CNRS de una estrategia robusta para responder a las necesidades de las comunidades en materia de plataformas para el análisis de datos a gran escala y de instaurar una política de gestión, de valorización y de perdurabilidad de los datos.

1. PNSO, July 4th2018 : https://cache.media.enseignementsup-recherche.gouv.fr/file/Recherche/50/1/SO_A4_2018_EN_01_leger_982501.pdf
2. ***Turning FAIR into reality, Final report and action plan from the European Commission expert group on FAIR data, European commission*** **(2018) :** https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf
3. ***White paper on data at the CNRS, overview of the situation and practices, Calculation and Data Mission, January 2018:***
http://www.cocin.cnrs.fr/IMG/pdf/livre_blanc_donne_es_2018.pdf

2 | PRINCIPIOS GENERALES

Con la ley digital de 2016 denominada Ley Lemaire[4], el marco legal de los datos de la investigación producidos en un contexto de financiamiento con fondos públicos ha evolucionado hacia un principio de puesta a disposición en datos abiertos, pero un número relativamente significativo de científicos todavía no tienen conocimiento de ello.

Más allá de las obligaciones legales, la hoja de ruta del CNRS para la ciencia abierta[5], adoptada en noviembre de 2019, a raíz de los objetivos determinados por Europa y el MESRI en materia de ciencia abierta, ya incluye un capítulo «datos de la investigación». Hace de la disponibilidad de los datos de investigación uno de los motores de los futuros avances del conocimiento.

La multiplicación exponencial del volumen y de la diversidad de los datos de la investigación, los desarrollos metodológicos del Big Data y las nuevas posibilidades de análisis que ofrece la Inteligencia Artificial (AI) implica que el CNRS se dote, en cooperación con sus socios, de una política y de una estrategia que incorporen los principios FAIR así como el principio de datos abiertos (salvo excepciones legítimas).

El CNRS, al ser uno de los principales actores europeos de producción de datos de investigación, debe desarrollar una estrategia y una política resuelta y clara, así como servicios de datos de calidad al servicio de la investigación. El enfoque propuesto forma parte de la iniciativa de European Open Science Cloud (EOSC)[6]. Naturalmente, corresponde a la promoción de una «EOSC francesa», que será una contribución y el punto de entrada lógico en la iniciativa europea. También está en consonancia con el nivel internacional a través de la participación del CNRS en instrumentos muy grandes, sistemas de observación e infraestructuras de datos.

El Plan de Datos de Investigación está impulsado ante todo por las necesidades de los científicos y tendrá en cuenta los contextos de las diferentes disciplinas. La articulación entre el trabajo realizado en los institutos y una gobernanza transversal de los datos es un punto clave y estratégico para el éxito de la aplicación del Plan de Datos de Investigación.

Esto incluye varios elementos:

- el trabajo de los institutos, que tienen en cuenta la diversidad de enfoques y comunidades y desarrollan estrategias nacionales para organizar sus comunidades en torno a los datos;

- la voluntad de hacer que las comunidades menos avanzadas se beneficien de los conocimientos y la experiencia de las más avanzadas sin penalizarlas;

- el «goteo» desde las comunidades cuyos datos son objeto de investigación hasta las comunidades de «usuarios».

El plan de acción deberá tener en cuenta la necesidad de cambiar las prácticas y las mentalidades, así como el desarrollo de herramientas adecuadas para la gestión, la puesta en común, la conservación a largo plazo y la difusión de los datos de investigación de acuerdo con los principios FAIR. La cuestión específica de los recursos, en particular los recursos humanos, así como un plan de formación, es parte integrante del Plan de Datos de la Investigación.

