Digital Preservation: a New Approach from Computational Intelligence

Jose Antonio Olvera
TECNIO EASY Innovation Center University of Girona
Campus de Montilivi, E17071
Girona, Catalonia (EU)
+34 972 418478
joc7188@gmail.com

ABSTRACT
This research lays the foundations for a new object-centric digital preservation paradigm based in computational intelligence, where the digital objects manage their own preservation and budget, powered by a social network as an environment that enables their behavior under the policy that "preservation is to share". The mission of these objects is their long-term preservation, which entails being accessible and reproducible by users at any time in the future regardless of frequent technological changes due to software and hardware vendors' upgrades. This approach tries to meet the following digital preservation requirements: adaption to unexpected situations, scalability and efficient cost management.

Categories and Subject Descriptors
I.2.11**[Artificial Intelligence]**: Distributed Artificial Intelligence – *multiagent systems*;

I.6.8 **[Simulation and Modeling]**: Types of Simulation– *discrete event, visual.*

General Terms
Algorithms, Measurement, Experimentation

Keywords
Digital Preservation, Self-preservation, Social, Agents, Computational Intelligence

1. CONTEXT AND SITUATION
The Long Term Digital Preservation (LTDP), or in short Digital Preservation (DP), is increasingly in focus for businesses, public sector agencies, as well as scientists and citizens. The challenge in preserving valuable digital information – consisting of text, music, images, multimedia documents, web pages, sensor data, etc. generated throughout all areas of our society – is real and growing at an exponential pace. A recent study by the International Data Corporation (IDC) found that in 2010 was broken the zettabyte barrier of information created and replicated, and in 2011 was expected to surpass 1.8 zettabytes (1.8 trillion gigabytes), growing by a factor of 9 in just five years [6]. The LTDP of such information will become a pervasive as well as ubiquitous problem that will concern everyone who has digital information to be kept for long time, implying a shift in at least a couple of software and hardware generations. So far, only large memory institutions with expert knowledge and specialized tools have been able to tackle this problem. LTDP cannot be addressed by a single institution or nation. Libraries, archives, and other memory institutions share this challenge with each other and with individual collectors and creators (www.digitalpreservation.gov).

A recent review of DP research noted that despite twenty years of active research, there is still a lot of work to solve the core problems [9].The same report suggested to look for radically new approaches to DP to support high volumes of data, dynamic and volatile digital content, keeping track of evolving meaning and usage contexts of digital content, safeguarding trust , usability and understandability, integrity, authenticity and accessibility over time, as a model enabling automatic and self-organizing approaches to DP.

The level of automation in DP solutions is low. The preservation process currently has many manual stages but should be approached in a flexible and distributed way, combining intelligent automated methods with human intervention. The scalability of existing preservation solutions has been demonstrated to be poor. In addition, solutions have often not been properly tested against diverse digital resources or in heterogeneous environments.

Research in the DP domain has moved away from trying to find one ideal solution to the DP problem and has been focused on defining practical solutions for different preservation situations. These solutions have to exploit the expert knowledge of memory institutions, be based on industry standards and above all, be scalable and adaptable to disparate environments.

Approaches that fit the new trend is the "self-preserving objects" or the crowdsourcing or socialization of the digital preservation efforts. Although preservation measures are currently being performed by repositories, they ought to be performed by the objects themselves, so mechanisms are required to enable preservation management of objects, and for this to succeed, objects must be self-preserving. There will be different preservation pathways for different kinds of object. As said in the Objective ICT-2011.4.3 Digital Preservation (EU FP7 call 6 of

2012)[1], Self-preserving objects are seen by many as the 'Holy Grail' of preservation, but no individual Research team has the capacity to address this problem.

2. STATEMENT OF THESIS OR PROBLEM

This research tries to meet the following requirements involved in DP:

1. **Scalability**: The exponential growth of digitally born objects requires of scalable solutions from the technological point of view.
2. **Cost**: Associated to the exponential growth because there are limited resources to cope with DP.
3. **Uncertain future**: DP is about heuristics of what results we will get in the future, only.

