Una propuesta metodológica de relevamiento para iniciar proyectos de digitalización y preservación

BORREL, Marina / Universidad Nacional de La Plata – mborrel@fahce.unlp.edu.ar

FUENTE, María Virginia / Instituto de Investigaciones en Humanidades y Ciencias Sociales (IdIHCS). Universidad Nacional de La Plata – mvfuente@fahce.unlp.edu.ar

GONZÁLEZ, Claudia / Instituto de Investigaciones en Humanidades y Ciencias Sociales (IdIHCS). Universidad Nacional de La Plata – cgonzalez@fahce.unlp.edu.ar

» *Palabras clave: metodología de relevamiento, curaduría de datos, preservación digital, Humanidades, Ciencias Sociales.*

› Resumen

A partir de la revisión de casos externos, se presentan los debates y decisiones que se generaron en vista de la construcción de un instrumento de relevamiento para la planificación de proyectos de curaduría de datos y preservación digital en el IdIHCS/FaHCE.

En lo relativo a la revisión de estos casos externos, se analizan propuestas metodológicas resultantes de los principales proyectos europeos, preocupados por el armado de infraestructuras de información para la investigación. Asimismo, a nivel nacional, se toma en consideración el antecedente de la Plataforma Interactiva de Investigaciones en Ciencias Sociales (PLIICS), proyecto del CONICET.

En cuanto a la definición de una metodología de relevamiento de fuentes, adecuada a las características de trabajo propias de nuestras instituciones de investigación, se presentan los principales aspectos a considerar y los elementos requeridos en el instrumento a elaborar. El mismo deberá reflejar los principales problemas relativos a: valoración de la importancia para las investigaciones en curso, estado de procesamiento, interés de los investigadores en la visibilización de las fuentes, ponderación de acuerdo al uso, evaluación de las posibilidades legales de difusión en acceso abierto, etc. Todo lo anterior con la finalidad de priorizar proyectos que permitan preservar las fuentes en riesgo, y obtener información que facilite una planificación adecuada a las necesidades del ámbito local.

ISBN 978-987-3617-89-8

› *Introducción*

La curaduría digital, entendida como la gestión y preservación de los datos digitales a lo largo del tiempo, no solo facilita el acceso persistente a datos digitales confiables, auténticos, previniendo los riesgos de pérdida y obsolescencia de reproducción, sino que también propende a la mejora de la calidad de los datos en su contexto de investigación. Si bien el término "curador" proviene del ámbito de los museos, en el ámbito de la información digital también denota las funciones de protección, contextualización y la exposición efectiva de la información a un conjunto adecuado de usuarios. El término "curaduría digital" se acuña en 2001 y surge a partir del cruce entre trabajadores de la información (bibliotecarios, archiveros) y científicos, para quienes los términos "archivo" y "preservación" denotan actividades post-proyecto de investigación, con poca o ninguna conexión con la etapa de creación y fomento del reuso de los materiales por ellos generados. Por lo tanto, es esta nueva actividad la que conforma un nuevo campo de aplicación de conocimientos compartidos entre productores de datos y trabajadores de la información, donde el ciclo de vida de los datos es uno de sus ejes centrales (Lee & Tibbo, 2007).

Partimos de considerar que el concepto de construcción de una colección de fuentes o datos para la investigación en Humanidades y Ciencias Sociales, en tanto selección y mantenimiento de un cuerpo de conocimientos de temas específicos, podría ser equiparable a lo que se hace en muchas otras disciplinas y sectores, por ejemplo en los centros de datos para las ciencias oceanográficas. Partimos también de considerar que las Ciencias Sociales y Humanas deben participar en el diseño y construcción de la e-infraestructura científica de un país, entendida esta no solo como la combinación de hardware, software y comunicaciones, sino como un entorno de trabajo donde las capacidades de la computación se ponen a disposición de los investigadores en redes interoperables. Implica información, conocimientos, normas, políticas, herramientas y servicios que se comparten ampliamente a través de las comunidades de investigación, pero desarrollados con fines académicos específicos (National Science Foundation and the Library of Congress, 2003).

Por otro lado, en el caso específico de las Humanidades, se observa que en nuestro entorno, ya se deba a la propia especificidad de los materiales y/o a la manera en que trabajan los investigadores, tienden a generarse proyectos atómicos, anclados en el campo del *cultural heritage* o de los archivos históricos tradicionales, con prácticamente ningún antecedente a nivel nacional de participación en infraestructuras de información científica más amplias y enriquecidas. El presente trabajo tiene como primer objetivo revisar estos escasos antecedentes, a la vez que relevar aquellos proyectos internacionales que aportan elementos de interés para pensar esta problemática. Asimismo presenta aportes preliminares para la etapa inicial de proyectos de curaduría digital en humanidades, específicamente en lo que se refiere al relevamiento y selección de materiales.

› *Proyectos en curaduría de datos en Humanidades y Ciencias*

ISBN 978-987-3617-89-8

Sociales. Antecedentes

Antecedentes internacionales

Existen diversas organizaciones interesadas en el desarrollo de las Humanidades Digitales. Ofreceremos aquí en primer lugar, una lista de algunas de ellas, y luego, un intento de síntesis de los que consideramos sus aportes fundamentales, y sus coincidencias de concepciones y propuestas. No obstante, nos interesa destacar que existen muchas otras, incluso de más larga tradición en el tema, especializadas en otras disciplinas que en algunos casos han servido de referencia para el desarrollo de estas que nos ocupan. Consideramos que sería relevante dar cuenta también de sus experiencias, en oportunidad de un trabajo más extenso[1].