4. Digital Republic Law (2016):
www.legifrance.gouv.fr/jorf/id/JOR-FTEXT000033202746/
5. The CNRS Roadmap for Open Science: https://www.science-ouverte.cnrs.fr/wp-content/uploads/2019/11/CNRS_Roadmap_Open_Science_18nov2019.pdf
6. EOSC : https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

3 | OBJETIVOS

La indispensable evolución de las prácticas y las mentalidades es una cuestión de organización general de las actividades de investigación, a nivel individual, pero también y sobre todo a nivel colectivo. Las iniciativas en este sentido, llevadas a cabo por el CNRS, los institutos, las comunidades científicas o las infraestructuras de investigación, deben animar a todas las estructuras de investigación a adoptar «políticas de datos». El Plan de Datos de la Investigación del CNRS se basará, pues, en tres objetivos principales:

- ***Difundir una cultura de datos FAIR***

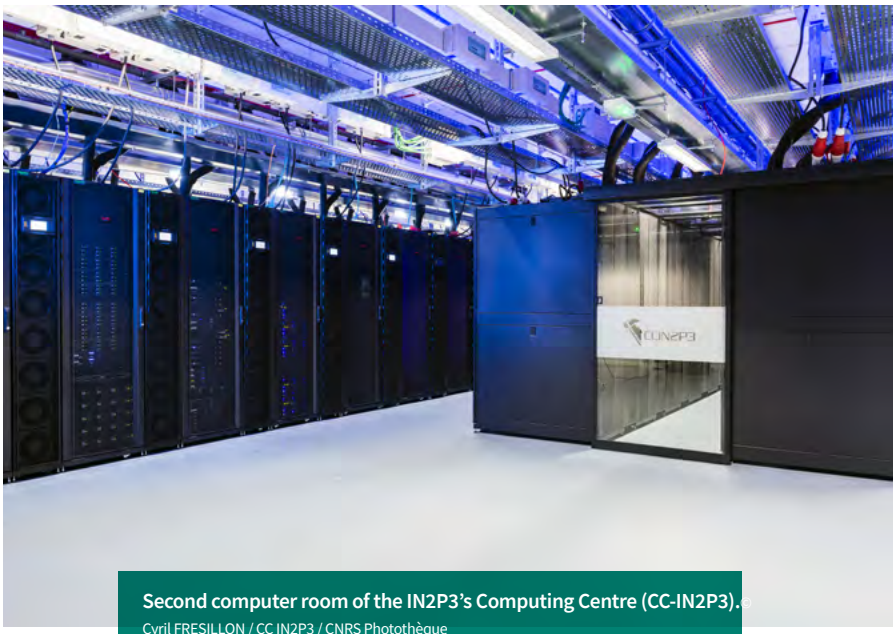
El objetivo general consiste en adoptar buenas prácticas con arreglo a los principios FAIR para acelerar el acceso y el cruce de datos y su uso/reutilización por parte de personas y máquinas, para contribuir a su fiabilidad (calidad y veracidad de los datos) y a la reproducibilidad de los resultados de la investigación. Ya desde el diseño de los proyectos de investigación, hay que tener en cuenta el ciclo de vida de los datos: desde su producción hasta su libre disponibilidad cuando sea posible (datos abiertos o restringidos), pasando por su almacenamiento, curación, análisis y modelización, y su archivo intermediario o incluso a largo plazo. Sin embargo, es evidente que la calidad y la reproducibilidad de los datos también dependen de otros factores que no se tienen en cuenta aquí, como la calidad de las fuentes o los procesos de análisis. Para acompañar esta difusión, hay que completar la oferta formativa, ya que la demanda es creciente.

- ***Dar a conocer los servicios y herramientas existentes***

Servicios, flujos de trabajo, repositorios de datos, portales de referencia y acceso, tecnologías, normas/estándares, etc. que facilitan a las comunidades el manejo de los datos de investigación a lo largo de su ciclo de vida y la aplicación de los principios FAIR. Dado que no todas las comunidades de investigación han alcanzado el mismo nivel de madurez en cuanto a hacerse cargo de sus datos y ponerlos a disposición de toda la comunidad, el Plan de Datos de Investigación del CNRS debe utilizarse como palanca siempre que sea posible.

- ***Apoyar la creación de nuevas prácticas, nuevos servicios y nuevas herramientas***

El archivo y la curación de los datos requieren una reflexión previa sobre qué datos conservar y cómo controlar su integridad, calidad y veracidad. El CNRS también podría apoyar y realizar experiencias piloto para la organización de plataformas distribuidas de almacenamiento, indexación y curación de datos, con un conjunto de servicios fundamentales.



