Evaluación de CIC Digital a través de NDSA Levels

Santiago Tettamanti, Marisa R. De Giusti, Ariel J. Lira

11 julio 2022

Contexto

CIC Digital

- Repositorio Institucional de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires, Argentina
- Preservar y dejar accesible en abierto toda la producción científico-tecnológica de la CIC
- En la actualidad cuenta con alrededor de 10.000 ítems
- https://digital.cic.gba.gob.ar/

Contexto

¿Qué es NDSA?

- Alianza Nacional para la Custodia Digital (National Digital Stewardship Alliance)
- Consorcio de organizaciones comprometidas con la conservación a largo plazo de la información digital
- Creado en 2010 por la Library of Congress
- 267 instituciones miembro

NDSA LEVELS

- Levels of Digital Preservation creado en 2013
    - Iniciativa de NDSA
- Ayudar a instituciones a evaluar su sistema de preservación digital

NDSA LEVELS

Matriz sobre la cual se definen **tareas** para asegurar la **preservación y el acceso** a largo plazo.

Estas tareas se dividen en cinco categorías distintas:

- Almacenamiento
- Integridad de los datos
- Control de la información
- Metadatos
- Contenido

Para cada tarea hay cuatro niveles de cumplimiento progresivos.

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

NDSA LEVELS - Matriz de evaluación

- Es progresiva:
  - Las acciones en el primer nivel son requisitos previos necesarios para aquellos en los niveles de superiores
  - O son actividades más urgentes a lograr
- A diferencia de los métodos tradicionales
  - Se detectan las acciones a implementar para lograr la preservación
  - Facilidad de aplicación
    - Las auditorías tradicionales suelen ser caras y de una complejidad alta
  - Permite la autoevaluación
    - No se requiere personal experto en preservación
  - Ideal para pequeños y medianos repositorios

NDSA LEVELS - Matriz de evaluación

Sin embargo...

- Ofrece una visión simplificada de las tareas de preservación
- No un listado exhaustivo de las mismas como en Nestor, TRAC o ISO 16363.
- Visión más optimista de la que correspondería a la situación real
- Ignoran otras facetas claves que sí cubre por ejemplo ISO 16363

Evaluación de CIC Digital

Evaluación con la matriz de preservación de NDSA del repositorio institucional de la CIC

Metodología

- Se marcó en la matriz para cada tarea o punto si el repositorio cumplía o no con lo propuesto.
- Con un color especial las respuestas afirmativas en la tabla, con otro color los puntos completos de manera parcial, y con otro las respuestas negativas.
- Se sumaron puntos por cada respuesta afirmativa y se obtuvo un puntaje sobre el total de tareas propuestas.

Evaluación de CIC Digital

Metodología

- La evaluación fue realizada de manera conjunta por varios integrantes de los distintos equipos que trabajan en CIC Digital.
  - Sucesivas reuniones
  - Debate de manera conjunta
  - Se fijaron las acciones a llevar adelante
- Equipo de infraestructura del repositorio
  - Copias de seguridad, chequeos de integridad, del mantenimiento y seguridad de los distintos servidores
- Equipo de desarrollo
  - Chequeos por software (por ej, virus, integridad), de la seguridad del sistema, de la base de datos, de la curación automática de los metadatos, etc;
- Equipo administrativo
  - Controlar el contenido que ingresa al repositorio, los tipos y cantidad de metadatos usados, la variedad de formatos;
- Encargados de la gestión del repositorio
  - Visión general, las políticas presupuestarias y las decisiones institucionales.
- En repositorios pequeños una misma persona puede pertenecer a varios equipos.

