Edición de archivos personales: el problema de la conservación del patrimonio textual ante la desmaterialización digital

Daniel Escandell Montiel

TEPPEL
Tecnología y poder en el pensamiento y las letras

HORIZON-MSCA-2021-SE-01-STAFF

Archivos, repositorios y discos duros

HORIZON-MSCA-2021-SE-01-STAFF

Las nuevas destrezas (que todavía no cubren desde las HHDD)

- La discrepancia entre las destrezas informáticas necesarias, las que se pueden aprender formalmente y las que se transmiten en titulaciones especializadas es notable
- Ausencia más notable: informática "forense"
  - Se enseña estilometría, lingüística forense (integrada típicamente en la computacional)... pero no informática forense
  - ¿Cómo se espera acceder, preservar y recuperar datos informáticos? ¿Quién va a *abrir los cajones* de los escritores de los autores cuando llegue el momento de desacralizar su creación?
  - ¿Saben los investigadores filológicos actuales recuperar la información de un ordenador? ¿Son incluso conscientes de que pueden recuperar archivos "perdidos", dañados o modificados como parte de una nueva crítica genética electrónica?

La complejidad del archivo contemporáneo

- ¿Cómo será el depósito del legado escritural de un Premio Cervantes en veinte años? ¿Cuánto papel habrá realmente (que no forme parte de una impostura)?
- La desmaterialización del archivo personal complica, más que simplifica, su preservación futura
  - Obsolescencia (de sistemas informáticos y del *hardware*)
  - Descentralización del archivo digital: discos en ordenadores, memorias en móviles o tabletas, servicios en la nube...
  - Desechabilidad del soporte: por lo general, no se conservan ordenadores ni dispositivos antiguos; simplemente, son sustituidos (por la confianza creciente en la nube)
  - Los contenidos digitales son más volátiles en realidad que el papel: corruptibles, fácilmente olvidables y susceptibles de ser considerados prescindibles (borradores, versiones preliminares, o documentos en progreso)

La multidimensión del archivo más personal

- Frente a la cultura epistolar impresa, la cultura de la mensajería digital es volátil y cuenta con escasa asociación afectiva
- La obsolescencia (y la falta de espacio en la memoria) hace que no sea raro borrar (a veces, impulsivamente) mensajes, correos o archivos adjuntados a través de sistemas de e-mail, mensajería, chat, etc.
- Muchas aplicaciones son deliberadamente volátiles (autoeliminan mensajes)
- El atomismo de las comunicaciones y el salto de un soporte a otro (conversaciones que pueden trasladarse de WhatsApp al e-mail, de ahí a Zoom, y vuelta a WhatsApp) complican la reconstrucción, preservación y estudio de las comunicaciones personales
    - Deben abordarse mediante estrategias más próximas al *big data* y sistemas claramente cualitativos que a la exploración cuantitativa o de inmersión en archivo clásica

