Factores económicos de la preservación documental digital: actualización 2021

Economic factors of digital preservation: 2021 review

Juan Voutssás M.

Atribución-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0)

Este es un resumen legible por humanos (y no un sustituto) de la licencia. Advertencia.

Usted es libre de:

Compartir — copiar y redistribuir el material en cualquier medio o formato

Adaptar — remezclar, transformar y construir a partir del material

La licenciante no puede revocar estas libertades en tanto usted siga los términos de la licencia

Bajo los siguientes términos:

Atribución — Usted debe dar crédito de manera adecuada, brindar un enlace a la licencia, e indicar si se han realizado cambios. Puede hacerlo en cualquier forma razonable, pero no de forma tal que sugiera que usted o su uso tienen el apoyo de la licenciante.

NoComercial — Usted no puede hacer uso del material con propósitos comerciales.

CompartirIgual — Si remezcla, transforma o crea a partir del material, debe distribuir su contribución bajo la la misma licencia del original.

Factores económicos de la preservación documental digital: actualización 2021/ Economic factors of digital preservation: 2021 review

COLECCIÓN
TECNOLOGÍAS DE LA INFORMACIÓN
Instituto de Investigaciones Bibliotecológicas y de la Información

Factores económicos de la preservación documental digital: actualización 2021/
Economic factors of digital preservation: 2021 review

Juan Voutssás Márquez

Universidad Nacional Autónoma de México
2022

Z701
V68

Voutssás Márquez, Juan.
Factores económicos de la preservación documental digital : actualización 2021 = Economic factors of digital preservation : 2021 review / Juan Voutssás Márquez. – México : UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información, 2022.
98 p. – (Tecnologías de la información)
ISBN: 978-607-30-5949-7

1. Preservación digital – Aspectos económicos. 2. Conservación de materiales – Aspectos económicos. I. Título. II. ser.

Diseño de la portada: Wendy Chávez
Primera edición: 14 marzo 2022

Instituto de Investigaciones Bibliotecológicas y de la Información
Circuito Interior s/n, Torre II de Humanidades,
pisos 11, 12 y 13, Ciudad Universitaria, C. P. 04510,
Alcaldía Coyoacán, Ciudad de México

ISBN:

Publicación dictaminada

Impreso y hecho en México

Contenido

Introducción 7

Desglose de los costos 13
- Costo de digitalizar 13
- Costo de editar 17
- Costo de registrar 20
- Costo de almacenar 24
- Costo de actualizar 34

Conclusiones 43

Anexo 1 47

Referencias bibliográficas 49

Economic factors of digital preservation: 2021 review

Introduction 55

Itemization of costs 61
- Digitizing cost 61
- Editing cost 65
- Registration cost 67
- Storage cost 71
- Updating costs 80

Conclusions 89

ANNEX 1 93

Bibliographic references 95

Introducción

Cuando se trata de dinero, todos son de la misma religión.
François-Marie Arouet "Voltaire", pensador francés del siglo XVIII, y hombre inmensamente rico.

La información digital en el mundo sigue creciendo incesantemente a pasos agigantados; según algunos estudios muy citados de la International Data Corporation (IDC) altamente citados (Ganz y Reisnel 2005 y 2012), el mundo produjo 0.13 Zettabytes de información en 2005; 1.23 en 2010; 2,9 en 2012; 8,6 en 2015; produciría poco más de 40 en 2020, y se pronostican 175 Zettabytes en 2025.[1] Todas estas cantidades se han ido acumulando anualmente a lo largo de las últimas décadas. La mayor parte de esa información es desechable a corto plazo, pero aun así queda una inmensa cantidad de ella que debe ser preservada para un futuro. Esa información es de todos los tipos imaginables: textos en libros, revistas, diarios y catálogos; fotografías, mapas, música, radio, cine y televisión; juegos, redes sociales, mensajería y chats, llamadas, eventos deportivos, análisis médicos, facturas, datos científicos, archivos gubernamentales, etcétera, en innumerables combinaciones y variados formatos digitales.

Toda esa información que requiere ser preservada tiene un responsable de hacerlo; al menos, en teoría debiese tenerlo. Muchos de ellos son las empresas particulares que producen y/o explotan

1 1 Zettabyte = 1,000 Exabytes = 1,000,000 Petabytes = 1,000,000,000 Terabytes = 1,000,000,000,000 Gigabytes = 1,000,000,000,000,000 Megabytes = 10^{21} bytes.

esa información: editores de libros y revistas, empresas musicales, de cine o entretenimiento; ofertantes de redes sociales, los grandes buscadores de la web, bancos, seguros y sistemas financieros, empresas comercializadoras de bienes y servicios, de aviación, de comunicaciones, etcétera. No obstante, queda todavía una buena parte que debe ser preservada por organizaciones públicas a quienes se les asigna esa responsabilidad: bibliotecas, archivos, museos, repositorios y otras que pertenecen a alguno de los sectores de gobierno: ejecutivo, legislativo, judicial; sistemas públicos educativos, de salud, de comunicaciones, culturales y de asistencia social, por mencionar algunos.

Los responsables de estas tareas del sector empresarial o privado por lo general están claramente asignados dentro de sus organizaciones y tienen ciertos tipos específicos de documentos a preservar, pero, sobre todo, cuentan de inicio con presupuestos para hacerlo. A diferencia de ellos, los responsables de estas tareas en los sectores públicos no siempre están claramente asignados, con frecuencia deben preservar variados tipos de información, y en repetidas ocasiones no cuentan con los presupuestos para esa tarea; al menos, no de inicio ni en cantidades suficientes. La principal razón de ello se debe a que los costos de la preservación digital no son evidentes para todos. Muchos siguen pensando que ésta se reduce a adquirir una cantidad suficiente de dispositivos de almacenamiento; entre esas personas, se encuentran con frecuencia los tomadores de decisiones y los financiadores de estos proyectos. El problema se agrava derivado de que los costos de la preservación digital son tangibles, mientras que sus beneficios son intangibles: es relativamente sencillo establecer sus costos numéricamente, pero no así sus beneficios.

En las últimas décadas, se ha dado a nivel mundial la tendencia de que exista cada vez más información pública, y que los ciudadanos tengan mayor acceso a ella. Por un lado, se han emitido globalmente numerosas leyes de transparencia y acceso a la información pública gubernamental y, por el otro, ha ocurrido una serie de movimientos mundiales en favor del acceso abierto a la información científica, académica, cultural, etcétera. Una parte

considerable de la información producida a nivel mundial mencionada anteriormente comprende precisamente todos estos materiales, mismos que por su naturaleza deben ser preservados en una alta proporción. Por este motivo, el número de "sujetos obligados" por las leyes de acceso a la información pública, así como el número de "organizaciones designadas" para colecta y preservación de información científica y académica en acceso abierto, datos abiertos, repositorios, etcétera, ha crecido inusitadamente en los últimos años. Hemos pasado de los Petabytes a los Exabytes y luego a los Zettabytes en tan solo un par de décadas. Cada vez hay más organizaciones públicas que son designadas para colectar y preservar parte de ese acervo documental en sus respectivas regiones y países: archivos nacionales, regionales o estatales; bibliotecas y archivos especializados de los poderes legislativos y judiciales; bibliotecas nacionales o grandes bibliotecas temáticas en arte, ciencias o humanidades, entre otras; repositorios nacionales o regionales de ciencia abierta, repositorios de datos, etcétera. Algunas de estas organizaciones ya estaban designadas para estas tareas, pero otras lo han sido recientemente. Independientemente de ello, la cantidad de información que deben preservar ha crecido exponencialmente en los últimos años y sigue creciendo cada día.

Pero preservar información documental digital cuesta, y no es tan solo asunto de comprar o rentar muchos dispositivos de almacenamiento; hay otros factores económicos que inciden en la preservación digital. Existen textos al respecto desde la década de los noventa, por ejemplo, los de Hendley (1998) y Ashley (1999). En esa época, se hacían además estudios comparativos entre costos de preservación en papel, microficha y digital –por ejemplo, el estudio de Kingma (2000). A este respecto, Morris y Truskowsky (2003, 206) establecieron que en 1996 se llegó al punto de inflexión en el cual el almacenamiento en dispositivos electrónicos lograba ya mejor costo/beneficio que en papel. Desde el siglo pasado, se han realizado incontables estudios acerca de los costos de almacenamiento de documentos digitales, con numerosas perspectivas, diversos enfoques, y para todo tipo de acervos

y documentos. Se han creado modelos y metodologías de costeo, fórmulas para su cálculo, etcétera: Como ejemplos significativos, están los trabajos de Bote *et al.* (2012), Calhoun *et al.* (2019), Zeller (2010) y Rosenthal *et al.* (2012).

Todos esos planteamientos han ido evolucionando junto con los contextos tecnológicos, económicos, sociales, etcétera. En lo personal, desde el año 2009 traté los diversos factores que inciden en la preservación documental digital: tecnológicos, culturales, documentales, sociales, legales, y por supuesto económicos (Voutssás 2009). Dado que ya transcurrió más de una década desde entonces, nuevos elementos se han agregado a cada uno de esos factores, en especial los económicos. La cantidad de información global producida por año pasó de 1.23 Zb en 2010 a 40 Zb en 2020. Los costos de los dispositivos de almacenamiento y sus rendimientos han seguido mejorando incesantemente. Incontables empresas ofrecen hoy en día servicios de digitalización, edición, etcétera. El advenimiento de los servicios de almacenamiento en la nube –que prácticamente no existían entonces– ha agregado nuevos elementos económicos a la ecuación. Por lo anterior, la creciente necesidad de las organizaciones públicas de preservar cada vez más información documental digital, así como el cambio en los contextos y en especial las opciones y los factores económicos para ello. Es pertinente hacer ahora una nueva revisión de esos componentes para actualizar los considerandos en su análisis y toma de decisiones al respecto.

Cabe resaltar en este punto que hay matices y contextos propios de cada tipo de documento o datos a preservar y de la organización que los custodia, por lo que el análisis de costos no puede reducirse a una fórmula que se aplique universalmente. No es lo mismo preservar libros que artículos, documentos de archivo, sentencias jurídicas, filmes, análisis clínicos o datos provenientes de telescopios, por mencionar algunos. Si bien todos tienen principios y características comunes, cada uno de ellos presenta su propia problemática. Por ello es común encontrar en la literatura numerosos modelos de costos de preservación digital específicamente construidos por país, por organización, por tipo de documento,

etcétera. El propósito de este texto no es crear un modelo más de cuantificación, sino estudiar de forma integral características comunes de los documentos digitales que permitan establecer metodologías generales para la determinación de los costos desde un enfoque práctico, todo ello a la luz de los contextos emergentes asociados.

Cabe también en este punto establecer algunas definiciones para los fines de este texto. La *preservación digital* puntualiza y resalta hoy en día la estabilización del contenido y forma del documento; es decir, comprende a los objetos lógicos y su alcance siempre es el largo plazo (Voutssás y Barnard 2014). La American Library Association (ALA) la define como:

> […] la reinterpretación exacta de contenido autenticado a largo plazo por medio de la combinación de políticas, estrategias y acciones que garanticen el acceso futuro a contenidos digitales a pesar de los desafíos de la obsolescencia tecnológica y lo efímero de los soportes (ALA y ALCTS 2007).

El Consejo Internacional de Archivos (ICA) define preservación digital como

> […] el proceso específico para mantener los materiales digitales durante y a través de las diferentes generaciones de la tecnología a través del tiempo, con independencia de donde residan (ICA 2015).

La *conservación digital* tiene que ver con el soporte del documento digital; es decir, incluye los objetos físicos, y su alcance es siempre a corto-mediano plazo; es sinónimo de *mantenimiento*. Procede del concepto de conservación presentado en el Glosario de Terminología Archivística y Documental de la Sociedad Estadounidense de Archivistas (SAA): "[...] la estabilización de los materiales mediante [...] tratamiento físico para garantizar que sobrevivan en su forma original el mayor tiempo posible" (Pearce-Moses 2005).

A la fecha, no existe una recomendación o estándar que unifique la división en subcategorías o rubros de los costos de la preservación digital; cada autor divide los costos en secciones de forma arbitraria, y la separación entre rubros no es algo absoluto. Para su estudio en este documento, se consideraron cinco tipos de costos de la preservación digital: de digitalizar, editar, registrar, almacenar y actualizar. Ésta es una categorización personal y arbitraria.

Desglose de los costos

Costo de digitalizar

Estrictamente, la digitalización no forma parte de la preservación digital. En teoría, un documento digital es susceptible de ser preservado una vez que es creado o producido y no antes. Aunque en la actualidad la inmensa mayoría de información que se produce ya es digital de origen, aún quedan muchos fondos en bibliotecas, archivos y repositorios sobre soportes "tradicionales" que eventualmente se desea preservar en forma digital, y por ello es indispensable que de forma previa esos fondos sean digitalizados en cierto momento, por lo que se revisa aquí este rubro, ya que en caso de ser necesario debe incluirse para que quede completo el análisis de costos. Por lo tanto, aunque en realidad el proceso de preservación de un documento digital se requiere una vez que este documento existe, dado que con frecuencia en la práctica la preservación no puede desasociarse de la digitalización, se incluye aquí un análisis de este costo. Obviamente si no debe realizarse puede ser omitido.

Se entiende por *digitalizar* al proceso de convertir cierto documento que se encuentra en un formato "tradicional" o analógico –papel, casete de audio o video, película, etcétera– hacia un formato digital con objeto de poder almacenarlo, distribuirlo y que sea posible acceder a él en esta nueva forma. Un formato digital es aquel que hace una representación de un objeto documental en una forma totalmente numérica; esto es, con dígitos –de ahí el nombre– por medio de una abstracción o representación preestablecida y arbitraria de esos números.

El primer punto a considerar al establecer proyectos de esta naturaleza es "tamizar" de inicio el conjunto documental a digitalizar

a través de una serie de "criterios de digitalización" que sirven para corroborar que en realidad es necesario y conveniente digitalizar una cierta colección documental, o parte de ella, no incurriendo en costos ociosos, reiterativos o inútiles. Existen numerosos ejemplos de esos criterios que pueden ser adaptados, por lo que no se abundará más al respecto.[1] Dependiendo del tipo de documentos y de organización, este punto se denomina también "selección", "depuración previa", "pre-ingesta", "evaluación", etcétera. En el contexto de archivística, esto es parte de la "valoración" –*appraisal*–, proceso en el que se examina un fondo documental para determinar su valor para una determinada institución y durante cuánto tiempo. Este paso es particularmente sensible para aquellos que desean preservar datos masivos (*Big Data)*, ya que sus volúmenes pueden crecer a pasos agigantados. En los entornos de datos, este paso también se conoce como "limpieza de datos" o "verificación de datos".

El segundo punto a considerar al momento de digitalizar documentos consiste en establecer los parámetros de calidad del documento resultante. Estos parámetros deben ser cuidadosamente establecidos de inicio, pues al determinar la calidad del resultado se incide directa y sensiblemente en los costos. Desde siempre –y esto no ha cambiado con los años– mayor calidad en los documentos digitales implica mayor costo de ellos, y viceversa. Por esta razón, debe establecerse de antemano un cuidadoso equilibrio entre la calidad del documento y el costo de digitalizarlo. Esto tiene que ver directamente con el propósito del documento resultante: ¿Es para preservar o para distribuir? Si los documentos solo se van a usar para distribución, la calidad –y por ende los costos– pueden ser reducidos: páginas web, copias simples para usuarios, acervos efímeros, etcétera. Si el propósito es la preservación a largo

1 Un excelente documento que incluye los criterios para digitalizar colecciones de bibliotecas y archivos fue elaborado por IFLA e ICA (2002): ***Guidelines for digitization projects for collections and holdings in the public domain, particularly those held by libraries and archives.*** Está disponible en https://www.ifla.org/files/assets/preservation-and-conservation/publications/digitization-projects-guidelines.pdf.

plazo de esos documentos, la calidad debe ser alta; no conviene producir documentos digitales de baja calidad, pues es imposible agregarla posteriormente. Empero, esto hace que los tamaños de ficheros y por ende de costos se incrementen exponencialmente;[2] de ahí la importancia de lograr el equilibrio entre calidad y costo. Muchos proyectos de digitalización caen en cuenta al final del proceso de que obtuvieron un conjunto de documentos de baja calidad que no cumplen con los estándares o recomendaciones para ese tipo de documentos, y que la información que han construido es inútil con fines de preservación. Establecer parámetros por debajo de la calidad necesaria da como resultado acervos inutilizables a futuro; pero establecer parámetros de calidad excesivos para la digitalización eleva los costos enormemente y vuelve los proyectos a menudo inasequibles. He ahí la razón de buscar un adecuado equilibrio desde el inicio. Por lo general y al respecto de la calidad requerida, ya existen recomendaciones hechas por incontables organizaciones para prácticamente todo tipo de documento a digitalizar: libros, revistas, periódicos, negativos, audio, filmes, mapas, partituras, archivos, etcétera, considerando además su propósito, lo cual facilita esta tarea, que jamás debe soslayarse u omitirse. Con frecuencia sucede que el propósito de la digitalización contiene a la vez los dos objetivos: preservar y distribuir, por lo que la organización debe contemplar la existencia de dos copias de cada documento digital: una de alta calidad para preservar y otra de baja calidad para distribuir. En este caso, se crean primero copias de alta calidad, de las cuales se obtienen las copias de baja calidad y no a la inversa: ello es imposible.

---

2 Incremento exponencial: Una imagen digitalizada a 100 puntos por pulgada o dpi contiene 100 x 100 = 10,000 puntos por pulgada cuadrada; si se multiplica la calidad por dos, 200 dpi, el resultado es una imagen de 200 x 200 = 40,000 puntos; esto es, el cuádruple a guardar, no el doble. Si se vuelve a duplicar, a 400 dpi, contiene 400 x 400 = 160,000 puntos por pulgada cuadrada. Siendo el cuádruple de la resolución original, el resultado es 16 veces más grande.

