# An Outline of the Application of Agents to Digital Preservation and an Introduction to Self Preservation Aware Digital Objects

Jose Antonio Olvera and Josep
Lluis de la Rosa i Esteva
TECNIO - Centre EASY
Universitat de Girona
+34 972 41 84 78
joc7188@gmail.com

## ABSTRACT
This is dissertation on three possible applications of agents to digital preservation through the agentification of namely digital objects, services, and social networks. Their proofs of concept are explained.

## Categories and Subject Descriptors
I.2.11 Distributed Artificial Intelligence: Intelligent Agents

## General Terms
Algorithms, Experimentation.

## Keywords
Digital Preservation, Self-awareness, Social, Agents

## 1. INTRODUCTION

The challenge in preserving valuable digital information – consisting of text, video, images, music, 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 a total of 3,892,179,868,480,350,000,000 (that's roughly 3.9 trillion times a trillion) new digital information bits were created in 2008. In the future, the digital universe is expected to double in size every 18 months, according to the IDC report[1]. The Digital Preservation (DP) 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. DP 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).

The mission of the PRESERVA project (acknowledgements in section 7) is to make DP easy enough for individuals, companies and general institutions to perform so that they can preserve digital content, and at the same time help to reduce the cost and increase the capacity of memory institutions to preserve digital information for the long-term. PRESERVA will build and validate software agents for DP that can be integrated in existing and new preservation systems.

This approach will be of interest to industry [1] because while 70% or more of the digital universe is created, captured, or replicated by individuals — consumers and desk and information workers toiling far away from the data center — enterprises and institutions, at some point in time, have responsibility or liability for 85%. The PRESERVA project will raise the DP awareness of individuals through personal agent environments as a daring step to help companies and institutions commit to DP needs and legal requirements. The vision is to connect the digital assets to the future, keeping digital content "alive", that means, always ready to be access at any time in the future.

The fact is that today 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 [30]. In addition, solutions have often not been properly tested against diverse digital resources or in heterogeneous environments. Quisbert in his PhD thesis [30] suggested to look for radically new approaches to DP to solve core problems like the 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.

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 [28]. 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.

The research of PRESERVA and this paper as a first attempt, will lay the foundations for a new object-centric DP paradigm. It solves preservation issues involving complex digital objects (Self-Preservation Aware Digital Object –SPADO) by building new DP environments where objects become active actors with their own budget for attracting know-how and services.

Exploratory research required to develop SPADOs examines a number of synergetic areas, focusing on preservation of complex objects, multi-agent systems (MAS), computational ecologies,

**Cite as**: An Outline of the Application of Agents to Digital Preservation and an Introduction to Self Preservation Aware Digital Objects, José Antonio Olvera and Josep Lluís de la Rosa i Esteva, Proc. of 13th European Agents Systems Summer School (EASSS 2011), July, 11–16, 2011, Girona, Spain, pp. XXX-XXX. 

[1] http://www.storagenewsletter.com/news/miscellaneous/idc-digital-information-created

cloud computing, and social networking. The project will provide a proof of concept of the emergent behaviour of huge communities of SPADOs competing for the best DP services and the appropriate DP know-how available. The expected result of the new DP paradigm is the efficient preservation of the "*Data Deluge*" in line with the statement of Fran Berman [29] that "*ensuring that our most valuable information is available both today and tomorrow is not just a matter of finding sufficient funds*". Fran Berman, co-chair of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, also asserted that "[*digital preservation] is about creating a 'data economy' in which those who care, those who pay, and those who preserve are working in coordination*".

This paper focus on the most urgent issues that the new object-centric preservation paradigm requires from the self-organization point of view, by understanding whether it provides with scalability and costs management, and explore whether it is possible to benefit from resilience from losses provoked by frequent changes of formats (what Berman referred to as the shift of hardware and software). Part of the work will then be about defining the architecture for the new type of digital object and the smarter environment that will support its activities.

Let us first analyse the requirements of DP. We will do it by comparing its requirements to those of robotic rescue, robotic soccer, chess, and deep space probe, that are domains where agents can be applied. Many factors determine the difficulty of the preservation of Digital Objects (DO): heterogeneous MAS work in dynamic and even hostile environments in which new and complex scenarios constantly appear (Table 1). An example is: a change of format that provokes waves of DP demand to update DOs of older formats. To keep up with the same analogy to robotic rescue, we use the image of a "catastrophe" or a "tsunami" of a wave of format changes that weakens the life of DOs.

