Mesa 3: Datos de investigación en abierto

Ianko López Ortiz de Artiñano

Director técnico, Consorcio Madroño (España)

Humberto Blanco Castillo

Universidad del Rosario (Colombia)

Raúl Sifuentes Arroyo

Unidad de Automatización del Sistema de Bibliotecas, Pontificia Universidad Católica del Perú (Perú)

Rafael Port da Rocha

Proyecto Red de Datos de Investigación Brasileña (RDP), Universidade Federal do Rio Grande do Sul (Brasil)

Modera: **Malgorzata Lisowska Navarro**

Universidad del Rosario (Colombia)

Resumen extendido

Ianko López Ortiz de Artiñano (España): Hay una cuestión que relevante cuando hablamos del Consorcio Madroño y es que en su ADN está incluida la ciencia abierta; de hecho en el plan estratégico actual se establece como misión básica principal del Consorcio proporcionar una infraestructura de información que impulse la excelencia de la investigación en las instituciones miembro, contribuyendo al desarrollo de la ciencia abierta, a la transformación digital y al desarrollo sostenible. Mejorarando la experiencia de sus usuarios. Es decir que la propia misión del Consorcio ya incluye la ciencia abierta y asimismo la visión, porque lo que queremos es sr una organización líder y, entre otras cosas, buscamos la promoción de la ciencia abierta, la transformación

digital y el desarrollo sostenible. El Consorcio se creó en 1999, con siete miembros originales, todos con los objetivos de incrementar la producción científica de sus universidades, mejorar la calidad de los servicios y ahorrar costes, promover planes de cooperación entre sus miembros y la aplicación de nuevas tecnologías. Por lo que se refiere específicamente al primero acceso abierto, después a la ciencia abierta, se ha seguido un determinado determinado camino, que comienza en 2005 con la definición de e-ciencia, que es una red de repositorios institucionales de acceso abierto de tal manera que cada universidad miembro del consorcio y también otras instituciones mantienen sus repositorios de acceso de acceso abierto de contenidos pero el Consorcio Madroño los aglutina en una red de repositorios común, apoyada por la Comunidad de Madrid, tanto financiera como operativamente, que es **e-ciencia**. A lo largo del tiempo en estos últimos 15 años pues han ido sucediendo varios pasos, entrar en cada uno de ellos posiblemente sea una tarea muy exhaustiva, pero es especialmente relevante llegar a 2016, cuando se crea **e-cienciaDatos**, que es un repositorio institucional para almacenar datos yponer en abierto datos de investigación, un proyecto pionero en el ámbito español. Inmediatamente después, todos los servicios relativos a ciencia abierta se integran en un mismo portal llamado **investigaM**, que a su vez está inserto en la web del Consorcio Madroño. Este portal incluye tanto e-ciencia, como e-cienciaDatos, como la herramienta PaGoDa de PDG on line que permite a los investigadores la realización de planes de gestión de datos según las premises establecidas por Horizonte 2020 de la Comunidad Europea, ahora Horizonte Europa. También en 2017 el Consorcio Madroño emite una declaración de apoyo la ciencia abierta con recomendaciones para los investigadores, para las universidades y también para los gobiernos y las instituciones públicas donde, y esto es importante, se mencionan ya expresamente los datos de investigación. En la actualidad hay aproximadamente unos 570.000 registros en e-ciencia, en e-cienciaDatos unos 500 a 600 datasets, 850 PDG y seguimos creciendo en todos estos aspectos. Como comentaba anteriormente, todos estos servicios y toda la infraestructura de ciencia abierta del Consorcio Madroño, todos estos desarrollos se integran

