PRESERVACIÓN DE DATOS DE INVESTIGACIÓN DE CIENCIAS SOCIALES CON HERRAMIENTAS DE ANÁLISIS FORENSE DIGITAL

Teodoro Wilderbeek López del Castillo y Miquel Térmens

RESUMEN: La preservación de datos de investigación es una nueva tarea que han de asumir las bibliotecas universitarias a partir de los mandatos emitidos por las instituciones de financiación de los distintos proyectos de investigación. Esta tarea es complicada debido a que los datos de investigación presentan una elevada variabilidad entre disciplinas y, además, en la actualidad no se dispone de *software* especializado que facilite su preservación.

En el marco de una tesis doctoral que estudia este problema para las disciplinas de ciencias sociales, se presenta una solución consistente en el uso de herramientas *hardware* y *software* de análisis forense digital en coordinación con el *software* de repositorios DSpace, el más usado a nivel mundial. Esta solución permite el tratamiento de volúmenes medianos de datos (de gigabytes a terabytes) minimizando las operaciones de carácter manual.

Palabras clave: Preservación digital; datos de investigación; repositorios institucionales; análisis forense digital.

Keywords: Digital preservation; research data; institutional repositories; digital forensics.

1. INTRODUCCIÓN

Las agencias de financiación de la investigación de forma creciente están exigiendo que los proyectos a los que dan soporte depositen sus datos de investigación en repositorios abiertos que permitan su reutilización por terceros. Los responsables de estos proyectos deben dar cumplimiento a este nuevo requerimiento, pero se encuentran de forma generalizada con el hecho de que la gran mayoría de los actuales repositorios institucionales o temáticos no están preparados para acoger datos de investigación.

El almacenamiento, difusión y preservación de datos de investigación presenta características propias que lo diferencian de las actividades afines de la producción científica:

- Los datos de investigación raramente están en los formatos que se usan en la publicación científica, como PDF, LaTeX o Word, pero usan un gran número de formatos de tipo genérico o específicos de alguna disciplina.
- Una investigación casi siempre genera más de un fichero de datos y estos ficheros pueden tener un gran volumen, medible en GB o TB.
- Los ficheros pueden contener metadatos que sigan esquemas propios de una disciplina, de forma que la estructura de ficheros y metadatos asociados de una investigación puede variar mucho respecto a otra.
- Los datos pueden contener informaciones e carácter personal o reservado que por razones de confidencialidad o éticas no se pueden mostrar libremente.
- El uso de los datos de investigación será limitado, pues los usuarios potenciales se limitan a otros investigadores de la misma área y tema de interés.

Todas estas características aconsejan que los datos de investigación sean gestionados mediante *software* distinto al usado en los repositorios abiertos de comunicación científica. Para solventar esta necesidad, en los últimos tiempos han aparecido distintos *softwares* o portales especializados en la gestión de datos científicos; podemos mencionar el Interuniversity Consortium for Political and Social Research (ICPSR), Dryad, Dataverse, Figshare o Zenodo. El problema de estos *softwares* radica en que suponen la creación de un nuevo repositorio dentro de la institución, con un funcionamiento diferente al del repositorio institucional ya existente. Nuestra investigación pretende solventar este problema mediante una aproximación diferente: adaptar para albergar datos el *software* DSpace, el más usado a nivel mundial en repositorios institucionales, y gestionar las particularidades técnicas de los ficheros de datos mediante técnicas de análisis forense digital.

El análisis forense digital ya se está aplicando en bibliotecas y archivos para acceder a fondos patrimoniales digitales que han llegado en soportes y formatos obsoletos. Con el uso de este tipo de técnicas es posible acceder a cualquier soporte de información digital, procesar grandes volúmenes de información, discriminar qué ficheros se pueden recuperar y migrar a formatos más actuales y también detectar la existencia de datos personales o confidenciales. Por todo ello creemos que se trata de unas técnicas que pueden facilitar el tratamiento de los datos de investigación.

2. MATERIALES Y METODOLOGÍA

Se ha creado un modelo de adquisición, procesamiento, almacenamiento y consulta de datos de investigación mediante técnicas de análisis forense y la adaptación del *software* DSpace. El modelo creado es válido para la gestión de datos de investigación en acceso abierto de ciencias sociales y de humanidades, de carácter no estructurado. La solución utiliza *hardware* forense de bajo coste y *software* de código abierto, no sujeto a licencias de uso. El modelo se ha probado en un equipo prototipo y ha sido testeado con distintos tipos de ficheros.