Organizaciones relevadas

Hemos relevado las siguientes organizaciones:

American Council of Learned Societies (ACLS): federación de 72 organizaciones académicas sin fines de lucro de las Humanidades y Ciencias Sociales de los Estados Unidos. En 2004, nombró una Comisión Nacional sobre Ciberinfraestructura en las Humanidades y las Ciencias Sociales, y en 2006, la Comisión publicó el informe "Our Cultural Commonwealth", en el que destaca el potencial transformador de una infraestructura que conserve y comparta el patrimonio cultural de la sociedad de la que forma parte, y señala que para que esto sea posible deben superarse limitaciones como la pérdida, la fragilidad y la dificultad para el acceso a los materiales y documentos; como también las relativas a la propiedad intelectual, y la falta de recursos y estímulos para experimentar con ciberinfraestructura (http://www.acls.org/).

UK Data Archive (UKDA): gestiona la principal colección de datos para investigación de Humanidades y Ciencias Sociales del Reino Unido. Es uno de los principales creadores de directrices y buenas prácticas en la curaduría digital de datos en ciencias sociales. Desde 2006 publica una guía de buenas prácticas (http://www.data-archive.ac.uk/).

Research Information Network (RIN): organismo sostenido por el gobierno del Reino Unido a través de su Consejo de Educación Superior, sus siete Agencias de Investigación y sus tres Bibliotecas Nacionales. Tiene como objetivo mejorar y ampliar la creación y el uso de recursos de información y servicios que desarrollan los investigadores y las instituciones, con el fin de

[1] Algunas de ellas son: *Data Curation Centre* del Reino Unido, *Inter-University Consortium for Political and Social Research*, grupo de trabajo *Digital Preservation & Curation*, *King's College London Center for e-Research*, Proyecto InSPECT, *National Science Foundation*, *UK-Data Service*, y, por ejemplo, su registro de set de datos de Historia Oral llamado *Digital Preservation Coalition*. También proyectos autónomos: *Dissemination Information Packages for Information Reuse* (DIPIR), *Digital Preservation Europe*, *Knowlodge Exchange*, PARADIGM (*The Personal Archives Accesibles in Digital Media*), PREPARDE, *Digital Curation Centre y Australian National Data Service*.

ISBN 978-987-3617-89-8

apoyar el desarrollo de políticas y prácticas eficaces. En 2010 publicó el Informe "Open to all? Case studies of openness in research" en el que, entre otra información de gran interés, se destaca que a pesar de las limitaciones económicas para su implementación, cada vez son mayores los esfuerzos por publicar corpus anotados en acceso abierto (http://www.rin.ac.uk/).

Digital Humanities Data Curation (DHDC): es un proyecto de investigación colaborativa del MITH (Maryland Institute For Technology In The Humanities) que se enfoca en desarrollar información sobre la práctica de curaduría en Humanidades mediante recursos de aprendizaje en línea. Para esto han desarrollado una *Guía de Curaduría en Humanidades Digitales* (http://www.dhcuration.org/institute/).

Association of Research Libraries (ARL): es una asociación sin fines de lucro de 125 bibliotecas de instituciones de investigación de EE. UU. y Canadá, cuyos principios rectores son: acceso abierto y equitativo a la información, y el reconocimiento de las bibliotecas como agentes activos en el proceso de transmisión y creación de conocimiento. Su Informe "Safeguarding Collections at the Dawn of the 21st Century: Describing Roles & Measuring contemporary Preservation Activities in ARL libraries", de 2009, menciona algunos programas que trabajan sobre la preservación de materiales, por ejemplo: LOCKSS (Lots of Copies Keep Stuff Safe: http://www.lockss.org/), radicado en las bibliotecas de la Universidad de Stanford que ofrece a bajo costo herramientas para trabajar en preservación digital. Y PORTICO (http://www.portico.org/digital-preservation/) que trabaja con bibliotecas y editoriales en la preservación de revistas, libros digitales y otros contenidos de enseñanza. Participan de este programa 922 bibliotecas de 14 países, entre los que se encuentra Argentina con un alto porcentaje de Instituciones correspondientes a Universidades públicas (UNLP, UBA, UNC, UTN, entre otras) (http://www.arl.org/).

Canadian Association of Research Libraries (CARL): es una asociación que cuenta como miembros a 29 bibliotecas universitarias y 2 de instituciones gubernamentales. Su objetivo es enriquecer la investigación y la educación superior mediante el acceso amplio a la información por la vía de la comunicación académica y las políticas públicas de información (http://www.carl-abrc.ca/en.html).

Association of College and Research Libraries (ACRL): con más de 12.000 miembros, conforma una asociación profesional de bibliotecas académicas y de miembros individuales. Desarrolla tareas para cubrir las necesidades de las bibliotecas académicas y de los profesionales de la información en educación superior y para mejorar las condiciones de enseñanza, aprendizaje e investigación (http://www.ala.org/acrl/).

