Open Science with Open Source Software, Open Data and Open Access

Nuria Rico Castro *
Departamento de Estadística e Investigación Operativa
Universidad de Granada
nrico@ugr.es
Juan Julián Merelo Gervós
Departamento de Arquitectura y Tecnología de Computadores
Universidad de Granada
jjmerelo@ugr.es

Ángel Pablo Hinojosa
OpenXXI
E-mail Autor 3

1 Introducción

Si hay una actividad científica que actualmente se apoya en procedimientos estadísticos por encima del resto, esta es la investigación. La investigación, entendiéndola en un sentido amplio, se lleva a cabo en el seno instituciones especializadas pero también está presente en las experiencias más sencilla y particulares que proporcionan pequeños descubrimientos. La metodología, predominantemente expermiental, hace que el avance científico se impulse fundamentalmente gracias al uso de estándares, muchos de los cuales responden a análisis estadísticos más o menos complejos que permiten una visión clara de qué procesos están ocurriendo. El resumen y la visualización de la información recogida para la investigación está presente en los estándares habituales, donde el uso de herramientas estadísticas es prácticamente imposible de soslayar. La estadística es, pues, una herramienta clave en el proceso investigador en prácticamente cualquier área de conocimiento y permite hacer llegar los resultados a la sociedad.

Sin embargo, en el momento actual se está produciendo una transformación del paradigma en que se realiza el avance científico, encaminando la producción científica hacia un punto donde sean posibles la divulgación y transmisión de conocimiento de forma completa. No basta ya con mostrar el resultado final de la investigación en comunidades especializadas o en instituciones reconocidas. Actualmente se demanda que se establezcan cauces que propicien la autonomía a los individuos para la adquisición de conocimiento a través de la posibilidad de replicar la experimentación.

*Corresponding Author.

Esto no quiere decir que los resultados dirigidos a los círculos especializados no sean relevantes. La demanda consiste en completar el resultado final, de forma que todo el proceso investigador pueda ser compartido, desde los datos que se recogen hasta la conclusión final que de ellos se deriven. Esta es la idea subyacente de la *Ciencia Abierta*: hacer llegar a cualquiera que pueda estar interesado, el conjunto de datos, herramientas y conclusiones que hacen posible el avance científico.

Hacer ciencia abierta requiere el uso de estándares abiertos que permitan la replicación, reutilización y colaboración entre comunidades. La ciencia abierta, pues, entendida como ciencia a disposición de todos, necesita de herramientas sólidas al alcance de los investigadores: herramientas contrastables, revisables y compartibles. Para hacer ciencia abierta, se hace necesario el uso de software libre (*open source software*), la liberación de datos de la investigación (*open data*) y la publicación de resultados en artículos de acceso abierto (*open access*). Estos tres pilares son los que garantizan que pueda existir esa transmisión completa de conocimiento.

2 El software libre

Desde los años 80, el software libre y la filosofía asociada a él se han mostrado como uno de los principales impulsores de internet y las tecnologías asociadas a esta. Se llama software libre (free software) o software de fuentes abiertas (open source software) a aquel programa que se distribuye con una licencia que autoriza expresamente su copia, distribución, análisis, estudio, modificación y uso sin limitaciones. Adicionalmente, este software debe ser distribuido junto con los componentes necesarios para que los derechos permitidos por la mencionada licencia puedan ser ejercidos efectivamente, como suele ser el caso del código fuente del programa en cuestión.

En un principio, el desarrollo del software libre se extendió principalmente en el ámbito académico, dado que la mentalidad subyacente a este modo de distribución encaja muy bien con la filosofía de transparencia, apertura y crítica implícita en la publicación científica. A pesar de estos inicios, el mundo empresarial no tardó en ver el potencial comercial del software libre y a utilizarlo y producirlo. El uso de software libre posee para las empresas ventajas como menores costes de producción e implementación, mayor adaptabilidad y facilidad de soporte, o una mayor posibilidad de estandarización. Además, en muchos casos, la distribución del software usando licencias libres aporta ventajas estratégicas y competitivas, como muestra el ejemplo de Android y su dominio del mercado de los smartphones. Las instituciones públicas se han unido más recientemente a esta filosofía, en base a la idea de que aquello que se produce con dinero público debe ser de uso también público.

