

Almacenamiento distribuido en la digitalización de expedientes y automatización de procesos

Jornada de archivos del Consejo de la Judicatura Federal
CDMX, 20.09.22



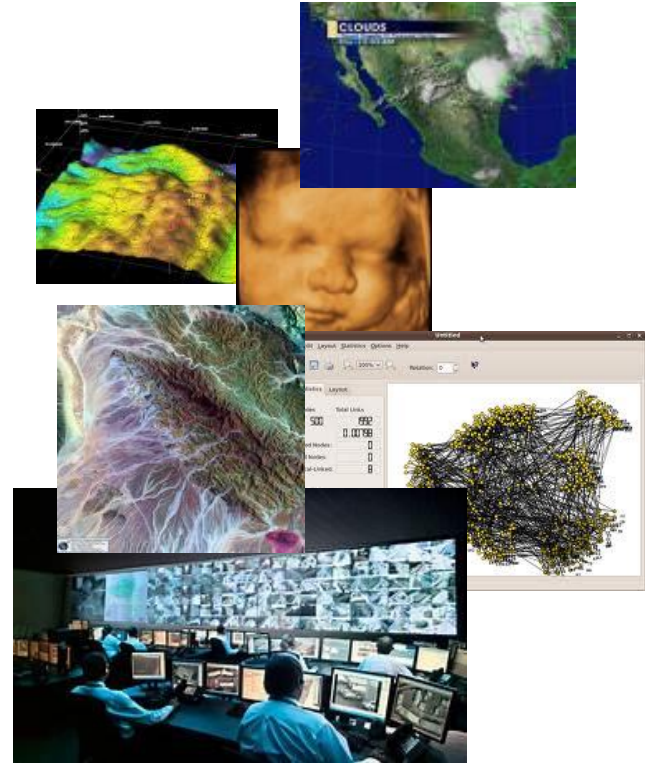
Dr. Ricardo Marcelín Jiménez
Departamento de Ingeniería Eléctrica
UAM-Iztapalapa

CONTENIDO

- 1. El contexto**
- 2. Los procesos de las organizaciones**
- 3. Los flujos de la información**
- 4. Los paradigmas del almacenamiento**
- 5. Los retos del almacenamiento secundario**
- 6. Nuestro trabajo**
- 7. Conclusiones y perspectivas**

1. El contexto (la crisis)

- Un crecimiento nunca antes visto en el volumen de la información generada.
- Nuevas tecnologías, paradigmas, aplicaciones, modos de trabajo
- Un aumento en la necesidad de intercambiar la información.



1. El contexto

- En el 2020 se generaba cada minuto del año, 500 horas de video en YouTube; 347,222 publicaciones en Instagram; 28 canciones en Spotify [1].
- 75% de la información es generada por individuos, el 80% de la información está bajo la responsabilidad de organizaciones. Se prevé un crecimiento exponencial en el tamaño del “universo digital” pasando de 33 Zettabytes (ZB) en 2018, a 175 ZB en 2025 (Kilo, Mega, Giga, Tera, Peta, Exa, Zetta, Yotta) [2].

1. Data never sleeps. <https://web-assets.domo.com/blog/wpcontent/uploads/2019/07/data-never-sleeps-7-896kb.jpg>
2. Reinsel, D., Gantz, J., and Rydning, J. (2018). The digitization of the world from edge to core. IDC White Paper.

2. Los procesos de la organización

El ciclo de vida de la información (p.ej. correspondencia u oficialía, trámite, concentración e histórico): los retos de la gestión: manejo de catálogos con metadatos definidos por la organización, definición de colecciones, soporte de roles y permisos de acceso, flujo de documentos, auditoría, preservación.



2. Los procesos de la organización

- ¿Cuánto tiempo invierte el personal en la recopilación de información interna, necesaria para el cumplimiento de sus funciones? (p.ej. Los compromisos en materia de rendición de cuentas, transparencia, leyes de archivos, o los procesos internos de la organización)
- los empleados ocupan casi el 20 % de su tiempo en la búsqueda de documentos e información (debido a que estos se encuentran dispersos, duplicados y no hay una forma ágil de recuperarlos)

informe de McKinsey & Co.