En el mercado se comercializa *hardware* y *software* especializado en tratamiento forense, como pueden ser las estaciones de trabajo FRED y el *software* FTK y EnCase. Sus capacidades son altísimas y son capaces de tratar todo tipo de ficheros almacenados en casi cualquier almacenamiento, pero su coste es muy elevado. Asimismo estas prestaciones avanzadas no son necesarias en la mayoría de bibliotecas. Por ello se optó por usar el paquete forense BitCurator, así como el *hardware* mínimo imprescindible, formado por un ordenador con 32GB de RAM, una *docking station* con capacidad para discos SATA e IDE, y una unidad de adquisición *write blocker* USB Wiebetech. El *software* DSpace por su parte necesita ser configurado para que permita la incorporación de ficheros mediante FTP. La solución que presentamos permite la adquisición de contenidos desde discos ópticos (CD y DVD), discos duros (internos de conexión IDE o SATA o bien externos con conexión USB) y memorias USB.

3. RESULTADOS

La solución propuesta parte de la evidencia de que el investigador no ingresará directamente los datos de investigación, como sí ocurre con los resultados de la investigación, como artículos e informes. La razón está en el volumen de los ficheros, que no pueden ser transmitidos desde un navegador con el protocolo HTTP. Por ello el investigador deberá realizar su entrega

mediante un soporte físico. A partir de este momento, los administradores del repositorio iniciarán las tareas para la incorporación de los ficheros en el sistema. Para ello se han previsto las siguientes fases de trabajo.

3.1. Preparación del AIP o paquete de ingestión

- Preparativos iniciales. Consisten en la recepción y comprobación del formulario de depósito y del soporte de almacenamiento con los ficheros, creación de la estructura de carpetas, escaneo y archivado del formulario, asignación de identificadores a los soportes, fotografía de los soportes y examen de los mismos.
- Captura de los soportes. El objetivo es realizar una copia correcta de los contenidos entregados y hacer el traspaso del soporte original al soporte de los administradores. Para ello se realizará la configuración del *hardware* y *software* de captura, se creará y verificará la imagen forense, se localizarán los datos confidenciales y, finalmente, se almacenarán los soportes originales.
- Examen y análisis del contenido. Después de montar la imagen forense, se procederá a realizar el control antivirus y, en el caso de existir, a la extracción de los datos confidenciales.
- Procesado de contenidos. Aquí se preparan los ficheros que se pondrán a disposición de los usuarios, realizando su inventario automatizado y su control de integridad a partir del cálculo de valores hash. En este bloque también se pueden realizar las migraciones de formatos, si así se ha establecido dentro de la política del repositorio.
- Preparación de paquetes AIP para su ingesta. En esta fase se recopilan los diferentes ficheros creados en los pasos anteriores: ficheros *raw*, ficheros para la consulta, metadatos e informes de las diferentes tareas realizadas. El conjunto se empaqueta mediante el protocolo BagIt, que permitirá una transmisión segura al repositorio.
- Ingesta en el repositorio. Si el paquete AIP creado con anterioridad tiene un peso menor a 4GB se procede a su ingestión en DSpace por métodos tradicionales, pero si el tamaño es mayor, como será lo común, la ingestión se deberá realizar por FTP, después de realizar las adaptaciones necesarias para que DSpace admita esta vía de entrada.

3.2. Almacenaje en DSpace

- El funcionamiento de DSpace como repositorio de datos no varía respecto a su funcionamiento habitual como repositorio de publicacio-

nes. Los datos podrán ser buscados gracias a los metadatos descriptivos marcados en Dublin Core.

3.3. Consulta de los ficheros de datos

- Debido al volumen de muchos de los ficheros de datos y a posibles restricciones a su acceso, el usuario no podrá recuperar directamente los ficheros que desee, sino que realizará una petición de consulta que le será entregada de forma diferida.
- Las peticiones de consulta de ficheros llegarán al administrador del sistema. Este genera un DIP o paquete de difusión con los ficheros y metadatos pertinentes y lo pone a disposición del usuario mediante la entrega de una clave de acceso a un espacio de almacenamiento restringido.

La fase A se apoya en el uso de técnicas y procedimientos habituales en análisis forense digital, ya en uso en algunas bibliotecas, acompañados de otros protocolos también conocidos, como BagIt. La fase B simplemente requiere disponer de un repositorio DSpace, con la única adaptación de que no servirá los ficheros elegidos por el usuario. Por último, la fase C consiste en la realización de pequeñas tareas manuales o programadas que extraerán los ficheros correspondientes de DSpace, eliminaran los que no se deban entregar al usuario, y montaran los resultados en un espacio temporal.