Aportes

En todas las organizaciones relevadas hallamos denominadores comunes de análisis y trabajo. Observamos un reconocimiento respecto de que el trabajo en Humanidades y Ciencias

ISBN 978-987-3617-89-8

Sociales en el entorno digital debe fortalecerse en:

1. La concientización de los investigadores respecto de la importancia de aprovechar e implementar las herramientas que el entorno digital ofrece para la búsqueda, la puesta en valor y el desarrollo de materiales, contenidos y trabajos. En primer lugar, y en muchos casos, esto implica conocer y aceptar nuevos instrumentos para la preservación y transmisión de las propias producciones (por ejemplo, superar la instancia del almacenamiento en las computadoras personales y de compartir la información vía correo electrónico, lo cual aparece señalado como malas prácticas).
2. La valoración de la Bibliotecas y su personal en lo referido a conocimientos y recursos específicos para capacitar y brindar herramientas para la búsqueda de información, la definición de técnicas de preservación y la creación de bases de datos que normalicen y de este modo simplifiquen la tarea del investigador.
3. La financiación para acompañar con recursos concretos la implementación de los cambios que representan los planteos de los puntos recién señalados. En este sentido, de los informes de las organizaciones surgen múltiples opciones: si bien por un lado se destaca la necesidad de mayor compromiso de las Instituciones de las que tanto investigadores como bibliotecas y centros de investigación forman parte, también se reconoce que es posible desarrollar estrategias como la inscripción en programas de financiamiento externos, o la asociación y formación de redes que unan esfuerzos y compartan los resultados.
4. La digitalización, que representa la opción más importante para acceder a los materiales, es un aspecto a ser trabajado para que deje de ser considerada un método adicional de preservación y de acceso a las colecciones.
5. El trabajo de desarrollo de los *metadata*, a fin de que especifiquen, técnica y administrativamente, las condiciones de trabajo en lo relativo a la digitalización de documentos.
6. La preservación de los objetos digitales, que merece una atención especial si se considera el cambio constante que estos experimentan, por lo que resulta necesario desarrollar soluciones que aseguren la integridad y la autenticidad de los materiales y que permitan su accesibilidad por largos períodos de tiempo.

Reconocemos en este relevamiento una valoración del trabajo de curaduría de datos, en tanto este implica el cruce de distintas acciones y procesos específicos, como: la descripción, el comentario y las notas (que agregan información a los *data*, mediante marcas o glosas para contextualizar el contenido), la creación y el desarrollo de colecciones (sobre la base del vínculos entre instituciones, proyectos o equipos), el almacenamiento (para lo que es necesario trabajar en la codificación de los materiales y en la provisión y el mantenimiento de sistemas que garanticen la estabilidad y el acceso) y la migración (el cambio de formatos y estándares exige que la curaduría desarrolle sistemas de migración regulares y a largo plazo para asegurar el uso y el acceso).

ISBN 978-987-3617-89-8

Antecedentes nacionales

En 2011 CONICET inició un proyecto denominado PLIICS, *Plataforma Interactiva de Información en Ciencias Sociales*. El objetivo de la PLIICS es resguardar la información de interés para la investigación del área que se encuentra dispersa o de difícil acceso, y está constituida por "objetos, bases de datos y fuentes de todo tipo y soporte". El propósito de su digitalización y catalogación reside en lograr un intercambio efectivo a fin de facilitar el trabajo entre disciplinas.

Como plataforma de trabajo se desarrolla en la modalidad de dos catálogos digitales: un catálogo de conjuntos de datos primarios y un catálogo de material multimedia de fuentes documentales.

El proyecto propone instrumentarse sobre la base de los catálogos centralizados por el CONICET, aunque se ocupará de referenciar a los archivos digitalizados institucionales distribuidos a lo largo de la Argentina.

En su inicio, el grupo de desarrollo se propuso realizar un relevamiento mediante una encuesta dirigida a los investigadores CONICET en la que se preguntaban aspectos generales sobre:

Uso: qué fuentes primarias usa el investigador (para relevar las tipologías generales, por ejemplo: documentos originales, libros antiguos, cartografía, fotografías, cartas, periódicos, grabaciones, observaciones de campo, etc.) y si estas están en formato analógico o digital. Por otro lado, se relevaba información sobre colecciones que el investigador usa, estén o no dentro de su institución.

Generación: qué tipos de datos primarios genera el investigador (se busca relevar tipologías generales, por ejemplo: bases de datos estadísticos, bases de datos documentales, entrevistas, historias de vida, registros de campo, objetos arqueológicos y antropológicos, etc.), y en qué grado de digitalización se encuentran. Asimismo, en qué formatos digitales los produce y con qué software, si tiene metadatos asociados y qué estándar usa.

Política de acceso: se le preguntaba al investigador qué tiempo considera que los datos necesitan estar embargados antes de publicarlos, y si ya están depositados en algún repositorio o archivo de acceso público.

Infraestructura de información y tecnológica de la propia institución: se le preguntaba sobre la existencia de archivos en su institución y el nivel de digitalización.

Los resultados de dicha encuesta pueden ser consultados en la página web de CONICET (CONICET, 2012).

Más recientemente, en 2014, se comienza a gestar en el ámbito del MINCyT el Programa de Digitalización y Acceso Abierto de Colecciones en Ciencias Sociales y Humanidades. Si bien se encuentra en etapa de conformación, aspira a tener alcance nacional y ser interinstitucional, aunque siempre restringido a los organismos que integran el Sistema Nacional de Ciencia, Tecnología e Innovación.

