BIREDIAL-ISTEC 2022
XI Conferencia Internacional Bibliotecas y Repositorios Digitales
Del 3 al 7 de octubre de 2022

Evaluación de un repositorio institucional a través de NDSA Levels: Caso CIC Digital

Tettamanti, Santiago; PREBI-SEDICI, Universidad Nacional de La Plata; CESGI, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires; santit@sedici.unlp.edu.ar.

De Giusti, Marisa R.; PREBI-SEDICI, Universidad Nacional de La Plata; CESGI, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires; marisa.degiusti@sedici.unlp.edu.ar.

Lira, Ariel Jorge; PREBI-SEDICI, Universidad Nacional de La Plata; CESGI, Comisión de Investigaciones Científicas de la Provincia de Buenos Aires; ariel.lira@sedici.unlp.edu.ar.

Palabras claves

Repositorio institucional; Preservación digital; Estándares de evaluación; Autoevaluación.
Institutional repositorio; Digital preservation; Evaluation standards; Self-evaluation.

Eje temático

Infraestructura tecnológica

Resumen

Este trabajo está dedicado a relatar la experiencia de autoevaluación de un repositorio institucional realizada utilizando NDSA, se inicia a partir de la exposición del contexto de los repositorios institucionales y los estándares utilizados para auditar la confiabilidad de los mismos en cuanto al mantenimiento del acceso y comprensión de sus contenidos a largo plazo. La propuesta de niveles de NDSA abarca diversas áreas del repositorio, es muy comprensible y brinda un primer acercamiento al tema de preservación más sencillo que otros estándares como la ISO 16363. El trabajo construye la matriz de resultados del test del repositorio CIC Digital así como una propuesta de acciones para poder mejorar los aspectos no cumplidos o parcialmente cumplidos.

Introducción

Contexto y antecedentes

Los repositorios institucionales de acceso abierto se han convertido en herramientas esenciales para la comunicación académica en la era digital, además de servir como una colección digital que agrupa la producción académica y científica de una institución, ofrece servicios sobre esta producción y asegura el acceso abierto y a largo plazo de sus obras. Para garantizar dicho acceso, la preservación digital juega un papel fundamental. Según Ferreras-Fernández (2010), la preservación digital se refiere a la conservación para asegurar la accesibilidad, la recuperación y el uso a futuro de los materiales digitales, y a las distintas técnicas y actividades dirigidas a alcanzar esos objetivos. Según la autora, para que un repositorio logre la perdurabilidad de sus objetos, no solo se necesitan esfuerzos técnicos sino que existen otros retos: legales (permiso de los autores), económicos e institucionales.

Térmens y Leija (2017) explican que una parte vital del proceso de preservación digital en un repositorio es poder valorar si los sistemas, actividades y técnicas de preservación que se utilizan son los correctos, para ello es necesario utilizar distintas herramientas y metodologías que sirven para evaluar el grado de cumplimiento con estándares y buenas prácticas existentes. Los autores del trabajo mencionado en este párrafo, señalan además a los sistemas de auditoría como una de las metodologías más extendidas para conseguir repositorios confiables, éstos permiten determinar si un repositorio de preservación es seguro, y por tanto, si es confiable, pero están diseñados para

evaluar sistemas de preservación existentes y no para la detección de acciones a implementar para lograr la preservación, objetivo común a muchas organizaciones pequeñas. Sumado a esto, el uso de auditorías tradicionales suele ser caro y de una complejidad alta que requiere personal experto para llevarlas a cabo. En ese sentido, las organizaciones pequeñas quizás prefieran otro tipo de sistemas de evaluación, con mayor facilidad de aplicación y con métodos que permitan la autoevaluación.

Sobre NDSA Levels

La NDSA (National Digital Stewardship Alliance) fue creada en 2010 como una iniciativa de la Library of Congress, se trata de un consorcio de organizaciones comprometidas con la conservación a largo plazo de la información digital. A la fecha cuenta con 267 instituciones miembro, entre las que se encuentran universidades, organizaciones gubernamentales, organizaciones sin fines de lucro, empresas comerciales y asociaciones profesionales.