Solutions today are not scalable enough, deal not properly with costs and no guarantee of success in the future. The prevailing paradigm is centralized, top-down, where institutions are the main players. I propose studying a change of paradigm, mainly bottom-up, where the digital objects self-preserve. I propose that their behaviors must be guided by computational intelligence with proven track record of adaptation to unexpected situations with strong resilience, scalability, and efficient cost management. To make this paradigm come true there are important research issues associated with providing with self-preserving capacities to the digital objects themselves, as well as a support environment where we will involve the end users (personal archiving) who will have a more important role with respect to the prevailing paradigm where only large institutions takes charge of preservation.

My thesis will study what **self-preservation** behaviors need the self-preserving digital objects, based in **computation intelligence (CI),** and related methods of cost management under their own **budget**, powered by a **social network** as an environment that enables their behavior under the policy that "preservation is to share". In this concept, digital objects become active actors in their own LTDP, here named the Self-Preserving Digital Object (SPDO) [4], which has a DP budget devoted to funding the replication of the objects and other operations such as format migration or moving through a social network of users; in all, a controlled environment where they will "live".

So the central question is: what if digital objects were self-preserved? I put forward this question from the perspectives that **preservation is to share** and that **bottom up** approaches are the right solution for the scalability and costs issues of digital preservation. In my thesis I will focus on their behavior, their architecture, and their environment.

The focus of this research is theoretical and practical at a time, with strong analysis with simulation and engineering application of the behaviors that have been found to work better. Figure 1 shows the roadmap of the present thesis, which is divided into three confluent areas to be explained in the following section.

[1] http://ec.europa.eu/research/participants/portal/page/cooperation?callIdentifier=FP7-ICT-2011-9

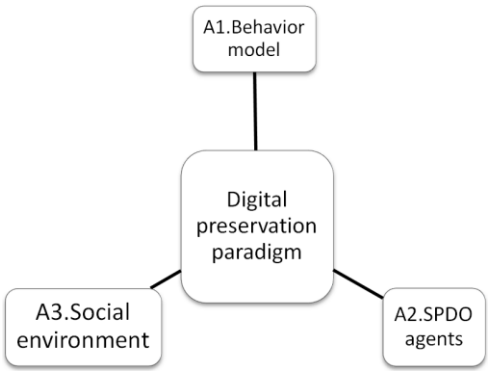

Figure 1: The roadmap of the thesis, divided into three confluent areas

3. PROBLEMS TO SOLVE

This self-preservation paradigm implies to solve several questions from three different points of view: the behavior model, the architecture of the self-preserving digital objects from the agents metaphor, and the social environment where they will live (operate).

3.1 Behavior model (A1)

The behavior of the objects must be simple and highly adaptable to an uncertain environment. LTDP is uncertain in the long term. A solution that today might be efficient might not work in a long future. Evolution has proved to be resilient and adaptive to unforeseen changes well into the future. That is why different CI techniques will be studied, yet I will focus on Evolutionary Computation (involving combinatorial optimization mechanisms and mechanisms inspired by biological evolution) and Swarm Intelligence (based on the collective behavior of decentralized and self-organized systems). These techniques solve problems in an implicit way through the collective behaviors and where the actors of the system cooperate.

3.2 SPDO architecture from the agents metaphor (A2)

It consists of an explicit and individualistic vision of how digital objects agents should be. In this area, the architecture of the agents will be designed following the agent metaphor, such as the **social** skills that determine how they interact with each other, how they must manage their **budget** (devoted to funding the replication of files and other operations such as format migration or copying themselves through the social network of users), what rules must be applied to determine its **mission**, and how they will manage their copies, that are distributed in the social network. At this level, digital objects cooperate with them, but also compete for attracting know-how from users and the best affordable services from tools such as format migration, metadata extraction or renewed storage under restrained budgets.