ISBN 978-987-3617-89-8

› *Proyectos en curaduría de datos en Humanidades y Ciencias Sociales. Problemáticas iniciales*

La curaduría digital atiende al ciclo de vida de los datos científicos desde el mismo momento de la creación. Por lo tanto, planificar proyectos en los que no se ha tenido oportunidad de intervenir tempranamente asesorando a los investigadores sobre cómo trabajar con los datos, es un aspecto que debemos estar dispuestos a resignar, al menos en esta etapa de nuestras instituciones en la que es prioritario comenzar a recuperar para preservar y dar acceso a conjuntos de datos y colecciones de fuentes que ya han sido generadas con mayor o menor pericia. Esto no significa que el trabajo con los investigadores sea dejado de lado, pero parece necesario iniciar el proyecto publicando algún tipo de contenido, para que luego se pueda incentivar al conjunto de investigadores a modificar prácticas de trabajo arraigadas.

Considerando que existen conjuntos de datos o colecciones de fuentes en el seno de nuestra institución que es de interés capturar para preservar para su uso futuro, el primer paso será planificar un relevamiento que permita identificarlas y obtener la información necesaria para realizar la selección y orden de prioridades para su procesamiento. Es menester realizar esto para dedicar recursos que garanticen su posterior preservación.

Problemáticas vinculadas a la selección

Revisión de la literatura

En el capítulo sobre el proceso de valoración (*appraisal*) y selección escrito por Ross Harvey (2007) para el Manual de Curaduría Digital, se reconoce el aporte que realizan tanto la Archivística tradicional como la Bibliotecología. Allí se define *appraisal* como el proceso de evaluación para determinar cuáles materiales serán retenidos en el archivo, cuáles serán retenidos durante un periodo específico y cuáles serán destruidos; mientras que define al proceso de selección, como la tarea de agregar materiales a la colección de una Biblioteca. Sin embargo, ambos procesos, que han sido ampliamente tratados para el mundo papel, son revisados por el autor con el nuevo sentido de curaduría digital. Así, mientras la práctica archivística tradicional aplica el criterio de valor administrativo, valor fiscal, valor legal, valor intrínseco, valor probatorio y valor informativo, y la Bibliotecología aplica valor probatorio, valor estético, valor comercial y valor de exposición, Harvey desarrolla 10 criterios que condicionan el proceso de selección: valor, condición física, recursos disponibles, uso, importancia social, derechos legales, cuestiones de formato, cuestiones técnicas de preservación, políticas y documentación que acompaña a los datos para que estos puedan ser accedidos en el futuro.

Al considerar que es la comunidad en la que está inserto el proyecto la que impone los requisitos de cantidad y naturaleza de los datos o fuentes seleccionados para la preservación, entiende que de la misma manera será la comunidad de usuarios quién defina el tipo de

ISBN 978-987-3617-89-8

información de contexto que se necesita preservar. Hay comunidades de determinadas disciplinas que necesitan más información de contexto que otras. Por lo tanto, Harvey considera de vital importancia emprender el proceso de selección con una clara comprensión de la comunidad de usuarios que originan los datos, la comunidad que los utiliza actualmente y quienes los podrán utilizar en el futuro. Otro aspecto que el autor considera que ha cambiado respecto al mundo impreso, donde la no preservación no suele causar pérdidas irreversibles, es el desarrollo técnico alcanzado hasta el momento en cuanto a la preservación de los datos digitales que nos lleva a seleccionar materiales en base a las posibilidades actuales. De manera similar, la selección de datos y materiales a preservar puede verse limitada por la legislación de propiedad intelectual de los diferentes países en un determinado momento.

En el mismo sentido, Niu (2014) destaca la importancia del *appraisal* como método de selección, a la vez que reconoce también la existencia de otros métodos como el muestreo estadístico y el análisis de riesgo. Su propuesta considera la metodología estructurada en función de tres aspectos centrales: los objetos que son susceptibles de ser valorados, los criterios con los que se realiza el *appraisal* y las decisiones que los curadores deben tomar. Respecto a lo que es objeto, enfatiza al igual que Harvey la importancia del contexto, pero aquí éste está visto de una manera quizá más amplia, lo que se denomina el *macro-appraisal*, donde lo que se tiene en cuenta son los recursos producidos por importantes productores o que son producidos en el marco de actividades importantes. Se realiza una valoración del contexto: contexto importante, contexto menor. Por otro lado son objeto del *appraisal* los propios recursos, lo que se conoce como *micro-appraisal* y aquí se debe revisar tanto el contenido, como la condición física y las características técnicas, por ejemplo obsolescencia del formato o de los medios de almacenamiento, dado que todo esto afecta directamente a las posibilidades de la preservación. Dentro del *micro-appraisal*, se contempla la posibilidad de realizar la valoración por partes específicas, es decir, se revisa cuáles son las propiedades significativas del recurso que se desean preservar. Otro objeto de análisis que plantea Niu (2014) es lo que tiene que ver con los metadatos y la documentación que acompaña al recurso, ya que esto afecta su valor y debe ser considerado en tanto costos y posibilidades de preservación. El autor manifiesta que en esto se pueden determinar dos niveles, en el primero, los metadatos y la documentación es útil para evaluar la autenticidad y la confiabilidad de los datos, pero no es vital para la preservación, uso y entendimiento de la información que proporciona el recurso. El segundo nivel es desde el punto de vista técnico de la preservación imprescindible, como también para el entendimiento intelectual. En el caso de las Ciencias Sociales, por ejemplo, son los libros de código, los procedimientos de muestreo, ponderación, etc.