Sin embargo este filosofía basada en la apertura, libertad y reutilización no se ha limitado al software, sino que se ha extendido a otros ámbitos muy diferentes. Tal es el caso, por ejemplo, de la producción en el ámbito cultural e intelectual, que ha asimilado los principios del software libre para dar lugar a la llamada Cultura libre y el movimiento Copyleft, con ejemplos tan famosos como la Wikipedia o las licencias Creative Commons.

Probablemente el último elemento en unirse a este ecosistema libre ha sido el llamado Open Data: principalmente desde el ámbito público, pero también en empresas privadas, se publica cada vez más información, más o menos estructurada, en las mismas condiciones de licencias abiertas y posibilidad de reutilización ya mencionadas. El rápido crecimiento de este campo está muy ligado al auge del llamado Gobierno Abierto, que cuenta con la Casa Blanca como su principal abanderado.

Las posibilidades de reutilización y estandarización, la interoperabilidad, los menores costes, la existencia de comunidades de usuarios y desarrolladores o la flexibilidad de adaptación son la principal causa de que prácticamente todos los protocolos y sistemas que permiten la misma existencia de Internet estén construidos sobre software libre.

2.1 El valor añadido del uso de software libre en el proceso investigador

En primer lugar, una de las características que hacen más relevante el papel del software libre para hacer ciencia abierta, es el hecho de que el uso de software libre favorece la reproducibilidad de los análisis llevados a cabo. Utilizar software libre permite que cualquier usuario pueda instalar y utilizar este mismo software sin restricciones ni condiciones externas. El programa podrá ser instalado, ejecutado, se verá su código fuente y cualquier investigador podrá conocer con detalle cuál es el proceso que el software ha realizado. Con total exactitud. Un software que no sea de código abierto realizará ciertas tareas, mediante la ejecución de un código que no puede ser examinado ni es, en general, es conocido. Aún pagando una licencia, lo cual no está al alcance de todos, este pago solamente da derecho a utilizar el programa, no a conocer el código que ejecuta el mismo.

En el caso de análisis estadísticos sobre datos experimentales, el uso de paquetes libres garantiza que los mismos resultados puedan ser alcanzados por otros investigadores, independientemente de sus posibilidades económicas.

Por otra parte, el uso de software libre evita el falseo, consciente o no, de resultados. Garantiza que el proceso que se lleva a cabo es el que se pretende o, en su caso, garantiza que puedan descubrirse eventuales errores o mejorarse el procedimiento implementado. El uso de software libre permite conocer el procedimiento computacional que se lleva a cabo con todo detalle, de forma transpa-

rente, siendo en numerosas ocasiones el propio software tanto herramienta como producto de investigación.

Destacamos también el hecho de que un software que soluciona un problema o ayuda en la dilucidación de una cuestión científica debe ser tan accesible a la comunidad como la conclusión a la que se llega y que se establece en el informe final. En no pocas ocasiones se implementan programas con nuevas técnicas, o bien con una combinación de técnicas conocidas, para el tratamiento de la información recogida. Poder aportar a la comunidad el software que ha sido implementado es tanto o más valioso que expresar la conclusión final.

Por último, notamos que un software que puede ser revisado, compartido, mejorado, instalado y compartido será fácilmente depurado y mejorado, permitiendo un avance real del conocimiento. Al utilizar software libre, el investigador se vale del trabajo de otros investigadores y en muchas ocasiones mejora o amplía este software, poniendo a su vez a disposición de la comunidad su trabajo y aportando conocimiento a la comunidad de la que se sirvió. Este es un proceso de retroalimentación que hace que los individuos se sirvan del conocimiento de la comunidad para construir más conocimiento y aportarlo a la comunidad.

3 Datos abiertos

La idea subyacente de los datos abiertos es que todos los datos que se generan a causa de los procesos de automatización e informatización de la información estén al alcance de la comunidad en un formato accesible, que permita el análisis, de forma que puedan ser tratados y puedan ser compartidos. Esta línea es la que recoge la Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno. En ella, se establece, fundamentalmente, que las instituciones públicas deben poner a disposición de los ciudadanos los datos que en ellas se generen. Con ello se persigue que el ejercicio de transparencia provoque en los servidores públicos una mayor eficacia, austeridad, imparcialidad y responsabilidad.

La iniciativa de apertura de datos en España toma cuerpo con la creación de un portal de datos abiertos (`http://datos.gob.es/`) en el año 2009, donde se da visibilidad a las iniciativas de apertura de datos en el territorio nacional.