Second computer room of the IN2P3's Computing Centre (CC-IN2P3).
Cyril FRESILLON / CC IN2P3 / CNRS Photothèque

4 | POLÍTICA DE DATOS

La política de datos del CNRS debe ser explícita y estar alineada con las necesidades de la investigación y los científicos. No todas las comunidades científicas están al mismo nivel de madurez. La puesta en común de datos ya está ampliamente desarrollada en algunas disciplinas; por ejemplo, los repositorios disciplinarios, a menudo internacionales, se han desarrollado desde hace varias décadas, especialmente en la astronomía. Otras comunidades aún no disponen de un sistema de tratamiento de sus datos, ni siquiera de una reflexión profunda sobre el tema.

En la actualidad, muchos datos se pierden irremediablemente con el paso del tiempo cuando podrían reutilizarse. El Plan de Datos de Investigación del CNRS debe basarse en los datos existentes siempre que sea posible. Desde 1972, la comunidad astronómica cuenta con un centro de puesta en común de datos, es el « Centre de données astronomiques de Strasbourg"[6]. Más recientemente, los SHS, a través de la TGIR Huma-Num[7], se han dotado de una herramienta de manejo de datos, desde su producción a su archivo a largo plazo y, dado el caso, a su difusión.

Teniendo en cuenta los principios FAIR y la necesidad de datos abiertos, la política de datos del CNRS debería incluir los siguientes puntos:

- Apoyar a las comunidades para animarlas a FAIRizar sus datos en una fase muy temprana, o al menos cuando estén a punto de almacenarlos o publicarlos. La cuestión de la gestión de datos no es sólo el volumen de almacenamiento necesario, sino también la forma en que se procesan los datos (curados, enriquecidos, estructurados) y se identifican con los servicios que permiten recuperarlos, reutilizarlos y compartirlos. La evaluación de las prácticas de investigación incorporará progresivamente criterios sobre la FAIRización de los datos.

- La recomendación de depositar los datos en un repositorio de acceso abierto. Esto es especialmente importante con los crecientes incentivos para depositar datos en repositorios gestionados por los editores de revistas, etc. Los repositorios no son necesariamente por disciplina, pero a la hora de considerar si son de confianza, se mira si sirven a su comunidad de destino (que puede ser más amplia que una disciplina) de acuerdo con las prácticas y expectativas de esa comunidad.

- El CNRS elaborará un directorio de repositorios de datos y servicios existentes para los investigadores, con el objetivo particular de avanzar hacia la certificación de los repositorios y servicios de datos.

Un repositorio debe tener un papel en la curación y preservación de los datos, y los principios FAIR son un objetivo en el contexto de la Ciencia Abierta. La certificación básica de CoreTrustSeal explicita los criterios de un repositorio «de confianza», lo que permite trabajar para mejorar las prácticas en función de los criterios, sin llegar necesariamente a presentar un expediente de certificación.

- La posibilidad de establecer períodos «razonables» de propiedad o embargo de datos (por ejemplo, de seis meses a dos años) teniendo en cuenta las prácticas de la disciplina.

- Fomentar la reutilización de los datos pertinentes, cuando existan, en lugar de la creación de nuevos datos.

La estrategia y la política del CNRS en materia de datos de investigación deben articularse con las del MESRI y sus socios del ESR. Por ello, el CNRS nombrará a un administrador o una administradora de datos, tal y como recomienda el PNSO, para que represente al CNRS en una red que está creando el MESRI.

El CNRS debe garantizar su participación en los foros nacionales, europeos e internacionales de debate sobre políticas de ciencia abierta, computación y datos de investigación. La participación del CNRS en la Alianza de Datos de Investigación (RDA) y en la Nube Europea de Ciencia Abierta (EOSC) forma parte de su política de datos de investigación, coordinada con el MESRI. El CNRS lidera el nodo francés de la RDA, con el apoyo financiero de MESRI. La EOSC incluye temas de computación, datos de investigación y ciencia abierta. A nivel internacional, se espera que el CNRS sea uno de los primeros miembros de la asociación EOSC a finales de 2020.