**Table 1. Comparison of the Requirements of Several Agent Applications and DP (adapted from [16])**

| Characteristic | Digital Preservation | Robo-Rescue | Robo-Soccer | Chess | Deep-Space Probe |
|---|---|---|---|---|---|
| *Number of agents* | **> 100,000** | > 1,000 | 11+ | 1 | <10 |
| *Homogeneity of agents* | **Heterogeneous** | Heterogeneous | Homogeneous | Homogeneous | Heterogeneous |
| *Control* | **Hierarchical /Distributed** | Hierarchical /Distributed | Distributed | Central | Central |
| *Similarity to reality* | **Total** | High | High | Low | Total |
| *Situation* | **Diverse** | Diverse | Simple | Simple | Diverse |
| *Actions* | **Varied** | Varied | Simple | Simple | Varied |
| *Information gathered* | **Diverse** | Diverse | Simple | Simple | Simple |
| *Representation* | **Hybrid** | Hybrid | Non-symbolic | Symbolic | Hybrid |
| *Emerging collaboration* | **Important** | Important | Early Stages | No | Average |
| *Real-time* | **months-years** | sec-min | milliseconds | min-hour | min-sec |
| *Inter-agent communication* | **Not perfect** | Very bad | Quite good | Total | Very good |
| *Resources* | **Highly heterogeneous** | Highly heterogeneous | Heterogeneous | Homogeneous | Subheterogeneous |
| *Logistics* | **Important** | Important | Irrelevant | Irrelevant | Impossible |
| *Short-term planning* | **Important** | Important | Important | + important | + important |
| *Long-term planning* | **Highly Important** | Important | Not important | + important | Important |
| *Scenarios* | **Complex** | Complex | Formations | Openings | Modes |
| *Hostility* | **Environment** | Environment | Opponent | Opponent | Faults |

Intelligent agents interact with alien cognitive agents, for example, human beings, with the need for monitoring and constant optimization of scarce resources, to solve the following:

- Scalability: An exponential growth in the number of digital objects with DP needs;
- Specialized knowledge on DP split among different institutions and users; and
- In many cases, results on DP cease to be verifiable after several years.

These problems, together with the rapid obsolescence of software and hardware because of frequent update of private vendors, make DP one of the most challenging application areas for MAS.

Additionally, agents have their own problems. Payne [19] claimed that agents are criticized as representing technology that is actively pursued in research labs but that rarely appears in deployed applications. In fact, many of the underlying technologies of intelligent agents have migrated into mainstream applications, where they are no longer referred to as "agents". Many research groups will revisit the evolution and application of intelligent agents and consider how they are shaping emergent technologies or becoming embedded within applications. Much of the research on MAS has provided formal proofs or proof-of-concept demonstrations (such as example systems or prototypes). It has provided only limited pragmatic support (systems, software, and tools) for the user community.

The problem of agents comes from the Artificial Intelligence (AI) field itself, which has come in and out of vogue many times in the past: It has been hyped and then, having failed to live up to the hype, discredited until being revived again. AI's (and agents') biggest enemy may be the promises made by its proponents—entrepreneurs looking for venture capital or academics who underestimate the challenge of meeting the needs of business users. Outside academia, AI successes in search and language technology, robotics, and the new "Web 3.0" applications are starting to enter industry in an exciting way [8]. This is a time of change in computing. As [7] claimed "New ideas such as cloud computing, social-networking sites are replacing search engines and news sites as people's favored homepages, and a new generation of applications are centering on large groups of people collaborating over an increasingly distributed network". Many of the entrenched models of AI will have to be re-thought as computers change from application devices to **social machines**. A potential phase transition in the nature of computing threatens to disrupt the entire computing field, including AI.