SPDOs will be ready to work with auction systems, with the responsibility for the search and selection of recovery and maintenance services at the best possible cost for its individual files. The digital objects will be responsible for damage level estimation, search for the best cost/quality ratio, integrity analysis, resource availability and the long-term view.

3.3 Social environment (A3)

This area defines the social network as an environment that enables the behavior of the self-preserving digital objects. It is

necessary to solve further questions: how the social network must technically be, what are the social behaviors that supports the task of preservation of the digital objects (rules of collaboration), what are the topologies of social networks that back better the digital objects and how to promote these topologies, the sign in and sign out management of the users of the network, and the definition of compensation mechanisms for users to promote user engagement.

Finally, the DO object expressed under the decisions of the three confluent areas, will be implemented using the **agent technology** in a prototype of DP and will be experimented with a real social network of users.

4. BACKGROUND AND RELATED WORK

The very first evidence of **object-centric paradigm**, was the Buckets [11] as aggregative, intelligent, WWW-accessible DOs that were optimized for publishing in Digital Libraries (DLs), that existed within the ―Smart Objects, Dumb Archives (SODA) Digital Library model of [10]. Buckets implement the philosophy that information itself is more important than the DL systems used to store and access information. Buckets were designed to imbue information objects with certain responsibilities, such as the display, dissemination, protection, and maintenance of their contents, as will be done in this research.

There are also projects following new digital preservation paradigms but with our same policy that **preservation is to share**. A known program is LOCKSS [12], that is an open-source, library-led digital preservation system built on the principle that "lots of copies keep stuff safe". The LOCKSS system allows librarians at each institution to take custody of and preserve access to the e-content to which they subscribe, restoring the print purchase model with which librarians are familiar. Using their computers and network connections, librarians can obtain, preserve and provide access to purchased copies of e-content. This is analogous to libraries' using their own buildings, shelves, and staff to obtain, preserve, and provide access to paper content. The common principle of my PhD thesis with this research is that they bet for a decentralized and distributed preservation. Among the projects presented in the last Personal Digital Archiving Workshop in Maryland on February 2013 (PDA 2013) we found the MUSE (Memories USing Email) system, which provides four novel types of cues to help spot interesting trends and messages in a large-scale email archive [7]. The possible benefits from tools like MUSE range from utilitarian ones such as summarizing work or backing up attachments, to reminiscence and remembering family events and grad school years with nostalgia, to reinforcing confidence, renewing relationships and playing memory games. They play a key role in the motivation for the personal archiving as a first distributed step for the DP at a large scale.

Regarding **cost management**, the work of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (BRTF) and its report *Sustainable Economics for a Digital Planet* (2010) [1] was a significant and novel addition to the literature on digital preservation and was the first systematic attempt to focus not just on the cost of managing information over time, but on the economic framework that is required to allow that to happen.

The idea of looking at the economics of preservation was further elaborated in the UK by the KRDS/I2S2 Digital Preservation Benefit Analysis Tools Project[2] that developed a toolkit for establishing the value chain and benefits analysis of digital preservation. The results of BRTF and KRDS were incorporated into the Digital Preservation Economic Sustainability Reference Model[3] (2011). The ESRM is looking at the whole economic life-cycle of digital assets, not just on actions once digital assets move into an archive. The present thesis will build on the economic life-cycle concept but approach it from the digital object perspective, rather than organisation level.

And last but not least, synergic works on **computational ecologies** [3][8] show how this approach might work, while the self-preserving DOs ask themselves how much preservation is necessary (appraise) and, according to a DP budget that would be regularly assigned to them, compete with each other for the preservation services. The self-preserving DOs might encapsulate the different versions they migrated to during their lives, in a sort of blog of their life, and their mission is to stay alive as long as possible. In this approach, being "alive" means being accessible, authentic, and readable, in the DP sense, creating an environment where DOs become *active actors* in DP *with their own budget for attracting DP know-how and services*. This is a shift of roles with respect to the prevailing DP paradigm, where users are the main actors; there has been recent research on new actors, such as preservation aware storage systems (IBM Haiffa) for the automation of DP services; but the DOs have never had such a role or responsibility before.