Para Niu (2014), la Institución debe manejar unos criterios generales que tienen que ver con el alineamiento respecto a la misión institucional y a la política definida para la colección. Pueden existir recursos valiosos que deban ser dejados de lado por no responder estrictamente con este alineamiento. El criterio clásico de valor también se sigue teniendo en cuenta. Por un lado hay un valor que tiene sentido para los productores, tal como el valor administrativo, fiscal o legal, donde los factores de autenticidad, integridad, confiabilidad y precisión son indispensables. Por otro lado, hay un valor que tiene sentido para los usuarios como valor evidencial o informativo,

ISBN 978-987-3617-89-8

donde los factores de utilidad, usabilidad y accesibilidad son claves. También hay factores como la singularidad de los datos, su diversidad o representatividad que afectan a su valor. Los datos que son imposibles de recoger o es muy costoso volver a hacerlo, son buenos candidatos a ser preservados. Otro criterio es el costo, que puede estar involucrado en la adquisición y el alojamiento, y que seguro lo está en la preservación y el procesamiento.

Problemáticas vinculadas al relevamiento

Revisión de la literatura

Dentro del amplio marco de trabajo que establece la llamada Auditoría de Información en las organizaciones, uno de los instrumentos que están dirigidos a los datos es la *Data Audit Framework Methodology*. Conocida como DAF, es un proyecto que ofrece herramientas metodológicas y de software preparadas para colaborar en la auditoría de datos, lo que facilita a las instituciones reconocer y ubicar la información para su posterior explotación. Esta metodología establece 4 pasos: la planificación de la auditoría, la identificación y clasificación de los conjuntos de datos, la evaluación de la gestión de los datos, la comunicación de los resultados y las recomendaciones de cambios. En el primer paso, se sugiere identificar a las personas que poseen conjuntos de datos dentro de la organización y de ser posible entrevistarlos, así como recoger documentación que antes de la visita *in situ* le permita reconocer el sistema de registración que se maneja en la organización. Los documentos pueden ser de tipos muy variados: registros existentes de datos, informes finales de proyectos, informes anuales, informes de investigación, publicaciones de resultados de investigación, manuales de procedimientos, documentación de sistemas, etc. Al finalizar la primera etapa, debe haber quedado definido quién será el responsable de llevar adelante la auditoría, haber obtenido la aprobación de quien corresponda para llevarla a cabo y acordar con los involucrados el tiempo que demandará. La etapa de planificación implica sensibilizar al entorno con el proceso que se realizará y planificar las solicitudes de requerimientos para ahorrar el tiempo de los involucrados.

En la etapa de identificación y clasificación de los conjuntos de datos, lo importante es establecer qué tipo de datos existen, quién los tiene y dónde, además de clasificarlos en función del valor que poseen para la organización. Para la identificación de estos datos se pueden utilizar distintas técnicas entre las que se encuentra el análisis de fuentes documentales, la realización de encuestas por escrito o entrevistas personales a los miembros de la institución, la preparación de un inventario con la información de los datos recopilados, entre otras. Tanto las entrevistas como las encuestas le permitirán al auditor conocer los datos existentes en la institución a través de sus propios creadores, de quiénes los usan y los mantienen. Lo relevante de las entrevistas a los distintos integrantes es la posibilidad de solicitarles que realicen una valoración de los documentos, es decir que los clasifiquen dentro de las categorías:

Vital: su gestión y protección debe ser prioridad en la institución (proyectos en curso, datos que apoyan la investigación o que se usan para proporcionar servicios a clientes).

ISBN 978-987-3617-89-8

Importante: si bien la institución es responsable de esos datos, corresponden a investigaciones que ya están finalizadas, no tienen ningún tipo de actualización.

Secundario: son los archivos de datos que la institución posee pero que no tiene obligación ni necesidad de preservar.

Esta primera valoración dará una de las pautas de selección cuando el volumen de datos sea superior a los que la institución podrá conservar, permitirá determinar costos de mantenimiento y seguridad necesarios.

La tercera etapa denominada evaluación de la gestión de los datos apunta a la recogida completa de la información referida al conjunto de datos, para ello se desarrolla un formulario abreviado de 16 campos, o uno extendido de 50 ítems diferentes a registrar.

La cuarta y última etapa consiste en reportar resultados y hacer recomendaciones. Aquí se espera que el auditor presente un informe final en el que conste una breve descripción de la organización que fue auditada, el perfil de los conjuntos de datos basado en el inventario y la clasificación realizados y, finalmente, las recomendaciones para mejorar la gestión de los datos.

› *Hacia la construcción de un marco de trabajo en curaduría de datos orientado al IdIHCS - Instituto de Investigación en Humanidades y Ciencias Sociales (FaHCE-UNLP/CONICET)*

Problema 1

La primera cuestión a considerar tiene que ver con que muchas de nuestras instituciones, en tanto órganos de gestión de ciencia, comparten estas dos áreas de conocimiento: Humanidades y Ciencias Sociales que son muy cercanas pero diferentes entre sí, a la vez que cada una de ellas presenta diferencias notables en su propio seno.

Los institutos de investigación de nuestro país trabajan en alguna de estas disciplinas o en varias de ellas, por lo que las características de la información que manipulan necesariamente presenta características muy diferentes. Tal es el caso del IdIHCS, que agrupa líneas de investigación históricas, de las letras, sociológicas, políticas, de la filosofía, etcétera.