Más allá de la información generada en el seno de la administración pública, los datos abiertos abarcan cualquier tipo de información cuantitativa y cualitativa susceptible de ser analizada, reutilizada y compartida. La liberación de datos consiste en hacer público y accesible un conjunto de datos que pueda ser tratado estadísticamente para establecer las conclusiones pertinentes, sin limitaciones de uso y cumpliendo unos esquemas de interoperabilidad. Esto último intenta evitar la proliferación de ingentes cantidades de conjuntos de datos que no puedan ser combinados en sistemas más grandes, donde está el verdadero

interés.

En la actualidad, la Comisión Europea ha lanzado un programa piloto que hace obligatorio el depósito en abierto de los datos de investigación producidos en el marco de los proyectos. Se trata sin lugar a dudas, de un primer paso para avanzar hacia la publicación de forma reutilizable de todos los productos de investigación generados.

3.1 Idoneidad de la apertura de datos de investigación

Como parte del proceso de producción científica, los datos no deben quedar relegados a un segundo plano, ya que es a partir de ellos que se establecen hipótesis, se elaboran teorías y se establecen conclusiones. Tan importante es compartir el resultado como compartir la información que lleva a establecer el mismo. Los datos experimentales son de una relevancia extrema: son la base de la investigación. Sin embargo su papel en la divulgación y en la publicación de resultados está prácticamente en la sombra. Como producto de investigación que son, deberían tener un papel protagonista y estar puestos a disposición de la comunidad científica bajo una licencia libre.

Al igual que el software libre, la publicación de los datos de investigación bajo licencia libre propicia la reproducibilidad de los análisis. Permite que cualquier investigador realice, con los mismos datos, tanto idénticos como diferentes análisis, de forma que los resultados que se deriven de estos pueden ser una fuente excepcional de testeo. La consistencia de una conclusión será tanto mayor en cuanto que pueda ser establecida por más de un investigador. Por otra parte, siempre pueden compararse análisis de diferente índole con el análisis original, propiciando el intercambio de experiencias, la apertura de debate, la discusión metodológica, etc.

Además, la liberación de los datos debe responder ante los esquemas de interoperabilidad. De esta forma, diferentes fuentes de datos pueden combinarse para obtener un sistema más completo de información que permita un análisis más profundo o más extenso y, por otra parte, se evita la replicación de experimentaciones en idénticas condiciones, ya que los datos de un experimento pueden ser utilizados para diferentes análisis. Esto cobra especial relevancia cuando la obtención de los datos es muy costosa. En tal caso, unificar la información de diferentes fuentes revierte en contar con conjuntos de datos más completos y evitar la duplicación de esfuerzos en la obtención de datos redundará en menores costes. Existen repositorios de datos de investigación, siendo una de las iniciativas más competitivas las que pueden verse en `http://www.re3data.org/` y `http://zenodo.org/`. Una herramienta para la gestión de datos de investigación en el marco de los proyectos de H2020 es `http://www.consorciomadrono.es/pagoda/`.

Excepciones a la idoneidad de publicación de los datos en abierto son poco

numerosas y responden principalmente a la protección derechos de propiedad industrial, protección datos personales, razones de seguridad, que el objetivo principal del proyecto se vea comprometido o bien que no se generen datos.

4 Acceso abierto

Un documento de acceso abierto es aquel que permite a cualquier usuario leer, descargar, copiar, distribuir, imprimir, buscar o enlazar su contenido de manera gratuita y sin ninguna restricción más allá de la que pueda suponer el medio técnico necesario para acceder a él. Los documentos de acceso abierto proporcionan una vía potencialmente rápida para el acceso y la difusión de la literatura científica. Así, la publicación en abierto promueve eliminar barreras económicas, legales y tecnológicas.

El modelo actual de publicación científica está sujeto a políticas de privacidad establecidas desde entidades privadas que gestionan el contenido, velando por la calidad del mismo, el formato y la difusión de los resultados científicos. Sin embargo este escenario está cambiando a grandes pasos y ya es obligatorio para los beneficiarios de H2020 depositar en abierto todas las publicaciones científicas que se elaboren en el marco de la investigación financiada por este programa y, para los beneficiarios del Plan Estatal, la obligación de depositar en acceso abierto se recoge en el artículo 37 de la Ley de la Ciencia, la Tecnología y la Innovación (`https://www.boe.es/boe/dias/2011/06/02/pdfs/BOE-A-2011-9617.pdf`).