en el portal investigaM, que es el portal de ciencia abierta. Específicamente, e-cienciaDatos se estructura como un sistema que está constituido por varias comunidades que agrupan los datasets de una de las universidades miembro. E-cienciaDatos ofrece por tanto el depósito y también la publicación de conjuntos de datos asignándole un identificador de objeto digital (DOI) a cada uno de estos datasets. La asociación de un dataset a un DOI facilita la verificación de los datos, la diseminación, su reutilización, el impacto y el acceso a largo plazo. Además el repositorio provee una cita normalizada para cada dataset, lo que contiene información suficiente para que éste pueda ser identificado y localizado en un momento dado incluyendo el DOI. Los datos, por tanto, están en acceso abierto. Como decíamos, se trata de un repositorio de datos multidisciplinar, que alberga conjuntos de datos científicos de los investigadores de las universidades miembro del Consorcio Madroño, para darles visibilidad, garantizar su preservación y facilitar su acceso y reutilización. Se basa en el sistema Dataverse, con una comunidad/dataverse por cada universidad, gestionada por la propia universidad, y comunidades internas por proyectos. También tiene un gestor de estadísticas. Además de esto, cuenta con el sistema de PDG on line que, como decía, es una herramienta que permite a los investigadores realizar, de manera tutorada, planes de gestión de datos según las premisas establecidas por Horizonte 2020, hoy Horizonte Europa.

JUAN CORALES (España): Voy a presentar el repositorio de datos e-cienciaDatos, lo haré muy rápidamente porque vamos justos de tiempo. E-cienciaDatos es un repositorio de datos muy útil y multidisciplinar. Pueden subir sus datos los investigadores de todas las universidades del Consorcio Madroño. Tenemos una comunidad por cada una de las universidades que forman parte del Consorcio y está dirigido a las necesidades de nuestros investigadores; principalmente, ofrece servicios como geolocalización, ya que la mayoría de los datasets están localizados, ofrecen por supuesto una dirección de correo de soporte, un manual, y otra característica que nos han pedido siempre los investigadores es que sus datos tienen que ser visibles. Por esta razón es recolectado por los principales harvesters de datos, como el

propio Dataverse de Harvard y otros; también, otra funcionalidad que necesitan nuestros investigadores, es que sea un repositorio válido tanto para las principales editoriales como para las agencias de financiación. Entonces nos hemos asegurado de que e-cienciaDatos pueda ser utilizado para almacenar los datasets de investigación; es decir, si alguna revista pide que estén en un repositorio temático. Si vemos ya un dataset en concreto, pues también damos otras herramientas interesantes para los investigadores, como estadísticas del dataset, por supuesto, lo que ha nombrado Ianko del DOI, que es indispensable, una cita estándar que se puede exportar en varios formatos y un potente gestor de versiones. Las estadísticas, aparte de ser importantes para los investigadores, también son importantes para nosotros como gestores de datos; también tenemos acceso a ellas en dos formas: una es la propia de Dataverse que nos da información sobre todo de cómo se usa e-cienciaDatos, a nivel de cómo ha ido evolucionando el uso de ficheros de datos, de las principales áreas temáticas y también tenemos unas estadísticas de Matomo, que nos da muchísima información sobre desde donde se accede, sistemas operativos, navegadores, muchísima más información de la que podemos navegar.

RAFAEL PORT DA ROCHA (Brasil): Nosotros pertenecemos al Grupo de Trabajo Rede de Dados de Pesquisa (GT-RDP) del Centro de Documentação e Acervo Digital da Pesquisa (CEDAP), que tiene como objetivo la digitalización y repositorios de datos de investigación. Tenemos un enfoque, un framework conceptual en el que se trabaja con el objetivo y la perspectiva de ser un repositorio digital confiable. Nos preguntamos inicialmente: ¿dónde está la información, los criterios, los requisitos que se tienen que identificar para planear y desarrollar el repositorio? Encontramos el modelo Open Archives y un mecanismo de certificación al que deberíamos someter el repositorio (FAIR, ACTDR ISO 16363 e ISO 14721). Nos propusimos entonces montar el proceso de desarrollo del repositorio para que todo lo que no se había conseguido lograr hasta entonces se comenzara a realizar con la idea de obtener un repositorio digital confiable. Pensamos los pasos de implementación del repositorio