Acción 1

Objetivo: Identificar claramente cuáles son las líneas de investigación que se han desarrollado en el Instituto en los últimos años. La decisión de cuántos años hacia atrás se tomarán es una cuestión de política que queda supeditada a las características de las diferentes instituciones. En el caso del IdIHCS, si bien es un instituto de reciente creación (2009), sus antecedentes se remontan a más de 50 años de investigación concentrada en el ámbito de la Facultad de

ISBN 978-987-3617-89-8

Humanidades de la UNLP. Considerar entre un mínimo de 5 años y un máximo de 10 años sería adecuado en esta primera etapa.

Recursos: Publicaciones resultantes emanadas de los proyectos de investigación, tesis y sería deseable tener acceso a los informes internos e informes parciales y finales que presentan los integrantes de los proyectos.

Problema 2

Una de las características de la manera de producción académico-científica actual en nuestras instituciones es que por lo general los resultados de investigación cobran formalidad en la etapa final de difusión, la que corresponde a la publicación en forma de libro, capítulo de libro, artículo o presentación a una reunión científica (todas ellas susceptibles de ser preservadas en los Repositorios Institucionales). Es en estas producciones donde se hace referencia a la información primaria, entendiendo como tal aquella sobre la cual cada tipo de investigación elaboró sus resultados y que en términos muy generales corresponden a uno de estos dos tipos: "fuentes usadas" y "datos producidos" en el marco de la investigación. Cuando una investigación usa fuentes, puede ser que estas sean analógicas o digitales, y que se las haya consultado en un espacio físico o virtual más o menos institucionalizado (archivo, biblioteca o museo). Otras veces, dichas fuentes están en lugares inaccesibles con acceso restringido por ser propiedad de personas o ser documentación inédita. En esas situaciones es cuando el investigador o grupo de investigación inicia un arduo trabajo que comienza con la gestión de permisos para actuar sobre la colección, continúa con un relevamiento exhaustivo que implica diversas tareas de organización y preparación de los materiales, que incluye, en la mayoría de los casos, la generación de imágenes digitales y una mínima descripción. Todo ese material producido es referenciado en las publicaciones que produce el grupo, pero por lo general no adquiere una formalidad de colección por estar almacenado en computadoras personales, con escasa descripción, sin acceso público y cabría suponer, que con alto nivel de vulnerabilidad. Un panorama similar presentan los datos producidos, ya que solo se publican los resúmenes, mientras que la totalidad de los datos que les dieron origen permanecen en las mismas condiciones antes descritas.

Acción 2

Objetivo: Una vez determinado el recorte disciplinar que en el pasado y en el presente caracteriza a la institución, lo adecuado parecería ser identificar las personas o grupos de personas que trabajan o trabajaron esas líneas de investigación, de manera que se las pueda encuestar o entrevistar para recabar información sobre las colecciones de fuentes o los conjuntos de datos de los que disponen. Se sugiere aquí conformar con claridad los diferentes grupos (pueden ser unipersonales), de manera que los instrumentos de relevamiento puedan ser preparados *ad-hoc* dependiendo las características disciplinares de personas y productos.

Recursos: Nómina de personal de la institución actual y pasada, nómina de proyectos de investigación acreditados clasificados por líneas de investigación con mención del equipo

ISBN 978-987-3617-89-8

interviniente.

Problema 3

Un proyecto de curaduría de datos dentro de una institución siempre debe atender cuestiones relacionadas con la legislación vigente. En Argentina, el 13 de noviembre de 2013 fue sancionada la Ley 26.899 de Repositorios Institucionales de Acceso Abierto (2013). En dicha norma, aún sin reglamentar, el art. 2 establece:

> Los organismos e instituciones públicas comprendidos en el artículo 1°, deberán establecer políticas para el acceso público a datos primarios de investigación a través de repositorios digitales institucionales de acceso abierto o portales de sistemas nacionales de grandes instrumentos y bases de datos, así como también políticas institucionales para su gestión y preservación a largo plazo.

En el artículo 3 se refuerza la idea sosteniendo que:

> Todo subsidio o financiamiento proveniente de agencias gubernamentales [...] destinado a proyectos de investigación científico-tecnológica que tengan entre sus resultados esperados la generación de datos primarios, documentos y/o publicaciones, deberá contener dentro de sus cláusulas contractuales la presentación de un plan de gestión acorde a las especificidades propias del área disciplinar.

Mientras que en el artículo 5 dice:

> Los datos primarios de investigación deberán depositarse en repositorios o archivos institucionales digitales propios o compartidos y estar disponibles públicamente en un plazo no mayor a cinco (5) años del momento de su recolección […]

Esta intención de hacer pública la información primaria presenta para el campo de las Humanidades y Ciencias Sociales dos aspectos diferentes que deben revisarse con cuidado. El primero tiene que ver con los derechos de autor de aquellos materiales de los cuales se han tomado reproducciones digitales o se piensa tomarlas para hacerlas públicas. En este sentido es imprescindible asegurar que no se está transgrediendo la legislación vigente y se cuenta con los debidos derechos de publicación. El segundo aspecto tiene que ver con esa característica particular de muchas disciplinas de las Ciencias Sociales que generan información a partir de fuentes humanas. La información de identidad de las personas que la proporcionan debe desvincularse de los datos no solo para cumplir con la legislación vigente sino también para no incurrir en problemas éticos, ya sea a nivel de convenciones internacionales, legislación nacional o normativa impuesta por las agencias de promoción de ciencia y técnica nacionales. El tema del tiempo en el que debe hacerse público es otra cuestión que debe ser transparente hacia el interior de la institución.