En 2013 la NDSA creó Levels of Digital Preservation (en 2019 fue lanzada la versión 2.0), un recurso para ayudar a aquellas instituciones o personas que deseen evaluar su sistema de preservación digital. NDSA Levels se compone de una matriz sobre la cual se definen tareas, actividades o normas que el repositorio o sistema debería llevar a cabo para asegurar la preservación y el acceso a largo plazo. Estas tareas se dividen en cinco categorías distintas:

- Almacenamiento y localización geográfica
- No alteración de ficheros/archivos e integridad de los datos
- Seguridad de la información
- Metadatos
- Formatos de ficheros/archivos

Por cada categoría se definen a su vez cinco niveles de cumplimiento de las tareas. La estructura general de la matriz es progresiva: las acciones en el primer nivel son requisitos previos necesarios para aquellos en los niveles de segundo a cuarto o son en sí mismas las actividades más urgentes a lograr. En términos generales, a medida que se avanza en cada uno de los niveles del Nivel 1 al Nivel 4, se está moviendo de la necesidad básica de garantizar la conservación de bits, a requisitos más amplios para realizar un seguimiento del contenido digital y poder garantizar que pueda estar disponible a través de períodos más largos de tiempo. En el anexo 1 se deja la traducción al español de la matriz de niveles de preservación digital.

Se debe tener en cuenta que la propuesta de NDSA se centra en la posibilidad de una autoevaluación por parte de los repositorios, ofrece una visión simplificada de las tareas de preservación y no un listado exhaustivo de las mismas como, por ejemplo, se puede encontrar en Nestor, TRAC o ISO 16363. En consecuencia, las respuestas pueden ofrecer una visión más optimista de la que correspondería a la situación real, ya que se ignoran otras facetas claves como la robustez institucional o la sostenibilidad económica (Término y Leija, 2017) que sí cubre por ejemplo ISO 16363.

La NDSA provee, además de traducciones a distintos idiomas de la matriz, una guía de evaluación en la que documenta cómo se debería evaluar un sistema con el uso de la misma, junto con una plantilla en formato de hoja de cálculo (NDSA Levels of Preservation Assessment Subgroup, 2019) a modo de ayuda para completar en qué medida nuestro sistema cumple con las tareas o requerimientos presentes en la matriz.

Evaluación de CIC Digital

Metodología

Se realizó una evaluación a partir de la matriz de preservación de NDSA del repositorio institucional CIC Digital de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires. Si bien se sugiere responder de manera afirmativa o negativa a las recomendaciones que se encuentran en la matriz, puede suceder que para alguno de los puntos planteados solo se haya completado una parte del mismo, o si el punto está dividido en subtareas, se hayan completado alguna de ellas, es decir, no se satisface el punto en su totalidad pero, a la vez ya se ha implementado alguna parte de lo recomendado. Para tener en cuenta dichos casos, se decidió por seguir la metodología propuesta por Térmens y Lejía, la cual consiste en marcar con un color especial las respuestas afirmativas en la tabla, pero con el agregado de marcar con otro color las respuestas en las que se ha completado el punto de manera parcial, y con otro las respuestas negativas. Las respuestas que se marcaron como completadas de manera parcial no sumaron puntos en el recuento final de resultados, solo sumaron puntos las respuestas afirmativas. De esa manera se puede apreciar rápidamente y a golpe de vista los resultados de manera general.

La evaluación de cada uno de los puntos planteados y el resultado final de cada uno fue tomado de manera conjunta por varios integrantes de los distintos equipos que trabajan en CIC Digital. Se realizaron sucesivas reuniones con al menos una persona encargada de alguna de las actividades involucradas en la preservación digital en el repositorio, o al menos que sean referenciadas por la matriz de evaluación. Entre los equipos se pueden mencionar al equipo de infraestructura del repositorio, responsable entre otras cosas de las copias de seguridad, chequeos de integridad, del mantenimiento y seguridad de los distintos servidores; el equipo de desarrollo, encargado de los chequeos por software (por ej, virus, integridad y formato de archivos), de la seguridad del sistema, de la base de datos, de la curación automática de los metadatos, etc; el equipo administrativo, que se ocupa de controlar el contenido que ingresa al repositorio, los tipos y cantidad de metadatos usados, la variedad de formatos; y los encargados de la gestión del repositorio, responsables de la visión general de este, las políticas presupuestarias y las decisiones institucionales. Puede darse el caso, en especial en repositorios pequeños que no poseen gran cantidad de personal disponible y que una misma persona realice actividades correspondientes a varios de estos equipos. En este trabajo, se debatió de manera conjunta acerca de el cumplimiento o no de las recomendaciones de la matriz, se realizó una evaluación de cada punto por separado y luego, en caso de no cumplimiento del requerimiento planteado, se fijaron las acciones a llevar adelante. Para algunas recomendaciones, por lo general las correspondientes al nivel 4 de la matriz, no fue posible plantear una solución viable para el repositorio, muchas veces esto ocurre por falta de recursos de infraestructura tecnológica o de presupuesto.