teniendo en cuenta los requisitos de repositorios digitales confiables. Un repositorio no es solamente un software y las normas de repositorios son muy importantes para la gestión de riesgo. ¿Dónde están los riesgos que hacen peligrar la continuidad de un repositorio? Los riesgos son tecnológicos pero también organizacionales. La investigación que llevamos a cabo abarcó DOI, Data Cite, Dataverse, infraestructura, archivemática y preservación digital, entre otros aspectos. Investigamos también la comunidad académica productora de los datos. Trabajamos mucho respecto a la comunidad productora de datos y encontramos que muchos investigadores desconocen el contexto de repositorios y repositorios de datos. Estudiamos Dataverse y DSpace y los trabajos de ingeniería de software y criterios de evaluación de software con el objeto de tener un cierto dominio, seguridad y argumentos para el uso de esas herramientas enfocadas en repositorios de datos. Los repositorios de datos son algo nuevo con respecto a los repositorios de documentos, que se conoce bien cómo funcionan, cómo se producen y se consumen los documentos, las lógicas de los artículos, las tesis, las tesinas y cómo se produce ese conocimiento. Con respecto a los repositorios de datos buscamos conocer cómo se produce ese conocimiento, esos datos, para no crear un repositorio que no sea usado. Como resultado de los primeros estudios y en conjunto con la Red Nacional de Investigación (Rede Nacional de Pesquisa) y con el IBICT en el lanzamiento de un proyecto de incubación de repositorios. Estamos apoyando a cuatro instituciones para el desarrollo de cuatro repositorios con esta perspectiva. Se colocaron ciertos requisitos, responsabilidad, autodirección, equipo de ciencias de la información y equipos de investigación.

HUMBERTO BLANCO CASTILLO (Colombia): La experiencia de la Universidad del Rosario arrancó con un diagnóstico de un grupo de investigadores, con el fin de conocer la forma en que ellos estaban generando sus datos, dónde los estaban almacenando y qué hábitos tenían sobre los datos, sobre su respaldo y demás. A partir de esto realizamos una encuesta y luego fuimos realizando diferentes ejercicios para obtener esta información, validar algunos supuestos iniciales y entender ese comportamiento que tenían los investigadores para poder

incluirlo en nuestra estrategia de gestión de datos. De aquí salió una trivia que llamamos «Datos de investigación: ¿ya eres un pionero?», una encuesta que llamamos «¡Es el momento de tus datos de investigación!». También hicimos un *focus group* con los investigadores que detectamos que ya estaban inmersos en algún proceso que requería de datos de investigación. Ya con esta información, lo que encontramos dentro de los hallazgos fue que un porcentaje alto de los profesores no incorporaba los planes de gestión de datos dentro de sus proyectos, que los datos estaban almacenados generalmente en su computador personal o en una memoria USB sin ningún respaldo o los tenían en la nube, muchos de sus buzones personales. Respecto a compartir los datos los compartían únicamente con sus redes de colaboración y tenían algunas inquietudes respecto a reusar datos de otros. Con esta información más o menos organizada se inició el despliegue de la estrategia de gestión de datos y el primer paso consistió en la creación de un marco normativo a través del cual se pudieran establecer lineamientos para preservar y gestionar estos datos. Así es como en 2019 creamos la política institucional para la gestión de datos de investigación, con el cual se promovía esta apertura de los datos, acogiéndonos al principio de «tan abierto como sea posible, tan cerrado como sea necesario». Dentro del documento se reconoce la importancia de los datos de investigación, se hace la definición de algunos principios, y se definen los compromisos de la institución y los compromisos de los investigadores. De parte la institución, a brindar el acompañamiento y proveer los mecanismos para el almacenamiento y la identificación de los datos, y por parte de los investigadores, proveer datos de calidad. El siguiente paso fue crear el repositorio de datos y los servicios asociados. Respecto a la implementación del repositorio, se realizó directamente con el equipo técnico que tenemos en el CRAI; el software que utilizamos, al igual que en las dos presentaciones anteriores en esta mesa, utilizamos Dataverse, porque se adaptaba mucho mejor a los requerimientos que habíamos planteado, en términos de los aspectos básicos de almacenamiento, explorar los datos, enlazarlos, citarlos, el tema del versionamiento, que también ya se mencionó, y además estábamos buscando una plataforma que fuera de código abierto, que contara con una

