K.4. Compartir datos (*data sharing*) en ciencia: contexto de una oportunidad

Por **Daniel Torres-Salinas**

1 octubre 2009

Torres-Salinas, Daniel. "Compartir datos (*data sharing*) en ciencia: contexto de una oportunidad". *Anuario ThinkEPI*, 2010, v. 4, pp. 258-261

Resumen: *El tema de compartir en acceso abierto los datos de investigación siempre ha estado presente entre la comunidad científica; sin embargo el momento tecnológico actual ha hecho resurgir este debate. Por ello en esta nota y a raíz de un número especial de la revista Nature sobre el tema repasamos las principales características de esta práctica. En primer lugar presentamos los argumentos a favor y en contra; a continuación se refleja como algunas agencias, como los National Institutes of Health (Estados Unidos), exigen a sus investigadores la obligación de poner los datos a libre disposición del resto de la comunidad. En este contexto se entiende que compartir datos maximiza la rentabilidad de la inversión pública en I+D. Para concluir se discute el rol que las bibliotecas universitarias pueden jugar en la gestión y preservación de los datos que se compartan, y como se presenta una excelente oportunidad de adquirir nuevas responsabilidades.*

Palabras clave: *Datos científicos, Data sharing, Acceso abierto, Bibliotecas universitarias, Preservación digital.*

Title: Data sharing *in science: the context of an opportunity*

Abstract: *The issue of sharing research data has always been present in the scientific community; however, the current state of technology has revived the debate. For this reason, and following a special issue of the journal Nature on the topic, we review the main features of this practice. First, we introduce the existing arguments for and against data sharing; in the next section we show how some agencies such as the National Institutes of Health (USA) already require their funded investigators to make their study data freely available to the rest of the scientific community. In this context, agencies understand that data sharing maximizes the return on public investment in R & D. In the conclusions we discuss the role that university libraries can play in data preservation, noting that we have an excellent opportunity to acquire new responsibilities.*

Keywords: *Research data, Data sharing, Open access, Academic libraries, Data curation.*

Introducción

RECIENTEMENTE, la revista *Nature* ha dedicado un número especial[1] a un tema relevante relacionado con la comunicación científica, el *data sharing*; es decir, la acción de compartir con otros colegas los ficheros de datos (o *raw data*) generados durante el curso de una investigación.

Por ejemplo en medicina consistiría en poner en línea todos los datos convenientemente organizados de los pacientes que han participado en un ensayo clínico, o en bibliometría, los ficheros .txt, .xls o .mdb de las publicaciones analizadas.

Es un concepto bastante fácil de entender que, aunque pueda parecer propio de la ciencia 2.0, cuenta con precedentes lejanos en el tiempo: así, **Galton** en 1901 afirmaba que no se debería dejar publicar estudios biométricos si previamente los datos no se depositaban en algún lugar para su consulta[2] (**Hrynaszkiewicz**, 2009).

> **"Estamos ante otra muy buena oportunidad: la especialidad denominada *data curation*"**

Asimismo, en lo que a repositorios se refiere, uno de los más veteranos, el *Protein Data Bank*[3], se inició en 1971, y una de las primeras revistas en exigir los datos como condición para la publi-

cación fue *Journal of biological chemistry* en 1983 (**Crawford** et al., 1996).

Por tanto estamos ante un viejo tema cuyo debate parece reabrirse periódicamente y en el que tal vez los documentalistas tengamos algo que decir. Por ello en esta nota expondremos brevemente algunas de las cuestiones que rodean a esta práctica científica.

A favor, en contra

Según **Nelson** (2009) la publicación en acceso abierto de los datos es percibida por la mayor parte de los investigadores como un ideal científico y algo beneficioso; sin embargo, no deja de ser una corriente muy minoritaria. En la literatura se ha dado buena cuenta de sus bondades ya que contribuye a reproducir, replicar y verificar resultados obtenidos por otros (**Renolls**, 1997), favorece la posibilidad de reutilización para otro tipo de análisis diferente al original (**Piwowar** y **Fridsman**, 2007), permite combinar diferentes archivos para realizar metanálisis (**Ramasamy** et al., 2008) y es un arma eficaz en la lucha contra el fraude.