6. CDS : https://cdsweb.u-strasbg.fr/index-fr.gml
This is a revealing example: the CDS's employees are made up of 1/3 astronomers, 1/3 IT specialists and 1/3 data curators.
7. www.huma-num.fr

5 | GOBERNANZA DE LOS DATOS

El CNRS debe dotarse de medios para actuar de forma coordinada en toda la dinámica, desde la computación hasta las publicaciones, pasando por los datos masivos, la larga cola de datos, las infraestructuras de hardware y software, los servicios a los usuarios, la referenciación y la documentación. Cualquier división sería un obstáculo para la mutualización y la transversalidad de las acciones.

La gobernanza de los datos debe tener en cuenta los tres aspectos siguientes:

Científico:
La reutilización de datos y el intercambio de buenas prácticas entre institutos pueden dar lugar a nuevos descubrimientos e incluso a nuevos temas de investigación.

Técnico:
Con el objetivo de promover y aplicar la gestión de datos de investigación FAIR, estimular las redes profesionales y los grupos interredes y definir a las nuevas profesiones.

Económico:
Se necesitarán nuevos recursos mancomunados, sobre todo humanos. Será necesario evaluar las consecuencias sobre los costos (incluido el ahorro a largo plazo) y sobre las posibles repercusiones socioeconómicas de los datos FAIR.

Además, habrá que definir con precisión «tan abierto como sea posible, cerrado como sea necesario», en consulta con la Dirección de Asuntos Jurídicos, la Dirección General de Innovación, la Delegación de Protección de Datos, el Responsable de Seguridad de la Defensa y la Dirección de Seguridad.

Al servicio de numerosas comunidades científicas, la gobernanza de los datos en el CNRS tendrá que trabajar con los socios de la ESR que contribuyen a numerosas infraestructuras y servicios de producción de datos. Este vínculo podría ser garantizado por el administrador o administradora de datos mencionado anteriormente y que representa al CNRS en una red interinstitucional que está creando el MESRI de acuerdo con el Plan Nacional de Ciencia Abierta. También será necesario trabajar en conjunto con proyectos y organizaciones internacionales en coordinación con los institutos. Estos últimos deberán participar intensamente en la gobernanza de los datos.

Será necesario establecer un «modelo económico sostenible» para apoyar todos los componentes del ecosistema de datos FAIR: curación, suministro, almacenamiento e incluso archivo de datos a largo plazo. El archivo y los servicios de datos tienen costos financieros y humanos que deben cuantificarse y tenerse en cuenta, sobre todo más allá de la duración de los proyectos que generaron los datos. La federación, la mutualización de ciertos servicios de datos, debería permitir optimizar y reducir estos costos. La definición de un modelo económico es también un punto clave de la participación en la EOSC, ya que condicionará la participación de los actores. Otro aspecto del modelo económico se refiere al costo[8] que supondría no hacer los datos según los principios FAIR, incluso en el ámbito de la innovación.

El interés de una gobernanza única de los datos, única en el sentido interinstitucional, es proporcionar recursos adicionales para las acciones transversales. Por lo tanto, debe contar con recursos humanos y financieros dedicados, para su propia capacidad de actuación, y para reforzar las estructuras en las que se asignarían estos recursos (centros de datos nacionales o regionales), INIST[9], CCSD[10], …) y promover la puesta en común de recursos y conocimientos expertos, como hace hoy MICADO[11] en la computación, y la DIST[12] para la ciencia abierta. Los mecanismos disciplinarios de los institutos no tienen por qué cambiar su funcionamiento.

Una nueva Dirección funcional para los datos abiertos de investigación, la DDOR, adscrita a la Dirección General de Ciencia (DGDS), se encargará de proponer y apoyar la aplicación de una política y una estrategia de datos abiertos en el CNRS, integrando todas las dimensiones del tema. En general, la DDOR define y aplica la estrategia de la ciencia abierta, ampliada a todas las cuestiones relacionadas con

