JOSEP LLUIS DE LA ROSA I ESTEVA, JOSE ANTONIO OLVERA CAÑIZARES

La preservación digital como asunto social: motivación al archivo personal

Preservar los recuerdos digitales de sí mismo o de la familia es una tarea sutilmente complicada: preservar un objeto digital no es lo mismo que preservar, por ejemplo, un libro o una fotografía, ni mucho menos como preservar un certificado, un diploma o un documento firmado en papel. Así como puedes poner un libro en un estante, una foto en una caja, un diploma enmarcarlo en una pared o una escritura de propiedad en un archivo y (si se mantienen secos y seguros) mirarlo 50 años después, en cambio lo mismo no ocurre con un objeto digital porque, en muchos casos, los materiales digitales son considerados más frágiles que los materiales físicos.

La increíble cantidad de material digital que hoy en día guardan las computadoras (y agravado por los dispositivos móviles) por su gran capacidad de almacenamiento conllevará una tarea de preservación difícil y costosa. Guardar un solo archivo digital para los próximos 100 años va a costar mucho trabajo y mucha suerte si no se toman precauciones. El problema ha crecido aún más con las fotos digitales: es mucho más sencillo tomar una foto en papel y guardarla en una caja de zapatos que almacenar fotos digitales durante muchos años, pero sin embargo la

Josep Lluis de la Rosa i Esteva (email: peplluis@eia.udg.edu)
Jose Antonio Olvera Cañizares (email: joseantonio.olvera@udg.edu)
TECNIO - Centre EASY, Agents Research Lab, VICOROB Institute, Universitat de Girona

Recibido: 06-10-2014. Aceptado: 17-10-2014
Citación: Rosa i Esteva, Josep Lluis de la; Olvera Cañizares, Jose Antonio (2014). "La preservación digital como asunto social: motivación al archivo personal". *Tábula,* n. 17, pp. 135-154

mitad de todas las fotos en 2008 ya se tomaban con cámaras digitales, y de esas la mayoría nunca sale del disco duro del ordenador. Tan difícil y complejo es el desafío de la preservación digital en general que, por su parte, los propietarios de ordenadores tienen sus propios problemas. Los altillos, armarios o trasteros están llenos de archivos que no se pueden recuperar, pilas de archivos guardados en disquetes zip, disquetes de 3½ y hasta en floppy de 5½ de los años 80. A falta de una solución clara, los expertos recomiendan a la gente que, en la medida de lo posible, copie y pase su material a CD y otros formatos actuales de resguardo. Pero no es suficiente porque hay falta de la necesaria automatización que reduzca los errores u olvidos de preservación si lo realizamos las personas mismas.

Figura 1: Foto de un escaparate de una tienda de fotos en 2012 avisando del peligro de tener las fotos solo en formato digital

Las cifras del volumen de información que nace en formato digital y no tiene contraparte en papel u otros soportes son asombrosas, se estima que 2,5 millones de personas en el mundo tienen cámaras digitales. Se toman quizá 375.000 millones de fotos cada año. Y nos encanta compartir esas fotos, cientos de millones de fotos se suben a la Web cada día. Además de compartirlas, muchas personas pueden querer preservar sus fotografías para futuros usos comerciales, artísticos o para su propia familia. Sabemos que es conveniente mantener múltiples copias de las fotografías que queremos preservar, incluyendo al menos una copia en un lugar fuera de nuestras casas. Así que la pregunta surge naturalmente: ¿el haber subido esas fotos a Internet puede ser considerado como parte de nuestra estrategia de preservación digital? De otro modo, ¿compartirlas es una buena estrategia de preservación?

Redes sociales y preservación digital

Profundicemos en el aspecto social de la preservación digital. Todavía no hace mucho tiempo que la memoria colectiva se construía a partir de unas pocas miradas particulares y que las instituciones públicas ejercían la salvaguardia, a través de los archivos o las bibliotecas. Pero cuando todos y cada uno de nosotros nos convertimos en cronistas del tiempo que nos toca vivir (reflejado en nuestra fotos, comentarios, escritos, etc.) desbordamos, por el exceso de información, la capacidad de los que hasta ahora habían gestionado la memoria histórica. Todo apunta a que la única solución posible es que nosotros mismos nos responsabilicemos del trabajo de salvaguarda pero, ¿cómo lo deberíamos hacer? Hasta ahora, la solución estaba en copiar las imágenes que se conservaban en discos duros locales o en servidores externos, eso sí, en el formato en que habían sido tomadas. Sin embargo, los expertos advierten que los formatos de los archivos en los que almacenamos las imágenes no sobreviven más de quince años porque el avance tecnológico los hace obsoletos, lo que significa que llegará un día en que no podremos mirar las imágenes tomadas tiempo atrás porque los lectores digitales del futuro no sabrán leer los archivos. Estamos ante un problema de alcance considerable.

El enfoque social será de máximo interés para las empresas porque mientras el 70% o más del contenido digital es creado, capturado, o replicado por los individuos, las empresas y las instituciones en algún momento tienen responsabilidad u obligación del 85% de dicho volumen 0. El desafío de preservar el patrimonio digital es real y está creciendo a un ritmo exponencial. Un estudio reciente realizado por la International Data Corporation (IDC) reveló que va creciendo en un factor de 9 en solo cinco años (ver

Figura 2) 0. Además, ya sucede que el 75% anual del crecimiento de la información digital en realidad es pura replicación, siendo solo el 25% de información nueva. Así pues, si la mayor parte de la información la generan los individuos es razonable asignarles responsabilidades de preservación digital, distribuir dicha responsabilidad a la población, socializar, en definitiva, dicha misión.

Así pues, la preservación del patrimonio personal es un asunto social, que cada vez está más presente en las empresas, los organismos del sector público, así como los investigadores, historiadores y que debiera, ahora más que nunca, en los mismos ciudadanos.