We think that, intelligent agents, whose social properties can automate social interactions and emulate ecosystems behavior, will have a lot to say in the future. Yet, where are all the agents? Hendler wrote in his blog in the late 1990s. He claimed that in the early 2000s many researchers believed that they were at a time where the large-scale deployment of agent-based computing was right around the corner, and many international research funding focused on making this deployment happen. Several magazines had hugely popular special issues on agents, and academic conferences on agent-based computing were popular. But in 2011, looking at what is hot on the Web and in IT development, many scientists wonder: Where are all the agents? And we wonder how can they be applied to DP?

This paper will show three approaches to an answer to those questions. It is structured as follows: section 2 is devoted to the introduction of the agentification of several DP actors; section 3 contains the agentification of digital objects, and the architecture of the SPADO (self-preservation aware digital objects); section 4 is the agentification of the preservation web services, and section 5 the agentification of the DP social networks. Section 6 will expose conclusions and future research towards a PhD.

## 2. AGENTIFICATION

We use the definition of agents as a *design metaphor*. The Agentlink Roadmap stated in 2005 [9] as a summarizing definition of agents, after two previous roadmaps dealing with more concrete definitions. Thus, we will design agents that suit the needs of DP, by introducing *agency* properties to the DP actors (objects, services, and people). This is called the *agentification*.

Agentification could be thought of as the encapsulation of agents inside existing systems, such as web services [19], robots [22], and objects. We will take other approach, encapsulating existing systems inside agent-oriented organizations.

In MAS design, traditionally there are two alternative design methodologies, called "top-down" and "bottom-up". In the top-down methodology, the design starts from the top with the assumption that resources are globally accessible by each subcomponent of the system, as in the centralized case. The specification is then defined in terms of the global system state and assumes that each individual component should be able to retrieve or estimate, with sufficient accuracy and within a reasonable time delay, resources that are local to other agents of the system. On the other hand, in the bottom-up methodology, the rules of agent interactions are typically designed in an *ad hoc* manner, although recent work has attempted to formalize the design process for some applications [15] by means of electronic institutions or virtual organizations. In systems designed starting from the bottom, the global state of all the components is assumed to be difficult to obtain, and the desired collective behavior is said to emerge from interactions among individual agents and between the agents and the environment.

Our approach of agentification is bottom up, about detecting which existing systems or entities need agent features, and then let the engineer decide what and how they are encapsulated. Agentification, then, is all about *what agent properties are worth*. Layers of soft and hard agency properties are added, as many as required by the application. They are added to a DO, to a DP service, or to the user. We propose three approaches for agentification:

a) The DOs to be preserved;

b) The DP resources: the Services;

c) The collective cognitive networks on DP: the Users;

For their design, a number of agent-oriented methodologies can be applied though there is as yet no robust agent-oriented language: MESSAGE-Ingenias, GAIA, PROMETEUS, TROPOS, MaSE, ADELFE, AMELIE, and even the electronic institutions might work, to name but a few. This paper focus on the emergent behavior of agentified objects, resources and users, but not on the design of a platform to implement them. This is future work.

These approaches answer three questions derived from Berman statements: **WHEN** (the need of preserving a DO and whether it is affordable) **is necessary to preserve,** and **HOW** (the solutions of the Users) to do **WHAT** (the DP Services) **is necessary to be done.**

The agents detected by the three agentification approaches might coexist: thus, the name to AOUS (Agentification of Objects, Users, and Services) comes up. Once we have agentified objects, users and services, here follows how much agency is needed for the DP application.

In this paper, the first approach analyzed is the agentification of DO, second the automation of DP social networks, and thirdly, the agentification of the DP services.

## 3. AGENTIFICATION OF DIGITAL OBJECTS

This agentification is supposed to have similar scalability results as the agentification of DP resources that we are going to see in section 5, and interesting resilience and DP cost management. However, the concept is radically different: digital objects themselves are agentified. This has little similarity to any approach in the literature on the information-ecological approach to digital libraries.

Synergic works on computational ecologies [23] [24] show how this approach might work, while agents ask themselves how much preservation is necessary (appraise) and, according to a sort of DP budget that would be regularly assigned to the DOs, compete with each other for the services to be preserved. Agents 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.

**New concepts** introduced by this approach are:

- **SPADO – the Self-Preservation Aware Digital Object**. It is responsible for its preservation, with its own budget for attracting know-how from users and services from tools such as format migration, metadata extraction or renewed storage. A SPADO has a complex structure of components and takes responsibility for being accessible, usable, and authentic at all times. It encapsulates copies of its older components in previous formats or devices so that it can reverse preservation paths that proved inadequate in the long term.