Matriz de evaluación NDSA Levels para CIC Digital

Aquí se presenta el análisis de preservación digital sobre el repositorio CIC Digital, realizado con la matriz de evaluación de NDSA. El color verde significa que la recomendación está satisfecha en su totalidad, el amarillo indica que gran parte del punto planteado fue resuelto pero que hay alguna parte de la recomendación que no fue satisfecha en su totalidad, y el color rojo muestra aquellos puntos que no fueron resueltos aún por el repositorio.

	Nivel 1 (Proteja sus datos)	Nivel 2 (Conozca sus datos)	Nivel 3 (Controle sus datos)	Nivel 4 (Repare sus datos)	Puntaje
Almacenamiento	Tener dos copias completas en ubicaciones separadas	Tener tres copias completas con al menos una copia en una ubicación geográfica distinta	Tener al menos una copia en una ubicación geográfica con amenaza de desastre diferente a las otras copias	Tener al menos tres copias en ubicaciones geográficas distintas, cada una con una amenaza de desastre diferente.	**0/4**
	Documentar todos los medios de almacenamiento donde esté almacenado el contenido	Documentar el almacenamiento y medios de almacenamiento, indicando los recursos y las dependencias que estos requieren para funcionar	Rastrear la obsolescencia del almacenamiento y los medios.	Maximizar la diversificación del almacenamiento para evitar puntos únicos de falla	
	Poner el contenido en soportes de almacenamiento estables		Tener al menos una copia en un medio de almacenamiento de diferente tipo	Tener un plan y realizar acciones para abordar la obsolescencia del hardware, software y medios de almacenamiento	
No alteración de archivos e integridad de los datos	Verificar que la información de integridad se ha proporcionado con el contenido	Verificar la información de integridad al mover o copiar contenido	Verificar la información de integridad del contenido en intervalos fijos	Comprobar la integridad de todo el contenido en respuesta a situaciones o actividades específicas.	**1/4**
	Generar información de integridad si esta no ha sido proporcionada con el contenido	Usar bloqueadores de escritura cuando se trabaja con medios originales	Documentar los procesos y resultados de verificación de información de integridad	Verificar la información de integridad en respuesta a eventos o actividades específicas	
	Se verifica virus en todo el contenido; se aísla el contenido en cuarentena según sea necesario	Hacer una copia de seguridad de la información de integridad y almacenar una copia en una ubicación separada del contenido	Realizar una auditoría de la información de integridad bajo demanda	Reemplazar o reparar el contenido dañado según sea necesario	
Seguridad de la información	Se determinan los agentes humanos y de software qué deben estar autorizados para leer, escribir, mover y eliminar contenido	Documentar a los agentes humanos y de software autorizados para leer, escribir, mover y eliminar contenido y aplicar estos cambios	Mantener los registros (logs) y se identifican a los agentes humanos y de software que realizaron acciones sobre el contenido.	Se realizan revisiones periódicas de acciones / registros (logs) de acceso	**1/4**
Metadatos	Crear un inventario de contenido, documentando también la ubicación de almacenamiento actual de estos	Almacenar suficientes metadatos para saber cuál es el contenido (esto podría incluir alguna combinación de aspectos administrativos, técnicos, descriptivos, de preservación y estructurales)	Determinar qué estándares de metadatos aplicar	Registrar las acciones de preservación asociadas con el contenido y cuándo ocurren esas acciones Implementa los estándares de metadatos elegidos	**3/4**
	Hacer una copia de respaldo del inventario y se almacena al menos una copia por separado		Encuentra y completa los vacíos en sus metadatos para cumplir con esos estándares		
Formatos de archivos	Documentar los formatos de archivo y otras características de contenido esenciales, incluido cómo y cuándo fueron identificados	Verificar los formatos de archivo y otras características de contenido esenciales	Monitorear la obsolescencia y los cambios en las tecnologías de las que depende el contenido	Realizar migraciones, normalizaciones, emulación y actividades similares que garanticen el acceso al contenido.	**4/4**
		Establecer relaciones con los creadores de contenido para fomentar la elección sostenible de archivos			
Puntaje global	**3/5**	**2/5**	**3/5**	**1/5**	**9/20**

