



Eje Temático 3. Herramientas para el desarrollo de la Ciencia Abierta en las bibliotecas y archivos

“LET’S MAKE DATA COUNT”: LOS DATOS COMO PRODUCTO DE LA INVESTIGACIÓN VITAL PARA LA CIENCIA ABIERTA

Mary Hirsch
DataCite

Resumen por: Miquel Àngel Plaza-Navas (URICI-CSIC)



Isabel Bernal, de la Oficina Técnica de Digital.CSIC, es la encargada de presentar esta ponencia. Indica que DataCite, la organización a la que pertenece Mary Hirsch, fue la primera en asignar DOIs a los datos que se obtienen durante el proceso de las investigaciones (*datasets*) y que, hasta fechas recientes, no habían sido tomados en consideración. Durante estos últimos años, DataCite ha venido trabajando en cómo realizar el seguimiento de los *datasets*, facilitar su búsqueda, su descripción y la medición de su impacto. Recuerda que

el CSIC, a través de URICI, es miembro de DataCite desde hace ya tres años, gracias a lo cual se asignan DOIs a los *datasets* que se depositan en el repositorio institucional Digital.CSIC. En esta ponencia, Hirsch presentará las últimas iniciativas de DataCite relacionadas con la medición del impacto de los *datasets*.

Mary Hirsch comienza diciendo que, en la actualidad, dentro del marco de la Ciencia Abierta se está dando cada día más importancia a resultados de investigación considerados como no tradicionales, como es el caso de los *datasets*. Es necesario disponer de la infraestructura necesaria para que puedan ser claramente identificados, atribuidos, recuperados y reutilizados. Explicará lo que DataCite está realizando en los siguientes cuatro aspectos: FAIR DATA, DATA USAGE, DATA CITATION y PID GRAPH.

Antes de ello realiza una breve descripción de DataCite. Se trata de una agencia de registro de DOIs y metadatos, una organización sin ánimo de lucro, con un equipo de trabajo de 9 personas y que cuenta con unos 180 miembros. Pueden asignar DOIs a metadatos y objetos, como *datasets*, que están depositados en sus respectivos repositorios. DataCite colabora también con otras organizaciones que asignan DOIs y con grandes infraestructuras, como la Nube Europea de la Ciencia Abierta (EOSC).

1) FAIR DATA.

Hay tres importantes retos para poder aprovechar los *datasets*. En primer lugar, hay que compartir esos datos de manera FAIR (*Findable, Accessible, Interoperable & Reusable*), crear vínculos entre ellos y otros objetos y, finalmente, construir servicios que faciliten su localización. En definitiva, se trata de proporcionar un marco para que los investigadores gestionen el resultado de sus propias investigaciones de una manera

eficaz que les permita encontrarlos y reutilizarlos fácilmente. Los DOIs y metadatos son un instrumento que permite que los *datasets* puedan cumplir con los principios FAIR. Una vez se consigue que estén en modo FAIR es factible pasar a comprobar cuál es su uso.

2) DATA USAGE

Hirsch habla del MAKE DATA COUNT, un proyecto de 3 años que finalizó en el 2019, con el objetivo de conseguir que los datos sean considerados como un producto de la investigación vital para la Ciencia Abierta. En colaboración con COUNTER (que provee normas para la contabilización del uso de los recursos electrónicos), el proyecto MAKE DATA COUNT ha desarrollado un “Código de Práctica para Métricas de Uso de Datos de Investigación.”

Hirsch lo describe como un paso importante en la contabilización del uso de los *datasets* y, además, útil para comprender cómo se están reutilizando. De esta forma, los repositorios también pueden compartir esta información (que en DataCite denominan “Vistas y Descargas”). De manera simplificada consiste en preparar informes siguiendo los requisitos de COUNTER, enviarlos a través de la DataCite Usage Reports API y, mostrarlos en Event Data. Lógicamente requiere que los repositorios dispongan de una infraestructura técnica adecuada para realizar este trabajo. Desde abril de 2019 las “Vistas y Descargas” se pueden ver en DataCite Search (portal para buscar DOIs y metadatos) cuando los repositorios origen permiten compartir esa información. DataCite también está desarrollando un *widget* para facilitar la inclusión de esa información en la página web de cualquier repositorio y así comprender mejor el esfuerzo que realizan estos en la tarea de compartir sus *datasets*.