- **Preservation paradigm with three actors**. This is a paradigm shift from a user-centric approach, where the user performs the roles of "caring", "paying for", and "curating" the DOs, to an object-centric approach where the object has the role of "caring" for itself, the users "pay" for its preservation and provide know-how for "curate" it, and the DP services compete to "preserve" it. The new role assignment will lead to more balanced preservation decisions on how, when, and what to preserve than the current paradigm.

- **Object-level preservation budget**. The SPADO handles its budget to ensure the best-in-breed DP services and to remain accessible and usable for the longest possible time. The budget is assigned by users in a simple appraisal: the more interest in this DO, the more budget it will receive and the more likely it will be preserved. This detailed budget assignment reduces the complexity of the big institutional budgets that need to be allocated on yearly basis. SPADO can supplement their budgets if they are useful for wide

audiences and thus increase their preservation chances. **SPADOs compete with each other for their preservation**. Research should be done on adequate DP investment and how SPADOs attract know-how from users to maximise their chance of remaining accessible. For materials that are not amenable to market provision and are at risk of loss —such as certain types of reports, maps, emails, Web-based materials, and digital orphans— public provision is necessary. We will explore how market channels can be used as efficient means of allocating resources for preserving many types of digital content, and the conditions when this is the case must be investigated for inclusion in the preservation ecosystems.

These concepts are in this research brought together to interact in an environment that will be adapted to real business needs recreated in simulation environments, where the emergent behaviour of the competitive and cooperative interaction of the resulting environment with the three actors needs to be studied for a full understanding of its possibilities, as there will be billions of objects competing for thousands of services and millions of users in real life application of SPADOs.


**Figure 1. Conceptual structure of a SPADO: it has multiple components that are preserved with [format] redundancy internally in the SPADO, which has some rules that determine its mission, and social skills that determine how it interacts, cooperates and competes for user know-how and the best-in-breed DP services under a constrained budget.**

## 3.1 Buckets

Buckets [14] were aggregative, intelligent, WWW-accessible digital objects that were optimized for publishing in Digital Libraries (DLs), that existed within the "Smart Objects, Dumb Archives (SODA) Digital Library model" of [11]. 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 SPADOs will do.

Before 2000, there were a number of projects that had similar aggregation goals to SPADOs and buckets, namely the Warwick Framework containers [9] and the following Flexible and Extensible Digital Object Repository Architecture (FEDORA) [2]. Interaction with FEDORA objects occurs through a Common Object Request Broker Architecture (CORBA) interface. Multivalent documents [26] appear similar to buckets at first glance. However, the focus of multivalent documents is more on expressing and managing the relationships of differing "semantic layers" of a document, including language translations, derived metadata, annotations, etc. The AURORA architecture defines a framework for using container technology to encapsulate content, metadata and usage [12], also developed in CORBA. Some of the mentioned projects are from the DP community, and others are from e-commerce and computational science. Most did not have the SODA-inspired motivation of freeing the information object from the control of a single server. The mobility and independence of buckets or SPADOs are not seen in other DP projects. Most DP projects that focused on intelligence or agency were focused on aids to the DL user or creator; the intelligence is machine-to-human based. In contrast to the state of the art, the SPADO paradigm is unique because the information object itself is intelligent, providing machine-to-machine (or, SPADO-to-SPADO and SPADO-to-Agentified Services) intelligence.

Buckets satisfy the autonomy condition for being considered agents or SPADOs, since buckets perform many computational tasks that are influenced by their individual preferences, though they only weakly satisfy the negotiation condition, since only a handful of transactions have actual negotiation. An example of such a transaction is the case when a bucket requests metadata conversion; there is a negotiation phase where the requesting bucket and the conversion server negotiate the availability of metadata formats. Another difference between SPADOs and buckets is that SPADOs are empowered by an intelligent environment made of agentified DP services and automated social networks, while with buckets the objects become "smarter" at the expense of the archiving services (which become "dumber"), as functionalities generally associated with archives are moved into the data objects themselves. Instead, in our research, an intelligent environment boosts the preservation success of SPADOs. Finally, buckets have not been deployed, even though they were developed 10+ years ago in the USA, presumably due to a lack of the proper computational environment to develop their emergent capabilities and due to not handling the DP cost. As [14] states: "The truly significant applications of that [at the moment] new breed of information objects remain undiscovered". We forecast that the same fate of buckets will not happen to SPADOs, as in our research we treat the creation of DP environments where SPADOs interact with agentified DP services and personal agents that work on behalf of people as the proper way to get people involved in such a self-organized paradigm, conveniently exemplified by complex digital objects from hospital and police records.