3. Los flujos de la información

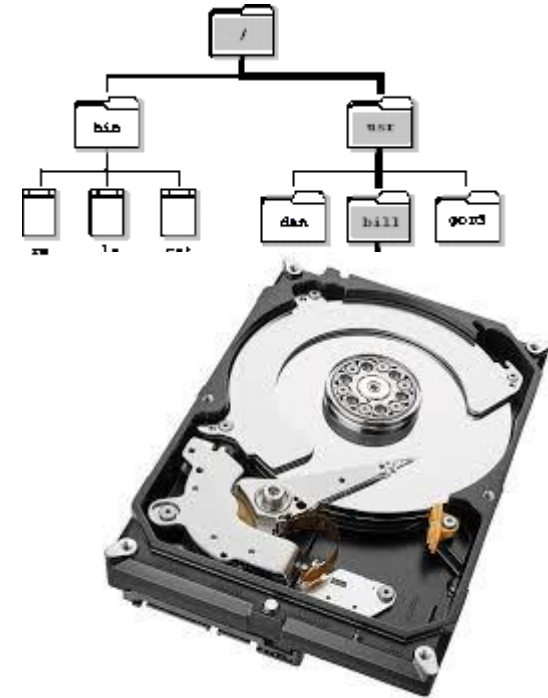
Los ecosistemas de almacenamiento: el **almacenamiento primario** asociado con la aplicación (PACS, DSpace, Alfresco, GRASS), de baja capacidad y baja latencia, en comunicación con un **almacenamiento secundario** (escalable y de mayor latencia).



4. Los paradigmas de almacenamiento

Archivos, Bloques y Objetos

Los dos primeros se conocen como sistemas de almacenamiento ***transaccional*** y se usan para modificar con agilidad la información registrada. En tanto, los sistemas por objetos son aquellos en los que la aplicación final desconoce el dispositivo donde su información es emplazada, porque esta puede migrar varias veces a lo largo de la vida del sistema, por lo tanto ésta no puede modificarse a menos que se lleve a un sistema transaccional.



4. Los paradigmas de almacenamiento

A cambio, los sistemas de almacenamiento por **objetos** pueden crecer con más facilidad incorporando nuevos dispositivos, lo que los hace la mejor opción para alojar la memoria de largo plazo de una organización.

Muchos especialistas recomiendan que el almacenamiento de largo plazo se base en un sistema distribuido, definido por software, que soporte requerimientos de crecimiento y disponibilidad, y que ofrezca una interfaz estándar con la que puedan acomodarse diferentes aplicaciones.

4. Los paradigmas de almacenamiento

Las diferencias se pueden explicar a partir de 2 elementos:

- La colección de documentos (tamaño y ubicación)
- Sus metadatos (elementos de descripción y ubicación)

Archivos: colección relativamente pequeña, metadatos con el mayor nivel de detalle, documentos y metadatos viven en el mismo dispositivo.

Bloques: colección un poco mayor, los metadatos son del mismo nivel de detalle que en el anterior, documentos y metadatos viven en dispositivos diferentes. Los documentos se fragmentan para agilizar el acceso a ellos.

4. Los paradigmas de almacenamiento

Los **archivos** y **bloques** se basan en el concepto de sistema de archivos. La mayoría de los usuarios están familiarizados con ellos, como FAT, NTFS y similares. Organizan los datos en archivos y carpetas en una jerarquía arborescente y le dan una ruta al archivo al mismo tiempo que conservan una pequeña cantidad de metadatos sobre el archivo.