La historia está documentada en libros, fotografías, películas, mapas, grabaciones sonoras, crónicas... Y, como se ha dicho y reiteramos, cada vez más estos materiales se producen digitalmente sin contraparte en papel o film u otro soporte. Los blogs, fotografías digitales, páginas webs y correos electrónicos creados hoy, si se preservan, proveerán testamento de nuestra historia mañana; este es nuestro patrimonio personal digital.

La preservación de dicha información se convertirá en un problema generalizado y omnipresente que preocupará a todos los que tengan información digital a preservar a largo plazo, superando al menos tres generaciones de software, es decir, tres nuevas versiones de los programas, por lo que los ficheros más antiguos serán irrecuperables y la versión de la aplicación que podía leerlos, ya no estará disponible o no será legible por los nuevos equipos (Rivera Donoso, Miguel Angel 2009).

Hasta la fecha, solo las instituciones públicas con los conocimientos técnicos y herramientas especializadas han sido capaces de hacer frente a este problema. Sin embargo, la preservación digital no puede ser abordada por una sola institución o nación debido al exceso de información. La solución es que nosotros mismos, los usuarios, nos responsabilicemos de preservarla, por lo tanto se trata de un deber social.

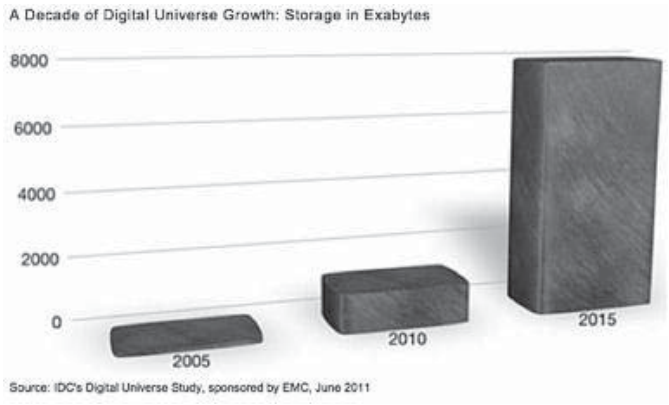

Figura 2: Crecimiento de la información digital creada, en Exabytes, extraído de Gantz, J. and Reisel, D. 2011

De hecho, hay indicios que nos indican que la preservación digital ya es incipientemente un hecho social y una actividad de colaboración, donde el 67% de los usuarios expertos buscan soluciones aportadas por otras instituciones, el 90% de estos usuarios consultan a colegas de confianza y el 83% consulta a los usuarios no expertos[1]. Como consecuencia de la naturaleza fragmentada del hoy todavía escaso conocimiento en preservación digital, se producen frecuentes intercambios de conocimiento en forma de soluciones prácticas de curación y preservación digital: el 83% de los usuarios expertos comparten sus soluciones con colegas y usuarios individuales, el 77% de ellos busca soluciones a través de la Web, el 60% visitan sitios webs especializados en preservación digital y el 20% contribuyen a dichos sitios webs.

Tecnología para el soporte de enfoques sociales para la solución de los problemas en preservación digital

En nuestro centro de investigación TECNIO EASY <www.easyinnova.com> trabajamos en distintos enfoques para dar solución a los problemas de preservación digital, con un enfoque social, y utilizando la tecnología de agentes inteligentes que son un tipo de software de inteligencia artificial que emula el comportamiento social de las personas y que tienen además las siguientes características: son autónomos, persistentes (se mantienen constantes hasta alcanzar un objetivo), como se ha dicho son sociales (se comunican con otros agentes), proactivos (asumen el pleno control de su conducta) y con comportamiento emergente (comportamiento inteligente, como consecuencia de una serie de acciones simples).

Nuestros principales objetivos en la investigación en preservación digital aplicando los agentes inteligentes como auténticas máquinas sociales son:

- Hacer bastante fácil la preservación digital para particulares y empresas pymes, y distribuir los esfuerzos de preservación antes de que los objetos y colecciones digitales lleguen a las instituciones públicas para ser preservadas.
- Reducir el coste y aumentar la capacidad tecnológica para preservar la exponencialmente creciente información digital a largo plazo.

Los agentes inteligentes se programan para compartir conocimiento, espacio, recursos, presupuesto y riesgo de preservación digital, la cual, de hecho, es toda una filosofía de preservación: COMPARTIR ES PRESERVAR, la cual precisamente es la filosofía que seguimos para llevar a cabo la preservación digital (De la Rosa, 2010) al igual que lo hacen otros proyectos existentes como LOCKSS (*Lots Of Copies Keep Stuff Safe*) 0 o MUSE (*Memories USing Email*) 0, entre otros proyectos interesantes. Y es que, dado que la preservación no tiene una solución única, necesitamos realizar las acciones de preservación de forma lo más participativa (social) pero automáticamente posibles para dar con ella.

Desde un enfoque de compartición entre expertos en preservación digital, que trabajan en los archivos y bibliotecas nacionales, regionales y locales, LOCKSS es un programa de código abierto, que permite a los bibliotecarios de cada institución preservar el acceso a la dirección de contenidos a los que se suscriban. Utilizando sus ordenadores y conexiones de red, pueden obtener, conservar y proporcionar acceso a copias adquiridas de contenido electrónico. Esto es análogo a como hacen las bibliotecas usando sus propios edificios, estantes, y el personal para obtener, conservar y proporcionar acceso al contenido en papel. El principio en común es que apuestan por una preservación descentralizada y distribuida, y desde una perspectiva socializada.