Descripción del cuadro

Se puede observar en base a los resultados que CIC Digital cumple casi con la mitad de las recomendaciones de la matriz, 9 de las 20 recomendaciones fueron marcadas como completadas, es decir, el 45% de la matriz. El nivel 1 se satisface en todos sus puntos excepto en dos; los niveles dos o tres tienen tres y cuatro puntos no satisfechos respectivamente, aunque por como se da la distribución de los mismos, el nivel dos cumple solo con la totalidad de

los puntos en dos categorías mientras que el nivel 3 lo consigue en tres; el punto 4 es el que tiene menos recomendaciones satisfechas, con solo la categoría “Formato de archivos” en donde todos los puntos se cumplen, esta baja performance en el último nivel es esperable si se tiene en cuenta que la estructura general de la matriz es progresiva y que por lo general las actividades de los primeros niveles son necesarias para cumplir con los niveles subsiguientes.

En lo que respecta a la categoría de “Almacenamiento”, CIC Digital posee copias de seguridad tanto en servidores propios, como en servidores que se encuentran en otra dependencia de la UNLP. Esta dependencia se encuentra en una ubicación geográfica distinta a la de los servidores del repositorio, y los servidores se encuentran por encima del primer piso de sus respectivos edificios (esto evitaría riesgos como la inundación), lo que permite cumplir tres de las recomendaciones propuestas para esta categoría, sin embargo, la distancia que separa las dos localizaciones con copias de seguridad no es lo suficientemente grande como para evitar que ambas sean afectadas por un desastre natural que involucre a toda la ciudad en donde se encuentran. En lo que refiere al control de la obsolescencia de los medios de almacenamiento, la renovación y mantenimiento del hardware se realiza pero de manera manual y esporádica, pero no hay un plan de renovación o de control de renovación regular del equipamiento, ni tampoco existe una planificación presupuestaria que lo contemple. Algo común a los puntos de esta categoría es la falta de documentación formal, no se cuenta con un plan de riesgos ni tampoco se encuentran documentados los distintos medios y equipos en los que se almacena el contenido y con los que se tiene en funcionamiento el repositorio.

La segunda categoría, “No alteración de archivos e integridad de los datos”, solo se cumple en su totalidad en el nivel 3. Se cuenta con un mecanismo de chequeo de integridad sobre el contenido, el cual es proporcionado por DSpace[1], el software en el que está desarrollado CIC Digital, con el cual a través de cronjobs[2] configuradas en el servidor, se realizan chequeos periódicos sobre la integridad de todo el contenido. No obstante, estos chequeos se realizan solo sobre la copia en funcionamiento del repositorio, no se ejecutan sobre ninguna de las copias de seguridad y, si bien es posible ejecutarlas bajo demanda en respuesta a acciones específicas, por ejemplo ante la ingesta de un ítem o la modificación de los metadatos de un recurso, solo se tiene programada de manera automática la ejecución ante alguno de estos eventos, se debe realizar de forma manual para el resto. Por esto último es que se marcaron en amarillo, es decir como punto parcialmente satisfecho, dos de las recomendaciones del último nivel. Por otro lado, no se realiza un análisis de virus sobre el contenido y en consecuencia no se cuenta con un mecanismo para aislar en cuarentena a un elemento sospechoso.

Sobre la categoría “Seguridad de la información” se cumple satisfactoriamente con el primer nivel, DSpace provee de un módulo de autenticación y autorización en el cual se pueden definir usuarios y grupos y distintos permisos sobre el contenido para esos usuarios/grupos. Por lo tanto es posible controlar y definir el tipo de acceso de las personas al contenido, es posible por ejemplo, otorgar al usuario final permiso de lectura sobre todo el contenido excepto el embargado o el privado, al usuario logueado y autenticado permiso de ingesta de items sobre determinadas colecciones, y al usuario administrador permiso de escritura sobre todo el contenido. A nivel de servidor se realiza lo mismo con los grupos y usuarios del sistema operativo utilizado. Si bien DSpace de alguna manera documenta los permisos aplicados dentro de su módulo de administración de autorización, y lo mismo ocurre con los permisos en el servidor en donde es posible visualizarlos al ingresar a los directorios correspondientes, estos no se encuentran formalizados ni centralizados en ningún documento institucional. Por otra parte, en lo que se refiere a logs y registros de cambios, el metadato provenance mantiene un pequeño registro de los usuarios que participaron en la ingesta de un ítem, tanto el encargado de la revisión como el que realizó el envío, sin embargo, si un usuario administrador

[1] https://duraspace.org/dspace/
[2] Procesos en segundo plano que se ejecutan a intervalos regulares. https://es.wikipedia.org/wiki/Cron_%28Unix%29

realiza modificaciones sobre el ítem desde la pantalla de edición para administradores, estos no quedan registrados. A su vez, en el servidor se registran los logs de acceso que provee el servidor web, pero no se revisan con regularidad, sólo en respuesta a eventos específicos, y sólo se mantienen las copias de los logs que tienen unos pocos días de antigüedad, el resto se elimina.