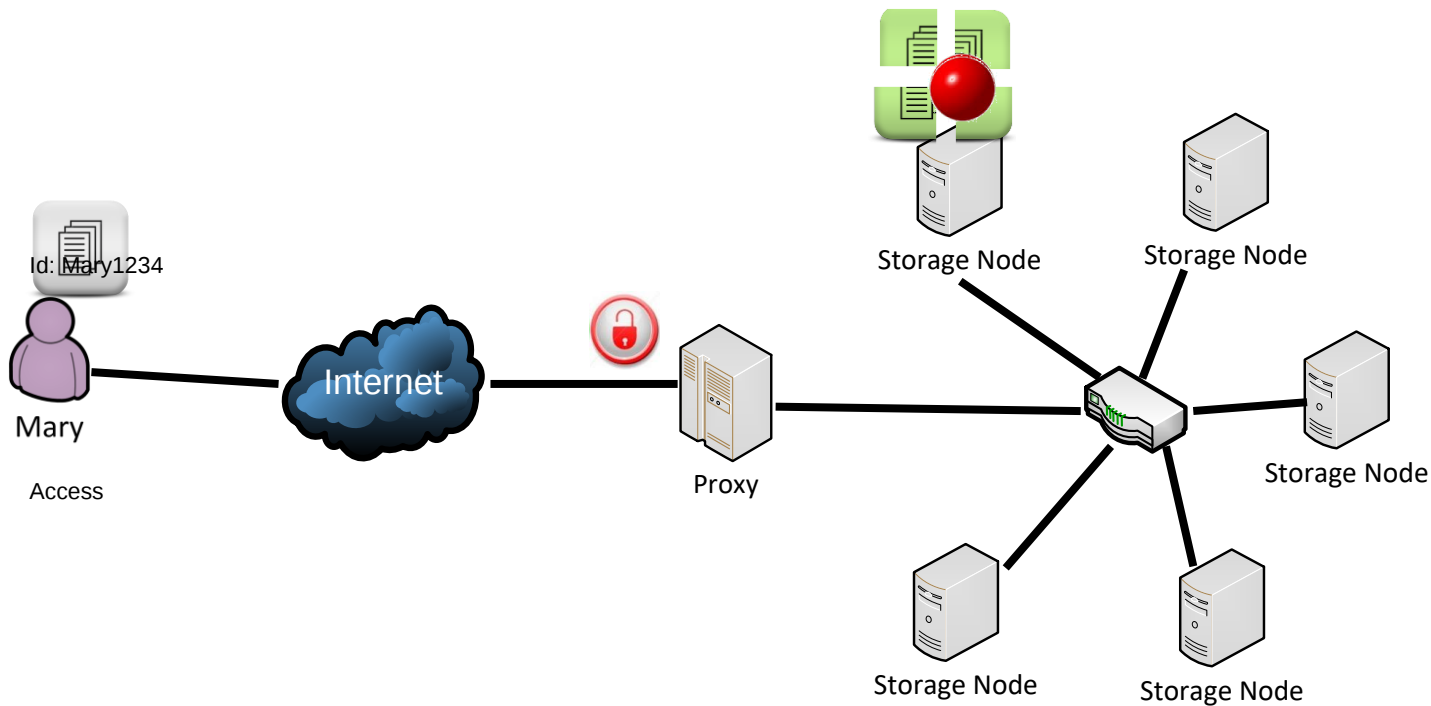
4. Los paradigmas de almacenamiento

Objetos: colección masiva, los metadatos deben contener información mínima, la estructura lógica es plana (y no jerárquica como en los otros casos), documentos y metadatos viven en dispositivos diferentes y los dispositivos de almacenamiento deben soportar escalamiento, disponibilidad, tolerancia a fallas.

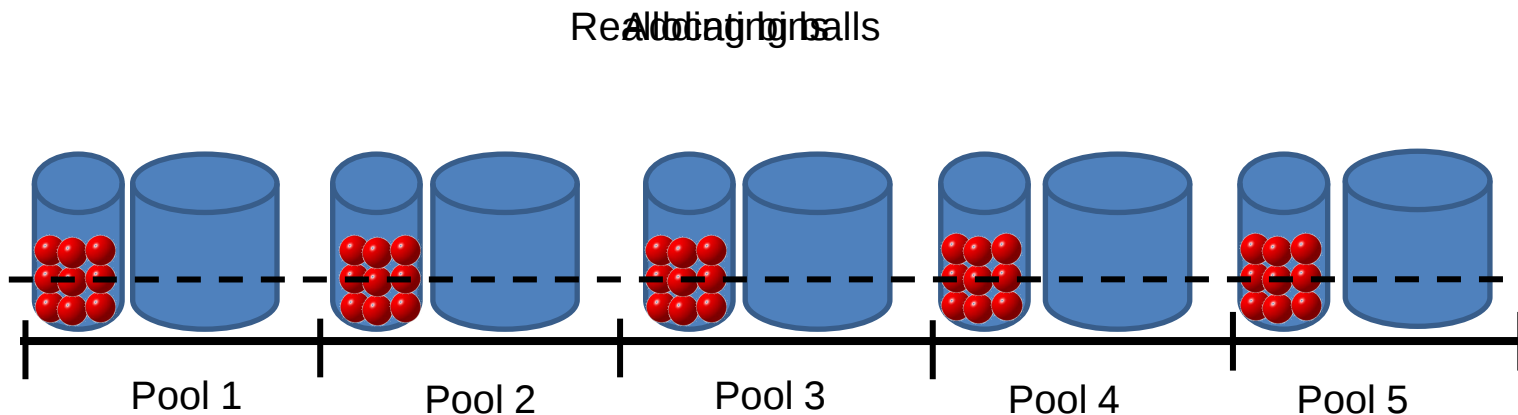
- Los sistemas definidos por software
- Independencia del proveedor
- Distribuidos, escalables, de alta disponibilidad
- Facilitan la gestión autónoma
- Ejemplos: CEPH, Lustre, Gluster, HDFS, Scality