Por otro lado, está la aproximación de la preservación del patrimonio personal. Entre los proyectos presentados en el Workshop Personal Digital Archiving

en Maryland, febrero 2013 (PDA 2013) encontramos a MUSE, que propone indicadores para detectar las tendencias y mensajes interesantes, dignos de ser preservados en un gran archivo de correo electrónico. Los posibles beneficios de herramientas como MUSE van desde utilidades tales como resumir el trabajo o copias de seguridad de los archivos adjuntos, a la reminiscencia y recuerdo de acontecimientos familiares y años escolares de graduado, los cuales ayudan a reforzar la confianza, la renovación de las relaciones y a jugar a juegos de memoria. Estos juegan un papel clave en la motivación para el archivo personal como un primer paso para la preservación digital distribuida a gran escala.

Finalmente, hay aproximaciones tecnológicas en forma de red social donde los usuarios comparten su espacio en forma de "community cloud", como son PYRAMID[2] o el proyecto DURAFILE[3] 0, del 7FP de la Unión Europea.

Con PYRAMID, una vez se seleccionan los archivos que se quieren preservar, tiene tres modos diferentes de funcionamiento en la distribución de copias de estos: se pueden hacer en local, en el cloud (en remoto) o con los amigos que el usuario haya declarado desde el propio programa (ver Figura 3). Sobre dicha red social se hace una búsqueda de los formatos más actuales para aplicar los cambios de formatos pertinentes para que estos archivos no queden obsoletos y puedan ser siempre legibles, por comparación de las soluciones de preservación digital aplicadas en los archivos digitales de los amigos. Además, los amigos comparten su espacio para preservar tus objetos digitales así como tú les ofreces en justa correspondencia lo mismo, siendo dicha compartición regulada por PYRAMID.

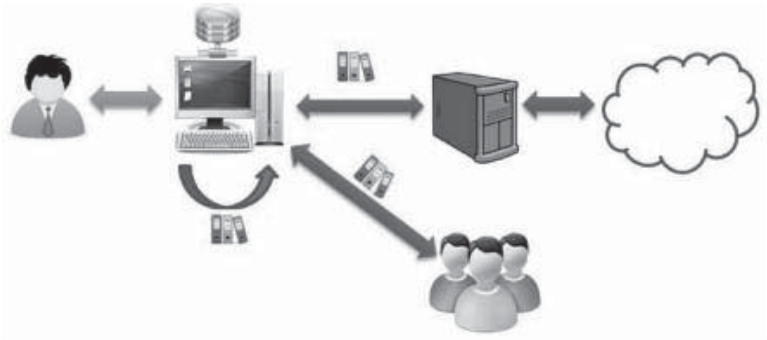

Figura 3: Esquema de funcionamiento de PYRAMID

Otra plataforma es DURAFIL, hecha en agentes inteligentes que intercambian el conocimiento de preservación digital entre usuarios expertos desarrollando búsqueda social en el protocolo ASKNEXT 0 con el objetivo de encontrar planes de preservación adecuados para los objetos digitales multimedia

obsoletos. Esto representa una solución completa para la preservación digital multimedia, que es extensible y flexible, siempre actualizado para que la información que contiene esté viva y sea accesible en tiempo real.

A continuación se describe el enfoque con el que trabajamos:

La preservación con amigos mediante objetos digitales inteligentes

Insistimos en que el paradigma prevaleciente de preservación digital basado en grandes instituciones que se encargan de la preservación y curación de documentos (objetos) digitales parece no ser lo suficientemente escalable ante el crecimiento exponencial antes mencionado. Esto es debido, a nuestro parecer, a que es un paradigma *top-down* donde los recursos e iniciativa de preservación se deciden desde las instituciones y se materializa en los objetos y colecciones digitales que proporcionan los usuarios. Con la socialización cambiamos de paradigma y lo hacemos *bottom-up* donde los mismos usuarios se encargan de dichas tareas y hasta los mismos objetos digitales son los que deben encargarse de su propia preservación y/o curación digital (ver Figura 4). Pasamos de un paradigma *institution-centric* hacia uno *object-centric* potenciado por la socialización de la preservación. Para hacer realidad este nuevo paradigma se debe realizar investigación aplicada en cuestiones asociadas a la autogestión de los objetos digitales mismos, a la forma que los servicios de preservación digital deben ser creados y gestionados, y en el entorno de apoyo (social) a la preservación donde implicaremos a los usuarios finales (personal archiving) de forma particularmente importante respecto al paradigma anterior donde solo grandes instituciones se encargaban de la preservación.

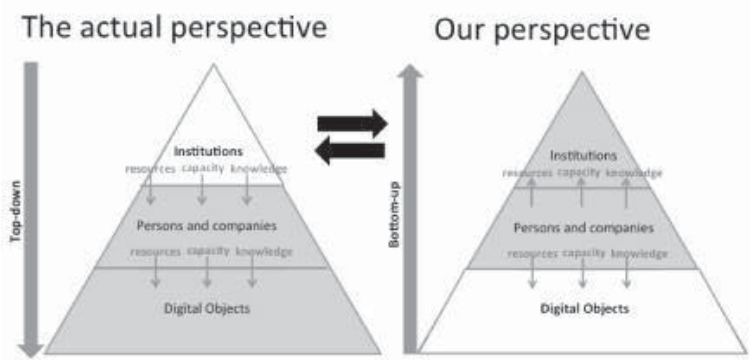

Figura 4: Comparación de la perspectiva actual con nuestra perspectiva

Como se ha indicado previamente, una manera de preservar es aprovechando los amigos que cada uno tenga en las redes sociales. La idea subyacente

es aprovechar sus recursos y compartir los tuyos para realizar la preservación. El sistema funciona de la siguiente manera: Se van realizando copias de los ficheros que un usuario quiere preservar en varios formatos automáticamente, para que estos formatos a medida que pasa el tiempo se actualicen para que se descarten los viejos y se incorporen los nuevos, y las copias que se realicen se distribuyan en diversos espacios, sean propios o ajenos, es decir, tanto en el disco duro del ordenador que los produce como en el de los amigos o en su espacio cloud que se tienen por ejemplo en Facebook o en Twitter. Entonces cada cierto tiempo, cuando se produzca una "catástrofe" (viene dada por cualquier tipo de problema de software, como virus, eliminación por error, cambios de formato o corrupción de archivos) lo que se intentará es recuperar los archivos afectados con el formato más adecuado gracias a este trabajo previo de realizar copias distribuidas en varios espacios.