Para ayudar a establecer estos conceptos de calidad de la información se recomienda revisar el estándar ISO/IEC 25012:2008 "Modelo General de Calidad de Datos para Formatos Estructurados en Sistemas Informáticos"; originalmente fue emitido para datos, pero con el tiempo se ha ido haciendo extensivo para otros tipos de información digital; en él se establecen quince características necesarias para la calidad de la información: exactitud, precisión, completitud, credibilidad, consistencia, actualidad, accesibilidad, exhaustividad, conformidad con normas, eficiencia, confidencialidad, trazabilidad, disponibilidad, portabilidad y recuperabilidad.[3] Al final de este texto se presenta un anexo con algunos ejemplos de sitios web que tratan recomendaciones al respecto de la calidad para diversos materiales documentales.

El tercer punto consiste en establecer e ir sumando los costos de los diversos insumos para la digitalización: la adquisición del equipo para digitalizar –escáneres, cámaras, tarjetas digitalizadoras de audio o video y probablemente algunos computadores, el *software* o programas necesarios para la tarea, el costo del personal dedicado a ello, y eventualmente ciertas instalaciones físicas. Esta suma dividida entre el número total de documentos a digitalizar, en un cierto tiempo, dará un costo unitario por documento y por el total de ellos. En este paso debe siempre ponderarse si la digitalización se hará *in situ*, en las instalaciones de la biblioteca o archivo con su propio personal y equipo, o en cambio será tercerizada a un proveedor al efecto. Ésta es una simple comparación de costo/rendimiento entre ambas opciones. Para ello debe obtenerse el gran total de la suma de los diversos componentes del costo de digitalizar para un cierto número de documentos en un proceso realizado en las instalaciones de la organización, y compararlo contra el costo total de un proveedor que realice ese proceso. En este paso es de suma importancia no incluir costos de otras etapas en una de las opciones sin incluirlos en la otra. Es decir, con frecuencia se solicita al proveedor que –además de

3 Si se desea abundar en el tema de otros estándares relacionados con la preservación documental archivística, véanse InterPARES (2008) y AGN (2015).

digitalizar– separe documentos encuadernados, edite y/o mejore el resultado digitalizado, le agregue ciertos metadatos, le aplique Reconocimiento Óptico de Caracteres (OCR), etcétera. La organización solicitante debe asegurarse de que se comparan procesos idénticos, y por lo tanto si se agregan estos requerimientos a lo solicitado al proveedor, deben costearse también en la opción realizada localmente. Con mucha frecuencia ocurre que se consideran costos de edición, optimización, catalogación, etcétera, en una de las dos opciones sin incluirlos en la otra, lo cual obviamente induce a sesgos y errores. Por ello es de capital importancia que la propuesta de un proveedor especifique claramente cuáles son los procesos adicionales a la pura digitalización, para que así puedan compararse procesos iguales y sus respectivos costos y beneficios. Otro factor adicional a considerar aquí es el tiempo de realización; por lo general un proveedor puede hacerlo de forma más rápida, y eso puede influir en la decisión final, al margen de los costos. La tendencia en los últimos años es que en grandes volúmenes se envíen los documentos a digitalizar con un proveedor es una opción más rentable; en pequeños volúmenes que se generan a lo largo del tiempo sucede lo opuesto. La Federación de Bibliotecas Digitales (DLF) y el Consejo de Bibliotecas y Recursos de Información (CLIR) elaboraron en 2017 un "calculador de costos de digitalización" (https://dashboard.diglib.org/) para ayudar a las organizaciones y personas sin mucha experiencia al respecto a calcular los costos para digitalizar documentos *in situ*. Si bien el calculador se encuentra todavía en fases experimentales, proporciona una idea práctica de los elementos requeridos para costeo de digitalización.

Costo de editar

En la práctica, todo documento digital que va a ser preservado requiere de cierta preparación; rara vez un documento tal como sale del procesador de texto, del escáner, o de cualquier otro origen reúne los requisitos para gestión y para preservación. Requiere de perfeccionamiento y de elementos agregados para ello; esto es, de edición.

Dado que es un término con varios significados, es conveniente aclarar en este punto qué se entiende por "edición". Para fines de este texto; debe entenderse aquí en su concepto amplio de revisión, pulimento, perfeccionamiento y preparación de un cierto texto u obra previa a su liberación o ingreso a un sistema para garantizar una calidad homogénea y estandarizada. Muchos programas informáticos procesadores de texto son denominados "editores de textos", lo cual incrementa la confusión. En efecto, un procesador de texto hace cierta parte de la edición del mismo, pero no conlleva todas las tareas que implica su completa edición. Editar un texto –en su acepción completa– significa que una vez que ha sido establecida la calidad deseada y/o mínima de los ítems a preservar, alguien tiene que verificar que todos los elementos que han sido estipulados para esa calidad existan en todos y cada uno de los documentos en un nivel adecuado. Y esta tarea no la puede realizar un simple programa "procesador de textos" o "editor de imágenes": es todo un proceso que implica una serie de pasos sucesivos y que es mayormente un quehacer intelectual realizado por una persona; por este motivo cuesta. Pero es un paso de suma importancia en la preservación, pues garantiza una calidad adecuada a lo largo de todos los documentos. Omitirlo pone en riesgo esa calidad y, como consecuencia, la preservación como un todo.

Los documentos digitales a preservar pueden provenir de dos fuentes de acuerdo a su origen: o fueron digitalizados a partir de un documento sobre soporte tradicional o analógico, o fueron creados digitales de origen. No importa: ambos requieren de revisión y perfeccionamiento.

Para comenzar, todo documento digital a preservar, cualquiera que sea su origen, debe ser una "reproducción digital fiel". La Federación de Bibliotecas Digitales (*Digital Libraries Federation*) estableció desde hace casi dos décadas que todos los documentos digitales –de origen o digitalizados– deben estar óptimamente formateados y reunir tres características indispensables: calidad, permanencia e interoperabilidad (DLF 2002). El *formateado óptimo* significa que el documento se encuentra codificado en un formato

no propietario y/o abierto, de larga duración y eficiencia para ese tipo de documento; el concepto de *calidad* va relacionado a su funcionalidad y propósito de uso ya mencionados; esto es, distribución o preservación; su *permanencia* se relaciona con la capacidad de ser accesible a largo plazo, y su *interoperabilidad* se refiere a que sus cadenas de bits estén codificadas de tal forma que sean altamente independientes de la plataforma computacional. En Archivística, las especificaciones para los documentos de archivo incluyen elementos de construcción y calidad semejantes pero todavía mayores, como las establecidas en MoReq2 (2002) o InterPARES (2005). Este último establece como un elemento de calidad indispensable de un documento de archivo desde su origen a la "confiabilidad" o "fehaciencia"; que a su vez es la suma de su "solvencia", "autenticidad" y "exactitud", las cuales conllevan que su información es precisa, correcta, veraz y pertinente; tiene identidad e integridad.[4]

Los posibles procesos para la edición de un documento digital son numerosos; dependiendo del tipo de documento y de su eventual proceso de digitalización, existen incontables actividades a realizar sobre ellos. Por ejemplo, los textos deben guardar ciertas formas en cuanto a su distribución espacial dentro de una página, con una o varias columnas, márgenes y renglones diferentes, con distintos tipos de letra, etcétera. Las imágenes –fotografía, negativos, páginas de diarios, etcétera– se recortan, se aclaran o contrastan, se les retiran rayones o manchas, se les aplica OCR, se les agregan marcas de agua o logotipos, etcétera. Los audios se mejoran, se les retira el ruido o *hiss*, se separan en canales, se traducen, etcétera; los videos y películas se "remasterizan" para mejorar imágenes o audio, retirarles fallas o defectos, agregarles color, añadirles subtítulos o créditos, etcétera.

Por supuesto, todas estas actividades de edición tienen un costo: se requieren para ello equipos, *softwares* y personal especializados. Al igual que con la digitalización, es necesario establecer

4 Confiabilidad = trustworthiness; solvencia = reliability; autenticidad = authenticity; exactitud = accuracy.

cuántas personas y tiempos unitarios se requieren para realizar esta tarea para el total de documentos, y agregarle el costo de los insumos. La suma total de todos los costos de esta actividad dividido entre el número de documentos nos dará el costo unitario de editar, así como el tiempo estimado para realizarlo. De forma semejante a la digitalización, existen las opciones de realizarlo en la organización o tercerizarlo con un proveedor. La comparación de costos y tiempos ayudará a tomar la decisión. Igualmente, es imperativo que los costos a comparar entre las opciones interna y con proveedor se hagan exactamente sobre las mismas especificaciones y requisitos para no sesgar el resultado. Debe tomarse en cuenta al momento de realizar este comparativo que, dado que ésta es una etapa que implica mayormente proceso intelectual, conlleva que el personal que la realiza esté altamente calificado para ello; si no cuenta con la experiencia previa, es indispensable un proceso de capacitación y entrenamiento del mismo el cual se traduce en un costo y una curva de aprendizaje adicionales.

Costo de registrar

Denominamos aquí *registrar* al proceso que permite poner a un ítem digital en condiciones para ser ingresado en un sistema informático de gestión que le permitirá eventualmente ingresar a un sistema informático de preservación. La diferencia con el anterior proceso, *la edición*, es que aquél cubre aspectos de la calidad técnica inherente a los documentos, mientras que *el registro* trata sus aspectos documentales. Este proceso contempla diversas actividades y niveles de profundidad; dependiendo de la organización y del tipo de documento, las actividades o pasos pueden consistir en recepción, inscripción o asiento; foliado, inventariado, catalogación, clasificación, descripción, etcétera. Obviamente cada institución aplica para ello sus propias reglas y especificaciones.

Cuando a un simple documento digital que cumple con el requisito de ser una "reproducción digital fiel" se le agregan metadatos, obtenemos un *objeto digital*. No se preservan documentos digitales:

se preservan objetos digitales. La diferencia fundamental estriba en que un documento digital es cualquier entidad documental[5] que ha sido creada de origen o convertida a una forma de representación basada en números bajo un cierto patrón arbitrario, con objeto de poder ser almacenada o transmitida por medio de dispositivos electrónicos. Un *objeto digital* –también llamado *objeto de información*– es cualquier entidad documental que ha sido codificada numéricamente bajo algún formato y ensamblada junto con algún conjunto de metadatos de tal forma que puede ser almacenada, buscada, encontrada y usada dentro de un sistema informático; cuando es necesario, contiene también los métodos o procedimientos para realizar operaciones sobre el objeto. Los ficheros con el contenido y sus metadatos correspondientes están entrelazados entre ellos física y/o lógicamente. Como puede verse, la diferencia entre un simple documento digital y un objeto digital estriba en los metadatos agregados: un texto producido en un procesador de palabra o una imagen recién salida de un escáner son documentos digitales, pero no son objetos digitales. Carecen de esa información adicional en forma de metadatos y procedimientos que activan la posibilidad de su preservación. Sin ellos, pueden almacenarse sin duda, pero no se podrán preservar. El concepto no existe solo en bibliotecas: el "Glosario de Terminología Archivística" SAA parte también del concepto de *objeto digital* y con la misma acepción: "[…] una unidad de información que incluye propiedades (atributos o características del objeto) y que puede incluir también métodos (medios para realizar operaciones con el objeto)" (Pearce-Moses 2005).

Todos los Sistemas Integrados de Gestión Bibliotecaria (*Integrated Library System* o ILS) y los Sistemas de Gestión de Documentos de Archivo (*Records Management Systems* o RMS) deben producir de origen objetos digitales y no simples documentos

5 Se entiende aquí por *entidad documental* todo tipo de manifestación de un hecho, idea o conocimiento: texto, imagen, sonido, etcétera, expresada en cualquiera de sus variantes: libro, revista, documento de archivo, fotografía, partitura, pieza musical, filme, página web, etcétera.

digitales; de hecho, ése es uno de los requisitos indispensables de los buenos sistemas de estos tipos.[6] Esto puede hacerse extensivo a los repositorios científicos y de datos.

Por lo anterior, resulta indispensable agregar metadatos a los documentos digitales. Cuando se piensa en metadatos, vienen a la mente aquellos usados para la descripción del documento, pues en efecto son los más comunes. Pero por lo general, cada tipo documental requiere con mayor o menor énfasis de otros metadatos adicionales más allá de la descripción. Si el documento será preservado a largo plazo, requiere de ciertos metadatos específicos para este propósito, sin los cuales ese proceso de preservación será imposible. Además, dependiendo del tipo de documento, pueden agregarse a ellos además de los metadatos descriptivos y de preservación metadatos tecnológicos, documentales, de contexto, de estructura, funcionales, jurídicos, administrativos, de procedimiento, de ubicación, de relación, de almacenamiento, de autenticidad, de privilegios o restricciones a su acceso y uso, de modificabilidad, de modularidad, de interoperatividad, de dinamismo, etcétera. Dependiendo de sus características y volumen, algunos metadatos pueden autocontenerse dentro de cada documento en sí, y otros pueden crearse en ficheros externos con sus correspondientes vínculos para no agrandar cada documento con metadatos idénticos. Inclusive, para algunos tipos de documentos se recomienda ya el uso de "metametadatos": datos acerca del origen de los metadatos y su compilación, con objeto de establecer su rigor, precisión, autenticidad, etcétera y, por ende, su confiabilidad. A la fecha, existen varios estándares y recomendaciones para

---

6 *Sistema Integrado de Gestión Bibliotecaria*. Conjunto de reglas y de recursos informáticos destinados a la automatización y administración de las diferentes actividades de una biblioteca relativas a colecciones, servicios o usuarios.
Sistema de Gestión de Documentos de Archivo. Conjunto de reglas que rigen la producción, almacenamiento, uso, mantenimiento y disposición de documentos de archivo de un cierto productor, además de las herramientas tecnológicas usadas para implementar esas reglas. Se le conoce también como "Sistema de Administración de Documentos de Archivo".

diversos metadatos para distintos documentos –METS, MODS, MIX por citar algunos–, y el más completo de todos es PREMIS (2016). No es objeto de este texto analizar la conveniencia de utilizar uno u otro. El punto central de todo lo anterior es que sin suficientes metadatos agregados para preservación en los documentos digitales, ese proceso se vuelve imposible y por ello no deben omitirse; obviamente agregar metadatos implica un costo para cada documento el cual debe ser calculado y contabilizado.

Algunos autores engloban el *costo de registrar* junto con el *costo de editar*; es decir, la edición también abarca el registro y agregado de metadatos al documento. Esto no es relevante, siempre y cuando se desagreguen y sumen adecuadamente todos y cada uno de los pasos y sus costos correspondientes sin omitir ninguno; como ejemplo al respecto, puede verse el Portal Académico del Consejo de Bibliotecas de Ontario, Canadá (Scholars Portal 2013). En Archivística, muchos de los procesos enumerados en este texto en los costos de editar y registrar son agrupados de otra forma; algunas veces denominados como procesos de *pre-ingesta* e *ingesta*;[7] en realidad cómo se agrupen no es lo más importante, sino estar conscientes de que hay pasos y etapas previas indispensables al ingreso de los documentos a sus sistemas de gestión y/o preservación, los cuales aseguran su calidad técnica y documental. Al igual que la etapa anterior, debe tomarse en cuenta que esta es una etapa que implica en gran medida proceso intelectual y experiencia, por lo que con cierta frecuencia implica procesos de capacitación y entrenamiento del personal que conllevan un costo y una curva de aprendizaje adicionales.

7 Aunque existen diversas definiciones, agrupamientos y matices, definimos aquí a la *ingesta* en un sistema archivístico como el proceso que da de alta oficialmente a los documentos de archivo en ese sistema; la *pre-ingesta* son todos los pasos previos en preparación a ese proceso.

Costo de almacenar

Éste es uno de los costos que ha experimentado los cambios más notables en la última década. Mark Kryder –exdirector tecnológico del fabricante de discos magnéticos Seagate– afirmó de forma semejante a la Ley de Moore[8] que la cantidad de almacenamiento de datos que pueden alojarse en una cierta área de un medio magnético se duplica cada 18 meses. Si bien esto no es ya del todo cierto, en efecto el crecimiento ha sido y sigue siendo notable. Las capacidades actuales de una sola unidad de disco para almacenamiento alcanzan varios Terabytes – 10^{12} bytes o billones de caracteres. El cambio en los costos es apabullante: hablando gruesamente, comprar 1 Gigabyte de almacenamiento magnético en disco costaba más de cien mil dólares en 1980; en el año 2000 ya costaba unos quince dólares; comprar esa misma capacidad cuesta en promedio poco menos de tres centavos de dólar en 2021.

Hasta hace una década, calcular el costo de almacenar para bibliotecas, archivos, repositorios e instituciones afines era una tarea relativamente sencilla. Típicamente consistía en calcular el número total de bytes a guardar derivado de un cierto número de documentos –considerando un incremento gradual a dos o tres años– y presupuestar en consecuencia el costo de adquirir una o varias piezas de discos duros suficientes para contener el número de registros en cuestión en el servidor de cómputo de la organización. La división del costo total de los discos entre el número de registros nos daba el costo unitario por cada uno de ellos. Por lo general, se repetía el ejercicio considerando como alternativa cintas, cartuchos o DVD en vez de los discos magnéticos para contar con una copia adicional de los documentos fuera de línea.