►►►

8. A study published by the European Community attempted to evaluate such costs: Cost-benefit analysis for FAIR research data Cost of not having FAIR research data - Study DOI: 10.2777/02999 : https://www.ouvrirlascience.fr/wp-content/uploads/2019/03/Cost-Benefit-analysis-for-FAIR-research-data_KI0219023ENN_en.pdf
9. CNRS Institute for Scientific and Technical Information: https://www.inist.fr/
10. Centre for Direct Scientific Communication: https://www.ccsd.cnrs.fr/
11. CNRS Computing–Data Mission
12. CNRS Scientific and Technical Information Department

los datos de investigación, incluidos los temas específicos de las infraestructuras digitales. El rumbo de la DDOR está marcado por la Hoja de Ruta de la Ciencia Abierta establecida en 2019 y por este Plan de Datos de Investigación del CNRS de octubre de 2020. Esta dirección, creada por la fusión de DIST y MICADO, proporcionará un marco para ayudar a tratar los temas relacionados con la apertura de las publicaciones científicas, la gestión y la puesta en común de datos de investigación, el problema de los datos masivos, su almacenamiento y las infraestructuras digitales. Los institutos del CNRS desempeñarán un papel destacado en el comité directivo de esta nueva Dirección. Este comité estará formado por un representante de cada instituto a nivel de toma de decisiones, principalmente los DAS (directores científicos adjuntos).

Además de este comité directivo y para tratar las cuestiones relacionadas con el cierre de datos y, en particular, la puesta en común de buenas prácticas que ayuden a diferenciar los datos abiertos de los que deben protegerse, se creará un grupo integrado por la dirección de la DDOR, el responsable de seguridad de la Defensa y representantes de la Dirección General de Innovación, la Delegación de Protección de Datos y la Dirección de Seguridad. Cuando sea necesario, se invitará a los miembros de este grupo a las reuniones del comité directivo.

6 | PLAN DE ACCIÓN

Aquí se esboza un plan de acción inmediata. No puede ser de ninguna manera exhaustivo en esta fase y tendrá que ser co-construido con los institutos. Una de las primeras tareas del comité directivo de la nueva dirección será revisar todas estas acciones, validarlas, proponer otras nuevas, priorizarlas y organizar su ejecución. Algunas de estas acciones están relacionadas con la adecuación de las normas y estándares (incluido el FAIR), otras con la política de incentivos de los institutos o con la práctica diaria de los investigadores.

Algunos puntos no se abordan deliberadamente en este texto, ya que necesitan más trabajo. Habría que estudiar un sistema de incentivos para animar a los investigadores a compartir sus datos. El repositorio de datos no es la única opción: las infraestructuras de federación o de publicación-suscripción también permiten compartir datos sin pasar por un repositorio. El incentivo para compartir debe ir acompañado de un incentivo para reutilizar los datos, con servicios de ingeniería adecuados para facilitar esta reutilización.

El plan de acción que aquí se propone es una ilustración concreta de los problemas inmediatos que se plantean en torno a la gestión de datos y que requieren un enfoque transversal, que no excluye tener en cuenta las especificidades de las prácticas disciplinarias y las políticas internas de los institutos.

1. Promover la aparición de buenas prácticas

Los principios FAIR son un hilo conductor que atraviesa las acciones y herramientas presentadas en esta sección, ya sea el Cat OPIDoR[13] implementado por el INIST, los planes de gestión de datos o los repositorios que integran enfoques de certificación. Los cuadernos de laboratorio son también una fuente para alcanzar los niveles de documentación exigidos a un enfoque FAIR.

Ayudar a las comunidades científicas a definir los elementos específicos de su gestión de datos.
Esto incluye la definición de los criterios de selección de datos para la FAIRización, teniendo en cuenta, entre otras cosas, el valor científico y el impacto en la investigación actual

y emergente, los principios utilizados por los archivistas y el hecho de que las observaciones pueden ser imposibles de reproducir. También es necesario apoyarlos en el desarrollo y mantenimiento de su marco disciplinario para compartir y FAIRizar los datos en un contexto internacional, especialmente a través de los institutos asociados a la gobernanza transversal del sistema. Apoyar a las comunidades significa ayudarlas a organizarse y estructurarse en torno a los datos y las prácticas de datos. Esta es una misión esencial de los institutos, que también aportan una visión nacional e internacional. El diálogo con las comunidades, la comprensión de sus necesidades y prácticas es, en efecto, el papel de los Institutos.

Cartografía de datos y herramientas y servicios para la gestión y el intercambio de datos.