As agents have not solved key problems yet in a way differently enough from existing mature less sophisticated technologies. Let DP give the chance to MAS? We will see how they could solve important problems of DP, with the initial requirement of being a scalable solution to the exponentially growing demand for DP, while offering the following:

- Freedom for content producers to create and publish content in a preservation-compatible manner
- Provision of digital repositories with tools for further automation of the preservation processes
- Seamless interoperation between content providers, repositories and end-users
- A shift of focus in DP from repository and preservation management systems to preservation-friendly DOs.

## 3.2 Preliminary Results

Two evolutionary computing approaches are taken, one from swarm intelligence similarly to the *shout and act* algorithm that is introduced in section 4, and another one from genetic algorithms.

In the first approach, the objects have descendents that split the preservation budget that they can use for their operations and descendents. Descendents might have a same or different format, regarding the site they would be operating. The operations like checksum, migration, version, and so on, charge to the budget, while operations like being accessed by users increase it. When any descendant run out of budget, it tries to go back to its ancestors site to get further DP budget and go on with its operations.

The proof of concept is implemented in Repast-Symphony, where DOs wander over a network of sites that emulate a file system of a social network of users. The simulation consists of 15 years of digital changes, with 3 catastrophes, one every 5 years, so that a significant proportion of DOs "die" at the year 15. That coincides with the fact that after 15 years, there are growing difficulties in accessing the digital assets.

The measure is the entropy of Shannon to know whether there is enough diversity of formats that provide the sufficient resilience to recover back to the former state after each catastrophe. Catastrophes consist in a sudden change of a ¼ of 1/3 of the sites, because of an update in their software that provoke massive changes and migrations in the format of the DOs that are there stored. Being resilient means the capacity of gaining back the lost entropy.

Every 15 years is an epoch, so 5 or more epochs are executed to get the average. Preliminary results show that in average there is a 15% resilience measured as the percentage of recovered entropy from the former state. This is true that after every catastrophe there is still loss of digital assets measured as progressively reduced entropy through the years. As Figure 2 shows, in average of every 5 epochs, the red line, there is about a 20% of resilience.

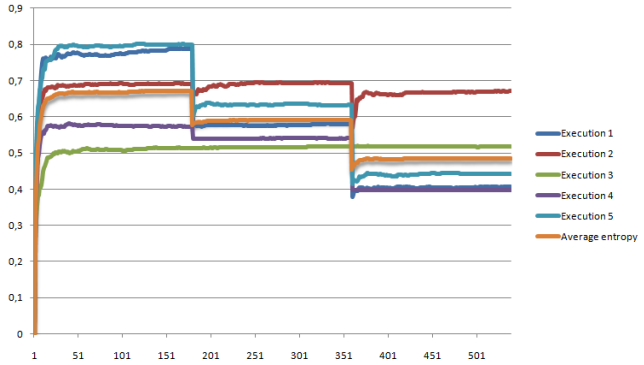


**Figure 2: Resilience of the swarm algorithm for SPADO**

A second approach is a genetic algorithm where the DOs genetic code is characterized by their formats. The DOs mutate (format migration) and cross each other (encapsulation of digital images into word, powerpoint or movies). Instead of entropy, a fitness function is used to measure how adapted is the population of DOs to the dominant conditions in a number of sites. For example, and just for illustration purposes, word 2003 was important in 2003-2006, while word 2007 spread from 2007 on, provoking a gradual decline of the objects that were (and still are) in word 2003 format.