Esto sigue siendo sin duda una opción válida hoy en día: la diferencia actual estriba en que partiendo del supuesto de que la organización posee y controla sus propios equipos informáticos, el

8 La Ley de Moore es un principio empírico establecido en abril de 1965 por Gordon E. Moore, cofundador de la empresa Intel, el cual establece que la capacidad de un procesador de cómputo se duplica cada 18 meses.

almacenamiento de grandes cantidades de documentos no se realiza a través de la adquisición de discos magnéticos individuales, sino de la adquisición de *clusters*; esto es, cúmulos o agrupamientos de discos, ya que este tipo de estructura reduce los costos por volumen. En esta modalidad, se asocia a un servidor uno o varios cúmulos de discos hasta lograr la capacidad de almacenamiento deseada. En esta estructura, existen ciertos factores que inciden sobre los precios en general: número de discos en cada cúmulo, la capacidad y velocidad de cada uno de ellos, y si son fijos o removibles.[9] No obstante, para fines de costeo en forma gruesa puede estimarse que en la actualidad, comprar un Gigabyte (GB) de almacenamiento en disco magnético cuesta 0.029 dólares. Para la información que no requiere estar en línea; es decir, no se necesita presente al instante en todo momento, sigue existiendo la opción de los cartuchos de cinta, denominados genéricamente LINEAR TAPE-OPEN (LTO); cada uno de estos cartuchos puede guardar en la actualidad 24 Terabytes (TB) de información, lo cual da un costo grueso de 0.022 dólares por GB. Si la información es comprimida dentro de ellos, pueden llegar a almacenar hasta 60 TB, lo cual baja los costos hasta 0.009 dólares por GB. Esta opción de cartuchos tiene el inconveniente de estar fuera de línea para su acceso y de ser más lenta, sobre todo si está comprimida, pero como almacenamiento masivo no instantáneo es muy rentable, pues cuesta nueve dólares por Terabyte. Las opciones de almacenamiento opto-magnético siguen existiendo, pero solo son rentables hasta cierta escala media en la cual siguen siendo prácticas; a gran escala dejan de ser atractivas. Un GB de almacenamiento en DVD o Blu-ray cuesta alrededor de 0.040 dólares; en CD-ROM cuesta 0.250 dólares. Las opciones de almacenamiento en memorias de estado sólido o USB y tarjetas SD son bastante más caras y por lo mismo

9 Una unidad de disco removible consiste en un soporte fijo en el cual se inserta y extrae un disco magnético especial como si fuera un cartucho o casete; esto permite que una sola unidad maneje una cantidad mucho mayor de información, no simultánea. Requiere de un operador humano o robótico para cambiarlos.

se usan muy poco para estos fines. Debe tomarse en cuenta que todos estos dispositivos tienen una vida útil de entre tres a cinco años y deben ser reemplazados por nuevos medios de reciente generación.

Como alternativa a la compra de almacenamiento, y como uno de los grandes cambios de la última década, el advenimiento en los últimos años de cada vez más servicios de almacenamiento de datos en la nube por parte de incontables proveedores ha ido creciendo y se ha convertido en una opción adicional e interesante, aunque no obligatoria. Para definirlo de forma simple, el cómputo en la *nube* consiste en un conjunto de recursos informáticos de equipo, programas, almacenamiento, procesamiento, comunicación, información, etcétera, que pueden ser rápida y ubicuamente suministrados por un proveedor como servicio vía una red y ampliamente escalados en función de las necesidades de un cierto usuario. Este concepto difiere de sus predecesores en que hasta antes de él, el modelo comercial del suministro de equipo de cómputo, *software*, comunicaciones, etcétera, fue manejado como la entrega de productos. En la nube, la provisión de estos insumos informáticos se otorga a través de la red *como un servicio* en lugar de como un producto, y es suministrado al usuario al igual que servicios comunitarios de electricidad, agua, o gas, pagando únicamente por lo que se consume; si se desea abundar en estos conceptos, véanse para más detalle Delgado (2013) y Voutssás (2013). En particular, dentro de los varios "modelos de servicio" en la nube, se encuentra el denominado "Almacenamiento como servicio" *(Storage as a Service* o STaaS*)*. En esta variedad de servicio, el usuario puede rentar a un proveedor la cantidad de almacenamiento electrónico que desee, accesible a través de la red, incrementándolo o disminuyéndolo al instante según sus necesidades.

El creciente número de proveedores ofertantes de este servicio, sus atractivos costos de inicio y sobre todo la facilidad de adquisición han hecho que en los últimos años el número de usuarios que lo consideran y adquieren haya crecido incesantemente. No obstante, es una opción que debe ser cuidadosamente estudiada

más allá de los números y costos, ya que conlleva otras fuertes connotaciones adicionales.

Para tener una idea, se describen a continuación los costos generales de algunos de los principales proveedores de este modelo de servicio STaaS de almacenamiento en la nube; los datos provienen de los sitios oficiales de cada proveedor y son los vigentes para fines del 2020:

1. Google Drive:
    - 15 GB: sin costo (30 GB para empresas)
    - 100 GB: 2 dólares mensuales
    - 1 TB: 10 dólares mensuales
    - 10 TB: 100 dólares mensuales
    - 20 TB: 200 dólares mensuales
    - 30 TB: 300 dólares mensuales

Rango: 0.010 a 0.020 dólares mensuales por GB

2. Amazon Simple Storage Service o S 3. El almacenamiento en la nube de Amazon tiene costos variables en función de cuatro variables: el tamaño de los documentos, el tiempo que se almacenen los documentos durante el mes, la frecuencia de acceso y transferencia de ellos, así como la administración y réplica de los documentos:
    - Primeros 50,000 GB: 0.023 dólares mensuales por GB
    - Siguientes 450,000 GB: 0.022 dólares mensuales por GB
    - Más de 500,000 GB: 0.021 dólares mensuales por GB

Rango: 0.021 a 0.023 dólares mensuales por GB

Estos costos pueden oscilar dependiendo de las cuatro características enunciadas anteriormente. Entre más volumen y menos frecuencia de acceso, los costos tienden a disminuir. Amazon ofrece además una opción de copias de seguridad extras a largo plazo con opción de recuperación de 1 minuto a 12 horas por 0.004 dólares mensuales por GB, y si tienen poco acceso –hasta dos veces al año– puede bajar hasta 0.00099 dólares mensuales por GB.

3. Microsoft OneDrive ofrece planes tanto para organizaciones, como a nivel personal. En casi todas las opciones sus costos van asociados a la compra de licencias de MS Office; en el nivel empresarial sus costos son:
    - 5 GB: sin costo
    - 50 GB: 2.4 dólares mensuales
    - 1 TB: 8.40 dólares mensuales
    - 5 TB: 14 dólares mensuales

Rango: 0.028 a 0.048 dólares mensuales por GB

4. IBM Cloud Object Storage tiene varios planes de precios de almacenamiento. Sus costos básicos son:
    - Hasta 500 GB: 0.021 dólares mensuales por GB
    - Más de 500 GB: 0.020 dólares mensuales por GB

Rango: 0.020 a 0.021 dólares mensuales por GB

Estos costos se consideran para datos con acceso y uso frecuente; si los datos son para archivo a largo plazo y no se usan con frecuencia, esto es, para respaldos, los costos van bajando hasta llegar a unos 0.007 dólares mensuales por GB.

5. Mega es una nueva versión del sitio Megaupload, el cual fue clausurado hace varios años por infringir derechos de autor; sus costos son:
    - 400 GB: 6 dólares mensuales
    - 2 TB: 12 dólares mensuales
    - 8 TB: 24 dólares mensuales
    - 16 TB: 36 dólares mensuales

Rango: 0.015 a 0.022 dólares mensuales por GB

Esta empresa impone además ciertos límites mensuales de la cantidad de transferencia de datos. Más allá de ellos hay que pagar sobreprecios.

6. DropBox es uno de los servicios de almacenamiento en la nube más antiguos, desde 2007. Su modelo de comercialización va dirigido a usuarios personales; sus costos son:
    - 2 GB: sin costo
    - 2 TB: 12 dólares mensuales
    - Mayor a 2 TB: 18 dólares mensuales

Rango: 0.006 a 0.009 dólares mensuales por GB

7. iCloud de Apple ofrece este servicio únicamente para usuarios poseedores de esta marca de equipos. Dado que existen muy pocos servidores de cómputo de esta marca a nivel de organizaciones, es también un servicio destinado principalmente a usuarios personales; sus costos son:
    - 50 GB: 1.4 dólares mensuales
    - 200 GB: 3.6 dólares mensuales
    - 2 TB: 12 dólares mensuales

Rango: 0.006 a 0.028 dólares mensuales por GB

8. Box es el servicio más antiguo de almacenamiento en la nube, data del 2005. Está dirigido también a usuarios personales; sus costos son:
    - 10 GB: sin costo
    - 100 GB: 5 dólares mensuales
    - Mayor a 100 GB: 15 dólares mensuales

Rango: 0.050 a 0.150 dólares mensuales por GB

Todos los costos anteriores no son absolutos; muchos de los proveedores ofrecen más de un tipo de plan, y por esta razón presentan algunas variaciones en función de volúmenes todavía mayores a los estipulados, frecuencia de uso, transferencia de datos, etcétera. Si la adquisición se hace conjuntamente con otros servicios en la nube, o con productos para manejo de datos masivos, los proveedores arman "paquetes" de servicios en bloque. No obstante, la lista anterior proporciona una buena idea de los costos promedio que se manejan actualmente por el almacenamiento en la nube, tanto a nivel de organizaciones, como de personas. A

pesar de su gran variabilidad, a partir de los rangos observados se encuentra que un costo promedio que sirva como base gruesa de inicio tanto a nivel personal como de organizaciones, oscila alrededor de 0.023 dólares mensuales por GB. Por supuesto puede hilarse más fino después de este número base dependiendo de las características deseadas por cada organización.

Si se comparan las opciones analizadas, puede verse que en realidad no existe una gran diferencia entre comprar el equipo de almacenamiento o rentarlo, pues, en cálculos gruesos, comprar un GB de almacenamiento cuesta en la actualidad entre 0.029 y 0.040 dólares, mientras que rentarlo cuesta entre 0.010 y 0.050 dólares, dependiendo del tipo de dispositivo y su velocidad. La diferencia fundamental entre ambos modelos que hace muy atractiva la renta estriba en que la compra se hace una vez cada tres a cinco años haciendo una erogación mayor mientras que el costo de renta se diluye al pagarlo mensual o anualmente. Otra diferencia importante entre ambas opciones consiste en que comprar almacenamiento sigue siendo una opción económica atractiva cuando la organización ya cuenta con la infraestructura informática necesaria para instalar y mantener ese almacenamiento: servidores informáticos de buena potencia, personal especializado para atenderlo, buena estructura de redes y telecomunicaciones, equipo auxiliar de acondicionamiento de aire, energía eléctrica ininterrumpida, seguridad física de acceso, etcétera. Si la organización no cuenta con todo esto, la inversión inicial en infraestructura sumada al costo de comprar el almacenamiento supera con mucho el costo de rentarlo. Por tanto, uno de los factores más importantes para decidir este aspecto consiste en si la organización ya cuenta con esa infraestructura informática adecuada. Cuando ya se tiene, en efecto la compra de almacenamiento masivo se reduce a la adquisición de los cúmulos de discos analizados, y los costos enunciados aplican. En caso contrario, la renta es la mejor opción económica.

Por lo anterior, debe tenerse en mente que el gran atractivo de la opción de renta en la nube no estriba por tanto en un gran ahorro en costos de almacenamiento en sí, sino precisamente en

la facilidad de adquirirlo de una manera simple e instantánea, haciéndolo crecer o disminuir según las necesidades de la organización, lo que evita casi en su totalidad el costo de adquirir y administrar una infraestructura informática y diluye su pago anual o mensualmente. Ése ha sido desde el inicio el gran argumento de venta de este tipo de servicios, y sin lugar a dudas esto es muy conveniente. Sumando infraestructura más almacenamiento, la renta en la nube ofrece una mejor proporción costo-beneficio al usuario al reducir la inversión directa de la organización en tecnología de cómputo y telecomunicaciones. No obstante, muchos autores coinciden en que esto es válido para lapsos cortos; para el largo plazo, los costos de los actuales servicios comerciales de almacenamiento en la nube no resultan tan económicos y rentables (Rosenthal *et al.* 2012, 7). El largo plazo es precisamente el caso de la preservación documental digital, por lo que los estudios comparativos de la adquisición contra la renta deben ser elaborados cuidadosamente proyectando para diversos periodos y escenarios.

Además, más allá de los costos, los servicios en la nube implican una serie de considerandos adicionales muy delicados que requieren insoslayablemente que la decisión no esté basada solo en factores económicos. No es objeto de este documento tratar todos los demás factores más allá de estos últimos que deben ser ponderados para una decisión de almacenamiento en la nube, ya que ese es un tema sumamente extenso y complejo; si se desea abundar en ellos, véase el apartado "Ventajas y desventajas del cómputo en la nube" en Voutssás (2013). Empero, para su comprensión se presenta aquí una lista breve:

- *Pérdida de control.* En esencia, el principal problema acerca del almacenamiento en la nube consiste en que las organizaciones pierden gran parte del control que normalmente ejercen sobre su información, en múltiples sentidos. Para algunas organizaciones –bibliotecas públicas, museos, etcétera–, esto no representa ningún problema grave, y con algunas medidas generales de seguridad informática puede compensarse. Para otras organizaciones, como los archivos,

este es precisamente el punto álgido de la conveniencia o no del uso de servicios en la nube, ya que comprometen los puntos cruciales sobre los que descansan los principios de la preservación archivística de documentos confiables y auténticos: su existencia, custodia y, en su caso, transferencia y destrucción efectiva de las diversas copias de los documentos de archivo.

- *Pérdida de la propiedad de los datos*. El establecimiento preciso de la propiedad de una organización sobre su información almacenada en la nube debe formar parte esencial del contrato de servicio. Algunos proveedores que colectan y almacenan "datos como servicio" para una organización se reservan el derecho de guardar parte de la información colectada. La mayoría de las redes sociales hacen lo mismo. En algunos servicios, existe por tanto la pérdida –o al menos la cesión parcial– de la propiedad de datos por parte del usuario. Las organizaciones, como bibliotecas y repositorios, deben cerciorarse siempre de mantener sus derechos de propiedad y de que el proveedor de la nube no adquiera inadvertidamente derechos de propiedad, concesión de licenciamientos ni uso alguno sobre la información de la organización. Este punto se relaciona muy de cerca con los aspectos de seguridad de la información y privacidad de datos personales, los cuales tienen sus propios y serios riesgos en la nube.
- *Pérdida de jurisdicción legal de los documentos*. En los servicios en la nube, los datos pueden estar almacenados en uno o varios servidores ubicados físicamente en muy diversas localidades en el mundo; esto significa que esos servidores están bajo la jurisdicción legal de otro país. Como ejemplo claro de este problema, se encuentra el caso del

sitio Megaupload.[10] Algunos países; por ejemplo, Canadá, ya han legislado acerca de la obligatoriedad de que archivos y/o datos sensibles de interés nacional se almacenen dentro de la jurisdicción de ese país, en la nube o no.

- *Fallas en el servicio en la nube*. En este tipo de servicios, la alta dependencia de la red se hace muy evidente; si no hay acceso a través de ella, nada existe. Por esa razón los "Acuerdos de Niveles de Servicio" (*Service-Level Agreements* o SLA) con el proveedor son sumamente importantes. Típicamente, un buen proveedor debe poder garantizar de forma aceptable de acuerdo con estándares internacionales que su servicio en la red estará disponible al menos un 98 por ciento, conocido como *uptime*, un tiempo medio entre fallas o MTBF bastante espaciado, tiempos de reparación cortos, buena atención a llamadas de ayuda o servicio, respaldos frecuentes y adecuados, alta seguridad, etcétera. No todos los proveedores en la nube brindan el mismo nivel de garantía de sus servicios, por lo que es necesario cuidar este aspecto.

El proyecto InterPARES de preservación archivística digital elaboró una "lista de verificación" muy completa al respecto de la opción de alojamiento de documentos de archivo electrónicos en la nube, con el fin de medir previamente el manejo y riesgo de los diversos elementos ya mencionados por parte de ese tipo de proveedores (InterPARES 2015).

Como puede verse, al margen de consideraciones puramente económicas, existen otros factores técnicos, legales y administrativos de suma importancia que obligan a realizar estudios muy serios y completos acerca de la mudanza de almacenamiento hacia la

10 Megaupload fue un sitio de alojamiento de archivos en la nube desde 2005 con sede en Hong-Kong. En 2012 el dominio fue clausurado intempestivamente por autoridades estadounidenses por el cargo de violaciones a derechos de propiedad intelectual. Sus usuarios no recuperaron nunca sus ficheros almacenados.

nube; la decisión no debe tomarse automáticamente o solo basada en criterios económicos. Entre más sensible es la información de una organización, más deben cuidarse los aspectos de migración hacia a la nube. No es lo mismo, por tanto, almacenar en ella fichas catalográficas de libros en una biblioteca, que expedientes médicos: toda información es importante para la organización que la preserva, pero existen ciertos tipos mucho más susceptibles a fallas o errores.

Costo de actualizar

Todos los soportes de la información digital –discos, cintas, DVD, etcétera– sufren un deterioro físico como cualquier tipo de material. Durante muchos años, se invirtieron todo tipo de recursos para tratar de extender la duración de los soportes digitales. Después de cierto tiempo, se cayó en cuenta de esto era un objetivo que además de interminable era estéril; el problema real no es la poca duración de los soportes digitales. En realidad, no están hechos para durar eternamente; no tienen por qué. Esto no significa que su duración no sea un problema, por supuesto lo es. El punto central es que el problema principal no se encuentra ahí, ya que esa situación puede solucionarse con cierta facilidad con alguna de las técnicas destinadas al efecto. El verdadero problema está en la *obsolescencia tecnológica* como el mayor riesgo para la preservación y futuro acceso a la información digital. Esta obsolescencia es un problema mucho más amplio que no solo afecta los soportes de la información, sino también a sus dispositivos lectores o reproductores; a las aplicaciones o programas informáticos que operan o administran la información y a los sistemas operativos que controlan al computador. Por si esto fuera poco, los formatos en los que esa información está codificada digitalmente también sufren de obsolescencia y caducidad. Esto significa que a cualquier información sobre un soporte físico le llegará la obsolescencia más rápido de lo que el medio se deteriora en sí. Por ello, la duración de los soportes pasó a un segundo plano de importancia y preocupación.