4. Los paradigmas de almacenamiento



4. Los paradigmas de almacenamiento



5. Los retos del almacenamiento secundario

- Un sistema masivo que puede ser el soporte de diferentes formas de trabajo y que mantenga abierta la posibilidad de hacer diferentes tipos de tratamiento sobre la información que se preserva. Basado en el paradigma de objetos, SDS
- Los NFR del almacenamiento masivo:
 - Escalabilidad, para administrar la complejidad
 - Interoperabilidad para soportar la heterogeneidad
 - modularidad para poder evolucionar y reestructurarse con facilidad
 - Confiabilidad para soportar la tolerancia a fallas, integridad y confidencialidad
 - Disponibilidad para recuperar oportunamente



6. Nuestro trabajo

Aleph: un sistema de almacenamiento secundario, basado en el paradigma de objetos.

Una interfaz que lo conecta con aplicaciones de terceros, como DSpace, Moodle, PACS.

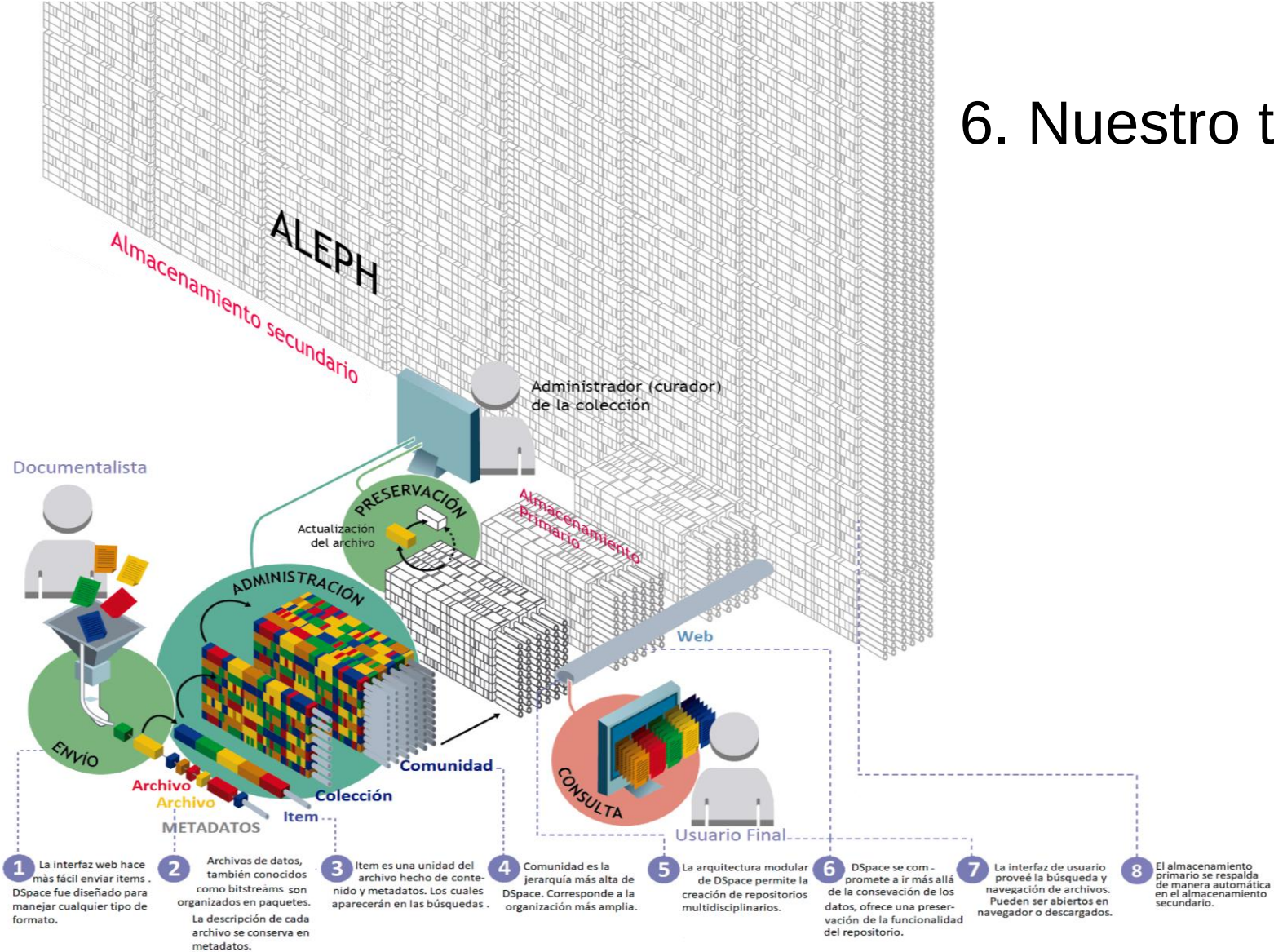
Una aplicación propia para el manejo colaborativo de expedientes.