Es esencial cartografiar los datos producidos por las unidades de investigación supervisadas por el CNRS. Actualmente se está elaborando mediante un cuestionario que la DIST está distribuyendo paulatinamente a los directores de las unidades y que permitirá identificar los lugares de producción de datos (infraestructuras y/o experimentos de laboratorio), los lugares de almacenamiento de datos, la cantidad de datos producidos, las políticas para compartir y conservar los datos dentro de las unidades, etc. También será necesario mapear la oferta existente de estructuras y servicios de gestión de datos y metadatos, tanto dentro del CNRS como de sus socios en el ESR. Esta acción está en curso en los repositorios y servicios de los que es responsable el CNRS. Este trabajo también debería permitir identificar, en una segunda fase, las necesidades de los investigadores de cada instituto.

Data Management Plan (DMP).

La obligación, en las convocatorias europeas para presentar proyectos y ahora en las del ANR[14], de elaborar un plan de gestión de datos para cualquier proyecto de investigación presentado es una buena oportunidad para animar a los investigadores a planificar las condiciones de almacenamiento, curación y difusión de sus datos a lo largo de su ciclo de vida. Es importante apoyarles en este proceso. El INIST contribuye a ello con la herramienta DMP OPIDoR[15], que facilita la redacción de estos planes, y con la organización de la «gira OPIDoR» en todas las regiones. Este apoyo debe basarse en las comunidades científicas y los componentes locales, que podrían ser coordinados por una red profesional del CNRS. La generalización y adopción de un plan de datos estandarizado para todos los proyectos (excluyendo ANR y Horizonte Europa) facilitaría la aplicación de los principios FAIR y la buena gestión de datos. El plan de gestión de datos evolucionará para facilitar su redacción utilizando todas las posibilidades tecnológicas que permitan la extracción y reutilización de datos de los diferentes sistemas de información. Se denominan PMD «automáticos», es decir, que se rellenan y utilizan automáticamente.

Identificadores perennes para conjuntos de datos de investigación.

Es esencial desarrollar el uso de identificadores perdurables para los datos, el software y las publicaciones. El INIST contribuye a esta acción proporcionando DOIs (Digital Object Identifiers) a los datos de investigación para el CNRS y, más ampliamente, para el ESR como agencia nacional de Datacite. Actualmente hay más de un centenar de usuarios -en el sentido de estructuras de investigación en el CNRS- y en el ESR, como el INRAE, el Ifremer, el ESRF (European synchrotron research facility), el CDS (Centre d'astronomie de Strasbourg), el Observatorio Midi Pyrénées, el CDSP (Centre de données socio-politiques), etc. Software Heritage ha hecho recientes propuestas de software[16].

Cuaderno electrónico de laboratorio (CLE).

Más allá de sus funciones relativas con la propiedad intelectual o la prueba de la integridad científica, el cuaderno de laboratorio puede constituir una herramienta de gestión de datos con vistas a su conservación y difusión, a través del propio cuaderno o de los enlaces que contiene. Es necesario posicionarse a favor del principio de adopción del cuaderno de laboratorio electrónico en el CNRS (en sustitución de los cuadernos de papel) y formular recomendaciones/obligaciones sobre las características de la herramienta adoptada, en términos de conservación, trazabilidad, perdurabilidad y accesibilidad de los datos y de la herramienta. Habrá que estudiar las soluciones elegidas por nuestros socios franceses (Inserm, Ifremer, Inrae, Cnes, etc.) a nivel europeo e internacional. Un grupo de trabajo dirigido por la Misión de pilotaje y relaciones con las delegaciones regionales y los institutos (MPR) está realizando esta labor en el CNRS.

Certificación de los sistemas de gestión de datos de investigación (en particular el CoreTrustSeal[17]).

La certificación de repositorios y servicios de datos, citada como objetivo en el Plan Nacional de Ciencia Abierta, garantiza que un centro de datos es «de confianza», examinando la forma en que implementa toda la cadena de datos, desde la ingestión hasta la difusión y la preservación. También puede entenderse en el contexto de las redes de centros de datos, por ejemplo, las de los Centros Temáticos de IR Data Terra[18], o las de la infraestructura europea CLARIN[19]. El CNRS podrá apoyarse en las actividades de apoyo para la certificación instauradas por el Nœud National RDA France[20].