La obsolescencia tecnológica está relacionada con dos principios asociados estrechamente entre ellos: la *permanencia* y la *accesibilidad*. La permanencia sí tiene relación con la ya mencionada duración del soporte: tiene que ver con que las cadenas de bits con la información sigan existiendo, esto es, permanezcan a lo largo del tiempo en buen estado sobre su soporte. Esto significa que el soporte y las características que propiamente guardan la información se conserven en buen estado: óxido férrico en discos o cintas magnéticos, motores de dispositivos, superficies reflejantes en discos ópticos, etcétera. Si las cadenas de bits sobre un cierto soporte no permanecen físicamente a lo largo del tiempo, cualquiera que sea la causa, la información no existirá. Como se estableció anteriormente, la *conservación* o *mantenimiento digital* tiene que ver con que los soportes u objetos físicos de un documento digital permanezcan en buenas condiciones a lo largo del tiempo, y debido a que son perecederos, su alcance es siempre de corto-mediano plazo. Pero el segundo elemento de la obsolescencia tecnológica, la accesibilidad, ha ido cobrando cada vez más importancia en los últimos años, y ha superado al problema de la permanencia. La accesibilidad tiene que ver con que la información –habiendo permanecido– pueda ser accedida a lo largo del tiempo; esto es, leída, reinterpretada y desplegada correctamente por medios tecnológicos. Esto significa que los soportes todavía sigan siendo legibles por medio de un dispositivo lector, que exista un programa informático que pueda leer la información en el formato en que haya sido codificada y que además puede volver a presentarla a un usuario con su apariencia típica. Por ejemplo, una hoja de cálculo elaborada en Lotus 1-2-3[11] guardada en un disquete de 5.25 pulgadas. Si la información está ahí todavía y en buen estado, ha tenido permanencia. Pero para poder acceder a ella se requiere de un lector de disquetes de esa medida correctamente acoplado a un computador, y además un programa informático que pueda reconocer correctamente esa información en su formato Lotus y desplegarla nuevamente al usuario, todo lo cual

11 Lotus 1-2-3 fue una hoja de cálculo muy popular en la década de los ochenta.

insoslayablemente involucra a un sistema operativo totalmente antiguo y descontinuado. Si no se cuenta con todos esos elementos perfectamente acoplados, no hay ya accesibilidad, a pesar de que haya habido permanencia, y por tanto el documento no se ha preservado. Como puede verse, no es cuestión solo de un soporte durable que ha tenido *permanencia*; se requiere del concurso de muchos otros componentes para que haya *accesibilidad*.

Existen varias técnicas ampliamente tratadas por numerosos autores y organizaciones para contender con la obsolescencia tanto en su parte de permanencia, como de accesibilidad: las cuatro principales fueron enunciadas originalmente por el Consejo de Bibliotecas y Recursos de Información (CLIR) de la Unión Americana por Garrett (1996), y siguen vigentes a la fecha con muy pocos cambios; de la más sencilla a la más compleja son: 1) réplica 2) re-copia 3) migración 4) emulación. No es objeto de este texto analizarlas técnicamente con detalle; lo que nos interesa son sus costos, por lo que simplemente se describen brevemente para contextualización.

La *réplica* consiste en la creación y almacenamiento de varias copias de la información en sitios distintos. Si existe una única copia de la información digital, en caso de falla, daño o accidente del soporte debido a desastres naturales, la información es altamente susceptible de perderse. Crear y almacenar varias copias de la información en diversos lugares reduce ese riesgo e incrementa las probabilidades de que cierta información sobreviva al tiempo y los eventuales percances. La *re-copia* –también llamada *refrescado*, *renovación* o *rejuvenecimiento*– consiste en la simple técnica de copiar los registros electrónicos con cierta periodicidad hacia otros soportes más nuevos, más "frescos" y de mayor capacidad. En esta técnica se copia el documento digital *a imagen*, sin modificación alguna. Por ello, se entiende que ni las plataformas que operan los documentos ni sus formatos internos cambian; los documentos simplemente se trasladan desde un soporte hacia otro considerado mejor, no obsoleto y generalmente de mayor capacidad: de un CD-ROM a un DVD, de una memoria de estado sólido a un cartucho, etcétera. Esta técnica pretende resolver la *permanencia*

evitando que los soportes de los documentos lleguen a deteriorarse físicamente por envejecimiento o uso, así como actualizar la tecnología de esos soportes con otra más nueva y por tanto más disponible. La *migración*, a diferencia de las anteriores, no se queda en un simple copiado de medios, sino que va más allá: implica el cambio de elementos internos de plataformas, programas y/o formatos. En esta técnica sí se cambia parte de la tecnología que opera internamente a los documentos; por ejemplo, cambios de versiones de los documentos tipo doc, xls o pdf; cambios de formatos de imágenes, de audio, video; cambios de bases de datos, etcétera. El propósito primordial de esta técnica es salvaguardar la integridad de los objetos digitales manteniendo la capacidad de los usuarios de acceder a ellos a lo largo de las cambiantes y diversas generaciones tecnológicas. Por su naturaleza, este proceso por lo general consume mucho más tiempo y recursos que la re-copia. Para abundar más en las recomendaciones acerca de los formatos, se recomienda ver el sitio acerca de la sostenibilidad de estos de la Biblioteca del Congreso de los Estados Unidos (Library of Congress s.d.). Finalmente, la *emulación* pretende reproducir la funcionalidad de un sistema informático obsoleto que ya no existe o ya no funciona. El ejemplo más conocido de esta técnica son los antiguos juegos electrónicos de video, como los originales de Nintendo o Atari. Estos pueden ser emulados en una computadora personal contemporánea; no es exactamente el mismo programa antiguo el que vemos en el actual computador: es un nuevo programa emulador que replica el funcionamiento del anterior para que funcione y se perciba igual. Cuando se usa la opción de MS-DOS que se encuentra en los sistemas Windows, en realidad el primero no existe como sistema operativo en la computadora; Windows se encarga de emular o replicar su funcionamiento para que su uso y percepción por parte del usuario sean semejantes a aquel. Prácticamente todos los CD-ROM con datos producidos en su "época de oro" durante la década de los noventa requieren de la emulación de sus entornos tecnológicos para poder ser leídos en la actualidad. Durante los últimos años, se ha considerado la conveniencia de que múltiples piezas de información se

"encapsulen" junto con todo su entorno tecnológico que permita explotarlas. De hecho, el principio básico de los documentos XML es precisamente el de encapsular cierta información con todo el entorno documental que requiere para ser reinterpretada.

Como puede observarse, las técnicas de conservación forman parte insoslayable de la preservación documental digital, y lo que es de nuestro interés: conllevan un costo. De acuerdo con su grado de complejidad, cada una de estas técnicas va implicando un costo de menor a mayor.

Para calcular los costos actuales de la *réplica*, se requiere definir de antemano el número de copias adicionales que se desea guardar. Aun cuando el almacenamiento principal se lleve a cabo en la nube –donde el proveedor crea sus propios respaldos–, es indispensable que la organización guarde al menos una copia de toda la información bajo su propia custodia; jamás debe dependerse al 100 por ciento de los respaldos de un proveedor externo. El cálculo se hace de manera semejante al costo de almacenar ya mencionado: calculando el precio de compra de un conjunto de ciertos dispositivos suficientes para el total de número de bytes a guardar. La diferencia consiste en que las réplicas por lo general no estarán en línea en discos magnéticos adosados a un computador, sino en dispositivos externos tipo DVD, cartuchos LTO, etcétera, lo cual reduce en cierta medida su costo. Recuérdese siempre que por cuestiones de seguridad es indispensable que las réplicas no residan en la misma instalación física que el almacenamiento principal; deben ubicarse en una instalación externa segura fuera del alcance de extraños.

Cada cierto tiempo, las réplicas requieren de su proceso de *recopia* o *refrescado* para garantizar que los dispositivos se mantienen "frescos" y con tecnología aún vigente. Cada dispositivo usado para almacenar la información tiene una vida útil indicada por el fabricante, independientemente de que se use con frecuencia o no; es decir, no son eternos. El uso frecuente de medios de almacenamiento para respaldos en discos magnéticos, opto-magnéticos re-escribibles, cintas o cartuchos, reduce el tiempo de su vida útil. Además, toda tecnología tiene una vigencia después de la cual se

hace obsoleta, sale del mercado y por esa razón se dificulta cada vez más conseguir medios o soportes, dispositivos lectores, refacciones, etcétera. Todos los discos ópticos, las cintas y cartuchos actuales son tecnologías con varias generaciones de desarrollo; no son ya los dispositivos originales. Por ejemplo, los cartuchos de cintas LTO van actualmente en su octava generación desde su arribo en 1990. Todas las primeras versiones de CD-ROM, cintas, cartuchos, etcétera, son prácticamente ilegibles, no por falta de permanencia, sino de accesibilidad; deben, por tanto, re-copiarse cada ciertos años. Otro ejemplo representativo de ello son los casetes de audio y video, cuya vida tecnológica fue relativamente breve; toda la información contenida en ellos que no ha sido recopiada se encuentra hoy en grave riesgo. La ventaja de este proceso es que cada nueva generación de dispositivos tiene un costo relativamente menor que la precedente. Pero el punto central es que no pueden guardarse por largos periodos sin refrescarlos o se volverán inaccesibles. El costo del refrescado de los dispositivos de almacenamiento no se presupuesta cada año en las organizaciones, pues no se realiza anualmente, pero es indispensable contemplarlo cada cierto lapso; un promedio de unos cinco años.

La *migración* conlleva un costo mayor a las técnicas anteriores. Dado que implica el cambio de elementos internos de plataformas, programas, y/o formatos; es decir, la tecnología que opera internamente a los documentos, hay que sumar el costo de este proceso al de los dispositivos para almacenamiento. Cada vez que hay un cambio de plataforma informática en un organización, es imperativo realizar este proceso, pues por lo general implica diferentes sistemas operativos, programas y aplicaciones informáticas, manejadores de bases de datos, y con frecuencia formatos documentales; por lo mismo debe contemplarse en estos casos la migración y sus costos asociados. Cuando el cambio de plataforma es radical –es decir, cambian las marcas de los equipos o programas informáticos–, este proceso además de imperativo se vuelve urgente, pues introduce un riesgo a la información almacenada en relación con la nueva debido al cambio sustancial de estructuras. Aun cuando no exista un cambio radical de plataforma,

es común que en las organizaciones se vayan dando actualizaciones periódicas a nuevas versiones de equipos, sistemas operativos, programas, etcétera, lo cual no es tan drástico como el cambio de marcas o productos informáticos, pero de todas formas va introduciendo cambios sutiles en los componentes que gradual y acumuladamente van afectando las características de la información guardada y guardable. Igualmente, los cambios eventuales de versiones a formatos de documentos por parte de los proveedores van introduciendo modificaciones graduales a sus estructuras documentales: doc, xls, pdf, tiff, mp4, etcétera. Debido a ello, es necesario hacer un estudio periódico de los cambios sufridos por las estructuras informáticas de la organización cada cierto tiempo; el promedio también es unos cinco años. Al costo del proceso de migración, hay que agregarle el costo de la adquisición de nuevos dispositivos de almacenamiento, pues por lo general se aprovecha el evento de la migración para *refrescar* estos dispositivos.

Finalmente, la última técnica de la *emulación* conlleva el costo de desarrollo de los nuevos entornos informáticos que simulen el manejo de la información en sus plataformas anteriores. Estos desarrollos implican por lo general un alto costo que va más allá de las posibilidades económicas y técnicas de las organizaciones promedio, por lo que generalmente se realiza adquiriendo desarrollos provenientes de terceros. El costo, en este caso, consiste en la compra de esos productos informáticos emuladores o como otra opción más de servicios en la nube: en ella se encuentra el modelo o variedad conocido como Emulation as a Service (EaaS) o "Emulación como Servicio". Como su nombre lo indica, el servicio consiste en proporcionar al usuario una plataforma informática que se comporte como alguna requerida por él y que por lo general se encuentra ya descontinuada. Utiliza componentes de emulación ya desarrollados por el proveedor que se interconectan con funciones modernas propias de la web para flujos de trabajo con propósitos de preservación digital (Von Suchodoletz y Rechert 2014).

Debido al alto costo de hacer desarrollos propios en las organizaciones y a que un emulador imita solo a una cierta plataforma muy específica, existe todavía en la actualidad un gran debate

a nivel de los preservadores acerca de su conveniencia y utilidad a largo plazo. En la gran mayoría de los casos relacionados con información documental para preservación, se usa esta técnica para poder acceder una vez más a documentos que ya no tienen accesibilidad por su obsolescencia tecnológica y, una vez logrado, reformatearlos a versiones más nuevas y operables en plataformas recientes, como es el caso de la reconversión de versiones muy antiguas de documentos provenientes de procesadores de texto, hojas de cálculo o presentadores, bases de datos, pdf, formatos de imágenes, audio o video, etcétera. Un ejemplo muy ilustrativo de ello es la versión actual número 11 de los documentos pdf de Adobe Acrobat; por lo general, su lector –el Acrobat Reader– puede acceder a documentos hasta cuatro o cinco versiones atrás, dependiendo de cómo se hayan guardado esos documentos en cuanto a compatibilidades retrospectivas, pero muy rara vez se puede acceder con este *software* a documentos de versiones previas.[12] Un emulador es una opción viable para acceder a ellos y, en su caso, reconvertirlos, pero su rentabilidad económica depende del número de documentos a actualizar, la importancia de los mismos, lo atrasado de las versiones, etcétera. En la gran mayoría de ocasiones, su uso no es rentable económicamente. No obstante, en ciertas ocasiones la rareza, unicidad o importancia de los documentos involucrados hace imperativa la incursión en esta técnica al margen de sus costos, por lo que debe tenerse en mente como una eventual opción y, en su caso, considerar presupuestarla y adquirirla.

Dado que el costo de actualizar no está presente nunca al momento de crear una colección digital nueva, tiende a ser olvidado, pero inexorablemente aparecerá cada cierto tiempo en los costos de preservación, y afectará al presupuesto de la organización con cierta periodicidad, por lo que es necesario tenerlo en mente para

12 La excepción a esta regla son los documentos guardados en el formato pdf-1-A, el estándar ISO 19005-1:2005 (Gestión de documentos - Formato de archivo de documentos electrónicos para conservación a largo plazo, Parte 1: Uso de PDF 1.4 (PDF/A-1)) https://www.iso.org/standard/38920.html.

incluirlo oportunamente en el costo total de la preservación digital cuando sea pertinente, evitando sorpresas hacia los directivos administradores.

De todo lo anterior, se desprende que el costo total de preservar documentos digitales consiste en la suma de todos los costos parciales que aquí han sido desglosados bajo un cierto agrupamiento hecho en lo personal; todo ello proyectado para un cierto contexto y entorno específico de documentos y un cierto lapso de tiempo.

Es de suma importancia no olvidar que muchas de estas etapas requieren de personal calificado y entrenado para realizarlas de manera adecuada, por lo que si el personal implicado en ellas no cuenta de antemano con esas características, es indispensable en todos las estimaciones de costos de proyectos incluir el correspondiente a la capacitación y entrenamiento del personal, ya sea en las respectivas fases, o como una capacitación global.

Conclusiones

La cantidad de información digital crece en el mundo a ritmos inéditos, y una buena parte de ella requiere de ser preservada. Cada vez más bibliotecas, archivos y repositorios están siendo designados para esa tarea conforme se requiere que más ciudadanos accedan a información académica, científica, pública gubernamental, etcétera. No obstante, todavía existe la percepción en muchas organizaciones de que la preservación digital es un proceso con costos ínfimos, y que una vez puesto un documento en una computadora, el costo de mantenerlo es mínimo y poco significativo. Nunca debe permitirse que esta percepción permee las capas directivas de la organización. El primer paso para la preservación digital consiste en hacer conciencia a todos niveles de su importancia y valor institucional, pero también, como toda preservación, de que lleva costos embebidos. Este problema se agudiza en las organizaciones públicas, ya que con frecuencia estos costos no son contemplados en sus presupuestos, o son incluidos de manera marginal. Para muchas de estas organizaciones, éstas son nuevas obligaciones, y con frecuencia no contaban con los recursos humanos, tecnológicos o económicos para ello, o al menos, no de inicio ni en cantidades suficientes. Esto se debe principalmente a que los costos de la preservación digital no son evidentes para todos. Muchos siguen pensando que esta tarea se reduce a adquirir una cantidad suficiente de dispositivos de almacenamiento; esto sucede con frecuencia entre los tomadores de decisiones y los financiadores de estos proyectos. El problema se agrava derivado de que los costos de la preservación digital son tangibles, mientras que sus beneficios son intangibles: es relativamente sencillo establecer sus costos numéricamente, pero no así sus beneficios, lo cual dificulta su justificación.

Se ha presentado a lo largo del texto una visión actualizada de los principales componentes que conforman el costo de la preservación digital, divididos en cinco rubros: costo de digitalizar, editar, registrar, almacenar y actualizar. Se ha tratado de hacer un análisis acerca de esos costos puesto al día a la luz de las opciones emergentes.