6. Nuestro trabajo

Fondo UAM es un sistema basado en software libre que

- Maneja documentos a lo largo de su vida
- Agiliza la búsqueda
- Herramienta adaptable a un clasificador de documentos
- Evita repetir información al tener un solo expediente
- Se puede conectar con otros sistemas de su tipo para soportar la “cosecha de datos”
- Aceptar las condiciones de acceso de la organización

6. Nuestro trabajo



6. Nuestro trabajo

Proponemos una conexión entre el RI, con un sistema de almacenamiento secundario, con ello se garantiza el crecimiento de la colección y su preservación.

Cada documento que se recibe es automáticamente respaldado en el almacenamiento secundario de una forma transparente para el usuario

Cuando un usuario realiza una consulta sobre el catálogo y se determina que requiere un documento previamente almacenado, el sistema establece si el archivo se encuentra disponible en el disco primario o si debe recuperarse del secundario y restaurarse. Todo esto de manera transparente para el usuario final.

6. Nuestro trabajo

- Provee un servicio ágil para un número importante de usuarios concurrentes.
- Administra diferentes colecciones y catálogos en el mismo servidor.
- Centraliza su documentación lo que le permite tener un mayor control.

6. Nuestro trabajo

- Se garantiza una preservación documental de largo plazo, manteniendo sus archivos históricos organizados.
- Evita la duplicación de documentos
- El desacoplamiento de los almacenamientos primario y secundario ofrece una alta disponibilidad y escalabilidad de forma sostenible.

6. Nuestro trabajo (el catálogo)

Etapas actuales

- Instalación de Repositorio Institucional (RI)
- Definición de los metadatos de la colección SG
- Puesta en producción del almacenamiento secundario y conexión con el RI
- Demostramos que es posible mover ítems entre colecciones y modificar sus metadatos, para soportar los requerimientos de la ley general de archivos

Etapas siguientes

- RI para cada unidad académica, interconectados entre si
- Módulo de admón. de metadatos, listas de RI, acuses
- Desarrollar un API para personalizar el uso del RI en cada unidad académica
- Extender las capacidades y funciones del almacenamiento secundario
- Desarrollar un cliente de escritorio para respaldo personal

6. Nuestro trabajo

- Documentación catalogada
- Flujo de trabajo entre los usuarios para mantener roles como publicador, editor y autorizador
- Capacidades de cosecha de metadatos para funcionar como un repositorio central interconectado con otros repositorios
- Búsquedas incluso en el contenido de los documentos con extensión PDF, doc.
- Aplicación complementaria del Fondo UAM
- Administración de listas en plantilla de metadatos
- Transferencias de documentos entre colecciones y modificación de metadatos para soportar los requerimientos de la ley general de archivos.
- Historial de transferencias de documentos para auditoria de procesos
- Conexión del Fondo UAM con el sistema de almacenamiento

6. Nuestro trabajo (el Aleph)

Etapas actuales

- Almacenamiento y recuperación de información a través de distintos tipos de aplicaciones
- Administración de cuotas de almacenamiento por departamento o área
- Gestión de roles y usuarios por departamento o área
- Aplicación complementaria de escritorio para almacenar, consultar y recuperar información

Etapas siguientes

- Uso del sistema de almacenamiento para compartir archivos y carpetas
- Realización de consenso para operaciones sensibles en el sistema (desactivar usuarios, enviar archivos a papelera de reciclaje,...)
- Interconexión de sistema para salvaguardar la información ante una situación de desastre

7. Conclusiones y perspectivas

- La información digital es el activo intangible más importante de las organizaciones
- Las organizaciones deben desarrollar e implantar soluciones para garantizar la preservación de su información sobre un horizonte de largo plazo, acoplando los procesos de la organización con requerimientos de seguridad, disponibilidad y escalabilidad.



Gracias

rmarcelin@izt.uam.mx