Este camino supone la ruptura con un círculo vicioso donde los resultados de la investigación subvencionada con fondos públicos se ceden a revistas científicas y bases de datos y el acceso desde la institución pública a los resultados supone un coste económico. Este círculo implica que las instituciones públicas inviertan en la investigación y a su vez deban volver a invertir, dependiendo de proveedores privados, para poder acceder a los resultados de la investigación que han subvencionado. Por lo tanto, la publicación en abierto genera un menor coste para las instituciones a la vez que promueve la salida a la luz de la llamada *literatura gris*. Esta es el conjunto de resultados que quedan recogidos en revistas científicas editadas en instituciones, tesis doctorales en papel, bases de datos institucionales, datos científicos que quedan almacenadas de forma local en los ordenadores y los servidores de la institución, ponencias, etc. Es, en resumen, el conjunto de resultados de investigación que en principio queda inaccesible a la comunidad y que en muchas ocasiones no son suficientemente explotados.

No son pocas las trabas con las que el investigador se encuentra a la hora de depositar sus manuscritos en abierto. El principal inconveniente se deriva del hecho de que se cede el copyright de la obra a la editorial y puede encontrarse con que es necesario que transcurra un periodo de embargo, establecido por la editorial, antes de que se permita a los autores hacer el depósito en abierto

del manuscrito final revisado por pares. Esta forma de depósito en abierto es la llamada *vía verde*. Una vía alternativa, la denominada *vía dorada* consiste cubrir los costes de publicación y tener inmediatamente disponible el manuscrito en acceso abierto desde el momento de publicación.

4.1 Beneficios derivados de la publicación en abierto

Amén de la citada ruptura del ciclo de doble pago que vienen ejerciendo las instituciones que subvencionan la actividad investigadora, la publicación en abierto propicia otros muchos beneficios tanto para la sociedad como para los investigadores.

La publicación en abierto de los resultados científicos revierte directamente en la sociedad, proporcionando una mayor accesibilidad a la literatura científica. La demanda de este tipo de textos no es exclusiva de la comunidad científica vinculada a instituciones donde se produce la investigación, sino que puede ser de gran interés para estudiantes, profesores de enseñanzas no universitarias, periodistas, aficionados, y un largo etcétera. El libre acceso a las publicaciones derivadas de la investigación puede llevar a una transmisión de conocimiento mucho más amplia y ágil desde las instituciones a la sociedad. Además, el libre acceso propicia que se puedan comparar y seleccionar diferentes publicaciones sobre un tópico concreto. Dada la libertad para distribuir la información que tiene el usuario, se favorece la difusión de los resultados en todos los ámbitos.

Por otra parte, la publicación en abierto da una mayor visibilidad a los autores.

The citation advantage of open access articles `https://dspace.lboro.ac.uk/dspace-jspui/handle/2134/4089`

Acerca de los autores

Nuria Rico es profesora en el Departamento de Estadística e Investigación Operativa de la Universidad de Granada y Subdirectora de la Oficina de Software Libre de la Universidad de Granada. Como responsable en el área de gestión sobre software libre, ha desarrollado diferentes proyectos, entre los que destaca la puesta en marcha del portal de transparencia y OpenData de la Universidad de Granada. Su línea de investigación se centra en el tratamiento estadístico de datos de forma interdisciplinar.

Juan Julian Merelo es Catedrático adscrito al Departamento de Arquitectura y Tecnología de Computadores de la Universidad de Granada así como Director de la Oficina de Software Libre. Sus líneas principales de investigación se enfocan en computación *soft computing* y estudio de métodos metaheurísticos y redes neuronales. Mantiene una intensa actividad de divulgación científica y producción en abierto, participando en foros nacionales e internacionales sobre software libre, ciencia abierta, datos abiertos y temas relacionados.

Ángel Pablo Hinojosa técnico especialista, analista programador y administrador web, miembro fundador de la empresa OpenXXI. Su trayectoria profesional está ligada al estudio y desarrollo de software, liberación, estudio de licencias y formación en TIC. Cuenta con una larga lista de intervenciones en seminarios, conferencias, cursos y talleres de programación en los lenguajes Python y Scratch y de formatos abiertos como LaTeX o XML, de liberación de proyectos de software, licencias de software libre, aspectos legales del software de fuentes abiertas, marco jurídico en social media y de introducción a las redes sociales.