comunidad robusta y que permitiera, digamos, «tocar» el código de una manera más fácil para el desarrollo de plugins y que contara con buenos estándares de interoperabilidad; por estas razones seleccionamos Dataverse. Luego, una vez que seleccionamos el software que íbamos a utilizar, debíamos definir la estructura a través de la cual íbamos a almacenar los datasets dentro del repositorio y, luego de explorar varias opciones, la que mejor se ajustaba a la forma como íbamos a desplegar el servicio era a través de proyectos: es decir cada dataverse corresponde a un proyecto de investigación y dentro de él se colocan cada uno de los datasets. También era importante definir cuál iba a ser el modelo de administración y para esto se determinó que el investigador principal, previo una capacitación que se le entrega, respecto a la administración interna del repositorio, iba a ser el administrador del dataverse, pues él ya iba a generar los permisos para cada uno de los investigadores o colaboradores para que subieran su información. Esto porque dependiendo de cada proyecto, la forma de gestionar los datos va a ser diferente o bien los proyectos pueden estar conformado por diferentes equipos, por diferentes grupos y por lo tanto pues variaría ese modelo de de permisos. En noviembre de 2019 pudimos poner a disposición de la comunidad el primer dataverse en abierto, llamada «Leishmania in the Américas DB». Dentro de este proceso de implementación, también se realizó la implementación de algunos plugins, como los de visualización de estadísticas, la visualización de hojas de datos tabulares y, en este caso, por ejemplo, los servicios de datos de geolocalización. Cuando hicimos la migración a la versión 5 tuvimos algunos temas respecto al despliegue, porque Dataverse cambió el modelo de visualización de estos datos. Otro aspecto importante fue la adquisición del DOI para la identificación de los datasets; en este caso nos suscribimos a Dataverse como miembros y esto permite, una vez que se configura, que los DOI se asignen directamente sobre los datasets una vez que se publican y se registra directamente ante DataCite y, además, esta información queda indexado directamente en Google Datasets, entonces esto nos pareció una de las cosas muy ventajosas para mostrar a nuestros investigadores. Respecto a los servicios, diseñamos e implementamos servicios que nos permiten soportar

algunas fases del ciclo de vida de los datos; por ejemplo, la primera fase, que es la fase de planeación, estábamos apoyando a los investigadores con la construcción del plan de gestión de datos y realizamos un acompañamiento a los profesores para la construcción de este documento; respecto a la etapa de almacenamiento y difusión ya se implementó un servicio de capacitación y asesoría, como les mencioné anteriormente, primero para uso del repositorio pero también hay aspectos de sensibilización acerca de la importancia de la apertura de los datos, el uso de los datos de investigación y la publicación de los datos.

RAÚL SIFUENTES ARROYO (Perú): Voy a compartir la experiencia de la Pontificia Universidad Católica del Perú en datos abiertos. Menos mal que los colegas que me han antecedido ya presentaron algunos conceptos relacionados a esto, porque como corre el tiempo, he sacado algunas cosas que ya se han tocado. Vamos a ver por qué escogimos Dataverse, cuáles son las características actuales que tiene el repositorio y cuál es la relación que tiene con el repositorio institucional, qué proyectos están en marcha y cuáles son los futuros pasos. y Nosotros escogimos Dataverse, en realidad, por una experiencia previa que antecede más o menos al año 2013, cuando lanzamos el repositorio institucional de la Pontificia Universidad Católica del Perú. Hicimos también una instalación de Dataverse porque en ese momento un instituto de investigación buscó el apoyo de la biblioteca para poder hacer la implementación y ellos conocían muy bien Dataverse, tenían experiencia ya por haber utilizado esto en Harvard. Sin embargo, después del lanzamiento del repositorio el Dataverse que que tuvimos se tuvo que dejar de lado, porque la persona que movió todo este tema salió de la universidad y se fue a otro trabajo, entonces nosotros nos quedamos con el tema de Dataverse por allí, digamos un poquito con la miel en la boca pero ya lo habíamos experimentado, ya habíamos hecho la instalación y todo. Pasaron seis años para que otro instituto de investigación nos busque y nosotros le ofrecimos, como una solución, el Dataverse para lo que ellos estaban buscando. Entonces en 2019 hicimos el lanzamiento del Dataverse porque necesitábamos tener esa