Esta idea la hemos implementado a través de una plataforma robusta de simulación (Time Machine, TiM) en la que dotamos a los objetos digitales de inteligencia, con comportamientos de autopreservación basado en varias técnicas de inteligencia computacional (inteligencia artificial centrada en el estudio de mecanismos adaptativos para permitir el comportamiento inteligente de sistemas complejos y cambiantes), con el objetivo de que no se pierda ningún contenido digital 00.

El proyecto aborda también las dificultades vinculadas a la conservación a largo plazo de información digital, relacionadas con la recuperación de datos en el proceso de preservación digital, mediante el desarrollo de un nuevo concepto orientado a objetos y colecciones y que está compuesto por servicios de recuperación y de intercambio de información. En este concepto, los objetos digitales se convierten en actores activos en su propia preservación digital a largo plazo (LTDP) mediante una nueva definición del objeto digital a preservar, el *Cost Aware Digital Object* (CADO), que dispone de un presupuesto asignado para competir con otros CADO y negociar con servicios de preservación y curación digitales en la búsqueda de su propia preservación digital a largo plazo. La optimización de la gestión del presupuesto de los CADO ha sido recientemente estudiado en (Olvera J. A., Carrillo P. y De la Rosa J. Ll 2014).

La innovación para satisfacer las necesidades de preservación digital en este nuevo paradigma está integrada en un entorno con características de microofertas. Representa un cambio de paradigma respecto a la actual práctica de preservación digital donde las instituciones y los usuarios son los actores principales y gestionan la preservación digital de forma macro, de forma planificada aunque ineficiente que resulta en la postergación habitual de la preservación de contenido digital hasta que urge preservarlo, y en este momento tardío se producen las mayores pérdidas por obsolescencia porque es demasiado tarde. Así pues, la propuesta disruptiva de este enfoque es que el modelo de costes de la preservación

digital a nivel macroeconómico (gestionado por grandes organizaciones) se cambiará a un modelo a nivel microeconómico (gestionado por los objetos digitales mismos) mucho más escalable y eficaz, que requiere además de la utilización de monedas virtuales específicamente diseñadas para la preservación digital, y que denominamos PRESERVAS, de manera que 1 PRESERVA es igual al coste de preservación en euros constantes para preservar digitalmente un objeto digital durante cien años. La asignación de presupuesto a los objetos se realizan en mili-PRESERVAS 0.

Demostración del valor de la compartición

A continuación se muestra un set de experimentos, realizados con la TiM, para intentar mostrar el valor de la compartición. El comportamiento de los objetos digitales se basa en la inteligencia de enjambre, la topología de red utilizada es *Small Wold* 0, de 100 usuarios, y el espacio que deciden compartir estos es el 25%, 50%, 75% y 100%. Todos los detalles de los experimentos realizados se describen en el apéndice. Los resultados obtenidos se muestran a continuación.

La Figura 5 muestra los resultados para la topología *Small World* con una adopción de software del 33% de CADO (consistente en migraciones masivas de formato motivadas por el cambio de los soportes de los proveedores de software que esperan que los objetos en formatos antiguos se actualicen a los nuevos formatos). Se muestra la evolución de la entropía media y se ha ejecutado un gran número de veces para tener estabilidad estadística. Utilizamos la entropía de Shannon para la evaluación de los resultados: cuanto mayor sea el valor de la entropía, mejor es la capacidad de preservación, de acuerdo con la estrategia de migración de formato (para más detalles ver apéndice). Dicha medida de entropía tiene capacidad predictiva de la durabilidad *obsolescence-free* de los objetos digitales.

Si observamos la Figura 5, que muestra la evolución del valor de la entropía para cada uno de los experimentos ejecutados (espacio compartido del 25%, 50%, 75% y 100%), a simple vista se ve que cuanto más espacio comparten los usuarios de la red, mayor es el valor de la entropía, es decir, presuntamente mejor es la capacidad de preservación del sistema.

Respecto a la capacidad de recuperación o resiliencia, como se muestra en la Tabla 1 (recuperación relativa después del efecto de obsolescencia producido por cada adopción de nuevo software), a medida que pasa el tiempo se van sucediendo las nuevas adopciones de software y medimos el porcentaje de recuperación respecto a cada nueva adopción de software, que vemos que en casi todos los casos va aumentando.

En el caso de compartición del 25% hay casi plena capacidad de recuperación. En los otros casos de 50% 75% y 100%, la entropía disminuye después de

cada nueva adopción de software, recordemos que debido al efecto de obsolescencia provocado por el nuevo software, y como buen indicio, el sistema muestra fuertes signos de recuperación de la entropía. La manera de medir la capacidad de recuperación es el porcentaje que obtenemos de la entropía al final de la emulación con respecto a la entropía en el primer pico (antes de la primera adopción de software): cuando el espacio compartido es del 25% hemos obtenido una tasa de recuperación del 99,72%; para una compartición del 50% tenemos un porcentaje del 92,48%, para la compartición del 75% un porcentaje del 91,75% y para la compartición total (100%) un porcentaje del 89.50%. En todos los casos la recuperación es vigorosa respecto a dejar que los contenidos entren en obsolescencia.