Si bien los costos de almacenamiento por unidad han seguido disminuyendo en la última década, no se observa ya el incremento tan notable en capacidades de dispositivos como en las dos décadas anteriores, ni la disminución tan espectacular en sus costos. Por tanto, la relación dinero/cantidad almacenada sigue mejorando, pero ya no a las tasas de tiempos anteriores; cada vez cuesta más trabajo a los fabricantes optimizar lo ya existente con las tecnologías actuales. En la práctica, esto significa que la cantidad de información a preservar crece más rápido que la eficiencia monetaria de los dispositivos para hacerlo.

El almacenamiento en la nube ha surgido como la nueva opción más notoria en los últimos años para este propósito pero, como ha sido analizado, en el largo plazo –que es el término en el que funciona la preservación– los modelos de comercialización de empresas particulares no son una solución tan rentable económicamente; muchos autores coinciden en ello. Además de ello, en muchas organizaciones, como los archivos, el almacenamiento en la nube introduce una serie de otros cuestionamientos que exigen revisiones y análisis adicionales muy cuidadosos para su adopción Por esta razón, están surgiendo iniciativas para construir grandes centros de almacenamiento digital por parte de entidades académicas, gubernamentales, etcétera, buscando mayores economías que los modelos comerciales de almacenamiento actuales, a la vez que se construyan como centros de muy alta confiabilidad.

Si bien la preservación digital se ve afectada por diversos factores, por supuesto el económico es de capital importancia en todo proyecto de esta naturaleza. Es necesario por tanto que el responsable de la preservación documental digital: bibliotecario, archivista u otro, se familiarice gradualmente con los diversos costos relativos a esta función para que pueda estar en capacidad de

calcular, interpretar correctamente y presentar de manera adecuada estos costos en el plan y presupuesto de su organización. Es indispensable que él pueda describir una correcta relación costo/beneficio de la función de preservación para que esta pueda ser plenamente percibida por los directivos institucionales y, en su caso, por los eventuales patrocinadores. El secreto del éxito que subyace en ello es el de poder convertir proyectos técnicamente viables en proyectos también económicamente viables. Esto requiere de un delicado balance que todo responsable de la preservación documental digital debe ir desarrollando con el tiempo, el estudio y la experiencia. Gran parte del problema consiste en que numerosos proyectos son presentados todos los días basados solo en factores tecnológicos, los cuales –siendo muy válidos desde ese enfoque– corren el riesgo de quedarse sin financiamiento al correr del tiempo. Esto no significa de manera alguna que el enfoque principal de los proyectos de preservación digital deba ser el económico y descuidar los factores técnicos, pues esto conduce inexorablemente a proyectos muy baratos pero de dudosa calidad que después de un tiempo deberán ser desechados, y que además pueden poner en riesgo futuro al material a preservar. Por eso, debe procurarse desde el diseño y durante el desarrollo de este tipo de proyectos un equilibrio cuidadoso entre beneficios satisfactorios, costos razonables y tecnología adecuada, a la vez que una presentación atractiva y convincente de todo ello.

Para lograrlo, como ha podido observarse, el conocimiento y estudio de todos los componentes asociados a los costos por parte de los responsables de la preservación digital, su manejo en forma integral y equilibrada siguiendo una metodología de aplicación, aunados a un poco de experiencia al respecto, maximizan las probabilidades de éxito en este tipo de proyectos al mantener los costos en niveles razonables y explícitos para todos en la organización y, lo más importante de todo, maximizan la permanencia a largo plazo de los acervos documentales digitales. Concluyo con una reflexión de Stewart Brand que resume todo lo anterior, y que a pesar de tener ya dos décadas sigue siendo válida: "[...] El almacenamiento digital es fácil; la preservación digital no lo es.

Preservar significa mantener la información almacenada catalogada, accesible y usable en soportes actuales, lo cual conlleva constante dinero y esfuerzo" (Brand 1999, 47).

Anexo 1

Algunos ejemplos de sitios con recomendaciones y estándares para digitalización, calidad, etcétera.

Para imágenes, audio y video:

- Smithsonian Institution. *Digitization Standards for Images, Audio and Video.* (2004) https://siarchives.si.edu/what-we-do/digital-curation/digitizing-collections.
- South Carolina Department of Archives & History. Electronic Records Management Guidelines. Digital Imaging, March 2008, Version 2 https://core.ac.uk/download/pdf/49235519.pdf.
- *Standards Related to Digital Imaging of Pictorial Materials* (2004). K.A. Peterson (Comp.) Library of Congress, Prints and Photographs Division https://www.loc.gov/rr/print/tp/DigitizationStandardsPictorial.pdf.

Para diarios:

- *National Digital Newspaper Program (NDNP), pdf Specification* (2008) https://www.loc.gov/ndnp/guidelines/NDNP_202123TechNotes.pdf.
- Skinner, Katherine; Schultz, Matt (2014). *Guidelines for Digital Newspaper Preservation Readiness*. Atlanta, Ga: Educopia, https://educopia.org/wp-content/uploads/2018/06/Guidelines_for_Digital_Newspaper_Preservation_Readiness_0.pdf.

Para arte digital:

- Erpanet: The Archiving and Preservation of Born-Digital Art Workshop (2004). *Briefing Paper for the Erpanet Workshop on Preservation of Digital Art* https://www.erpanet.org/events/2004/glasgowart/briefingpaper.pdf.

Referencias bibliográficas

Todas las referencias electrónicas verificadas como existentes y exactas al 2 de marzo de 2021.

American Library Associations (ALA) y Association for Library Collections and Technical Services (ALCTS). 2007. http://www.ala.org/alcts/resources/preserv/defdigpres0408.

Archivo General de la Nación (AGN). 2015. *Recomendaciones para proyectos de digitalización de documentos*. México: Archivo General de la Nación. https://www.gob.mx/cms/uploads/attachment/file/146401/Recomendaciones_para_proyectos_de_digitalizacion_de_documentos.pdf.

Bote, Juanjo; Belén Fernández-Feijoo y Silvia Ruiz. 2019. The cost of Digital Preservation: A methodological analysis. *Procedia Technology*, vol. 5 (2019): 103-111. https://doi.org/10.1016/j.protcy.2012.09.012.

Brand, Stewart. Escaping the digital Dark Age. 1999. *Library Journal*, vol. 124, núm. 2 (1999): 46-49. https://rense.com/general38/escap.htm.

Calhoun, Patrick, David Akin, Brett Zimmerman y Henry Neeman. 2019. Large scale research data archiving: Training for an inconvenient technology. *Journal of Computational Science*, vol. 36 (2019). https://doi.org/10.1016/j.jocs.2016.07.005.

Delgado, Alejandro. "La Nube". *Legajos*, Boletín del Archivo General de la Nación. México, AGN, 7ª época, año 4, núm. 16 (abril-junio 2013): 107-122. http://iibi.unam.mx/p_a/archivistica/AGN%20legajos16-delgado.pdf.

Digital Libraries Federation (DLF). 2002. *Benchmark for Faithful Digital Reproductions of Monographs and Serials.* The Digital Library Federation Benchmark Working Group, 2001-2002. http://www.diglib.org/standards/bmarkfin.htm.

Gantz, John y David Reinsel. 2012. *The Digital Universe in 2020.* International Data Corporation (IDC). Patrocinado por EMC Corporation. https://www.slideshare.net/arms8586/the-digital-universe-in-2020.

___ 2007. *The Expanding Digital Universe.* An IDC White paper. (International Data Corporation). Patrocinado por EMC Corporation. March 2007. https://web.archive.org/web/20090612013506/http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf.

Garrett, John. 1996. *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information*. CLIR Commission on Preservation and Access and RLG. https://clir.wordpress.clir.org/wp-content/uploads/sites/6/pub63watersgarrett.pdf.

Hendley, Tony. 1998. *Comparison of Methods & Costs of Digital Preservation*. British Library Research and Innovation Report 106. http://www.ukoln.ac.uk/services/elib/papers/tavistock/hendley/hendley.html.International Council on Archives (ICA) (2015). ICA Terminology Database. Entrada por "preservación digital" http://www.ciscra.org/mat/mat/term/3062.

The International Research on Permanent Authentic Records in Electronic Systems (InterPARES). 2015. *Checklist for Cloud Service Contracts. Final Report.* https://interparestrust.org/assets/public/dissemination/NA14_20160226_CloudServiceProviderContracts_Checklist_Final.pdf.

The International Research on Permanent Authentic Records in Electronic Systems (InterPARES). 2008. *International Standards Relevant to the InterPARES 3 Project - General Study 04.* Sherry Xie y Joanna Hammerschmidt.http://interpares.org/ip3/display_file.cfm?doc=ip3_canada_gs04_international_standards.pdf.

___2005. *The Long-term Preservation of Authentic Electronic Records.* http://www.interpares.org/%5C/book/index.cfm.

ISO/IEC 2 5012:2008. *Software engineering - Software product Quality Requirements and Evaluation (SQuaRE)-Data quality model.* https://www.iso.org/standard/35736.html.

Kingma, Bruce. 2000. The Costs of Print, Fiche, and Digital Access: The E arly C anadiana O nline P roject. *D-Lib Magazine* (f ebrero 2000), vol. 6, núm. 2. http://www.dlib.org/dlib/february00/kingma/02kingma.html.

Library of Congress (s.d). *Sustainability of Digital Formats* https://www.loc.gov/preservation/digital/formats/.

Model Requirements for the Management of Electronic Records (MoReq2). 2002. 2nd version, European Commission: Bruselas, Bélgica. Glosario. https://cdn.ymaws.com/irms.org.uk/resource/resmgr/moreq/presentations/docubarcelona_moreq2_mvpalma.pdf.

Morris, Robert y B. J. Truskowski. 2003. The evolution of storage systems. *ibm Systems Journal*, vol. 42, núm. 2 (2003): 205-217. DOI: 10.1147/sj.422.0205.

Pearce-Moses, Richard. 2005. *A Glossary of Archival and Records Terminology.* Chicago: Society of American Archivists. Entradas por "digital object" y "conservation".

Premis-Data Dictionary for Preservation Metadata, Version 3.0. 2016. Library of Congress. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

Rosenthal, D avid; D aniel R osenthal, E than M iller, I an A dams, Mark Storer y Erez Zadok. 2012. The Economics of Long-Term Digital Storage. *The Memory of the World in the Digital Age: Digitization and Preservation.* Santa Cruz, California. https://web.stanford.edu/group/lockss/resources/2012-09_The_Economics_of_Long-Term_Digital_Storage.pdf.

Scholars P ortal. 2013. A S ervice of t he O ntario C ouncil of U niversity L ibraries. h ttps://learn.scholarsportal.info/all-guides/handling-digital-archives/concepts/#SIPAIPDIP.

Von Suchodoletz, Dirk y Klaus Rechert. 2014. *Emulation as a Service – Framework for Curation and Rendering of Complex Digital Objects.* Cu rate G ear 2 014. h ttps://ils.unc.edu/digccurr/curategear2014-talks/von-sucholodetz-curategear2014.pdf.

Voutssás, Juan. 2009. *Preservación del Patrimonio Documental Digital en México*. México: UNAM, Centro Universitario de Investigaciones Bibliotecológicas. http://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/L49.

Voutssás, J. y A. Barnard (coords.). 2013. Documentos de Archivo en la Nube: evolución y problemática. *Legajos*, Boletín del Archivo General de la Nación. México. Año 5, núm. 17. (Jul.-Sept. 2013): 77-140. http://iibi.unam.mx/voutssasmt/documentos/legajos17_nube_corto.pdf.

___. 2014. Glosario de preservación archivística digital versión 4.0 México: UNAM, Instituto de Investigaciones Bibliotecológicas y de la Información. http://ru.iibi.unam.mx/jspui/bitstream/IIBI_UNAM/L93/1/glosario_preservacion_archivistica_digital_v4.0.pdf.

Zeller, Jean-Daniel. 2010. Cost of digital archiving: Is there an universal model? *Eight Conference on Digital Archiving*, abril 2010, Ginebra. https://regarddejanus.files.wordpress.com/2010/05/costsdigitalarchiving-_jdz_eca2010.pdf.

Existe además una bibliografía acerca de sitios con recomendaciones para calidades en proyectos de digitalización: *Digital Conversion – Documents & Guidelines. A Bibliographic Reference* (2009). Está disponible en http://www.digitizationguidelines.gov/guidelines/Guidelines_Bibliography-2009rev.pdf.

Economic factors of digital preservation: 2021 Review

Introduction

When it comes to money, everyone is of the same religion"
FRANÇOIS-MARIE AROUET "Voltaire", 18th Century French thinker, and an immensely wealthy man.

Digital information in the world keeps growing incessantly by leaps and bounds; according to some highly cited International Data Corporation or IDC studies – Ganz & Reisnel (2005) and (2012), the world produced 0.13 Zettabytes of information in 2005, 1.23 in 2010, 2.9 in 2012; 8.6 in 2015; it would produce just over 40 in 2020, and 175 Zettabytes are predicted in 2025.[1] All of these amounts have been accumulating annually over the past few decades. Most of such information is disposable in the short term, but even so, there remains an immense volume of it that must be preserved for the future. Obviously it comes in every conceivable type: catalogs, texts in books, magazines, and newspapers; photographs, maps, music, radio, film and TV, games, social networks, messaging and chats, calls, sporting events, medical analyses, invoices, scientific data, government files, and so on, in countless combinations and varied digital formats.

All the information which needs to be preserved has someone responsible for doing so; at least in theoretical terms it should have one. Many of them are in private companies producing and/or exploiting such information: book and magazine publishers,

1 1 Zettabyte = 1,000 Exabytes = 1,000,000 Petabytes = 1,000,000,000 Terabytes = 1,000,000,000,000 Gigabytes = 1,000,000,000,000,000 Megabytes = 10^{21} bytes.

music, film or entertainment companies, social network providers, the big search engines on the web, banks, insurance and financial systems, companies marketing goods and services, aviation, communications, and so on. However, there is still a large part that must be preserved by public organizations assigned with this responsibility: libraries, archives, museums, and repositories, as well as others belonging to one of the government sectors: executive, legislative, and judicial; public education, health, communications, cultural, social assistance systems, to mention a few.

Those responsible for these tasks in the corporate or private sector are usually clearly assigned within their organizations and have specific types of documents to preserve, but most of all, they have budgets to do so from the outset. Unlike them, those responsible for these tasks in the public sector are not always clearly assigned, frequently have to preserve a variety of types of information, and often do not have the budgets for the task, at least not initially or in sufficient quantities. The main reason for this is that the costs of digital preservation are not obvious to everyone. Many people still think of digital preservation as simply acquiring a certain number of storage devices; among these people are often the decision-makers and funders of these projects. The problem is magnified by the fact that the costs of digital preservation are tangible, while its benefits are intangible: its costs are relatively easy to establish numerically, but its benefits not so.

In recent decades, there has been a worldwide trend towards increasing the availability of public information and citizens' access to it. On the one hand, numerous laws on transparency and access to governmental public information have been issued globally, and on the other hand, there have been a series of worldwide movements in favor of open access to scientific, academic, cultural, etc. information. A considerable part of the aforementioned worldwide content includes precisely all these materials, which by their nature, must be preserved in a high proportion. For the same reason, the number of "obliged subjects" by the laws of access to public information as well as the number of "designated organizations" for the collection and preservation of scientific and

academic information in open access, open data, repositories, etc., has grown unusually in the last few years. We have gone from Petabytes to Exabytes and then to Zettabytes in just a couple of decades. Increasingly more public organizations are designated to collect and preserve part of this documentary heritage in their respective regions and countries: national, regional or state archives; specialized libraries and archives of the legislative and judicial branches; national libraries or large thematic libraries in art, science or humanities, among others; national or regional open science repositories, data repositories, and so on. Some of these organizations were already designated for these tasks, but others have only recently been appointed. Regardless of this, the amount of information they must preserve has grown exponentially in recent years and continues to do so every day.

But preserving digital documentary information does cost, and it is not only a matter of buying or renting many storage devices; there are other economic factors affecting digital preservation. There have been texts on this subject since the 1990s, for example Hendley (1998), Ashley (1999). At that time there were also comparative studies between paper, microfiche and digital preservation costs – e.g. Kingma (2000). In this regard, Morris & Truskowsky (2003: 206) established that in 1996 the inflexion point was reached at which storage on electronic devices already achieved better cost/benefit than on paper. Since the last century, countless studies have been conducted on the costs of digital document storage, with numerous perspectives, diverse approaches, and for all types of collections and documents. Costing models and methodologies have been created, formulas for their calculation, etc.: Bote *et al.* (2012), Calhoun *et al.* (2019), Zeller (2010), Rosenthal *et al.* (2012), as representative examples.

All these approaches have obviously evolved along with the technological, economic, social, etc. contexts. Personally, since 2009 I have dealt with the various factors affecting digital document preservation: technological, cultural, documentary, social, legal, and of course economic factors – (Voutssás, 2009). Now that more than a decade has passed since then, new elements have

been added to each of these factors, especially the economic ones. The global amount of information produced per year increased from 1.23 Zb in 2010 to 40 Zb in 2020. The costs of storage devices and their performance have continued to improve steadily. Countless companies now offer digitization, editing, etc., services. The advent of cloud storage services –virtually scarce at that time– has added new economic elements to the equation. Given the above: the growing need of public organizations to preserve more and more digital documentary information, as well as the change in the contexts and especially the options and economic factors for this purpose, it is now pertinent to review these components again in order to update the considerations in their analysis and decision making in this regard.

It should be noted at this point that there are nuances and contexts specific to each type of document or data to be preserved and to the organization that holds them, so the cost analysis cannot be reduced to a formula that applies universally. Preserving books, articles, archival documents, legal rulings, films, clinical analyses or data from telescopes, to name just a few, is not the same. Although they all have common principles and characteristics, each of them presents its own problems. It is therefore common to find in the literature numerous models of digital preservation costs specifically constructed by country, by organization, by type of document, etc. The purpose of this text is not to create yet another quantification model, but to study in perspective and in a comprehensive way common characteristics of digital documents to enable the establishment of general methodologies for the determination of costs from a practical approach, all this in the light of the associated emerging contexts.