13. Cat OPIDoR : https://opidor.fr/reperer/
14. 2. GENCI, the French Very Large-Scale Computing Centre, will soon be making a DMP compulsory for those requesting to use national computing resources.
15. DMP OPIDoR : https://dmp.opidor.fr/
16. Identifiers for software: https://www.softwareheritage.org/2020/05/26/citing-software-with-style/
17. *CoreTrustSeal : CoreTrustSeal Requirements* v02.00-2020-2022 (doi:10.5281/zenodo.3638211)
18. Data Terra : https://www.data-terra.org/
19. CLARIN : https://office.clarin.eu/v/CE-2013-0095-B_checklist-v7_3_1.pdf
20. RDA France : https://www.rd-alliance.org/groups/rda-france

2. Fomentar la aparición de nuevas herramientas

Algunas comunidades producen datos sin un DMP (como en el caso de experimentos de laboratorio) y no tienen una «tradición» de cuadernos de laboratorio. Estos datos, que a menudo están mal referenciados, suelen perderse definitivamente después de las publicaciones correspondientes, cuando el doctorando que los produjo se va o cuando cambia el equipo de almacenamiento. En términos más prosaicos, actualmente no existe una oferta de almacenamiento genérico para los laboratorios, y mucho menos de referenciación o FAIRización. Si un laboratorio se pregunta «¿dónde puedo almacenar mis datos?», la mayoría de las veces no obtiene respuesta y acaba equipándose o, en el mejor de los casos, se le ofrecen soluciones específicas ad hoc. ¿Cómo podemos iniciar y fomentar un proceso de gestión de estos datos?
Una herramienta electrónica, el «directorio de investigación» del laboratorio o el cuaderno de laboratorio informatizado, podría proporcionar un enlace con el servidor que aloja los datos, fomentar el uso de formatos de datos interoperables e imponer una documentación mínima de los datos con vistas a su posible difusión. Dicha herramienta debe ser sencilla, interoperable (y barata) y poder servir, si es necesario, como base para un futuro DMP o para facilitar un depósito en un repositorio al final de la investigación.

Una alternativa podría ser un repositorio institucional del CNRS para los numerosos datos denominados de « cola larga», que suelen ser muy diversos y de pequeño volumen, pero cuyo volumen y diversidad crecen rápidamente con el desarrollo de las técnicas de imagen y los sensores distribuidos. La cuestión de este repositorio debería plantearse en coordinación con la actual iniciativa del MESRI, que está estudiando la viabilidad de un repositorio nacional de datos «sencillos» a través de un grupo de trabajo dirigido por J.C. Desconnets del IRD. Por su parte, el INRAE, el IRD, el CIRAD, etc. ya han optado por la solución de un repositorio institucional, mientras que otras instituciones han hecho otras elecciones; el objetivo es buscar la coherencia y la flexibilidad nacional para simplificar el trabajo de depósito, referenciación y puesta a disposición de los datos, respetando las diferentes prácticas de investigación.

Este repositorio genérico de datos de investigación de cola larga, si se creara, tendría que complementar los sistemas existentes, ya sean infraestructuras de investigación, las más grandes con capacidad para desarrollar sus propias herramientas en un marco nacional y/o internacional, o infraestructuras de datos que se encarguen de la FAIRización de los datos, a menudo temáticos, y teniendo en cuenta los sistemas creados localmente por las universidades o a nivel regional. Es necesario asociar estrechamente la computación, el almacenamiento y el procesamiento de los datos, especialmente cuando se producen volúmenes muy grandes. Los centros nacionales pueden desempeñar un papel privilegiado, como ya es el caso del CC-IN2P3[21] que es un centro temático dedicado a la comunidad IN2P3. También son capaces de ofrecer recursos y servicios informáticos y de datos con las capacidades necesarias para un procesamiento y análisis cada vez más pesados.

Algunos TGIR archivan datos a largo plazo y los distribuyen tras un periodo de embargo, sobre todo en astronomía y ciencias de la Tierra. Otros no almacenan ni curan los datos. En la mayoría de los casos queda por desarrollar la posibilidad de poner los datos de la investigación a disposición de los científicos que no utilizan los instrumentos y sistemas de observación que produjeron los datos, en particular para las aplicaciones inter y/o transdisciplinarias.