4. DISCUSIÓN Y CONCLUSIONES

La solución propuesta no se ha probado en un entorno operativo, por lo que no se puede considerar como directamente aplicable. En una situación real posiblemente se podrían realizar adaptaciones que simplificaran algunos de los subprocesos previstos, bien sea porque estos no sean necesarios en un entorno determinado o bien porque se pudiera programar y automatizar su ejecución. Uno de los aspectos a estudiar es la mejora del procedimiento de consulta final de los datos.

Más allá de estas salvedades, creemos que la adaptación de *software* bien conocido y usado como DSpace y la aplicación de nuevas metodologías de trabajo como las del análisis forense digital pueden representar una buena línea de actuación que resuelva las necesidades de instituciones de tamaño mediano. La alternativa aquí presentada consigue automatizar y sistematizar el depósito y tratamiento de volúmenes medianos de datos de investigación (de gigabytes a terabytes) minimizando las operaciones de carácter manual.

5. AGRADECIMIENTOS

Este trabajo se realizó en el marco de una tesis doctoral sobre la preservación de datos científicos. Ha contado con el soporte del proyecto «El acceso abierto a la ciencia en España: evaluación de su impacto en el sistema de comunicación científica» (Plan Nacional, ref. CSO2014-52830-P). También ha recibido soporte del GRC *Cultura i continguts digitals*: *aspectes documentals*, *polítics i econòmics*.

6. BIBLIOGRAFÍA

BARRERA-GOMEZ, J. & ERWAY, R. (2013). *Walk this way*: *detailed steps for transferring born-digital content from media you can read in-house*. Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2013/2013-02.pdf

DALLMEIER-TIESSEN, S. et al. (2014). Enabling sharing and reuse of scientific data. *New Review of Information Networking*, 19(1), pp. 16-43. http://dx.doi.org/10.1080/13614576.2014.883936

DOWNS, R. R. & CHEN, R. S. (2010). Self-assessment of a long-term archive for interdisciplinary scientific data as a trustworthy digital repository. *Journal of Digital Information*, 11(1). https://journals.tdl.org/jodi/index.php/jodi/article/view/753

EUROPEAN COMMISSION (2016). *Guidelines on open access to scientific publications and research data in Horizon* 2020 *(version* 3.1*)*. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

FERRER-SAPENA, A., PESET, F. & ALEIXANDRE-BENAVENT, R. (2011). Acceso a los datos públicos y su reutilización: open data y open government. *El Profesional de la Información*, 20(3), pp. 260-269. http://dx.doi.org/10.3145/epi.2011.may.03

GENGENBACH, M. J., CHASSANOFF, A., & OLSEN, P. (2012). Integrating digital forensics into born-digital workflows: the BitCurator project. *Proceedings of the American Society for Information Science and Technology*, 49(1). http://dx.doi.org/10.1002/meet.14504901343

GÓMEZ, N. D., MÉNDEZ, E., & HERNÁNDEZ-PÉREZ, T. (2016). Datos y metadatos de investigación en ciencias sociales y humanidades: una aproximación desde los repositorios temáticos de datos. *El Profesional de la Información*, 25(4), pp. 545-555. http://dx.doi.org/10.3145/epi.2016.jul.04

JOHN, J. L. (2008). Adapting existing technologies for digitally archiving personal lives: digital forensics, ancestral computing, and evolutionary perspectives and tools. *5th International Conference on Preservation of Digital Objects (iPRES)*. http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf

JOHN, J. L. (2012). *Digital forensics and preservation*. http://dx.doi.org/10.7207/twr12-03

KIRSCHENBAUM, M. G., OVENDEN, R., & REDWINE, G. (2010). *Digital forensics and born-digital content in cultural heritage collections*. Washington, DC: Council on Library and Information Resources. http://www.clir.org/pubs/reports/pub149/pub149.pdf

LEE, C. A. et al. (2012). BitCurator: tools and techniques for digital forensics in collecting institutions. *D-Lib Magazine*, 18(5/6). http://dx.doi.org/10.1045/may2012-lee

REILLY Jr., B. & WALTZ, M. E. (2014). Trustworthy data repositories: the value and benefits of auditing and certification. En: *Research data management: practical strategies for information professionals*. West Lafayette, Indiana: Purdue University Press, p. 109-126.

WILDERBEEK, T. & TÉRMENS, M. (2015). Creación de unidades de análisis forense en bibliotecas. *El Profesional de la Información*, vol. 24(1), pp. 44-54. http://dx.doi.org/10.3145/epi.2015.ene.06

WOLVERTON, M. (2016). Digital forensics in the library. *Nature*, 534, pp. 139-140. http://dx.doi.org/10.1038/534139a

WOODS, K. & LEE, C. A. (2015). Redacting private and sensitive information in born-digital collections. *Archiving* 2015, 6, pp. 2-7.