Con respecto a la categoría "Metadatos", se utiliza un esquema propio, mezclando elementos de distintos esquemas como dc, dcterms y un esquema propio. Todos los metadatos se encuentran inventariados en la base de datos y con dos copias de seguridad, lo que permite cumplir con el primer nivel de esta categoría. DSpace además provee de algunos metadatos administrativos y técnicos , que guardan información de la fecha de ingreso del contenido y del usuario que realizó la carga (por ejemplo el metadato dc.provenance) y la revisión del ítem, además de los metadatos descriptivos. Se realizan revisiones por parte de usuarios administradores dedicados a todas las ingestas de ítems al repositorio para comprobar que el contenido cumpla con el formato propuesto para el perfil de metadatos de CIC Digital. También se realizan otras revisiones una vez ingresado el contenido, como tareas de curación[3] destinadas al control de calidad de los metadatos y distintas reformulaciones al esquema de metadatos para agregar nuevos metadatos o modificar los existentes con el objetivo de que estos se adapten lo mejor posible al contenido ingresado. A pesar de proveer metadatos como el provenance que registran ciertos cambios en el contenido, DSpace no provee de metadatos que se ajusten a algún estándar de preservación existente, o que registren cambios en el ítem durante todo su ciclo de vida; en CIC Digital tampoco se ha implementado ninguno.

Por último, la categoría "Formato de archivos" es la única en la que se cumplen satisfactoriamente con todos los noveles. Desde la ingesta de los ítems se definen los formatos de archivos que se permiten cargar en el repositorio. Estos formatos se actualizan para siempre ofrecer los formatos más utilizados y con el mayor soporte dependiendo del tipo de recurso ingresado. Para el control de obsolescencia de los formatos se realizan chequeos y consultas manuales sobre los datos y si algún objeto digital es ingresado en una versión obsoleta de alguno de los formatos de ingesta, se lo transforma a la versión adecuada. Por ejemplo, en el caso de los pdfs, además de realizarse la extracción automática de texto, se realizan chequeos y normalizaciones para que todos los pdfs se encuentren en formato PDF/A.

Propuesta de mejora

Siguiendo la distribución de las recomendaciones de la matriz de NDSA, se presenta una propuesta de mejora a cada uno de los puntos no satisfechos por CIC Digital, agrupados por categoría:

- **Almacenamiento y localización geográfica**

  Como primera medida sería deseable que el backup que está en una localización geográfica distinta (en la Facultad de Ciencias Jurídicas y Sociales de la UNLP), se encuentre además en una red distinta a la del resto de las copias. Así, si llega a haber alguna falla que afecte a toda la red de la UNLP, el backup sería todavía totalmente accesible, sin necesidad de un acceso físico al servidor en donde se guardan.

  El siguiente paso en la prevención de amenazas y riesgos sobre los datos sería el almacenamiento de una de las copias de seguridad en una ciudad distinta de la que se almacenan las restantes, y que esta ciudad no sea afectada por las mismas amenazas ambientales (por ejemplo, un desastre natural) que la primera. Una solución viable para este caso sería el almacenamiento de la copia de seguridad en algún servicio en la nube,

[3] Tareas que permiten de forma sencilla y extensible gestionar operaciones de rutina sobre el contenido en un repositorio. https://wiki.lyrasis.org/display/DSDOC6x/Curation+System

por ej Glacier Deep Archive de Amazon, un servicio de almacenamiento a largo plazo que se autoproclama durable, seguro y de bajo costo y que permite cumplir con las leyes vigentes en cuanto a privacidad y seguridad de los datos. En este servicio, los datos son almacenados en cintas y se replican en 3 zonas de disponibilidad distintas dentro de la infraestructura de Amazon lo que asegura una disponibilidad mayor al 99%.