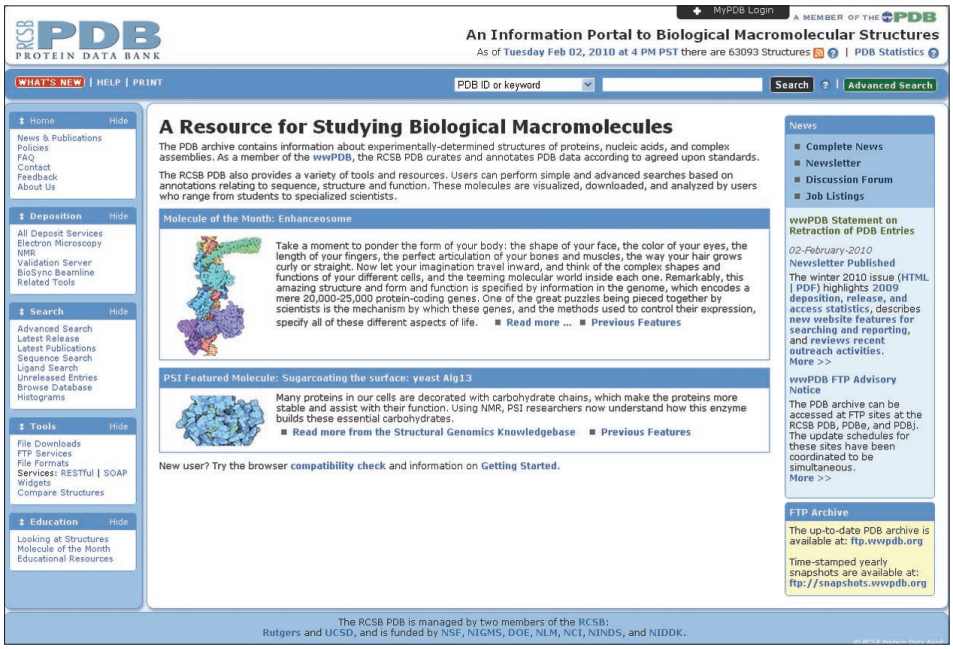

Figura 1. Protein data bank, http://www.rcsb.org

Y si esto aún no nos convence, produce además una aceleración en la citación e incrementos en el número de citas del 70% (**Piwowar** et al, 2008). Sin embargo pese a la percepción positiva y a la existencia de un buen puñado de argumentos, los científicos no se animan y el investigador prefiere mantener sus ficheros en su disco duro hasta que un día éste dé "error fatal" y desaparezcan para siempre.

Claro, si se analiza la cuestión, no les falta razón, ya que por ahora son muchas las dudas en torno al tema. La cuestión básica de "¿dónde deposito mis datos?" aún no está resuelta, ya que apenas existen repositorios ni infraestructuras y, además, en el caso de que se crearan, las incertidumbres sobre el destino de los datos son enormes. La escasa protección en caso de apropiaciones indebidas, la falta de reconocimiento a la hora de ser citado o la posibilidad de que otros hagan *papers* a costa de nuestro esfuerzo son cuestiones a resolver.

"Para llevar a cabo el *data sharing* la comunidad científica necesita el equivalente digital de las bibliotecas actuales"

Pero principalmente no olvidemos que los datos son el motor de las publicaciones y éstas, para el científico, son un bien intercambiable por dinero en forma de tramos o de financiación y nadie en su sano juicio va a poner a disposición de desconocidos su pequeño capital.

Por todo esto el *data sharing* como práctica generalizada implica no sólo un cambio cultural sino también unas reglas de juego bien establecidas; y aunque puede generar beneficios colectivos también puede provocar perjuicios individuales, y al final son estos últimos los que decantan la balanza.

Compartir datos por decretazo

Sin embargo ya hay quien se está tomando el tema en serio y obliga a compartir datos si no se quiere hacer voluntariamente. Esta posición se entiende desde el punto de vista de las administraciones públicas ya que financian la investigación, y el *output* de sus proyectos no debe ser sólo un conjunto de resultados y conclusiones si pueden ser más.