Matriz de evaluación para CIC Digital

	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	

Matriz de evaluación para CIC Digital

Categoría Almacenamiento:

- Copias de seguridad tanto en servidores propios, como en servidores en otra dependencia de la UNLP.
    - En una ubicación geográfica distinta a la de los servidores del repositorio
    - Permite cumplir tres de las recomendaciones de la categoría
    - Distancia no muy grande, ambas serían afectadas por un desastre natural en toda la ciudad
- Servidores por encima del primer piso (esto evitaría riesgo por inundación).
- Control de obsolescencia, renovación y mantenimiento del hardware de manera manual y esporádica.
    - Pero no hay un plan de renovación o de control de renovación regular del equipamiento,
- Falta de documentación formal
    - No se cuenta con un plan de riesgos
    - No hay documentación del equipamiento

Matriz de evaluación para CIC Digital

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 especificas.	**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 especificas	
	Se verifica virus en todo el contenido; se aisla 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	

Matriz de evaluación para CIC Digital

Categoría Integridad de los datos

- Chequeo de integridad automático periódico sobre el contenido,
  - Proporcionado por DSpace, ejecutado a través de cronjobs
  - Se guarda en base de datos con copia de seguridad
  - No se ejecutan sobre ninguna de las copias de seguridad
  - Solo se tiene programada de manera automática la ejecución ante algunos eventos
    - Se debe realizar de forma manual para el resto
- No se realiza un análisis de virus sobre el contenido
  - No se cuenta con un mecanismo para aislar en cuarentena a un elemento sospechoso.

Matriz de evaluación para CIC Digital

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,	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	**3/4**
	Hacer una copia de respaldo del inventario y se almacena al menos una copia por separado	técnicos, descriptivos, de preservación y estructurales)	Encuentra y completa los vacíos en sus metadatos para cumplir con esos estándares	elegidos	

Matriz de evaluación para CIC Digital

Categoría “Seguridad de la información”

- Autenticación y permisos
  - DSpace provee de un módulo de autenticación y autorización con usuarios, grupos y permisos
    - Controlar y definir el tipo de acceso de las personas
    - Distintas autorizaciones y permisos para distintos tipos de usuarios
  - A nivel de servidor se realiza lo mismo con los grupos y usuarios del sistema operativo utilizado.
  - Los permisos no se encuentran centralizados en ningún documento institucional.
    - DSpace y el servidor permiten listar los permisos y autorizaciones administradas

- Logs y registros de cambios,
  - El metadato **provenance** mantiene un pequeño registro
    - El encargado de la revisión y el que realizó el envío
    - No quedan registrados cambios de los administradores.
  - En el servidor se registran los **logs de acceso**
    - No se revisan con regularidad, sólo en respuesta a eventos específicos.
    - Se mantienen las copias de los logs que tienen unos pocos días de antigüedad.

Matriz de evaluación para CIC Digital

Categoría “Metadatos”

- Esquema propio en CIC
  - Mezclando elementos de distintos esquemas
- Inventariados en la base de datos y con dos copias de seguridad
  - Permite cumplir con el primer nivel de esta categoría.
- DSpace provee metadatos administrativos y técnicos (dc.provenance, dc.date.accessioned)
  - Fecha de ingreso del contenido, el usuario que realizó la carga y la revisión del ítem, etc.
  - Además de los metadatos descriptivos.
- Revisiones de todas las ingestas por parte de administradores
  - Comprobar que el contenido cumpla con el perfil de metadatos de CIC Digital.
- Otras revisiones
  - Tareas de curación para el control de calidad
  - Reformulaciones al esquema de metadatos para que se adapten lo mejor posible al contenido ingresado.
- No hay metadatos que se ajusten a algún estándar de preservación existente
  - No se registran cambios durante todo el ciclo de vida del ítem

Matriz de evaluación para CIC Digital

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 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.	**4/4**
Puntaje global	**3/5**	**2/5**	**3/5**	**1/5**	**9/20**

Matriz de evaluación para CIC Digital

Categoría “Formato de archivos”

- Se cumplen satisfactoriamente con todos los niveles.
- Desde la ingesta de los ítems se definen los formatos de archivos permitidos.
    - 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 se realizan chequeos y consultas manuales.
    - Por ejemplo, se chequea que todos los pdfs se encuentren en formato PDF/A.
- Faltaría como mejora algún control automático
    - Cronjobs o tareas de curation podrían ser opciones

Propuesta de mejora

CIC Digital cumple casi con la mitad de las recomendaciones de la matriz

- 9 de las 20 recomendaciones fueron marcadas como completadas
- El cumplimiento es menor al avanzar en los niveles