También, se debería documentar todos los procesos que participan en la creación de los backups, junto a los soportes y sistemas de almacenamiento disponibles y elaborar un plan de gestión de riesgos adecuado que incluya casos en donde los distintos sistemas de almacenamiento y backups queden inaccesibles.

- **No alteración de archivos e integridad de los datos**

  DSpace provee de una tarea de curación para el análisis de virus sobre los ítems, pero se debe ejecutar de forma manual. Se propone, primero realizar un análisis sobre la tarea de curación para saber si realmente cumple su propósito y luego, si esto se cumple, automatizar su ejecución para que se ejecute cada cierto período de tiempo sobre todo el contenido del repositorio, mediante alguna cronjob. También se podría ejecutar ante cada ingesta de un nuevo ítem al repositorio, durante el proceso de revisión. Algo parecido se podría implementar para la ejecución del chequeo de integridad para que, además de ejecutarse periódicamente sobre todo el contenido y sobre un ítem específico durante la ingesta del mismo, se ejecute ante otro tipo de eventos, como la edición de un ítem, y también se debería ejecutar este chequeo sobre las copias de seguridad del contenido, para lo cual a priori se tendría que descomprimir el contenido de la copia de seguridad y luego ejecutar el chequeo de integridad de DSpace desde la consola del servidor en donde se encuentre la copia, lo que implica también la instalación en ese servidor de las herramientas correspondientes (Java, DSpace, etc).

- **Seguridad de la información**

  Una de las opciones para mejorar el seguimiento y el registro de quien ha realizado qué acción con qué archivo es la de la incorporación de metadatos de preservación al repositorio. Si bien se cuenta con el metadato provenance, este no registra todas las acciones que se efectúan sobre un ítem y su sintaxis no es la adecuada. Para la correcta inclusión de los metadatos de preservación necesarios y su correcto uso con la sintaxis correspondiente, se debería mirar la propuesta del diccionario de datos PREMIS. Por el lado del servidor, serviría como una primera medida mantener los backups de los logs de acceso por un tiempo mayor al que se mantiene actualmente, al menos se podría almacenar la información de registros de los últimos 15 días.

- **Metadatos**

  Con respecto a la protección y localización de los metadatos, DSpace (y por lo tanto CIC Digital) ya provee un inventario y una clara localización del almacenamiento al guardar los metadatos en una tabla dedicada dentro de la base de datos. Esto a su vez también permite la separación del almacenamiento y las copias de seguridad de los metadatos en distintas localizaciones si se realizan periódicos backups de la base de datos y estos backups se trasladan a otro servidor, el cual puede estar incluso en una localización geográfica distinta al servidor de base de datos.

  Si bien DSpace crea automáticamente metadatos, por ejemplo el metadato provenance, con información administrativa sobre quien crea el contenido, la fecha de creación y de última modificación, y quien realizó

una edición sobre los metadatos de un ítem, esa información no se registra por metadato sino que se registra sobre el ítem como una unidad y tampoco se mantiene un histórico o historial de estos cambios, por lo que a veces se puede saber quien fue la última persona o la última vez que el ítem sufrió modificaciones, pero no se puede saber sobre qué metadato se hicieron ni tampoco cuáles fueron las modificaciones previas a esa y quién las hizo. A su vez, la sintaxis utilizada para guardar la poca información de preservación con la que se cuenta, no se adecua a ningún estándar de metadatos de preservación conocido. La implementación y adopción de los correctos metadatos de preservación, en particular adaptar el perfil actual de metadatos, y modificar alguno de los esquemas utilizados, para emular la propuesta del diccionario PREMIS de Metadatos de Preservación. Si bien PREMIS es realidad una propuesta de unidades semánticas (objetos, agente, derechos y eventos) que propone un esquema jerárquico de metadatos, tal vez pueda plantearse adaptar el diccionario de datos de PREMIS a un esquema plano, utilizando alguno de los esquemas usados por el perfil de metadatos de CIC Digital, por ejemplo dcterms. Esto solucionaría los problemas de falta de información en cuanto a modificaciones sobre los metadatos y el formato con que esa información se conserva.

- **Formatos de archivos**

  Se implementó una tarea de curación que chequea si los pdfs de los ítems se encuentran en formato PDFA, la ejecución periódica de esta tarea ayudaría a tener un control más automatizado sobre los formatos de los archivos. Se podría implementar de manera similar otras tareas que chequeen que el resto de los formatos se encuentren vigentes y que se pueda ejecutar esa tarea de manera automática.