Las administraciones también tendrían que reclamar los datos generados, que deberían poder ser utilizados por otros científicos del mis-

mo sistema público. Asimismo se reciclarían todos esos proyectos que no alcanzaron los resultados esperados pero cuyos datos sí pueden ser de interés y, en última instancia, se podrían evitar investigaciones duplicadas ahorrando dinero.

"Los datos son el motor de las publicaciones y éstas son para el científico un bien intercambiable por dinero"

En fin, las agencias pagan y tendrían que darse cuenta que los datos no son de los científicos que ejecutan los proyectos sino de ellas, que para eso ponen el dinero sobre la mesa.

Aunque esta política pueda parecer exagerada, agencias y organismos, nacionales y supranacionales se están poniendo en marcha (**Fukasaku**, 2007). El ejemplo más claro es el de los *National Institutes of Health* (*NIH*)[4], que desde 2003 exigen a todos los proyectos financiados con más de 500.000 US$ que compartan sus datos.

El plan de los *NIH* es muy sencillo: los investigadores al presentar la solicitud deben incluir un plan para compartir los datos generados por el proyecto. Además los *NIH* no han dejado solos a los investigadores y han creado diversos repositorios como el *GenBank*[5], *Protein Cluster*[6] y *PubChem*[7].

Si a esta política le sumamos la reciente, relativa a que todas las publicaciones sufragadas por los *NIH* deben ponerse en acceso abierto (**Martínez**, 2008), podemos sospechar cuál puede ser el siguiente paso a unos años vista.

Se puede concluir que si los investigadores no se animan a compartir de forma natural, lo mejor es actuar con políticas de decretazo como la de los *NIH*, lo que pensado en frío no deja de ser un poco triste.

Una buena oportunidad

Dejando tristezas a un lado, si estas políticas se extendieran y al final los argumentos a favor prevalecieran sobre los argumentos en contra, el tema nos presentaría un buen puñado de problemas técnicos bastante estimulantes. Y es que colgar datos no es igual que colgar ppts o compartir enlaces. Nos encontramos con información mucho más compleja, con especificidades propias de cada especialidad, a veces sujeta a leyes de protección de datos (por ejemplo de pacientes), con formatos múltiples (numéricos, textuales, multimedia…; sas, html, raw…), que requerirían pautas de normalización y presentación para su depósito, sistemas de recuperación más complejos y más amigables, y una conservación de los datos a largo plazo.

Figura 2. National Institutes of Health, http://nih.gov/

A todo esto habría que sumar unas normas éticas y un contexto legal para proteger a los depositantes y por supuesto encontrar quien corra con los costes de las infraestructuras y formación de los científicos. En fin, toda una serie de cuestiones que no se resuelven en dos días ni en dos años.

Está claro por tan-

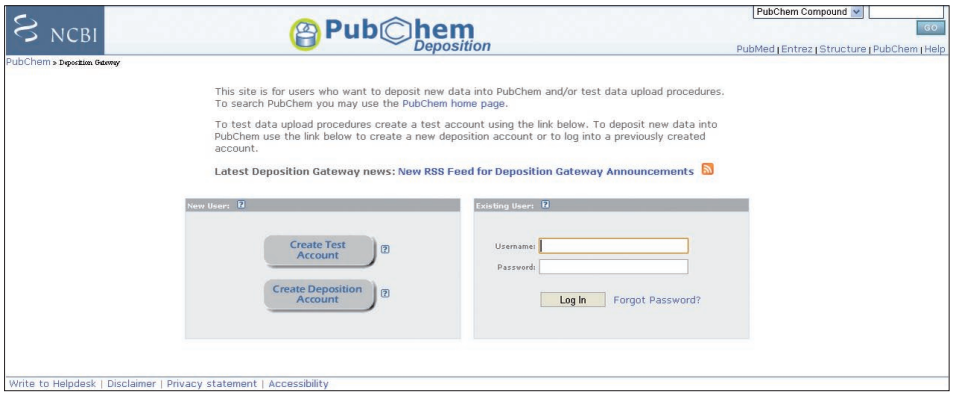

Figura 3. Pubchem deposition, http://pubchem.ncbi.nlm.nih.gov

to que en toda esta historia puede haber algo positivo para nosotros. Si las grandes ganadoras del *open access* son las bibliotecas universitarias como entidades encargadas de tutelar los repositorios, con los repositorios de datos puede o debe ocurrir lo mismo. Y quizá deberían ser los profesionales de la información los encargados de comenzar a resolver los problemas reseñados y abonar el terreno.