It is also worthwhile at this point to establish some definitions for the purposes of this text. Nowadays, *digital preservation* emphasizes and highlights the stabilization of the content and form of the document; i.e., it comprises the logical objects, and its scope is always the long term. The ALA –American Library Association– defines it as: "[...] digital preservation combines policies, strategies and actions to ensure the accurate rendering of authenticated

content over time, regardless of the challenges of media failure and technological change." – (ALA & ALCTS, 2007). The ICA –International Council on Archives– defines *digital preservation* as: "[…] the specific process of maintaining digital materials during and through the different generations of technology over time, regardless of where they reside." – (ICA, 2015). On the other hand, *digital conservation* has to do with the support of the digital document, i.e., it comprises physical objects, and its scope is always short-medium term; it is synonymous with *maintenance*. It comes from the concept of *conservation* as presented in the SAA Glossary of Archival and Records Terminology: "[…] the stabilization of materials through… physical treatment to ensure that they survive in their original form as long as possible." – (Pearce-Moses, 2005).

To date, there is no standard which unifies the division of digital preservation costs into sub-categories or items; each author divides costs into sections arbitrarily, and the separation between groups is not absolute at all. In this paper, five types of digital preservation costs were considered for study: digitizing, editing, registration, storing, and updating. This is a personal and arbitrary categorization.

Itemization of costs

DIGITIZING COST

Strictly speaking, digitization is not part of digital preservation. In theory, a digital document is susceptible to preservation once it is created or produced and not before. However, although nowadays the vast majority of information produced is already born digital, there are still many collections in libraries, archives and repositories on "traditional" media that may be preserved in digital form, and therefore it is essential that these collections be previously digitized at some point in time, which is why this step is reviewed here, since it should be included if necessary to complete the cost analysis. Therefore, although in reality the preservation process of a digital document is required once the document exists, given that in practice preservation often cannot be disassociated from digitization, an analysis of this cost is included here. Obviously, if it is not to be performed, it can be omitted.

Digitizing is defined here as the process of converting a certain document which is in a "traditional" or analog format –paper, audio or video cassette, film, etc.– into a digital format, in order to be able to store, distribute and access it in this new form. A digital format is one that represents a documentary object in a totally numerical form, i.e., with digits –hence the name– by means of a pre-established and arbitrary abstraction or representation of those numbers.

The first point to consider when establishing projects of this nature is to "sift" from the very beginning the set of documents to be digitized through a series of "digitization criteria" which will

help to corroborate whether it is really necessary and convenient to digitize a certain documentary collection –or part of it– without incurring in idle, reiterative or useless costs. There are many examples of such criteria that can be adapted, so we will not go into further detail here.[1] Depending on the type of documents and organization, this step is also referred to as "selection", "screening", "pre-ingest", "value assessing", etc. In the context of archival practice, this is part of "appraisal"; the process in which a body of records is examined to determine its value for a certain institution, and for how long. This stage is particularly sensitive for those who wish to preserve Big Data, as their volumes can grow by leaps and bounds.

The second point to consider when digitizing documents is to determine the quality parameters of the resulting digitized one. These parameters must be thoroughly established from the beginning, since determining the quality of the produced document has a direct and significant impact on costs. Since always –and this has not changed over the years– higher quality in digital documents implies higher costs, and vice versa. Therefore, a careful balance must be established beforehand between the quality of the document and the cost of digitizing it. This has to do directly with the purpose of the resulting document: Is it for preservation or for distribution? If the documents are only to be used for distribution, the quality –and therefore the costs– can be reduced: web pages, simple copies for users, ephemeral collections, and so on. But if the purpose is the long-term preservation of these documents, the quality must be high; it is not convenient to produce digital documents of low quality, since it is impossible to add it later. However, this causes file sizes and therefore costs to increase

1 An excellent document including criteria for digitizing library and archival collections was prepared by *IFLA and ICA (2002). Guidelines for digitization projects for collections and holdings in the public domain, particularly those held by libraries and archives* https://www.ifla.org/files/assets/preservation-and-conservation/publications/digitization-projects-guidelines.pdf.

exponentially;[2] hence the importance of achieving the right quality/cost balance. Many digitization projects realize at the end of the process that they have obtained a set of low-quality documents not matching the standards or recommendations for a certain type of documents, and that the information they have built is useless for preservation purposes. Setting parameters below the necessary quality results in unusable collections in the future; but setting excessive quality parameters for digitization raises costs enormously, making the projects often unaffordable. This is why it is important to ensure that the right balance is struck from the outset. Generally and regarding the required quality, there are already recommendations made by countless organizations for virtually every type of document to be digitized: books, magazines, newspapers, negatives, audio, films, maps, music scores, archives, etc., considering also its purpose, facilitating this task, which should never be overlooked or omitted. It often happens that the purpose of digitization contains both objectives: to preserve and to distribute, so the organization must consider the existence of two copies of each digital document: a high-quality copy to preserve and a low-quality copy to distribute. In this case, high-quality copies are obviously created first, from which the low-quality copies are obtained, and not the other way around: this is impossible.

To help establish these concepts of information quality, it is recommended to review the ISO/IEC 25012:2008 Standard "General Data Quality Model for Structured Formats in Information Systems"; it was originally issued for data, but over time it has been extended to other types of digital information; it establishes fifteen features necessary for information quality: accuracy, precision, completeness, credibility, consistency, timeliness,

2 Exponential increase: An image digitized at 100 dots per inch or dpi contains 100 x 100 = 10,000 dots per square inch; if the quality is multiplied by two, 200 dpi, the result is an image of 200 x 200 = 40,000 dots; that is, quadruple to save, not double. If doubled again to 400 dpi, it contains 400 x 400 = 160,000 dots per square inch. Being the quadruple of the original resolution, the result is 16 times larger. This is what exponential means.

accessibility, exhaustiveness, standards compliance, efficiency, confidentiality, traceability, availability, portability, recoverability. For more information on other standards related to the preservation of archival records, see InterPARES (2008). In addition, at the end of this text an annex is provided with some examples of websites that include recommendations on quality for various documentary materials.

The third point consists of establishing and adding up the costs of the various inputs for digitization: the acquisition of the equipment to digitize: scanners, cameras, audio or video digitizing boards; probably some computers; the software or programs necessary for the task, the cost of the personnel dedicated to it and their training, and eventually certain physical facilities. This sum divided by the total number of documents to digitize, in a certain time, will give a unit cost per document and for the total of them. In this step it must always be considered whether the digitization will be done on-site, in the library or archive facilities with its own staff and equipment, or will be outsourced to a supplier. This is a simple cost/performance comparison between the two options. To do this, the grand total of the sum of the various components of the cost of digitizing a certain number of documents in a process performed on the organization's premises must be obtained and compared against the total cost of a vendor performing that process. In this step it is of utmost importance not to include costs of other stages in one of the options without including them in the other. In other words, the supplier is often asked to –in addition to digitizing– separate bound documents, edit and/or enhance the digitized output, add certain metadata, perform Optical Character Recognition or OCR, and so on. The requesting organization must ensure that identical processes are being compared, and therefore if these requirements are added to what is requested from the supplier, they must also be included in the local option. Very often, costs for editing, optimization, cataloging, etc., are considered in one of the two options without including them in the other, which obviously leads to biases and errors. It is therefore of paramount importance that a vendor's proposal clearly specifies which

processes are additional to pure digitization, so that equal processes and their respective costs and benefits can be compared. An additional factor to consider here is the completion time; usually a supplier can do it faster, which may influence the final decision, apart from the costs. The trend in recent years is that –in large volumes– outsourcing the documents to be digitized with a vendor is a more cost-effective option; in small volumes being generated gradually over time the opposite is true. The "Digital Library Federation" – DLF and the "Council on Library and Information Resources" – CLIR developed in 2017 a "digitization cost calculator" https://dashboard.diglib.org/ to help organizations and individuals without much experience in this regard to calculate the costs for digitizing documents on-site. While the calculator is still in experimental stages, it provides a practical idea of the elements required for digitization costing.

EDITING COST

In practice, any digital document to be preserved requires some preparation; very rarely does a document as it comes out of the word processor, scanner, or any other source meet the requirements for management and preservation. It requires refinement and added elements for that purpose, namely, editing.

Since it is a term with multiple meanings, it is convenient to specify at this point what is meant by "editing" for the purposes of this text; it should be understood here in its broad concept of revision, improvement, refinement and preparation of a certain text or work prior to its release or entry into a system to ensure a homogeneous and standardized quality. Many word processing software are referred to as "text editors", which adds to the confusion. Indeed, a word processor does some of the editing of a text, but it does not carry out all the tasks involved in full editing. The edition of a text or a record –in its broadest sense– means that once the desired and/or minimum quality of the items to be preserved has been established, someone has to verify that all the elements

that have been stipulated for such quality exist in each and every document at an adequate level. This task cannot be performed by a simple "word processor" or "image editor" software: it is a comprehensive process which involves a series of successive steps and is mostly an intellectual task performed by an individual, hence it is costly. But it is an extremely important step in preservation, as it ensures adequate quality throughout all documents. Omitting to do so jeopardizes such quality and consequently the preservation as a whole.

Digital documents to be preserved may come from two sources depending on their origin: either they were digitized from a document on "traditional" or analog supports, or they were already born digital from origin. It is irrelevant: both require revision and improvement.

To begin with, any digital document to be preserved, whatever its origin, must be a "faithful digital reproduction". The Digital Libraries Federation – DLF established almost two decades ago that all digital documents –whether original or digitized– must be optimally formatted and meet three indispensable characteristics: quality, permanence and interoperability – (DLF, 2002). *Optimal formatting* means that the document is encoded in a non-proprietary and/or open format of long duration and efficiency for that type of document; the concept of *quality* is related to its functionality and purpose of use already mentioned; i.e., distribution or preservation; its *permanence* is related to the ability to be accessible in the long term, and its *interoperability* refers to its bit chains being encoded in such a way that they are highly independent of the computational platform. In archival environments, specifications for records include similar but even stronger elements of construction and quality, such as those established in MoReq2 (2002) or InterPARES (2005). The latter establishes "trustworthiness" as an indispensable quality element of a record from its origin, which in turn is the sum of its "reliability", "authenticity" and "accuracy", all of which imply that its information is accurate, correct, truthful and pertinent; it has identity and integrity.

The possible processes for editing a digital item are numerous; depending on the type of document and its eventual digitization process, there are countless activities to be performed on it. For example, texts must keep certain forms in terms of their spatial layout within a page, with one or several columns, several margins and spacing, with different fonts, etc. Images –photographs, negatives, newspaper pages, etc.– are cropped, clarified or contrasted; scratches or stains are removed, OCR is applied, watermarks or logos are added, etc.; audios are enhanced, hiss removed, separated into channels, translated, etc.; videos and movies are "remastered" to improve images or audio, remove flaws or defects, add color, insert subtitles or credits, and so on.

Obviously, all these editing activities have a cost: they require specialized equipment, software and personnel. As with digitization, it is necessary to establish how many people and how much time per unit are required to perform this task for the total number of documents, and to add the cost of all the inputs. The total sum of this activity divided by the number of documents will give us the unit cost of editing, as well as the estimated time required to perform it. Similar to digitization, there are options to do it onsite or outsource it to a supplier. The cost and time comparison will help to make the decision. Likewise, it is imperative that costs to be compared between the internal and outsourced options be made on exactly the same specifications and requirements, so as not to bias the result. It should be taken into account when making this comparison that –since this is a stage that involves mostly intellectual process– the personnel who perform it must be highly qualified to do so; if they do not have already the previous experience and skills, a training process is essential, which entails an additional cost and a learning curve.

Registration cost

We denominate here *registration* the process enabling a digital item with conditions to be entered into a management system, which

will also allow it to eventually enter a preservation system. The difference with the previous process, *editing*, is that the former covers aspects of the technical quality inherent to the documents, while *registration* deals with their documentary aspects. This process involves various activities and levels of depth; depending on the organization and the type of document, the activities or steps may consist in: reception, inscription or filing, foliation, inventory, cataloging, classification, description, etc. Obviously, each institution applies its own rules and specifications.

When metadata is added to a simple digital document which complies with the requirement of being a "faithful digital reproduction", a "digital object" is obtained. Digital documents cannot be preserved: digital objects are preserved. The fundamental difference is that a digital document is any documentary entity[3] that has been originally created or converted to a form of representation based on numbers under a certain arbitrary pattern, in order to be stored or transmitted by means of electronic devices. A *digital object* –also called *information object*– is any documentary entity that has been numerically encoded under some format and assembled together with some set of metadata in such a way that it can be stored, searched, found, and used within a computer system; when necessary, it also contains the methods or procedures to perform operations on the object. Computer files with content and their corresponding metadata are physically and/or logically intertwined with each other. As can be seen, the difference between a simple digital document and a digital object lies in the added metadata: a text produced in a word processor, or an image fresh from a scanner are digital documents, but they are not digital objects. They lack the additional information in the form of metadata and procedures which enable them to be preserved. Without them, they can certainly be stored, but they cannot be preserved.

---

3 *Documentary entity* is understood here as any type of manifestation of a fact, idea or knowledge: text, image, sound, etc., expressed in any of its variants: book, magazine, record, photograph, music score, film, web page, etc.

The concept does not exist only in libraries: the SAA "Glossary of Archival Terminology" departs too from the concept of digital object and with the same meaning: "[...] a unit of information that includes properties (attributes or characteristics of the object) and may also include methods (means of performing operations on the object)" – (Pearce-Moses, 2005).

All "Integrated Library Management Systems" – ILS and "Records Management Systems" – RMS[4] must produce digital objects from their origin and not simply digital documents; in fact, this is one of the essential prerequisites of good systems of these types. This certainly can be extended to scientific and data repositories.

Therefore, it is indispensable to add metadata to digital documents. When thinking about metadata, usually those used for the description of the document come to mind, since they are indeed the most common ones. But each documentary type requires more or less emphasis on additional metadata beyond the description. If the document is to be preserved in the long term, it requires certain specific metadata for this purpose, without which the preservation process would be impossible. Furthermore –and also depending on the type of document– in addition to the descriptive and preservation ones, other metadata may be added, such as technological, documentary, contextual, structural, functional, legal, administrative, procedural, locational, relational, storage, authenticity, about privileges or restrictions on access and use, modifiability, modularity, interoperability, dynamism, and so on. Depending on their characteristics and volume, some metadata can be self-contained within each document itself, and others can be created in external files with their corresponding links so as not to enlarge each document with identical metadata. For

4 Integrated Library Management System. A set of rules and computer resources intended for the automation and administration of the different activities of a library related to collections, services or users.
Records Management System. A set of rules governing the creation, storage, use, maintenance and disposition of records of a certain creator, as well as the technological tools used to implement those rules.

some types of documents, the use of "metametadata" is already recommended: data about the origin of the metadata and its compilation, in order to establish its rigor, accuracy, authenticity, etc., and therefore its reliability. To date, there are several standards and recommendations for various metadata for different documents –METS, MODS, MIX– to name a few, the most comprehensive of all being PREMIS (2016). It is not the purpose of this text to analyze the convenience of using one or the other. The central point of all the above is that without sufficient metadata added for preservation in digital documents, this process will become impossible; therefore it should not be omitted; obviously adding metadata implies a cost for each document which must be calculated and accounted for.

Some authors include registration cost together with editing cost; i.e., the latter also includes the registration and addition of metadata to the document. This is not relevant, as long as each and every step and its corresponding costs are properly disaggregated and added up without omitting any of them; an example of this different grouping can be seen at the Scholars Portal (2013) of the Library Council of Ontario, Canada. In archival environments, many of the processes listed in this text as costs of editing and registration are grouped differently, often referred to as "pre-ingest" and "ingest";[5] actually how they are grouped is not the most important issue, but to be aware that there are steps and stages prior to the entry of documents into their management and/or preservation systems which will ensure both their technical and documentary quality. As with the previous stage, it should be taken into account that this is a step involving a great deal of intellectual process and experience, so it often involves staff training, which entails an additional cost and a learning curve.

---

5 Although there are various definitions, groupings and nuances, we define *ingest* into an archival system here as the process that formally enters the records into that system; pre-ingest are all the previous steps in preparation for that process.

Storage cost

This is one of the costs that has undergone the most notable changes in the last decade. Mark Kryder –former chief technology officer of magnetic disks manufacturer Seagate– stated in a manner similar to Moore's Law[6] that the amount of data storage which can be accommodated in a certain area of a magnetic medium doubles every 18 months. While this is no longer entirely true, indeed the growth has been and continues to be remarkable. Current capacities of a single disk drive for storage reach several Terabytes – 10^{12} bytes or trillions of characters. The change in costs has been staggering: roughly speaking, buying 1 Gigabyte of magnetic disk storage costed over a hundred thousand dollars in 1980; buying that same capacity in 2000 costed about fifteen dollars; buying it costs on average just under three cents on the dollar in 2021.

Until a decade ago, calculating the cost of storage for libraries, archives, repositories and related institutions was a relatively simple task. Typically it consisted of calculating the total number of bytes to be stored derived from a certain number of documents, considering a gradual increase over two or three years, and accordingly budgeting the cost of acquiring one or more pieces of hard disks enough to contain the number of records in question on the organization's computer server. Dividing the total cost of the disks by the number of records provided the unit cost. The procedure was usually repeated by considering tapes, cartridges or DVDs as an alternative to magnetic disks in order to have additional offline copies of the documents.