Conclusión

La matriz de autoevaluación propuesta por NDSA permite a un repositorio determinar el estado de un repositorio en materia de preservación digital de manera sencilla y relativamente rápida. Basta con un diagnóstico y una mirada crítica sobre la documentación, implementación y políticas del repositorio por parte del personal experto para poder detectar las fortalezas y falencias de este. Este esquema es ideal para instituciones o repositorios que no poseen los recursos o la capacidad (tanto económica como de personal), de contratar una auditoría externa experta. Funciona como un primer paso de evaluación en el camino hacia la certificación en preservación pero no es un reemplazo a los métodos existentes, como el cumplimiento de las distintas normas ISO u otros sistemas de auditoría, que son la siguiente etapa a la que debería apuntar un repositorio. Por estas razones fue que se decidió realizar esta evaluación sobre CIC Digital, un repositorio de tamaño mediano al que nunca se le había realizado ningún análisis en términos de preservación digital. No se debe pasar por alto que NDSA Levels por su sencillez y su carácter subjetivo (por ser una autoevaluación), no provee un listado exhaustivo de las tareas de preservación y tiende a tener una visión optimista del estado del repositorio (Térmens y Leija, 2017).

Si bien CIC Digital cumple de manera satisfactoria casi la totalidad del nivel 1, tiene varias falencias en el resto de los niveles, especialmente en lo que se refiere a la documentación de los distintos procesos y recursos de almacenamiento, planeamiento de la seguridad y riesgo, además de que el seguimiento de las modificaciones a los archivos y metadatos se podría mejorar considerablemente. Se cuenta con un buen nivel en la cantidad y calidad de las copias de seguridad, pero un faltante en la documentación de la localización y la forma de restaurar esas copias y la carencia de un plan integral que indique qué hacer ante cada situación de riesgo. Esto último debería ser algo en lo que el repositorio se debería centrar y que no debería demandar un costo muy alto, tanto en tiempo como en dificultad. Quizás incluso el plan integral de riesgos es viable en el corto plazo para el repositorio, a diferencia del resto de medidas que se deberían tomar sobre las copias de seguridad, por ejemplo, la creación de otra copia de seguridad en una localización con amenaza de desastre diferente, que implicaría un costo de inversión en

infraestructura o la contratación de servicios en la nube. Otras acciones a implementar que implican solo ejecuciones de tareas ya implementadas, como el análisis de virus, o el chequeo de calidad de los metadatos y el chequeo sobre el formato de los archivos, no demandarían un esfuerzo tecnológico significativo (bastaría con la configuración de algunas cronjobs y tareas de curación) y serían realizables en el corto plazo.

Bibliografía

National Digital Information Infrastructure and Preservation Program (NDIIPP) of the Library of Congress. (n.d.). About the NDSA. National Digital Stewardship Alliance - Digital Library Federation. Retrieved April 25, 2022, from https://ndsa.org/about/

National Digital Information Infrastructure and Preservation Program (NDIIPP) of the Library of Congress. (n.d.-b). Levels of Digital Preservation. National Digital Stewardship Alliance - Digital Library Federation. Retrieved April 25, 2022, from https://ndsa.org/publications/levels-of-digital-preservation/

Levels of Preservation Revision Working Group, Kussmann, C., National Digital Stewardship Alliance (NDSA), Graham, W., Atkins, W., Reich, A., & Walker, P. (2019, October). 2019 LOP Implementation Guide and Working Definitions. https://osf.io/nt8u9/

NDSA Levels of Preservation Assessment Subgroup. (2019). Using the Levels of Digital Preservation as an Assesment Tool. https://doi.org/10.17605/OSF.IO/QGZ98

Leija, David; Térmens, Miquel. (2019). Traducción de Niveles de Preservación Digital NDSA 2019: Traducción al Español de Versión 2.0. APREDIG - Asociación Iberoamericana de Preservación Digital.