La referenciación es un punto complementario a las estrategias de almacenamiento. Se trata de disponer de una plataforma de software de servicios que facilite la visibilidad y el acceso a los conjuntos de datos a través de herramientas con capacidades de búsqueda fáciles de usar. La calidad y relevancia de los metadatos es fundamental para garantizar una buena accesibilidad a los datos.

3. RH - Evaluación y gestión de las carreras

El CNRS debe desarrollar una respuesta coordinada a las nuevas necesidades en términos de conocimientos técnicos, recursos humanos y reconocimiento de estas nuevas actividades transdisciplinarias sobre los datos de investigación. También sería conveniente crear un sistema de valores que anime a los investigadores a publicar sus datos.

Hacer un inventario de todas las funciones que desempeña el personal en la FAIRización de los datos y de las competencias requeridas en estas funciones, con el fin de alimentar y desarrollar los marcos de evaluación de los investigadores y los marcos de trabajo informáticos (Referens[22]). Esta acción conducirá a identificar nuevos perfiles y nuevas profesiones.

Dirigir la contratación gracias a los nuevos perfiles identificados y así beneficiarse de recursos humanos adicionales para apoyar la gestión y el intercambio de datos. Los recursos humanos son principalmente informáticos (administradores de datos), pero también pueden ser investigadores (científicos de datos). La necesidad inmediata de ingenieros es muy fuerte. Puede facilitar el intercambio de experiencias y esfuerzos.

21. The CNRS Institute of Nuclear and Particle Physics Computing Centre: https://cc.in2p3.fr/
22. Referencing list of Jobs in Research and Higher Education (REFERENS) : https://data.enseignementsup-recherche.gouv.fr/pages/referens/

Acompañar un cambio profundo en la evaluación de los científicos. Será necesario actuar para que las inversiones en apoyo de la FAIRización y la puesta en común de datos se tengan en cuenta en la evaluación. Se trata de tareas de interés colectivo como compartir datos en modo FAIR, la participación en la definición de prácticas y normas a nivel disciplinario e interdisciplinario, la participación en su aplicación como experto científico, la participación en las tareas científicas necesarias para compartir datos en las infraestructuras de investigación y las infraestructuras de datos. A nivel individual, debe valorarse la producción y puesta en común (cuando sea posible) de datos y, más en general, de objetos digitales (software) FAIR, en los proyectos de investigación y en la práctica, la práctica del DMP, la reutilización de los datos en cuanto permita nuevos avances en el conocimiento.

23. DoRANum : https://doranum.fr/
24. CNRS Mission for Transversal and Interdisciplinary Initiatives

4. Formación

La formación y la comunicación son elementos esenciales de este plan de acción. El primer paso es desarrollar competencias en materia de datos FAIR, tanto en lo que se refiere a la gestión de datos (principios FAIR de gestión, puesta en común, archivo intermedio y permanente) como a la ciencia de los datos (manejo, tratamiento y análisis de los datos de investigación). El INIST participa en esta acción con la formación a distancia a través de la plataforma DoRANum[23] y en persona, sobre la gestión de datos, el intercambio, el archivo y el plan de gestión de datos.

La cultura de los datos debe desarrollarse a través de la formación continua de todo tipo de personal para sensibilizarlo sobre el tema, hacer que conozca y aprenda a utilizar las normas y, en general, las distintas herramientas relacionadas con los datos (DMP, repositorios, lenguajes, etc.).

Es importante identificar y compartir los éxitos, para convencer a las comunidades que aún no están muy involucradas de que compartan sus datos. Su concienciación se basará en el intercambio de buenas prácticas y lecciones aprendidas de la aplicación de FAIR y la definición de los límites legítimos de los datos abiertos. Un aspecto es la recopilación de ejemplos concretos y guías de aplicación. Es necesario adoptar el mismo tipo de medidas para la aplicación de los DMP. Estos intercambios pueden tener lugar en el marco de ámbitos disciplinarios (institutos) o a través de redes profesionales (MITI[24]), y animar al personal implicado en el intercambio de datos a participar en la RDA.

© AdobeStock/ra2studio

CNRS
3, rue Michel-Ange
75794 Paris Cedex 16
+331 44 96 40 00
www.cnrs.fr/en

Print : CNRS IFSeM
2022