3) DATA CITATION

Hirsch menciona que cuando se normalicen las citas de *datasets* en las publicaciones se facilitará su identificación y reutilización, y permitirá que los investigadores puedan obtener reconocimiento por haberlos compartido. Es de esperar que cada vez sea algo más normal, pero en la actualidad es todavía difícil encontrar publicaciones que contengan este tipo de citas.

DataCite, en colaboración con Crossref, ha trabajado en facilitar la conexión de los objetos a nivel de metadatos, vinculando los *datasets* con la publicación. En el momento de incluir un *dataset* en el repositorio, los investigadores deben indicar cuál es la publicación relacionada con esos datos. Y, en el repositorio, cuando se le va asignar el DOI a ese *dataset* hay que asignar en los metadatos cuál es el DOI de la publicación relacionada.

Scholix, es una iniciativa internacional que intenta establecer un marco para el intercambio de información sobre conexiones entre publicaciones y *datasets*, y ofrece recomendaciones a los repositorios sobre los metadatos necesarios para conseguir esa vinculación. Los metadatos indispensables son: *RelatedIdentifier*, *RelatedIdentifierType* y *RelationType*. Existen muchos tipos de *RelationType* de manera que se requiere una buena planificación. En el portal DataCite Search se mostrará esta información como *Citation*, *Reference* o *Relation* dependiendo del *RelationType* que se establezca. Esta información está recogida en Event Data, un servicio desarrollado por DataCite en colaboración también con Crossref que recoge los *Linking events* (relaciones entre dos DOIs) y los *Usage events* (vistas y descargas del contenido relacionado con el DOI)

Como resumen, Hirsch indica que toda esta infraestructura que ha explicado en los tres puntos anteriores permite el desarrollo de nuevas prácticas en la Ciencia Abierta con el objetivo de dar acceso a métricas normalizadas abiertas y transparentes.

4) PID GRAPH

En la última parte de la ponencia, Hirsch presenta PID Graph, un concepto en el que DataCite está trabajando. Se centra en destacar cómo se está trabajando (proyecto FREYA coordinado con EOSC) para que los Identificadores Persistentes (PIDs) de diferentes tipos (DOIs, ORCIDs, etc.) sean un componente central en la Ciencia Abierta. Si los investigadores, las instituciones, las publicaciones, los *datasets*, etc. estuvieran todos conectados a través de esos diferentes PIDs, se podrían realizar gráficos de gran interés. DataCite ha desarrollado *GraphQL API* para obtener ese tipo de gráficos

Al finalizar hubo un breve turno de preguntas. El investigador Emilio Cano (CENIM) comenta que para conseguir el objetivo de compartir los *datasets* de manera abierta es necesario que toda la infraestructura explicada en la ponencia esté mucho más extendida. Y, sobre el posible mal uso de los *datasets* (por ejemplo, no reconocer adecuadamente el origen de esos datos) pregunta si hay mecanismos para poder evitarlo. Hirsch contesta que es lógico esa preocupación, pero que DataCite lo que recopila son los metadatos, pero no controla el uso que se realice de esos datos. Isabel Bernal, de Digital.CSIC, interviene diciendo que es un tema en el que hay que avanzar y en el que se requiere la colaboración de todos (investigadores, repositorios, organizaciones como DataCite) para hacer que los datos, a través de metadatos bien usados, cada vez sean más buscables, reutilizables, citables, etc. y, por tanto, más complicado el hacer alguna infracción. En la actualidad ya se está avanzando mucho en la parte técnica de los metadatos y PIDs, lo que es un avance en contra de ese posible mal uso.

Isabel Bernal, finalmente, pregunta a Hirsch si DataCite ha realizado algún tipo de análisis con el contenido que ya existe en DataCite Search para ver si se cita más de *dataset* a *dataset*, de *dataset* a publicación o de publicación a *dataset*, es decir, para ver si se está cambiando la práctica cultural de que los autores empiecen a citar *datasets*. Hirsch le responde que, hasta el momento sí que han visto que hay muchos vínculos de *dataset* a *dataset*, pero todavía no hay ningún análisis global sobre todos los tipos de relaciones.



Enlace a la [presentación y grabación de la ponencia](#).