Térmens, M., & Leija, D. (2017). Auditoría de preservación digital con NDSA Levels. Profesional De La Información, 26(3), 447–456. https://doi.org/10.3145/epi.2017.may.11

Ferreras-Fernández, Tránsito. (2010). Preservación digital en repositorios institucionales: GREDOS. https://www.researchgate.net/publication/223905922_Preservacion_digital_en_repositorios_institucionales_GREDOS

Biblioteca Nacional de España. (n.d.). Diccionario de Datos PREMIS de Metadatos de Preservación. PUBLICACIONES DE LA BIBLIOTECA NACIONAL DE ESPAÑA. Retrieved May 2, 2022, from http://www.bne.es/es/Micrositios/Publicaciones/PREMIS/

Anexo 1

Traducción al Español de la matriz de preservación digital de NDSA

Área Funcional	Nivel			
	Nivel 1 - (Conocer su contenido)	Nivel 2 - (Proteger su contenido)	Nivel 3 - (Controlar su contenido)	Nivel 4 - (Mantener su contenido)
Almacenamiento	Tener dos copias completas en ubicaciones separadas Documentar todos los medios de almacenamiento donde este almacenado el contenido Poner el contenido en soportes de almacenamiento estables	Tener tres copias completas con al menos una copia en una ubicación geográfica distinta Documentar el almacenamiento y medios de almacenamiento, indicando los recursos y las dependencias que estos requieren para funcionar	Tener al menos una copia en una ubicación geográfica con amenaza de desastre diferente a las otras copias Tener al menos una copia en un medio de almacenamiento de diferente tipo Rastrear la obsolescencia del almacenamiento y los medios	Tener al menos tres copias en ubicaciones geográficas distintas, cada una con una amenaza de desastre diferente Maximizar la diversificación del almacenamiento para evitar puntos únicos de falla Tener un plan y realizar acciones para abordar la obsolescencia del hardware, software y medios de almacenamiento
Integridad	Verificar que la información de integridad se ha proporcionado con el contenido Generar información de integridad si esta no ha sido proporcionada con el contenido Se verifica virus en todo el contenido; se aísla el contenido en cuarentena según sea necesario	Verificar la información de integridad al mover o copiar contenido Usar bloqueadores de escritura cuando se trabaja con medios originales Hacer una copia de seguridad de la información de integridad y almacenar una copia en una ubicación separada del contenido	Verificar la información de integridad del contenido en intervalos fijos Documentar los procesos y resultados de verificación de información de integridad Realizar una auditoría de la información de integridad bajo demanda	Verificar la información de integridad en respuesta a eventos o actividades específicas Reemplazar o reparar el contenido dañado según sea necesario
Control	Se determinan los agentes humanos y de software qué deben estar autorizados para leer, escribir, mover y eliminar contenido	Documentar a los agentes humanos y de software autorizados para leer, escribir, mover y eliminar contenido y aplicar estos	Mantener los registros (logs) y se identifican a los agentes humanos y de software que realizaron acciones sobre el contenido.	Se realizan revisiones periódicas de acciones / registros (logs) de acceso
Metadatos	Crear un inventario de contenido, documentando también la ubicación de almacenamiento actual de estos Hacer una copia de respaldo del inventario y se almacena al menos una copia por separado	Almacenar suficientes metadatos para saber cuál es el contenido (esto podría incluir alguna combinación de aspectos administrativos, técnicos, descriptivos, de preservación y estructurales)	Determinar qué estándares de metadatos aplicar Encuentra y completa los vacíos en sus metadatos para cumplir con esos estándares	Registrar las acciones de preservación asociadas con el contenido y cuándo ocurren esas acciones Implementa los estándares de metadatos elegidos
Contenido	Documentar los formatos de archivo y otras características de contenido esenciales, incluido cómo y cuándo fueron identificados	Verificar los formatos de archivo y otras características de contenido esenciales Establecer relaciones con los creadores de contenido para fomentar la elección sostenible de archivos	Monitorear la obsolescencia y los cambios en las tecnologías de las que depende el contenido	Realizar migraciones, normalizaciones, emulación y actividades similares que garanticen el acceso al contenido

Traducción de Niveles de Preservación Digital NDSA 2019: Traducción al Español de Versión 2.0. (Leija, D; Térmens, M, 2019).

Anexo 2

Resumen biográfico de los autores

Santiago Tettamanti:

Licenciado en Informática, egresado de la Facultad de Informática de la UNLP. Actualmente se desempeña como ayudante alumno en las cátedras de Programación Concurrente y Sistemas Distribuidos y Paralelos de la misma facultad. Desde fines de 2017 integra el equipo de desarrollo de los repositorios SEDICI y CIC-Digital en PREBI-SEDICI.

Marisa De Giusti:

Es doctora en Ciencias Informáticas e Ingeniera en Telecomunicaciones de la UNLP. Dirige PREBI-SEDICI de la UNLP y CESGI centro CIC, ambos dedicados a la gestión de la información académica y científica. Como Investigador Principal de CIC sus temas cubran el área de ciencia abierta y repositorios institucionales.