This research is a continuation of the first series of studies applying computational ecologies to DP, as noted in previous work in [5], in which was used Swarm Intelligence for DP. In that context, the DOs were preserved by a swarm of preservation robots that were searching for DOs with obsolescence problems in a file system. The result was an enhanced scalability in conducting DP with an exponential growth in the number of DO (doubling every 18 months). Differing from that model [5], in this PhD thesis, the DOs themselves seek self-preservation.

5. RESEARCH METHODOLOGY

I'm in the beginning of the PhD that is expected to be finished on June 2016. During the completion of this thesis, I will follow a standard research methodology. It consists of a state of the art from the articles published on the topic. After this initial state of the art, I will work in the three confluent areas explained above (A1, A2 and A3) separately, that I have some preliminary results in A1 with [4] . The work will be done on a simulated environment where results are obtained, analyzed and made partial publication of these results in high-level journals and international conferences on the topic. The method of experimenting the several behaviors of computational intelligence A1, the cost management A2, and social behavior A3 will be experimented with a number of social networks similarly to the method used by Nelson in [2] . As a result of this research, the state of the art will be extended. Finally a prototype will be implemented in a real environment in order to assure the DP requirements explained above.

6. DISSERTATION STATUS

In [4], we proposed a new method for DP utilising the self-preservation of DOs by creating DOs designed for fighting for their own preservation explained above. We demonstrated the

[2] http://beagrie.com/krds-i2s2.php

[3] http://unsustainableideas.wordpress.com/economic-sustainability-ref-model-page/

effectiveness of this strategy through experimentation with a prototype of digital obsolescence resulting from several (3) waves of new software adoption. This study is with Nelson's one of the first experimental works in this field illustrating how self-preserving DOs with simple behaviours can provide the ability to preserve their digital information.

We demonstrated that resilience was acquired by DOs through their self-preserving behaviour when the objects show swarm intelligence under constrained DP budgets. We observed how SPDOs partly recover after several software adoption waves that are caused by frequent format changes by software vendors (every 5 years on average). Our results indicate that resilience increases with swarms of DOs, yielding sound results of 91.09% readability at the rate of 10% software adoption. However, a long research path remains in the development of the ability to cope with the standard rate of 33% new software adoption because we obtained a promising yet insufficient 56.06% recovery at this adoption rate; for a 50% new software adoption rate, we obtained only a 23.63% recovery.

Figure 2 shows the dissertation plan of this thesis and remarks what is the current state. Actually we have implemented a robust platform of simulation of the self-preservation paradigm explained in this paper. Now we are experimenting A1 (behavior models) and A3 (social environments) separately in order to obtain preliminary conclusions of each area. The idea is to finish the research in the three areas between the end of this year and beginning of the next year, and then start to implement the real prototype during the following year and finally exploit the results with a real social network of users.

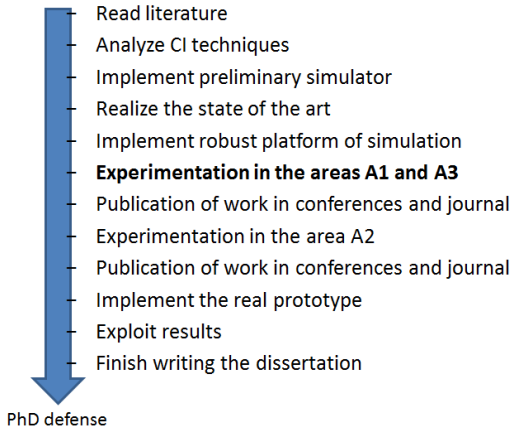

Figure 2: Dissertation plan

7. EXPECTED CONTRIBUTION

We expect to obtain with our platform of simulation, the 99% of readability of the digital objects at the rates of 10%, 33% and 50% of new software adoption waves along cycles of 3 times 5 years (that is 20 years of simulation). And with the real prototype it is intended to achieve the requirements that have been exposed, necessary for DP.