Acción 3

Objetivo: Antes de iniciar cualquier tipo de relevamiento que involucre el tiempo de las personas, es aconsejable tener definida la política que regirá al sistema. Dicha política deberá ser

ISBN 978-987-3617-89-8

una declaración amplia, ya que previo al relevamiento se desconocen muchas características del escenario, pero a la vez muy estricta en cuanto a cuestiones de normas que no se transgredirán. Además deberá definir claramente quiénes podrán aportar datos primarios y qué tipo de características en términos legales deberán poseer esos datos. Sería el momento de hacer explícito que el compromiso de la institución respecto a la ingesta para el sistema de preservación será gradual y que se realizará priorizando criterios de valor científico o histórico, singularidad, no replicabilidad o alto costo de replicabilidad, calidad en la información de contexto que acompaña a los datos (metadatos y documentación).

Recursos: Normativa internacional, nacional e institucional que rige aspectos de propiedad intelectual, cuestiones éticas y políticas de acceso a la información. Modelos de políticas de otras instituciones.

Problema 4

La generación de proyectos nuevos dentro de las instituciones implica la articulación de diferentes actores. Por un lado están los gestores encargados de delinear y aprobar las políticas y conseguir los recursos. Están también los proveedores de datos, con todas las características diferenciales por disciplina mencionadas en el problema 1 y mayor o menor grado de interés en disponer del tiempo que se le pueda solicitar para trabajar sobre sus datos. Por otro lado se debe armar una infraestructura de publicación que necesariamente requiere de la participación de otro tipo de actores: bibliotecarios que trabajen los metadatos, informáticos, diseñadores de usabilidad. Es deseable que todo el entramado que se necesita poner en movimiento se apoye en estructuras existentes. Tal como se sostiene en un trabajo anterior (González, Pené y Unzurrunzaga, 2013) en muchas instituciones hay un *know how* interesante en las Bibliotecas, ya que son organizaciones que han tenido que resolver problemas técnicos, revisar aspectos legales, conseguir recursos financieros y articular recursos humanos para lograr construir sistemas de información de similares características como son los Repositorios Institucionales de acceso abierto. Ya se ha manifestado también que los conocimientos y modos de hacer de los bibliotecarios y archiveros son similares a los que se usan en la curaduría digital, aunque deben enriquecerse con el aporte de los investigadores. Son ellos quienes brindan el contexto de producción del dato y proyectan el significado que puede tener para otros académicos. En otro sentido, también se considera que de no dar participación a profesionales de apoyo en esta actividad de intermediación se corre el riesgo de desviar recursos humanos de investigación para el armado de este tipo de infraestructuras. En el caso de la Facultad de Humanidades de la UNLP, la Biblioteca de la institución ha desarrollado sendos proyectos de gestión de información de alto interés para nosotros: el repositorio institucional Memoria Académica (http://www.memoria.fahce.unlp.edu.ar/) y el archivo de autores destacados ARCAS (http://arcas.fahce.unlp.edu.ar:9090/arcas/portada). Este último en colaboración con personal del IdIHCS.

Acción 4

ISBN 978-987-3617-89-8

Objetivo y recursos: De lo expuesto hasta aquí se deriva que lo adecuado será integrar un grupo de trabajo que incorpore diferentes perfiles profesionales a la vez que articule claramente la participación de cada sector dentro de la Institución. Por el momento, en lo que concierne a la etapa de relevamiento, lo principal es plantear un buen nivel de comunicación con los investigadores, lo que implica que la coordinación general de toda la etapa de relevamiento se concentre en un solo sector. La propuesta en este caso es que sea el sector de apoyo a la investigación del IdIHCS. Sin embargo, para la elaboración de los instrumentos se requiere la conformación de un grupo en el que participen profesionales con conocimiento diversos: aspectos legales en el uso de la información (bibliotecarios o editores), cuestiones técnicas vinculadas a los formatos digitales, migraciones y software (informático) y especialistas en el armado de cuestionarios y entrevistas (sociólogos).

Problema 5

Para iniciar el trabajo de relevamiento será necesario preparar un plan y seguir una metodología que ordene las acciones. Asimismo, será necesario definir en qué medida pueden planificarse en simultáneo ciertas acciones y en qué momentos es necesario establecer una escala de prioridades. Por ejemplo, será necesario decidir, más allá de los trabajos y el diálogo previo que se haya mantenido con los investigadores, el momento de inicio del relevamiento en función de la disponibilidad de las personas para responder consultas o saldar dudas que surjan durante este proceso. Por otro lado, como hemos observado en las experiencias de muchas de las Organizaciones presentadas en el marco contextual, un criterio de inicio a considerar puede ser evaluar el grado de riesgo de pérdida de los materiales a relevar. Este análisis podrá sostenerse mediante la observación tanto de las líneas como de los equipos de trabajo. Al arribar a estas definiciones, recién se podrá generar un sistema de entrevistas y encuestas.

Acción 5

Objetivo: en esta primera etapa de relevamiento el objetivo es identificar y organizar un "plano general" de las características de los materiales existentes; especificar su formato de origen (papel, audio, fílmico, digital); su grado de relevancia para los equipos de trabajo o institutos, y su análisis de riesgo.

Recursos:

1. Grilla o tabla que considere tanto la materialidad de los datos relevados, como sus datos de filiación institucional, personal a cargo de su generación, referencias y valoración de su grado de relevancia. Como referente: formulario DAF adaptado a las particularidades de los casos locales.
2. "Informantes": miembros de equipos; Personal de apoyo abocado a la tarea de diseño de las herramientas de relevamiento, y al proceso de entrevistas.