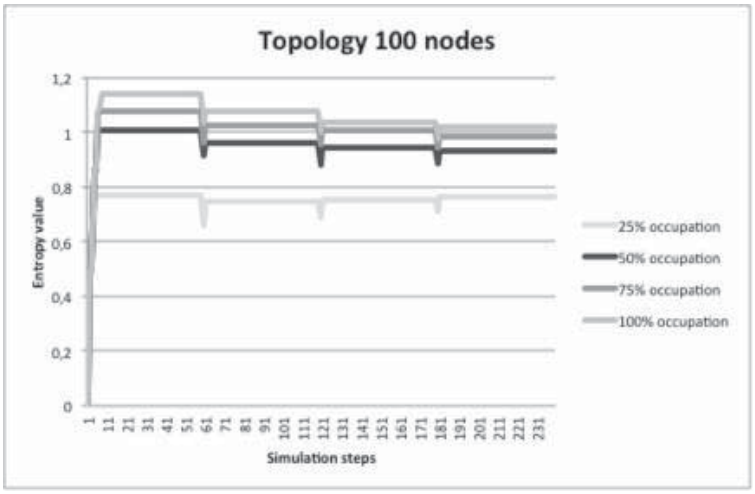

Figura 5: Evolución de la entropía a lo largo de las simulaciones ejecutadas

Espacio local que donan usuarios de la red	Adopción de software 1	Adopción de software 1	Adopción de software 1
25%	78,91%	111,90%	132,25%
50%	51,73%	74,47%	84,34%
75%	53,75%	75,87%	68,86%
100%	41,60%	52,51%	73,78%

Tabla 1: Porcentaje relativo de recuperación después de las adopciones de software

Conclusiones

Después de la experimentación realizada podemos concluir lo siguiente: cuanto más espacio comparten los usuarios de la red, mejor es la capacidad de preservación del sistema. Esta demostración la hemos realizado con un sistema de preservación *object-centric* que es la más adecuada para medir dicho efecto benéfico de la socialización de la preservación social. Pretendemos hacer una extensión de dichos resultados para demostrar los beneficios de la preservación digital como asunto social que sirvan para motivación al archivo personal, la compartición como método de preservación y la adopción de nuevos paradigmas object-centric más escalables y eficientes en coste de preservación digital.

Respecto a la capacidad de recuperación o resiliencia, a medida que van sucediendo las nuevas adopciones de software el porcentaje de recuperación en cada nueva adopción va aumentando casi sin excepciones. Esto es un buen indicio, el cual indica que a medida que pasa el tiempo el sistema va mejorando en cuanto al tratamiento curativo y predictivo de la obsolescencia, pudiendo llegar a estabilizarse y no verse afectado por adopciones de software futuras.

Finalmente, comentar que si nos fijamos en la capacidad de recuperación, los mejores resultados se obtienen cuanto menor es el porcentaje de compartición de los usuarios de la red. La explicación seguramente viene dada porque no se han podido hacer suficientes copias de los objetos digitales, lo cual hace que se vean menos afectados por las primeras adopciones de software.

Bibliografía

A. Trias i Mansilla and J. Ll. de la Rosa i Esteva. Asknext: An Agent Protocol for Social Search, ISSN 0020-0255, Information Sciences 190, 144–161, Elsevier, May 2012 http://dx.doi.org/10.1016/j.ins.2011.12.012

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small world' networks. Nature, 393:440–442, June 1998.

De la Rosa J. Ll. i Olvera J. A. (2012), First Studies on Self-Preserving Digital Objects, 15è Congrés Internacional de l'Associació Catalana d'Intel·ligència Artificial (CCIA 2012).

De la Rosa, J. Ll., Trias, A., del Acebo, E., Aciar, S., and Quisbert, H. (2009). Crew Intelligence Systems for Digital Objects Preservation, SIAAS-09 – 2nd Swarm Intelligence Algorithms and Applications Symposium, April 6-9, 2009.

De la Rosa, J. Ll., Trias, A., Ruusalepp, R., Aas, F., Moreno, A., Roura, E., Bres, A., and Bosch, T. 2010. Agents for Social Search in Long-Term Digital Preservation, The Sixth International Conference on Semantics, Knowledge and Grid, SKG 2010, Nov 1-3, Ningbo, China.

Gantz, J. and Reisel, D. (2011). Extracting Value from Chaos. Disponible en: www.emc.com /collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf

Hangal S. Reshaping reminiscence, web browsing and web search using personal digital archives. Ph.D. thesis, Computer Science Department. Stanford University, 2012. DOI= http://suif.stanford.edu/~hangal/hangal-thesis.pdf

Jiménez León, A. (2006). Preservación digital vs. obsolescencia de la información. Apertura, 3, 101-107.

Jin, X., Jiang, J. and De la Rosa J. Ll. 2010. PROTAGE: Long-Term Digital Preservation Based on Intelligent Agents and Web Services, ERCIM News Vol: 80, pp: 15- 16, January 2010.

Josep Lluis de la Rosa, Johan E. Bengtsson, Raivo Ruusalepp, Ann Hägerfors, and Hugo Quisbert, Using Agents for Long-Term Digital Reservation the PROTAGE Project Book Series Advances in Soft Computing Publisher - International Symposium on Distributed Computing and Artificial Intelligence 2008 (DCAI 2008), Vol. 50/2009 pp: 118-122, ISSN 1615-3871(Print) 1860-0794 (Online) , Ed. Springer Berlin / Heidelberg, DOI 10.1007/978-3-540-85863-853

Olvera J. A. 2013. Digital Preservation: a New Approach from Computational Intelligence, Joint Conference on Digital Libraries 2013, JCDL Doctoral Consortium 2013, July 22-26, Indianapolis, Indiana, USA. Available at http://www.ieee-tcdl .org/Bulletin/current/papers/olvera.pdf

Olvera J. A., Carrillo P. i De la Rosa J. Ll, Combinatorial and Multi-Unit Auctions Applied to Digital Preservation of Self-Preserving Objects, CCIA 2014, October 22-24, Barcelona, Spain.

Reich V. and Rosenthal D.S.H. LOCKSS (Lots Of Copies Keep Stuff Safe), Presented at Preservation 2000: An International Conference on the Preservation and Long Term Accessibility of Digital Materials, December 7-8, 2000, York, England. Also published in *The New Review of Academic Librarianship*, vol. 6, no. 1, 2000, pp. 155-161.