experiencia. ¿Por qué nos metimos en ese tema? Porque la legislación peruana así como ya exige tener en las instituciones repositorios de tesis o repositorios institucionales, también en algún momento ya va a empezar a exigir que las instituciones como las universidades tengan repositorios de datos, entonces queríamos tener esa experiencia, tuvimos el aliado pertinente para ese momento y lo lanzamos en septiembre del 2019. después de haber analizado todos los programas que hay en el mercado escogimos Dataverse porque para nosotros tenía los metadatos muchos más desarrollados en las áreas que nosotros tenemos en la universidad, que son las ciencias sociales y humanidades, que es muy fuerte en nuestra universidad, astronomía, astrofísica, ciencias de la vida y espaciales, y además revistas. Es decir, cumplía con nuestros requerimientos y además vimos que había una creciente comunidad en español. Me alegra ver que ahora hay una nueva instalación en Ecuador; en Perú tenemos dos oficialmente, la de la Universidad Católica y un Centro Internacional de la Papa en Colombia, hay otra en Chile; vemos que en Brasil hay varias instalaciones, hay en México también y en el Caribe una universidad de las Antillas también. Una de las cosas por las que nos elegimos Dataverse también fue que se integraba con el OJS. ¿Qué características actuales tiene? Ya mencioné que fue lanzado en septiembre del 2019, tenemos actualmente 6 repositorios, 92 conjuntos de datos, la mayor parte de estos conjuntos de datos son entrevistas realizadas por este Instituto de Opinión Pública que a lo largo de su historia ha venido haciendo encuestas electorales y encuestas sobre temas importantes para la coyuntura del país, algunas veces encargadas por terceros y otras veces por la misma motivación de la universidad. Actualmente tenemos 541 archivos y utilizamos como ID persistente el handle; estamos utilizando Amazon Web Services para el software alojado ahí y algo que les quiero comentar es que tanto el repositorio de datos como el repositorio institucional, las revistas y todo todo lo que tenga que ver en realidad con la ciencia abierta está gestionado bajo el Sistema de Bibliotecas de la universidad que tiene a su cargo los servidores y además el servicio profesional para brindar los servicios desde el punto de vista de las bibliotecas. Entonces tenemos ahí un área de informática especializada, que es

el área que yo dirijo para todos estos temas. Otra de las características actuales es que estamos indizados en Google Dataset Search, como mencionó mi colega de la Universidad del Rosario, ellos utilizan el Data Cite y eso les permite entrar directamente en Dataset Search. Nosotros tenemos que actualizar, creo que es una primera vez se tiene que ir mandar un archivo XML a Google Dataset para la indexación. Hasta el día de ayer hemos registrado 5.598 descargas de los seis repositorios que mencioné. Hay una relación con el repositorio institucional, pues todos los datasets están cosechados por el RI porque el RI es la fuente principal de metadatos para el nodo país que vendría a ser el CONCYTEC, que es un aportante de LA Referencia. Esos metadatos se cosechan del repositorio institucional y se enriquecen según la legislación vigente nacional. Tenemos una legislación que nos exige la presencia de ciertos metadatos. ¿Cuáles son los proyectos en marcha y a futuro? Ahora mismo tenemos Mapas del Perú: se están creando bases de datos de mapas del siglo XIX o XX que les sirven a varios investigadores para poder estudiar justamente cómo han cambiado los mapas; hay mapas mineros, mapas de petróleo, mapas de carreteras, de caminos, etc. Incluso recuerdo que hace poco nos pidieron unos mapas para resolver un conflicto de tierras que se había originado en una zona del norte del país. Tenemos el Archivo Digital de Lenguas Peruanas que son estudios, audios, videos, transcripciones de documentos fonéticos, diccionarios, y es el único archivo digital de las lenguas nativas del Perú. Ha sido llevado a cabo por los lingüistas de la universidad y también otras investigadores. Tenemos bases de datos de poetas y eso más o menos representaría 50 datasets y 800 archivos. Para terminar, estamos elaborando los archivos bibliográficos en formato RIS de todas las revistas para ponerlos on line y para que generen interés en investigación bibliométrica de todas nuestras revistas. También estamos pensando poner datos transaccionales estadísticos y de gestión de las bibliotecas porque nos piden mucho las personas que quieren hacer estudios de usuarios, y también vamos a integrar las revistas de la universidad. Además algo que está muy cerca es que se va a incluir los datos en una futura política de ciencia abierta en la que vamos a mirar la experiencia de la Universidad del Rosario que nos presentó este año en

algún evento hace unos pocos meses. Como ustedes pueden ver, tenemos datos de la investigación y también para generar investigación; también datos crudos que buscamos que generen investigación como en el caso, por ejemplo, de los registros bibliográficos de las revistas.