This is still undoubtedly a valid calculation today; the current difference lies in the fact that, assuming that the organization owns and controls its computer equipment, the storage of large quantities of documents is not done anymore through the acquisition of individual magnetic disks, but through the acquisition of

6 Moore's Law is an empirical principle established in April 1965 by Gordon E. Moore, co-founder of Intel, which states that the capacity of a computer processor doubles every 18 months.

clusters, i.e., sets of grouped disks, since this type of structure reduces costs by volume. In this modality, one or several disk clusters are associated to a server until the desired storage capacity is achieved. In this structure there are certain factors that affect prices in general: the number of disks in each cluster, the capacity and speed of each of them, and whether the drives are fixed or removable.[7] However, for coarse costing purposes it can be estimated that at present, buying one Gigabyte – GB of magnetic disk storage costs 0.029 dollars. For information which does not need to be online, i.e., does not need to be instantly present at all times, there is still the option of tape cartridges, generically called LTO – Linear Tape-Open; each one of these devices can currently store 24 Terabytes – TB of information, which gives a gross cost of 0.022 dollars per GB. If the information is compressed inside them, they can store up to 60 TB, which lowers costs to 0.009 dollars per GB. Obviously this cartridge option has the drawback of being offline for access and slower, especially if compressed, but as non-instantaneous mass storage it is very cost-effective, meaning 9 dollars per Terabyte. Opto-magnetic storage options still exist, but are only cost-effective up to a certain medium scale where they remain practical; at large scale they cease to be attractive. One GB of storage on DVD or Blu-ray costs around 0.040 dollars; on CD-ROM it costs 0.250 dollars. Storage options on solid-state or USB flash drives and SD cards are considerably more expensive and therefore rarely used for these purposes. It should be noted that all of these devices have a useful life of three to five years and must be replaced by fresh media from newer generations.

As a different alternative to storage acquisition, and being one of the big changes of the last decade, the advent in recent years of more and more data storage services in *the cloud* by countless providers has been growing and has become an additional

7 A removable disk drive consists of a fixed carrier into which a special magnetic disk is inserted and removed like a cartridge or cassette; this allows a single drive to handle a much larger, non-simultaneous amount of information. It requires a human or robotic operator to exchange them.

and interesting, though not mandatory, option. To define it simply, *Cloud Computing* consists of a set of IT resources of hardware, software, storage, processing, communications, information, and so on, which can be quickly and ubiquitously delivered by a provider as a service via a network and extensively scaled according to the needs of a certain user. This concept differs from its predecessors in the sense that until before it, the business model for the provision of computer equipment, software, communications, etc., was handled as the delivery of products. In the cloud, the provision of these computing inputs is delivered over the network as a service rather than as a product, and is provided to the user in the same way as community services such as electricity, water, or gas, paying only for what is consumed. In particular, among the various "service models" in the cloud, there is the so-called "Storage as a Service" or STaaS. In this type of service, users can hire with a provider the amount of electronic storage they want, accessible via the network, increasing or decreasing it instantly according to their needs.

The growing number of providers offering this service, its attractive start-up costs and, above all, the ease of acquisition have led to a steady increase in the number of users considering and acquiring this service in recent years. However, it is an option that should be carefully studied beyond the numbers and costs, as it entails other strong additional connotations.

To provide an idea, the general costs of some of the main providers of this STaaS cloud storage service model are described below; the data come from the official sites of each provider and are those in effect at the end of 2020:

1. *Google Drive*:
    - 15 GB: no cost (30 GB for businesses)
    - 100 GB: 2 dollars per month
    - 1 TB: 10 dollars per month
    - 10 TB: 100 dollars per month
    - 20 TB: 200 dollars per month
    - 30 TB: 300 dollars per month

Range: 0.010 to 0.020 dollars per month per GB

2. *Amazon Simple Storage Service* or *S3*. Amazon cloud storage has variable costs based on four variables: the size of the documents, the length of time the documents are stored during the month, the frequency of access and transfer of the documents, and the management and replication of the documents:
    - First 50,000 GB: 0.023 dollars per month per GB
    - Next 450,000 GB: 0.022 dollars per month per GB
    - More than 500,000 GB: 0.021 dollars per month per GB

Range: 0.021 to 0.023 dollars per month per GB

These costs can vary depending on the four characteristics listed above. The more volume and less frequency of access the costs tend to decrease. Amazon also offers an option for extra long-term backups with a 1-minute to 12-hour recovery option for $0.004 per GB per month, and for very little access –up to twice a year– it can go as low as $0.00099 dollar per month per GB.

3. *Microsoft OneDrive* offers plans for both organizations and individuals. In almost all the options its costs are associated with the purchase of MS Office licenses; at the enterprise level its costs are:
    - 5 GB: no cost
    - 50 GB: 2.4 dollars per month
    - 1 TB: 8.40 dollars per month
    - 5 TB: 14 dollars per month

Range: 0.028 to 0.048 dollars per month per GB

4. *IBM Cloud Object Storage* has several storage pricing plans. Its basic costs are:
    - Up to 500 GB: 0.021 dollars per month per GB
    - More than 500 GB: 0.020 dollars per month per GB

Range: 0.020 to 0.021 dollars per month per GB

These costs are considered for data with frequent access and use; if the data is for long-term archiving and is not used frequently, i.e., for backups, the costs drop to about 0.007 dollars per GB per month.

5. *Mega* is a new version of the *Megaupload* site which was shut down several years ago for copyright infringement; their costs are:
    - 400 GB: 6 dollars per month
    - 2 TB: 12 6 dollars per month
    - 8 TB: 24 6 dollars per month
    - 16 TB: 36 6 dollars per month

Range: 0.015 to 0.022 6 dollars per month per GB

This company also imposes certain monthly limits on the amount of data transfer. Beyond these limits, there are additional charges to be paid.

6. DropBox is one of the oldest cloud storage services, since 2007. Its marketing model is aimed at personal users; its costs are:
    - 2 GB: no cost
    - 2 TB: 12 dollars per month
    - More than 2 TB: 18 dollars per month

Range: 0.006 to 0.009 dollars per month per GB

7. *Icloud* from Apple offers this service only for users who own this brand of equipment. Since there are very few computer servers of this brand at the level of organizations, it is also a service intended mainly for personal users; its costs are:
    - 50 GB: 1.4 dollars per month
    - 200 GB: 3.6 dollars per month

- 2 TB: 12 dollars per month

Range: 0.006 to 0.028 dollars per month per GB

8. *Box* is the oldest cloud storage service, dating back to 2005. It is also aimed at personal users; its costs are:
   - 10 GB: no cost
   - 100 GB: 5 dollars per month
   - More than 100 GB: 15 dollars per month

Range: 0.050 to 0.150 dollars per month per GB

All the above costs are not absolute; many of the providers offer more than one type of plan, and therefore present some variations depending on volumes even greater than those stipulated, frequency of use, data transfer, etc. If the acquisition is made in combination with other cloud services, or with products for Big Data management, vendors assemble "bundles" of services as a block. Nevertheless, the above list provides a good idea of the average costs currently being driven by cloud storage, both at the level of organizations and individuals. Despite its great variability, from the ranges observed it is found that an average cost to serve as a coarse starting base for both personal and organizational level ranges around 0.023 dollars per month per GB. Obviously, it can be refined beyond this base number depending on the desired characteristics of each organization.

Comparing the analyzed options, it can be seen that there is not really a big difference between purchasing storage equipment and renting it, since –in rough calculations– buying a GB of storage currently costs between 0.029 and 0.040 dollars, while renting it costs between 0.010 and 0.050 dollars, depending on the type of device and its speed. The fundamental difference between the two models which makes renting very attractive is that the purchase is made once every three to five years, making an initial larger outlay, while the rental cost is diluted by paying monthly or annually. Another important difference between the two options is that buying storage is still an attractive economic option when

the organization already has the necessary IT infrastructure to install and maintain the storage: powerful computer servers, specialized personnel, good network and telecommunications structure, auxiliary air conditioning equipment, uninterruptible power supply, physical access security, etc. If the organization does not have all this in place, the initial investment in infrastructure plus the cost of buying the storage far exceeds the cost of renting it. Therefore, one of the most important factors in deciding this issue is whether the organization already has adequate IT infrastructure in place. When it already has it, in effect the purchase of mass storage is reduced to the acquisition of the aforementioned disk clusters, and the above indicated costs apply. Otherwise, renting is the best economic option.

Therefore, it should be kept in mind that the great attractive of the cloud rental option does not lie in a great saving in storage costs per se, but precisely in the ease of acquiring it in a simple and instantaneous way, making it grow or shrink according to the needs of the organization, avoiding almost entirely the cost of acquiring and managing an IT infrastructure and diluting its annual or monthly payment. This has been from the beginning the great selling point of this type of services, and with no doubt this is very convenient. By adding infrastructure plus storage, cloud rental offers a better cost-benefit ratio to the user, reducing the organization's direct investment in computing and telecommunications technology. However, many authors agree that this is valid for short periods of time; for the long term the costs of current commercial cloud storage services are not as economical and profitable – (Rosenthal *et al*. 2012, 7). The long term is precisely the case of digital document preservation, so comparative studies of acquisition versus rental must be carefully elaborated projecting for different periods and scenarios.

In addition, beyond costs, cloud services involve a series of very delicate additional considerations that inevitably require the decision to be based not only on economic factors. It is not the purpose of this paper to discuss all the other factors beyond the latter that must be weighed for a cloud storage decision, as that is

an extremely large and complex topic; however, a brief list is presented for purposes of comprehension:

Loss of control. In essence, the main problem with cloud storage is organizations losing much of the control they normally have over their information, in multiple ways. For some organizations –public libraries, museums, etc.– this is not a serious problem, and some general IT security measures can compensate. For other organizations, such as archives, this is precisely the crucial point of the convenience or otherwise of using cloud services, since they compromise the key points on which rest the principles of archival preservation of trustworthy and authentic records, their existence, custody, and when necessary the transfer and effective destruction of the various copies of the records.

- *Loss of data ownership.* The precise establishment of an organization's ownership of its information stored in the cloud should be an essential part of the service contract. Some providers that collect and store "data as a service" for an organization reserve the right to keep some of the information collected. Most social networks do the same. In some services there is therefore a loss, or at least partial transfer, of ownership of data by the user. Organizations as libraries and repositories should always ensure that they maintain their ownership rights and that the cloud provider does not inadvertently acquire ownership rights, licensing or any use of the organization's information. This point relates closely to the issues of information security and personal data privacy, which have their own serious risks in the cloud.
- *Loss of legal jurisdiction of documents*. In cloud services, data may be stored on one or several servers physically placed in many different locations around the world; this meaning that those servers are under the legal jurisdiction of other country. A clear example of this problem was the case of

the *Megaupload* site.[8] For this reason, some countries –such as Canada– have already legislated that sensitive files and/or data of national interest must be stored within the jurisdiction of the country, being in the cloud or not.

- *Failures in the cloud service.* In this type of services the high dependency on the network becomes very evident; if there is no access through it, nothing exists. For this reason, Service-Level Agreements or SLAs with the provider are extremely important. Typically, a good provider must be able to guarantee in an acceptable way according to international standards that its network service will be available at least 98%, known as *uptime*, a well-spaced *Medium Time Between Failures* or MTBF, short repair times, good support for help or service calls, frequent and adequate backups, high security, and so on. Not all cloud providers offer the same level of guarantee of their services, so it is necessary to take care of this aspect.

The InterPARES project about digital records preservation developed a very complete "checklist" regarding the option of hosting electronic records in the cloud, in order to measure in advance the management and risk of the various elements already mentioned by such providers – (InterPARES 2015).

As can be seen, apart from purely economic considerations, there are other technical, legal and administrative factors of utmost importance that make it necessary to conduct very serious and comprehensive studies before moving storage to the cloud; the decision should not be made automatically or only based on economic criteria.

The more sensitive an organization's information is, the more care must be taken when migrating it to the cloud. It is therefore

8 *Megaupload* was a cloud-based file hosting site based in Hong Kong since 2005. In 2012 the domain was abruptly shut down by U.S. authorities on charges of intellectual property rights violations. The site users never recovered their stored files.

not the same to store in the cloud catalog cards of books from a library as it is to store medical records: all information is important for the organization that preserves it, but there are certain types of information that are much more susceptible to failures or errors.

UPDATING COSTS

All digital information supports –disks, tapes, DVDs, etc.– suffer physical deterioration as any other type of material. For many years, all kinds of resources were invested to try to extend the life of digital media. After some time it was realized that this was a goal that was not only endless but also sterile; the real problem is not the short life of digital media. In fact, they are not made to last forever; they do not need to. This does not mean that their durability is not a problem; obviously it is. The central point is that the main problem is not there, since that situation can be solved quite easily with some of the techniques designed for that purpose. The real problem lies in technological obsolescence as the greatest risk for the preservation and future access to digital information. This obsolescence is a much broader problem that not only affects the information carriers: it also affects their reading or playing devices, the applications or computer programs which handle or manage the information and the operating systems that control the computer. As if this were not enough, the formats in which this information is digitally encoded also suffer from obsolescence and caducity. This means that any information on a physical medium will become obsolete faster than the medium itself deteriorates. For this reason, the duration of the media has taken a secondary level of importance and concern.

Technological obsolescence is related to two closely associated principles: *permanence* and *accessibility*. Permanence is related to the aforementioned duration of the medium: it has to do with the bit strings containing the information to continue to exist, i.e., to remain in good condition over time on their support: this means that the carrier and the characteristics that store the information

itself must be kept in good condition: ferric oxide on magnetic disks or tapes, device motors, reflective surfaces on optical disks, etc. Obviously, if the bit strings on a certain support do not physically remain over time, whatever the cause, the information will not exist. This means that the carrier and the characteristics that store the information itself must remain in good condition. As stated above, *digital conservation* or *maintenance* has to do with the physical objects or supports of a digital document remaining in good condition over time, and because they are perishable their scope is always the short-medium term. But the second element of technological obsolescence, *accessibility*, has become increasingly important in recent years, overtaking the problem of permanence. Accessibility has to do with the fact that information –having remained– is able to be accessed over time; that is, read, reinterpreted and displayed correctly by technological tools. This means that the carriers are still readable by a reading device, that there is a computer program that can recognize the information in the format in which it was encoded, and that can present it again to a user in its typical appearance. For example, a spreadsheet produced in Lotus 1-2-3[9] saved on a 5.25 inch floppy disk. If the information is still there and in good condition, it has had permanence. But to access it requires a floppy disk reader of the right size properly attached to a computer, as well as a computer program that can correctly read that information in its Lotus original format and display it back to the user, all of which inevitably involves an operating system totally old and discontinued. If all these elements are not perfectly coupled, there is no longer *accessibility*, even if there has been *permanence*, and therefore the document has not been preserved. As can be seen, it is not only a matter of a durable medium that has had permanence; it requires the participation of many other components for there to be accessibility.

There are several techniques widely discussed by numerous authors and organizations to contend with obsolescence, both

9 Lotus 1-2-3 was a very popular spreadsheet through the eighties.

permanence and accessibility: the four main ones were originally enunciated by the American "Council on Library and Information Resources" or CLIR by Garrett (1996), and are still in force today with very few changes; from the simplest to the most complex they are: 1) replication 2) re-copying 3) migration 4) emulation. It is not the purpose of this text to analyze them technically in detail; what interests us are their costs, so they are simply described briefly for contextualization:

Replication consists of creating and storing several copies of the information in different places. If there is only one copy of the digital information, in case of failure, damage or accident of the support due to natural disasters, the information is highly susceptible to be lost. Creating and storing several copies of the information in different places reduces this risk and increases the probability that certain information will survive time and possible mishaps. *Re-copying* –also called *refreshing, renovation*, or *rejuvenation*– consists of the simple technique of copying electronic records periodically to newer, "fresher" media with greater capacity. In this technique, the digital document is copied as an image, without any modification. It is therefore understood that neither the platforms that operate the documents nor their internal formats change; the documents are simply transferred from one medium to another that is considered better, not obsolete and generally of greater capacity: from a CD-ROM to a DVD, from a solid state memory to a cartridge, etc. This technique aims to solve the problem of permanence by preventing document media from physically deteriorating due to aging or use, as well as to update the technology of those media with newer and therefore more available technology. *Migration* technique, unlike the previous ones, is not just a simple copying of media, but goes further: it involves changing the internal elements of platforms, programs and/or formats. In this method, part of the technology that operates internally to the documents is modified; for example, changes in versions of doc, xls or pdf documents; changes in image, audio and video formats; changes in databases, etc. The primary purpose of this technique is to safeguard the integrity of digital objects while maintaining the ability

of users to access them throughout the ever-changing and diverse technological generations. By its nature this process is generally much more time and resource consuming than re-copying. For further recommendations on format recommendations, see the Library of Congress (n.d.) site on format sustainability. Finally, *emulation* aims to reproduce the functionality of an obsolete computer system that no longer exists or no longer works. The best known example of this technique are old electronic video games, such as the original Nintendo or Atari games. These can be emulated on a contemporary personal computer; it is not exactly the same old program which is seen on the current computer: it is a new emulator program that replicates the operation of the old one so that it works and looks the same. When using the MS-DOS option found in Windows systems, the former does not actually exist as an operating system on the computer; Windows takes care of emulating or replicating its operation so that its use and perception by the user is similar to that one. Virtually all data CD-ROMs produced in their "golden age" during the 1990s require emulation of their technological environments to be readable today. Over the last few years, the desirability of multiple pieces of information being "encapsulated" together with their entire technological environment allowing them to be exploited has been considered. In fact, the basic principle of XML documents is precisely to encapsulate certain information with all the document environment it requires to be reinterpreted.

As can be seen, conservation techniques are an unavoidable part of digital document preservation, and what is of interest to us: they have a cost. According to their degree of complexity, each of these techniques implies a cost from lower to higher.