Rivera Donoso, Miguel Angel (2009). Directrices para la creación de un programa de preservación digital. Serie Bibliotecología y Gestión de Información N° 43, Marzo 2009. ISSN: 0718 – 1701

Apéndice
Comportamiento de los CADO con inteligencia de enjambre

Las soluciones de Preservación Digital (PD) se clasifican en dos grandes estrategias: la preservación del entorno tecnológico (emulación) y superar la obsolescencia de los formatos de archivo (migración). Es el segundo enfoque en el que nos hemos centrado en simular las migraciones de formato con la técnica de inteligencia de enjambre (*Swarm Intelligence*).

Se trata de una primera serie de estudios con los paradigmas emergentes o ecologías computacionales aplicados a la PD, como se señala en nuestro trabajo previo en el SIAAS de 2009 (De la Rosa, J. Ll., Trias, A., del Acebo, E., Aciar, S., and Quisbert, H.). , donde utilizamos inteligencia de enjambre para la PD. En ese trabajo los CADO fueron preservados por un enjambre de robots de preservación que buscaban CADO con problemas de obsolescencia en un sistema de archivos. El resultado fue una mayor escalabilidad en la realización de PD vs. un crecimiento exponencial del número de CADO (se duplica cada 18 meses).

A diferencia de (De la Rosa, J. Ll., Trias, A., Ruusalepp, R., Aas, F., Moreno, A., Roura, E., Bres, A., and Bosch, T. 2010 y De la Rosa, J. Ll., Trias, A., del Acebo, E., Aciar, S., and Quisbert, 2009), en este estudio, los CADO son ellos mismos los que buscan su propia preservación. Por lo tanto, interpretamos la migración de los formatos de los CADO como la replicación de archivos en diferentes formatos para ser distribuidos en las redes P2P para tratar de preservarlos. La analogía enjambre utilizada para esta simulación es: los CADO son las hormigas, los diferentes usuarios que están en la red representan el lugar donde hay comida (mantendrá vivas las copias de objetos digitales) y las computadoras de los usuarios son hábitats. Estos objetos digitales (hormigas) buscan la preservación moviéndose a través de la red y haciendo copias de sí mismos (descendientes) en diferentes formatos (Figura 6). En este trabajo, los CADO son fotos o archivos de vídeo para simplificar la ilustración.

Los CADO tienen un presupuesto de PD (al que nos referiremos como el *presupuesto*) dedicado a financiar la replicación de archivos y otras operaciones, como la migración de formato o moverse a través de la red (social) de los usuarios, y una parte de él a su descendientes. Operaciones como checksum, la migración y otras gastan presupuesto, mientras que operaciones como el acceso de los usuarios lo aumentan. Cuando algún descendiente se queda sin presupuesto, este trata de volver al sitio de su antecesor para obtener más presupuesto y luego seguir con sus operaciones.

En esta prueba habrá copias de CADO en diferentes formatos para acabar manteniéndolos vivos contra catástrofes que probablemente suceden (siguiendo la imagen de PD como el problema de rescate mencionado anteriormente). Las catástrofes se definen como un cambio masivo de formatos de los CADO que sufren

durante toda su vida. También hacemos desaparecer un porcentaje de CADO para simular que ya no se pueden leer debido a los cambios tecnológicos que ocurren en dichas catástrofes.

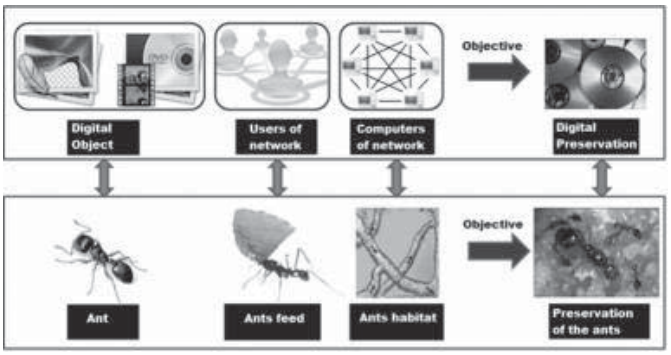

Figura 6: Modelo de enjambre propuesto para la preservación digital

En el modelo, una red de usuarios donde los nodos son los equipos de los usuarios y las conexiones entre los nodos determinan qué usuarios son amigos de quién, lo que resulta en una red social enfocada en PD (Jin, X., Jiang, J. and De la Rosa J. Ll. 2010) 00. Los CADO viajan a través de la red P2P que se construye a partir de la red social y hacen copias de sí mismos para poder ser preservados. Las copias llevarán un vínculo al objeto padre porque, como se explica en la sección 2, corresponden a un mismo objeto. Cada nodo, que representa el equipo de un usuario, solo podrá leer un único formato de imagen de vídeo. Los formatos varían de más antiguo a más reciente para simular lo que ocurre con los archivos que tenemos en una computadora, donde solo puede ser leído un formato concreto por el software específico instalado en el ordenador. Los archivos en formatos antiguos terminan por no ser legíbles porque en ese equipo no son compatibles con las nuevas versiones de software.

Otros detalles de las simulaciones son explicados a continuación.

Configuración

Los CADO y sus comportamientos

Los CADO en nuestros experimentos son archivos de vídeo o imagen con un formato representado por los valores numéricos de 1 (antiguo) a 5 (el más nuevo) y con un presupuesto de PD. El Presupuesto de PD se compone de sentimientos (*feelings*), para que paguen los servicios PD con los sentimientos. Un sentimiento es una unidad monetaria de PD definida como un acceso de un usuario (cuando el

usuario abre un fichero demuestra un interés por este). En el futuro los sentimientos serán suministrados por evaluaciones de los usuarios o incluso tendrán la conversión en monedas legales. Cuando el presupuesto se agota, el CADO debe volver al usuario al que pertenece para ser recargado; en nuestros experimentos el presupuesto inicial es de cualquier tamaño inferior a 60 sentimientos (una estimación que hemos hecho para permitir que los CADO se puedan distribuir a través de la red P2P).