El editorial de *Nature* (2009) no puede ser más explícito al respecto: la comunidad científica, para llevar a cabo el *data sharing*, necesita el equivalente digital de las bibliotecas actuales, es decir, alguien que preserve y haga accesible todos esos datos, y se apunta directamente a las bibliotecas universitarias (como instituciones) y al *data management* (en tanto que rama del conocimiento) como los pilares sobre los que se debe apoyar el futuro del *data sharing*.

Ante estas afirmaciones no voy a apuntar qué es lo que se debe o no enseñar en las facultades de Documentación, simplemente quiero resaltar que estamos ante otra muy buena oportunidad: la especialidad denominada *data curation*.

Notas

1. *http://www.nature.com/news/specials/datasharing/index.html*

2. Texto original de **Francis Galton** (*Biometrika*, n. 1, 1901): "*'I have begun to think that no one ought to publish biometric results, without lodging a well-arranged and well-bound manuscript copy of his data in some place where it should be accessible*".

3. *http://www.rcsb.org/pdb/home/home.do*

4. *http://grants.nih.gov/grants/policy/data_sharing/*

5. *http://www.ncbi.nlm.nih.gov/Genbank/index.html*

6. *http://www.ncbi.nlm.nih.gov/sites/entrez?db=proteinclusters*

7. *http://pubchem.ncbi.nlm.nih.gov/*

Referencias

Crawford, Susan Y.; **Hurd, Julie M.**; **Weller, Ann C.** "From print to electronic: the transformation of scientific communication". Medford: ASIS, 1996.

Fukasaku, Yukiko. International initiatives in data sharing: OECD, Codata and Gicsi, 2007. *http://www.aepic.it/conf/viewabstract.php?id=269&cf=10*

Hrynaszkiewicz, Iain; **Altman, Douglas**. "Towards agreement on best practice for publishing raw clinical trial". *Trials*, 2009, v. 10, n. 17. *http://www.trialsjournal.com/content/10/1/17*

Martínez, Luis-Javier. "Más acceso abierto". *Observatorio de Martinej*, 2008. *http://martinej.wordpress.com/2008/01/24/mas-acceso-abierto-nih/*

Martínez-Uribe, Luis; **Macdonald, Stuart**. «Un nuevo cometido para los bibliotecarios académicos: data curation». *El profesional de la información*, 2008, v. 17, n. 3, pp. 273-280.

Data's shameful neglect. *Nature*, 2009, v. 461, n. 7261, p. 145.

Nelson, Bryn. "Empty archives". *Nature*, 2009, v. 461, n. 10, pp. 160-163.

Piwowar, Heather; **Fridsma, Douglas B.** "Examining the uses of shared data". *Nature preceedings*, 2007. *http://precedings.nature.com/documents/425/version/3*

Piwowar, Heather; **Day, Roger S.**; **Fridsma, Douglas B.** "Sharing detailed research data is associated with increased citation rate". *Plos One*, 2007, v. 3, e308.

Ramasamy, Adaikalavan; **Mondry, Adrian**; **Holmes, Chris C.**; **Altman, Douglas G.** "Key issues in conducting a meta-analysis of gene expression microarray datasets". *Plos medicine*, 2008, v. 5, n.9, e184.

Rennolls, Keith. "Science demands data sharing". *BMJ*, 1997, v. 315, n. 7106. *http://www.bmj.com/archive/7106/7106l7.htm*

Roba-Stuart, Óscar. "Archivos de datos en línea para ciencias sociales". *El profesional de la información*, 2003, v. 12, n. 5, sept.-oct., pp. 400-410.