8. THE NOVELTY OF THE RESEARCH

I will make a contribution to a change of paradigm, mainly bottom-up, where the digital objects self-preserve and with the novelty that their behaviors must be guided by computational intelligence software, and bottom-up cost management, of proven track record in the adaptation to unexpected situations, scalability, and efficient cost management.

We will try to achieve the objectives of EU FP7 call 6 of 2012 (mentioned before), that is to preserve digital content in a more effective and cost-efficient manner while protecting its authenticity and integrity, significantly reducing the loss of irreplaceable information, and ensuring it may be reused in the future. The special targeting are the technologies and systems for intelligent management of preservation. This research should help to support high scale LTDP through innovative technologies that embed reasoning and intelligence in the content itself, and also activities may cover solutions to manage obsolete information under affordable costs.

9. ACKNOWLEDGMENTS

This research is partly funded by the Spanish MICINN (Ministerio de Ciencia e Innovación) projects TIN2010-17903 *Comparative approaches to the implementation of intelligent agents in digital preservation from a perspective of the automation of social networks*, IPT20120482430000 (MIDPOINT) *Nuevos enfoques de preservación digital con mejor gestión de costes que garantizan su sostenibilidad*, the EU DURAFILE num. 605356, FP7-SME-2013, BSG-SME (Research for SMEs) *Innovative Digital Preservation using Social Search in Agent Environments*, as well as AGAUR 2012 FI_B00927 awarded to José Antonio Olvera and the grup de recerca consolidat CSI-ref.2009SGR-1202.

10. REFERENCES

[1] Berman F. and Lavoie B., the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Sustainable Economics for a Digital Planet, April 2010. DOI= http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

[2] Cartledge C.L., Nelson M.L. *Analysis of graphs for digital preservation suitability*. Procs. 21st ACM Conference on Hypertext and Hypermedia, Toronto, Ontario, Canada, June 13-16, 2010. pages 109-118

[3] de la Rosa J. L., Hormazábal N., Aciar S., Lopardo G., Trias A., and Montaner M. 2011. A Negotiation Style Recommender Based on Computational Ecology in Open Negotiation Environments, ISSN: 0278-0046, IEEE Trans. on Industrial Electronics 58 (6) 2073-2085, June 2011

[4] de la Rosa J. L. and Olvera J.A. *First Studies on Self-Preserving Digital Objects*. Artificial Intelligence Research & Dev., Procs 15th Intl Conf. of the Catalan Assoc. for Artificial Intelligence, CCIA 2012, Vol.248, pp: 213-222, 2012, Alacant, Spain.

[5] de la Rosa, J. L., 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

[6] Gantz, J. and Reisel, D. (2011). Extracting Value from Chaos. DOI= http://idcdocserv.com/1142.

[7] 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

[8] Hogg, T. and Huberman, B.A. 1991. Controlling chaos in distributed systems, IEEE Trans. Syst. Man Cybernetics 21 (6) 1325, 1991

[9] Hugo Quisbert. On long-term digital preservation information systems : a framework and characteristics for development. PhD Thesis. Luleå University of Technology. DOI= http://epubl.ltu.se/1402-1544/2008/77/LTU-DT-0877-SE.pdf

[10] Maly, K., Nelson, M. L., &Zubair, M. 1999. Smart objects, dumb archives: a user-centric, layered digital library framework. D-Lib Magazine, 5(3), 1999. DOI= http://www.dlib.org/dlib/march99/maly/03maly.html

[11] Nelson M. 2001, Buckets: Smart Objects for Digital Libraries, PhD thesis, Old Dominion Univ.

[12] 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.