Hay muchas mejoras posibles

- Algunas realizables en el corto plazo
- Propuesta de mejora por cada categoría

Propuesta de mejora

Almacenamiento y localización geográfica

- Backup en una red distinta a la del resto de las copias.
    - En caso de falla que afecte a toda la red de la UNLP, el backup sería accesible.
- Backup en una ciudad distinta de la que se almacenan las restantes
    - No sea afectada por las mismas amenazas ambientales
    - Almacenamiento en algún servicio en la nube, por ej Glacier Deep Archive de Amazon.
- Documentación de los procesos que participan en la creación de los backups, soportes y sistemas de almacenamiento disponibles
- Elaborar un plan de gestión de riesgos
    - Con pasos a seguir ante cada potencial problema

Propuesta de mejora

No alteración de archivos e integridad de los datos

- Tarea de curación para el análisis de virus sobre los ítems
  - Analizar si cumple su propósito
  - Automatizar su ejecución mediante cronjobs
  - Periódicamente y ante cada ingesta de un nuevo ítem al repositorio.
- Chequeo de integridad
  - Programar su ejecución ante eventos específicos (ya funciona sobre la ingesta)
  - Edición de un ítem y sobre las copias de seguridad del contenido,

Propuesta de mejora

Seguridad de la información

- Incorporación de metadatos de preservación al repositorio.
  - Mejorar el seguimiento y el registro de cambios en el ítem
  - Se cuenta solo con el provenance
    - Incompleto
    - Sintaxis inadecuada.
  - Evaluar la propuesta del diccionario de datos PREMIS.
- Backups de los logs de acceso en el servidor por un tiempo mayor al actual
  - Almacenar la información de registros de los últimos 15 días.

Propuesta de mejora

Metadatos

- DSpace crea metadatos con información administrativa
  - Quien crea el contenido
  - La fecha de creación y de última modificación
  - Quien editó sobre los metadatos de un ítem
- No se mantiene un histórico o historial de estos cambios
- No registra cambios desde la edición administrativa
- La sintaxis utilizada no se adecua a ningún estándar de metadatos de preservación conocido.
- Implementación y adopción de los correctos metadatos de preservación
  - Modificar alguno de los esquemas utilizados por CIC Digital
  - Propuesta del diccionario PREMIS de Metadatos de Preservación.
  - Adaptar PREMIS a un esquema plano

Propuesta de mejora

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.
    - Mediante cronjobs
- Implementar otras tareas que chequeen que el resto de los formatos se encuentren vigentes.

Conclusión

La matriz de autoevaluación de NDSA

- Permite determinar el estado de la preservación digital en un repositorio de manera sencilla y rápida.
- Permite el autodiagnóstico 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
- No provee un listado exhaustivo de las tareas de preservación
- Tiende a tener una visión optimista del estado del repositorio.

Conclusión

Por esas características

- Recomendable para instituciones o repositorios que no poseen los recursos o la capacidad para una auditoría externa experta.
- Primer paso de evaluación en el camino hacia la certificación en preservación
- **No** es un reemplazo a los métodos existentes (normas ISO u otros sistemas de auditoría)
    - Siguiente etapa a la que debería apuntar un repositorio.
- Ideal para el caso de CIC Digital
    - Repositorio de tamaño mediano
    - Nunca se le había realizado ningún análisis en términos de preservación digital

Conclusión

CIC Digital

- Cumple de manera satisfactoria casi la totalidad del nivel 1
- Tiene varias falencias en el resto de los niveles
    - Especialmente documentación de procesos, recursos, planeamiento de la seguridad y riesgos.
- Viable en el corto plazo
    - Documentación de la localización y la forma de restaurar copias de seguridad
    - Plan integral de riesgos
    - Análisis de virus
    - Chequeo de calidad de los metadatos
    - Chequeo sobre el formato de los archivos
- A futuro
    - Creación de otra copia de seguridad en una localización con amenaza de desastre diferente
        - Costo de inversión en infraestructura o la contratación de servicios en la nube.

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

Bibliografía

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/