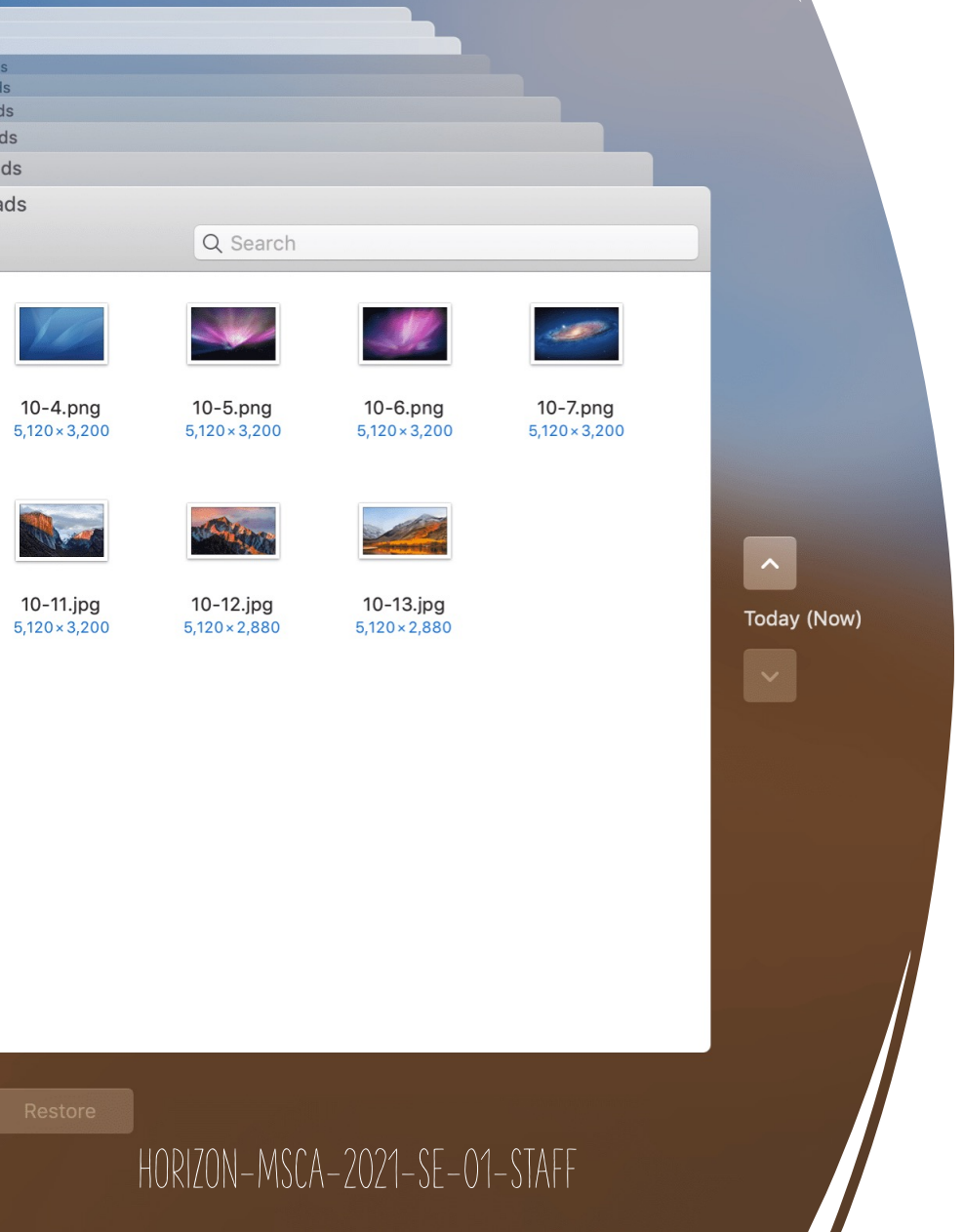

El caos del archivo inmaterial
HORIZON-MSCA-2021-SE-01-STAFF

El síndrome de Diógenes digital

- Ante los ojos del usuario, la información-basura digital no ocupa espacio ni necesariamente molesta
    - El usuario promedio difícilmente llega a los límites de almacenamiento de sistemas de correo, archivos en la nube, etc.
    - Aunque se detecta tendencia a la dificultad para borrar archivos, sean significativos (emocional o informacionalmente) o no
    - Pero cuando el usuario borra, corre el riesgo de borrar indiscriminadamente
    - Afecta también a usuarios concretos, que almacenan e-waste

Retos inherentes al sistema

- La inmaterialidad desistematiza los métodos de autoarchivo y clasificación
- Puede haber un entramado de contraseñas, identificadores de usuarios y otras *passkeys* por desentrañar
- El usuario emplea sistemas diversos, no necesariamente coherentes entre sí, de preservación de archivos, versiones previas, *work-in-progress*, revisiones, etc.
- Los propios sistemas operativos operan como pantalla que incluye mecanismos propios a través de sus parámetros de funcionamiento y tratamiento de archivos que, a su vez, pueden ser personalizados por usuarios mínimamente avanzados
- Es difícil tener una representación visual del archivo cuando este es inmaterial, su reproductibilidad es infinita, pero muchas veces parcial, y diseminada en dispositivos locales diversos y múltiples servicios en línea
- Las tipologías textuales (las funciones del *software*) imponen clasificaciones propias
- Partes del archivo se pierden para siempre por acciones ajenas al usuario (v.g. cierre de servidores)



Sistemas de preservación cronológica de archivos

- Ha habido pocas iniciativas, tanto en vertiente local como en servicios en línea, orientadas a usuarios finales
- Son sistemas orientados a recuperar versiones de archivos, directorios, etc. en cada versión generada y de forma cronológica, pudiendo recuperar versiones antiguas de archivos, compararlos y contrastar sus contenidos
- Suelen estar automatizados y borran por su cuenta las versiones más antiguas
- Forman parte de categorías de suscripción *premium* en servicios en línea

Orden y análisis en el archivo digital

HORIZON-MSCA-2021-SE-01-STAFF

Informática forense aplicada al estudio y preservación de datos

- Análisis de sistemas de archivo de datos (discos duros, discos, memorias, cintas...) físicos y métodos de restauración: la recuperación física de soportes
- Análisis de espacio de datos: el análisis de los espacios teóricamente vacíos de las memorias (esto puede incluir no solo datos persistentes, sino también datos volátiles; esto último de limitado interés para nosotros)
- Trazabilidad de redes, cookies, archivos temporales y otros archivos de librerías de sistema

Análisis masivo de datos: herramientas de referencia

- El procesamiento del lenguaje natural (PLN) con Python análisis de datos estadístico, con opción de apoyo en IA
- Puede ser conveniente procesar los archivos de sonido para convertirlos en transcripciones
- Tras la fase de extracción de contenidos pueden utilizarse técnicas comunes de tratamiento de *big data* textual, como:
  - Análisis de sentimientos
  - Análisis de temas
  - Análisis de opiniones
  - Extracción de entidades (individuos, organizaciones, lugares, etc.)
- Python: NLTK (Natural Language Toolkit), spaCY (PLN), scikit-learn (*machine learning*) y matplotlib (visualización) son herramientas comunes

Conclusiones

- No basta con echar un vistazo a la carpeta "Documentos" y leer algunas decenas de miles de emails
- Debe procederse con una labor de recuperación de archivos, revisión de soportes, y generación de datos tratables
- La preservación y determinación de lo relevante se sustenta más en herramientas que provienen de la informática forense y el análisis de *big data* sociológico/mercadotécnico que de la tradición de la lingüística computacional
- Herramientas de PLN e IA pueden ayudar a localizar y taxonomizar datos (por ejemplo, identificar rostros en un catálogo de miles de archivos de imagen, y extrapolar información mediante datos EXIF/GPS) que luego pueden ser tratados
- Los análisis clásicos de HHDD son una fase relativamente tardía del estudio del archivo inmaterial

HORIZON-MSCA-2021-SE-01-STAFF