Inicialmente, cualquier CADO pertenece a un usuario en particular, y vive en su sistema de archivos, su computadora, un nodo de la red P2P. El nodo donde nació digitalmente se llama el nido. Cuenta con los siguientes comportamientos que consumen presupuesto: hacer una copia (opción más cara), trasladarse a otros nodos de la red (opción intermedia) o permanecer en el nodo donde se encuentran (opción más barata). El CADO y sus descendientes con formatos migrados van de un usuario a otro por toda la red, en busca de presupuesto fresco y hacer copias digitales de ellos mismos si pueden permitírselo.

El coste de un CADO para vivir en un nodo (siguiendo con la analogía enjambre, el nido) que admite el mismo formato que el del CADO es menos costoso que si el nodo admite un formato diferente. Por último, estos tienen que saber si el CADO original o una de sus copias están vivos en la red.

Catástrofes

Regularmente hay una *catástrofe* digital que consiste en migraciones masivas de formato motivadas por el cambio de los soportes de los proveedores de software que esperan que los objetos en formatos antiguos se actualicen a los nuevos formatos. Los cambios se comunican a todos los usuarios de la red. También desaparece un porcentaje de los CADO. La catástrofe simula la actualización masiva para el nuevo software (por ejemplo, migrar de un Word 2003 a un Word 2007) como si un grupo importante de usuarios están convencidos de cambiar a un nuevo software y sus formatos propietarios o abiertos (siguiendo el ejemplo anterior, de .doc a .docx, por ejemplo). Asumimos que cada 5 años es el período típico de actualización de software, y por lo tanto las catástrofes se producen al mismo ritmo.

Ejecutamos varias veces el mismo patrón de catástrofes como una sucesión de adopciones de software a lo largo de la vida de los CADO y hacemos el promedio de las ejecuciones y sacamos conclusiones acerca de cómo se conservan los CADO, mientras afrontan varios cambios de formato. El patrón se almacena en un fichero donde cada línea contiene tres valores: el tipo de objeto (vídeo o imagen), el formato que cambia y el nuevo formato al que cambia. Una vez que el archivo se lee y los atributos necesarios para procesar los cambios de formato se actualizan, se advierte al usuario para actualizar el formato de los videos e imágenes.

Hemos llevado a cabo los experimentos bajo la intensidad más representativa de las catástrofes; 33% de los usuarios adoptan el nuevo software, por lo

que después de 3 catástrofes todos los usuarios han adoptado un nuevo software (esto es, en 15 años).

Los usuarios

Los usuarios tienen un servicio de preservación, representado por un porcentaje del empleo (0-100%) que describe el coste que los CADO pagarán por su preservación en el nodo del usuario (o nido) en cada paso de la simulación, y una estructura de datos que indica la proximidad social con los otros nodos como los usuarios que puedan ser susceptibles de ayudar en compartir esfuerzos de PD, por ejemplo, al compartir sus CADO. Por último, el nodo del usuario le dice los formatos de vídeo e imagen que son compatibles.

Los usuarios tienen una lista de contactos que se representa mediante nodos (nodos de otros usuarios) que están conectadas. Estos son las computadoras de sus amigos, lo que resulta en una red social donde los CADO pueden moverse. Los usuarios pueden actualizar los formatos de los CADO cuando cambian de formatos con la actualización de su software.

La medida de la resiliencia esperada

Utilizamos la entropía de Shannon (ecuación 1) para la evaluación de los resultados. Cuanto mayor sea el valor de la entropía, mejor es la capacidad de preservación, de acuerdo con la estrategia de migración de formato. El hecho de tener varias copias en formatos diversificados dará mayor resistencia y recuperación frente a las catástrofes (mejor capacidad de recuperación).

$$H(x) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

La ecuación 2 es la aplicación de la ecuación 1 para la PD, donde n es el número total de original de CADOs, j son los formatos que el CADO puede migrar (de 1 a 5), y *Pij* es el porcentaje de las copias de formato *j* del total de copias de un CADO original *i*.

En la ecuación 3 pi,j se describe, donde *k* son los formatos de 1 a 5 .

$$H(x) = \frac{-\sum_{i=1}^{n}\left(\sum_{j=1}^{5} p_{i,j} \cdot \log_2 p_{i,j}\right)}{n}$$

$$p_{i,j} = \frac{f_{i,j}}{\sum_{k=1}^{5} f_{i,k}}$$

La Tabla 2 muestra lo que nuestra entropía calcula: cuantas más copias y más formatos conducen a los más altos valores de la entropía, así esperando mayor capacidad de recuperación. El ejemplo 1 tiene más resistencia que el ejemplo 3 y mucho más que el ejemplo 2, ya que se han diversificado los formatos incluso aunque los ejemplos 1 y 2 tengan el mismo número de copias.

Ejemplo	Formatos	Copias	$p_{i,j}$	$p_{i,j\ log2}\, p_{i,j}$	H(x)
1	1	3	0,231	–0,4881	
	2	1	0,077	–0,2846	
	3	5	0,385	–0,5301	2,0758
	4	1	0,077	–0,2846	
	5	3	0,231	–0,4881	
2	1	13	1	0	
	2	0	0	0	
	3	0	0	0	0
	4	0	0	0	
	5	0	0	0	
3	1	0	0	0	
	2	1	0,333	–0,5283	
	3	2	0,667	–0,3899	0,9182
	4	0	0	0	
	5	0	0	0	

Tabla 2: Tres ejemplos aleatorios de entropía como medida de diversidad

Trabajo experimental

Configuración de los experimentos

Los experimentos se llevaron a cabo con el TiM, como se describe en el documento. En este set de pruebas hemos escogido como topología de red social a simular la

Small World, de Watts and Strogatz 0, ya que es una de las topologías de red más referenciada que se encuentra en la literatura, debido a que estas estructuras se caracterizan por tener nodos fácilmente alcanzables en pocos pasos y los nodos tienden a agruparse localmente, mucho más que si las redes se generaran al azar. Son, por lo tanto, las que más se asimilan a las redes sociales actuales.