To calculate the actual costs of *replication*, it is necessary to define beforehand the number of additional copies to be stored. Even if the main storage is carried out in the cloud –where the provider creates its own backups– it is essential that the organization keeps at least one copy of all the information under its own custody; it should never depend 100% on the backups of an external provider. Calculation is done in a similar way to the cost of

storage already mentioned: by estimating the purchase price of a set of certain devices sufficient for the total number of bytes to be stored. The difference is that the replicas will generally not be online on magnetic disks attached to a computer, but on external devices such as DVDs, LTO cartridges, etc., which reduces their cost to some extent. Always be reminded that for security reasons it is essential that the replicas do not reside in the same physical facility as the main storage; they should be located in a secure offsite facility out of the reach of outsiders.

From time to time, replicas require a *re-copying* or *refreshing* process to ensure that the devices are kept "fresh" and the technology is still current. Each device used to store information has a shelf life indicated by the manufacturer, regardless of whether it is used frequently or not; in other words, they are not eternal. The frequent use of storage media for backups on magnetic disks, rewritable opto-magnetic disks, tapes or cartridges reduces their useful life. In addition, all technology has a lifespan after which it becomes obsolete, goes out of the market and therefore it becomes more and more difficult to obtain media or supports, reading devices, spare parts, etc. All current optical discs, tapes and cartridges are technologies with several generations of development; they are no longer the original devices. For example, LTO tape cartridges are currently in their eighth generation since their arrival in 1990. All early versions of CD-ROMs, tapes, cartridges, etc., are virtually unreadable, not for lack of permanence, but for lack of accessibility; they must therefore be re-copied every few years. Another representative examples of this are audio and video cassettes, whose technological life was relatively short; all the information contained in them that has not been re-copied is today at serious risk. The advantage of this process is that each new generation of devices has a relatively lower cost than the previous one. But the central point is that they cannot be stored for long periods without refreshing, or they will become inaccessible. The cost of refreshing storage devices is not budgeted every year in a given organization, since it is not done annually, but it is essential to consider it every few years, on average about five years.

Migration entails a higher cost than the previous techniques. Since it involves changing the internal elements of platforms, programs and/or formats –i.e. the technology that operates internally to the documents– the cost of this process must be added to the cost of the storage devices. Whenever there is a change of computer platform in an organization, it is imperative to carry out this process, since it usually involves different operating systems, programs and computer applications, database management systems, and often document formats; therefore, migration and its associated costs must be considered in these cases. When the platform shift is radical –i.e. the hardware or software brands change– this process is not only imperative but also urgent, since it introduces a risk to the information stored in relation to the new one due to the substantial change of structures. Even though there is no radical change of platform, it is common for organizations to periodically update to newer versions of equipment, operating systems, programs, etc. This is not as drastic as a brand shifting on computer products, but it does introduce subtle changes in the components that gradually and cumulatively affect the characteristics of the information stored and saved. Likewise, the eventual changes of versions to document formats by suppliers gradually introduce modifications to their document structures: doc, xls, pdf, tiff, mp4, etc. For this reason it is necessary to carry out a periodic study of the changes undergone by the organization's IT structures, from time to time; the average is also about five years. To the cost of the migration process must be added the cost of acquiring new storage devices, since the *migration* event is usually used to refresh these devices.

Finally, the last technique of *emulation* involves the cost of developing new computer environments which simulate the management of information in their previous platforms. These developments usually involve a high cost that goes beyond the economic and technical possibilities of the average organization, so it is generally done by acquiring developments from third parties. The cost in this case consists in the purchase of these emulating software products or, as another option from cloud services, the

model or variety known as "Emulation as a Service" or EaaS. As its name suggests, the service consists of providing the users with a computer platform behaving like one required by them and which is usually already discontinued. It uses emulation components already developed by the provider interfacing with modern web-based functions for workflows for digital preservation purposes (Von Suchodoletz & Rechert 2014).

Due to the high cost of self-development in organizations and the fact that an emulator only imitates a certain very specific platform, there is still a great debate among preservation professionals about its convenience and usefulness in the long term. In the vast majority of cases related to digital documentary information for preservation, this technique is used to once again access documents that are no longer accessible due to technological obsolescence and, once achieved, to reformat them to newer versions operable on recent platforms, as in the case of the reconversion of very old versions of documents from word processors, spreadsheets or presenters, databases, pdf, image, audio or video formats, and so on. A very illustrative example of this is the current version number 11 of Adobe Acrobat pdf documents; usually its *Acrobat Reader* can access documents up to four or five versions back, depending on how those documents have been saved in terms of backward compatibility, but very rarely documents from previous versions can be accessed with this software.[10] An *emulator* is a viable option to access them and, if necessary, to reconvert them, but its economic profitability depends on the number of documents to update, their importance, the backwardness of the versions, and so on. In the vast majority of cases, their use is not economically profitable. However, on certain occasions the rarity, uniqueness or importance of the documents involved makes it imperative to use this technique regardless of its costs, so it should

10 The exception to this rule are documents stored in the pdf-1-A format, the ISO 19005-1:2005 standard (Document management - Electronic document file format for long-term preservation, Part 1: Use of PDF 1.4 (PDF/A-1)) https://www.iso.org/standard/38920.html.

be kept in mind as a possible option and, if necessary, consider budgeting for and acquiring it.

Since the cost of updating is never present at the time of creating a new digital collection, it tends to be forgotten, but it will inexorably appear from time to time in the preservation costs, and will affect the organization's budget with certain periodicity, so it is necessary to keep it in mind to include it in the total cost of digital preservation when appropriate, avoiding surprises for the administrators.

From all of the above it is clear that the total cost of preserving digital documents consists of the sum of all its partial costs that have been broken down here under a certain grouping made personally; all of it projected for a certain context and specific environment of documents and a certain time span.

It is of utmost importance not to forget that many of these stages require qualified and trained personnel to perform them properly, so –if the personnel involved in them do not have those characteristics beforehand– it is essential in all project cost estimates to include the corresponding personnel training, either in the respective phases, or as a global training.

Conclusions

The amount of digital information in the world is growing at unprecedented rates, and much of it needs to be preserved. More and more libraries, archives, repositories, etc., are being designated for this task as more citizens are required to access academic, scientific, public, governmental, etc., information. However, there is still a perception in many organizations that digital preservation is a very low-cost process, and that once a document is placed in a computer, the cost of maintaining it is minimal and therefore insignificant. Such perception must never be allowed to permeate into the management layers of the organization. The first step in digital preservation is to raise awareness at all levels of its importance and institutional value, but also –as with all types of preservation– of the costs involved. This problem is exacerbated in public organizations, as these costs are often not included in their budgets, or are only marginally included. For many of these organizations these are new obligations, and they often did not have the human, technological or economic resources to do so, or at least not initially or in sufficient quantities. This is mainly because the costs of digital preservation are not obvious to everyone. Many people still think that this task boils down to acquiring a sufficient number of storage devices; this is often the case among decision-makers and funders of these projects. The problem is aggravated by the fact that digital preservation costs are tangible, while its benefits are intangible: its costs are relatively easy to establish numerically, but its benefits are not, making it difficult to justify.

Throughout the text an updated view of the main components that make up the cost of digital preservation has been presented, divided into five headings: the cost of digitizing, editing, recording, storing and updating. An attempt has been made to analyze these costs in the light of emerging options.

Although storage costs per unit have continued to decrease in the last decade, the growth in device capacities is no longer as remarkable as it was in the previous two decades, nor is there such a dramatic reduction in their costs. Therefore the money/quantity ratio continues to improve, but no longer at the rates of earlier times; it is becoming increasingly difficult for manufacturers to optimize what already exists with today's technologies. In practice, this means that the amount of information to be preserved is growing faster than the monetary efficiency of the devices to do so.

Cloud storage has emerged as the most notorious new option in recent years for this purpose but, as has been discussed, in the long term –which is the timeframe in which preservation works– the commercialization models of individual companies are not as cost-effective a solution economically; many authors agree on this. In addition, in many organizations, such as archives, cloud storage introduces a number of other issues that require additional careful review and analysis before adoption. For the same reason, initiatives are emerging to build large digital storage centers by academic and governmental entities, etc., seeking greater economies than current commercial storage models, as well as being built as highly trustable centers.

Although digital preservation is affected by various factors, the economic factor is obviously of paramount importance in any project of this nature. It is therefore necessary that the responsible for digital preservation: librarian, archivist, or other, gradually becomes familiar with the various costs related to this function so that he/she can be able to calculate, interpret correctly and present these costs in the organization's plan and budget. It is essential that they be able to describe the cost/benefit of the preservation function so that it can be fully understood by institutional managers and, where appropriate, by potential sponsors. The underlying secret of success is to be able to turn technically viable projects into economically viable projects. This requires a delicate balance that every person responsible for digital documentary preservation must develop over time, study and experience. A significant part of the problem is that numerous projects are presented every

day based on technological factors alone, which –while very valid in that approach– face the risk of running out of funding over time. This in no way means that the main focus of digital preservation projects should be the economic one, neglecting technical factors, as this inexorably leads to very cheap projects of dubious quality that after a while will have to be discarded, and that may also put the material to be preserved at risk in the future. Therefore, a careful balance between satisfactory benefits, reasonable costs and appropriate technology, as well as an attractive and convincing presentation, must be sought from the design and during the development of this type of projects.

To achieve this, as it has been observed, knowledge and study of all the components associated with costs by those responsible for digital preservation, their integral and balanced management following an application methodology, together with a little experience in this regard, maximize the chances of success in this type of projects by keeping costs at reasonable and explicit levels for everyone in the organization and –most important of all– maximizing the long-term permanence of digital documentary collections. I conclude with a quote from Stewart Brand summarizing all of the above, which, even though it is now two decades old, is still valid: "[...] Digital storage is easy; digital preservation is not. Preservation means keeping the stored information cataloged, accessible, and usable on current media, which requires constant effort and expense" (Brand 1999, 47).

ANNEX 1

Some examples of sites with recommendations and standards for digitization, quality, etc.:

For imaging, audio and video:

- Smithsonian Institution. *Digitization Standards for Images, Audio and Video.* (2004) https://siarchives.si.edu/what-we-do/digital-curation/digitizing-collections.
- South Carolina Department of Archives & History. Electronic Records Management Guidelines. Digital Imaging, March 2008, Version 2 https://core.ac.uk/download/pdf/49235519.pdf.
- *Standards Related to Digital Imaging of Pictorial Materials* (2004). K.A. Peterson (Comp.) Library of Congress, Prints and Photographs Division https://www.loc.gov/rr/print/tp/DigitizationStandardsPictorial.pdf.

For newspapers:

- *National Digital Newspaper Program (NDNP), pdf Specification* (2008) https://www.loc.gov/ndnp/guidelines/NDNP_202123TechNotes.pdf.
- Skinner, Katherine; Schultz, Matt (2014). *Guidelines for Digital Newspaper Preservation Readiness*. Atlanta, Ga: Educopia, https://educopia.org/wp-content/uploads/2018/06/Guidelines_for_Digital_Newspaper_Preservation_Readiness_0.pdf.

For digital art:

- Erpanet: The Archiving and Preservation of Born-Digital Art Workshop (2004). *Briefing Paper for the Erpanet Workshop on Preservation of Digital Art* https://www.erpanet.org/events/2004/glasgowart/briefingpaper.pdf.

Bibliographic references

(All electronic references verified as extant and accurate as of March 2, 2021)

ALA – American Library Associations and ALCTS – Association for Library Collections and Technical Services. 2007. http://www.ala.org/alcts/resources/preserv/defdigpres0408.

Bote, Juanjo; Fernández-Feijoo, Belén; Ruiz, Silvia. 2012. The cost of Digital Preservation: A methodological analysis. In: *Procedia Technology*, vol. 5, 103-111. https://doi.org/10.1016/j.protcy.2012.09.012.

Brand, Stewart. 1999. Escaping the digital Dark Age. In: *Library Journal*, vol. 124, num. 2: 46-49. https://rense.com/general38/escap.htm.

Calhoun, Patrick; Akin, David; Zimmerman, Brett; Neeman, Henry. 2019. Large scale research data archiving: Training for an inconvenient technology. *Journal of Computational Science*, vol. 36 https://doi.org/10.1016/j.jocs.2016.07.005.

DLF - Digital Libraries Federation. 2002. *Benchmark for Faithful Digital Reproductions of Monographs and Serials.* The Digital Library Federation Benchmark Working Group, 2001-2002. http://www.diglib.org/standards/bmarkfin.htm.

Gantz, John; Reinsel, David. 2012. *The Digital Universe in 2020.* IDC. (International Data Corporation). Sponsored by EMC Corporation. December 2012. https://www.slideshare.net/arms8586/the-digital-universe-in-2020.

Gantz, John; Reinsel, David. 2007. *The Expanding Digital Universe.* IDC White paper. (International Data Corporation). Sponsored by EMC Corporation. March 2007. https://web.archive.org/web/20090612013506/http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf.

Garrett, John. 1996. *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information.* CLIR Commission on Preservation and Access and RLG. https://clir.wordpress.clir.org/wp-content/uploads/sites/6/pub63watersgarrett.pdf.

Hendley, Tony. 1998. *Comparison of Methods & Costs of Digital Preservation.* British Library Research and Innovation Report 106. http://www.ukoln.ac.uk/services/elib/papers/tavistock/hendley/hendley.html.

ICA - International Council on Archives. 2015. ICA Terminology Database. Entry by: "digital preservation". http://www.ciscra.org/mat/mat/term/145.

InterPARES – The International Research on Permanent Authentic Records in Electronic Systems. 2015. *Checklist for Cloud Service Contracts. Final Report.* https://interparestrust.org/assets/public/dissemination/NA14_20160226_CloudServiceProviderContracts_Checklist_Final.pdf.

InterPARES – The International Research on Permanent Authentic Records in Electronic Systems. 2008. *International Standards Relevant to the InterPARES 3 Project - General Study 04.* Xie, Sherry & Hammerschmidt, Joanna. http://interpares.org/ip3/display_file.cfm?doc=ip3_canada_gs04_international_standards.pdf.

InterPARES – The International Research on Permanent Authentic Records in Electronic Systems. 2005. *The Long-term Preservation of Authentic Electronic Records.* http://www.interpares.org/%5C/book/index.cfm.

ISO/IEC 25012:2008. *Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model.* https://www.iso.org/standard/35736.html.

Kingma, Bruce. 2000. The Costs of Print, Fiche, and Digital Access: The Early Canadiana Online Project. *D-Lib Magazine*, February 2000, vol. 6, num. 2. http://www.dlib.org/dlib/february00/kingma/02kingma.html.

Library of Congress (s.d). *Sustainability of Digital Formats.* https://www.loc.gov/preservation/digital/formats/.

MoReq2 - Model Requirements for the Management of Electronic Records. 2002. 2nd version, European Commission: Brussels, Belgium. Glossary Section. https://cdn.ymaws.com/irms.org.uk/resource/resmgr/moreq/presentations/docubarcelona_moreq2_mvpalma.pdf.

Morris, Robert; Truskowski, B.J. 2003. The evolution of storage systems. *IBM Systems Journal*, vol. 42, num. 2: 205-217. DOI: 10.1147/sj.422.0205.

Pearce-Moses, Richard. 2005. A Glossary of Archival and Records Terminology. Chicago: Society of American Archivists. Entries by: "digital object" and "conservation". https://dictionary.archivists.org/entry/conservation.html.

PREMIS – Data Dictionary for Preservation Metadata, Version 3.0, Nov. 2016. Library of Congress. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

Rosenthal, David; Rosenthal, Daniel; Miller, Ethan; Adams, Ian; Storer, Mark; Zadok, Erez. 2012. The Economics of Long-Term Digital Storage. *The Memory of the World in the Digital Age: Digitization and Preservation*. Santa Cruz, CA. https://web.stanford.edu/group/lockss/resources/2012-09_The_Economics_of_Long-Term_Digital_Storage.pdf.

Scholars Portal. 2013. A Service of the Ontario Council of University Libraries. https://learn.scholarsportal.info/all-guides/handling-digital-archives/concepts/#SIPAIPDIP.

Von Suchodoletz, Dirk; and Rechert, Klaus. 2014. *Emulation as a Service – Framework for Curation and Rendering of Complex Digital Objects.* Curate Gear 2014. https://ils.unc.edu/digccurr/curategear2014-talks/von-sucholodetz-curategear2014.pdf.

Voutssás-M., Juan. 2009. *Preservación del Patrimonio Documental Digital en México*. México: UNAM, Centro Universitario de Investigaciones Bibliotecológicas. 207. http://ru.iibi.unam.mx/jspui/handle/IIBI_UNAM/L49.

Zeller, Jean-Daniel. 2010. Cost of digital archiving: Is there an universal model? *Eight Conference on Digital Archiving*, April 2010, Genève. https://regarddejanus.files.wordpress.com/2010/05/costsdigitalarchiving-_jdz_eca2010.pdf.

There is also a bibliography of sites with recommendations about quality in digitization projects: *Digital Conversion – Documents & Guidelines. A Bibliographic Reference.* 2009. http://www.digitizationguidelines.gov/guidelines/Guidelines_Bibliography-2009rev.pdf.

Factores económicos de la preservación documental digital: actualización 2021/Economic factors of digital preservation: 2021 Review. Instituto de Investigaciones Bibliotecológicas y de la Información/UNAM. La edición consta de 100 ejemplares. Coordinación editorial, Anabel Olivares Chávez; revisión especializada, Valeria Guzmán González y Francisco González y Ortíz; corrección de pruebas, Carlos Ceballos Sosa; revisión de pruebas, Carlos Ceballos Sosa; formación editorial, Sonia Wendy Chávez Nolasco. Fue impreso en papel cultural de 90 g en los talleres de Gráfica Premier, S.A. de C.V., Metepec, Estado de México. Se terminó de imprimir en abril de 2022.