ISBN 978-987-3617-89-8

› *Conclusiones*

Hay una demanda cada vez mayor de promoción de proyectos que recuperen y hagan pública la información primaria que se ha utilizado en las investigaciones, que en forma de fuentes documentales o conjuntos de datos elaborados se hallan en manos de los grupos de investigación o investigadores individuales. Esto involucra tareas de descripción y tratamiento de los datos con un sentido diferente a los impuestos por la propia investigación, cuestión que hace que los grupos, por lo general, no estén dispuestos a hacerlo sin el apoyo económico y técnico de sus instituciones. Por lo tanto, es la curaduría digital la que se presenta como una nueva especialidad capaz de intervenir en el armado de esta e-infraestructura.

Estos proyectos implican poner bajo consideración los diversos intereses de la partes: las agencias de investigación y su preocupación por la llamada ciencia abierta, los investigadores y el celo respecto a su trabajo, los herederos de fuentes documentales interesantes, las bibliotecas y su interés en ampliar servicios para el sector investigación. Solo con la armonización de todas estas partes se podrán obtener resultados satisfactorios.

Si bien los saberes de la Bibliotecología y la Archivística aportan un conjunto de técnicas imprescindibles, es cierto que también surgen nuevos problemas derivados de que el producto de información que se pretende lograr es diferente: mucho más enriquecido que la biblioteca tradicional y mucho más integrado que el archivo tradicional en un ecosistema informativo multidisciplinar.

Finalmente, se deben establecer prioridades frente a la abundancia de materiales y se deben manejar con previsibilidad aspectos diversos como el riesgo de pérdida, la obsolescencia, los tiempos de compromiso de los investigadores hacia ese conjunto de datos o fuentes. Identificar las líneas de investigación y sus investigadores, definir la política del futuro sistema, formar un grupo de trabajo y construir un instrumento de relevamiento de los materiales disponibles, son los primeros pasos a dar. Solo una buena planificación puede garantizar buenos resultados.

› *Bibliografía*

CONICET (2012). *Resultados Encuesta PLIICS.* Recuperado de http://www.conicet.gov.ar/wp-content/uploads/2012/10/DEFINITIVO-REVISADO-Y-CONTROLADO-Resultados-Encuesta-PLIICS-Procesamiento-de-los-730-casos.pdf. el 10/10/2014

González, C., Pené, M. & Unzurrunzaga, C. (2013). Proyectos conjuntos entre grupos de investigación y bibliotecas para la preservación digital y la difusión. El caso de un archivo de fuentes primarias de autores destacados. *Actas de la III Conferencia Internacional sobre Bibliotecas y Repositorios Digitales* (*BIREDIAL'13*) *y del VIII Simposio Internacional de Biblioteca Digitales* (*SIBD'13*). 15 a 17 de Octubre de 2013, San José, Costa Rica: Consejo Nacional de Rectores de Costa Rica (CONARE), la Universidad de Costa Rica (UCR) y el Laboratorio Nacional de Nanotecnología

ISBN 978-987-3617-89-8

(LANOTEC) del Centro Nacional de Alta Tecnología (CeNAT). Recuperado de http://biredial2013.ucr.ac.cr/index.php/Biredial2013/ai/paper/viewFile/15/46 el 10/10/2014

Harvey, R. (2007). Appraisal and Selection. En Ross, S. & Day, M. (Eds.), *Curation Reference Manual.* Recuperado de http://www.dcc.ac.uk/resource/curation-manual/chapters/appraisal-and-selection el 10/10/2014

Jones, S., Ross, S., & Ruusalepp, R. (2009). *Data Audit Framework Methodology, draft for discussion.* Glasgow: HATII, University of Glasgow. Recuperado de http://www.data-audit.eu/DAF_Methodology.pdf el 10/10/2014

Lee, C. A. & Tibbo, H. R. (2007). Digital Curation and Trusted Repositories: Steps Toward Success. *Journal of Digital Information*, 8(2). Recuperado de https://journals.tdl.org/jodi/index.php/jodi/article/view/229/183 el 10/10/2014

Ley 26899 (2013). Creación de Repositorios Digitales Institucionales de Acceso Abierto, Propios o Compartidos, 32781 BO. Recuperado de http://repositorios.mincyt.gob.ar/recursos.php el 10/10/2014

Mayer, L. (2009). *Safeguarding Collections at the Dawn of the 21st Century: Describing Roles & Measuring Contemporary Preservation Activities in ARL Libraries.* Washington, DC: Association of Research Libraries. Recuperado de http://www.libqual.org/documents/admin/safeguarding-collections.pdf el 10/10/2014

National Science Foundation and the Library of Congress (2003). *It's about Time: Research Challenges in Digital Archiving and Long-term Preservation.* Recuperado de http://www.digitalpreservation.gov/documents/about_time2003.pdf el 10/10/2014

Niu, J. (2014). Appraisal and Selection for Digital Curation. *International Journal of Digital Curation.* 9(2), 65-82. Recuperado de http://www.ijdc.net/index.php/ijdc/article/view/9.2.65/370 el 10/10/2014

Walters, T. & Skinner, K. (2011). *New Roles for New Times: Digital Curation for Preservation*. Washington, DC: Association of Research Libraries. Recuperado de http://www.arl.org/storage/documents/publications/nrnt_digital_curation17mar11.pdf el 10/10/2014

ISBN 978-987-3617-89-8