Concretamente se ha utilizado una topología de 100 nodos, que representan a los usuarios de la red (ver Figura 7), con una adopción de software del 33% de CADOs. Las simulaciones son de un período de 20 años, donde se produce una catástrofe cada 5 años. Es el tiempo que los CADO muestran síntomas graves de obsolescencia y necesidad urgente de PD. Un paso de simulación es equivalente a un mes. Por lo tanto, habrá una catástrofe cada 60 pasos (1 steps x 12 meses x 5 años) y el total de la simulación es de 240 pasos (1 steps x 12 meses x 20 años). La Tabla 3 muestra los parámetros utilizados en las simulaciones.

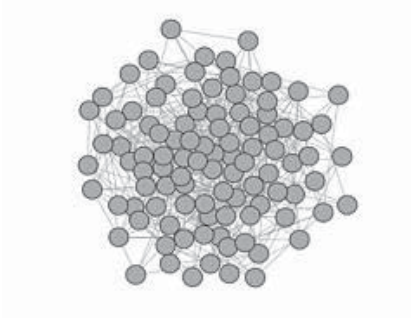

Figura 7: Topología de red utilizada

Parámetros	
Cambios de formato (3)	0,5,4; 1,4,3; 0,3,2
# DOs inicialmente asociados a un usuario	Entre 1 y 5
Coste de desplazarse de un nodo a otro	2
Coste de permanecer en un nodo (por paso de simulación)	1
Rango inicial de presupuesto	20 = 60
Coste de acomodación de un DO en un usuario ajeno	Entre 1 y 10
Servicio de preservación de un usuario	25%, 50% y 100%
Porcentaje de usuarios que adoptan nuevos formatos	33%

Tabla 3: Parámetros utilizados en la simulación

Con este prototipo hemos hecho un nuevo paso en la simulación que permite variar fácilmente los parámetros que intervienen en la emulación a través de interfaz gráfica de usuario para trabajar en ciclos de experimentación masivos. Además, proporciona la visualización en tiempo real de lo que está pasando para entender el comportamiento de los CADO. El análisis y diseño del prototipo se llevó a cabo siguiendo la metodología INGENIAS[4] y fue implementado en Java.

Como hemos mencionado en las secciones anteriores, el CADO tiene un presupuesto para llevar a cabo las operaciones (como copiar, mover o permanecer en un nodo). Esto se mide en unidades de sentimientos (el número de veces que el usuario experimenta cualquier tipo de sensación después de haber estado expuesto a una foto o un vídeo contenido en el OD). El presupuesto se recarga con una serie de sensaciones a medida que se visita. Supusimos en este experimento que las fotos se observan una vez al año cuando los usuarios desean recordar experiencias pasadas. En el prototipo la visita se produce cuando es válido el estado de la ecuación 4 (la probabilidad de ser visitado en un ciclo). Cada vez que el CADO es visitado, se recarga con una cuarta parte del presupuesto que tenía inicialmente en su nacimiento.

$$random < \frac{1}{year\_cycles}$$

La heurística detrás del presupuesto de PD es que cuanto más se pague más capacidad de preservación tendrá.

Otra característica es que cuando un CADO decide hacer una copia, divide su presupuesto en dos: una mitad dedicada al padre CADO y la otra mitad dedicada al hijo CADO. Esto se debe a que representan un mismo OD (encapsulan las diferentes versiones que emigran) y el presupuesto total del OD debe ser la misma, invariable, antes y después de hacer las copias. El formato más reciente de la copia se selecciona al azar de 1 a 5.

Estabilidad estadística

Se ha trabajado con precisión la estabilidad estadística mediante la introducción del criterio de la suma acumulada, donde en cada ejecución se compara la entropía promedio de las últimas ejecuciones con el promedio de la entropía de la ejecución actual: Si esta diferencia es menos de una épsilon a través de una ventana de 5 a continuación se termina la ejecución (véase la ecuación 5 y la Figura 8). Hemos implementado un método que parte de una ventana de cinco valores de estas diferencias, normaliza estos valores y calcula el porcentaje que

representa estos valores hasta el momento. Así, el número de ejecuciones finaliza cuando se alcanza la fiabilidad de $1 - \varepsilon = 6,.2\%$ o mayor.

$$S_n = \sum_{i=1}^{n} \left(\overline{x_n} - \overline{x_{n-1}} \right) < \varepsilon$$

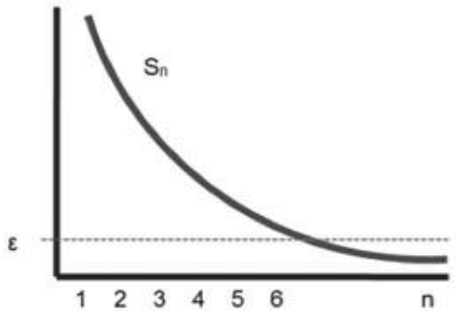

Figura 8: Cálculo de la estabilidad estadística usando la suma acumulada

Notas

[1] Del informe de PROTAGE, http://www.protage.eu

[2] http://www.pyramid.cat/

[3] http://durafile.eu/

[4] INGENIAS. 2005. Universidad Complutense de Madrid (GRASIA). http://grasia.fdi.ucm.es/ingenias


ESPAÑA
LIBRO
DE
FAMILIA
Sistema Nacional de Salud
PERMISO DE CONDUCCIÓN REINO DE ESPAÑA
ESPAÑA
¡A QUE ESTÁS CAMINANDO SIN PATINES Y ESTROPEANDO EL PARQUET!