The result of the proof of concept, facing the same type of 5 years catastrophe, is an interesting resilience of the whole ecosystem of DO as shown in the following picture:

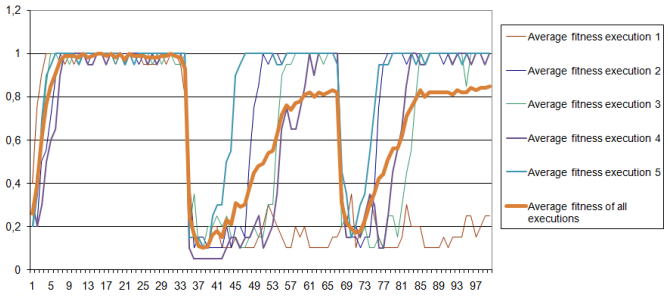


**Figure 3. Resilience of the genetic algorithm for SPADOs. The x-axis reflect the number of generations, 100 generations for 15 years of evolution and a catastrophe every 33 generations**

These two soft computing proofs of concept give us a hint of how to design the SPADOs that need to interact with the other SPADOs, and with those of the agentified DP services to obtain the DP resources to format migration, checksum, and other necessary operations for maximizing their chances of being preserved in a future.

Now, we need to discuss the other two types of agents that will cooperate with SPADOs.

# 4. AGENTIFICATION OF DP SOCIAL NETWORKS

DP relies on curation organization, technology, and resources. Organization implies people having broad knowledge on digital preservation; technology includes personnel with deep knowledge in digital preservation. This means that preservation knowledge resides in different levels of an organization. Moreover, any curator organization may be specializing in some part of the digital preservation process or have in-depth knowledge in a particular area of digital preservation. This means that the bodies (silos) of knowledge reside in different places.

Since DP has its roots in traditional archiving practice, there is a separation between the information producers, curators and consumers. However, IT is blurring these borders making the preservation process troublesome in terms of who and where should perform the actual active preservation activities.

At present the DP realm consists of scattered and fragmented "islands" of knowledge. A new quality level can only be achieved if these knowledge islands are connected one another in an environment that supports and facilitates active knowledge exchange. New and novel approaches to both general and specific DP problems can render this knowledge growth possible and bring about the necessary synergy effect. Furthermore, the proliferation of repository software (including: eprints.org, DSpace, Fedora, Greenstone, bepress, DAITSS and many others) has so far not delivered actual preservation solutions and organizations still need to decide how to select an appropriate repository option by considering the capabilities and limitations of each and the extent to which the repository software meets archival requirements and suits the digital content to be preserved.

Social networking and Web 2.0 are symptoms that information management is exploding and becoming ubiquitous. There is growing freedom, ever-increasing specialization of work, cheaper

information storage, and public availability of information through the Internet. All these facts are changing the information world, and this ever-growing universe of information requires proper management. It is no longer a problem of organizations, but of individuals. Many services sell information as an added-value service, and this might apply to DP as a solution to make skilled experts available, ready to share their knowledge with the vast majority of lay people, so far unaware of the need for DP so as to be very ignorant of the subject. It is a way to make bridges among the DP islands of knowledge.

An old vision of a profession of information pathfinders [20] becomes relevant today. Most of the work that bloggers—and in general, Web sites—do on the Internet is, in fact, connecting people with other resources and people. This is, in the end, a reference function. Wisely, libraries are trying to integrate and assimilate this "social networking" world. The changing reference world will produce big surprises and a permanent flow of innovation and new information agents outside the current network of information professionals, which is also a very promising and enriching trend, especially from our approach if agents were doing the job.

Thus, the use of social networks for DP seems not only plausible but even necessary, as DP is a new issue that is receptive to many alternatives and opinions, what we have called DP "recipes". The recipes might work fine with some users but not so well with others, so it is important to connect properly those users with similar needs who might benefit from the same recipes, or simply might better understand or apply the recipes. The fact is that today's strategies for DP are labor-intensive and often require specialist skills. To meet the demand, it is necessary to increase the automation and self-reliance of preservation solutions. However, users are the people who can best understand the needs of other users. The goal, again, is to match people and let them share. Thus, a network of people provides multiple links to solutions to others' needs, *crowdsourcing* the DP solutions [27]. We propose a further level of automation toward a fourth generation of search engines (Figure 3). This is because, still, interaction in social networks is human-made, requiring further levels of automation. From this level comes the need to assign at least one agent to every contributor or consumer, every user in the social network, to automate a certain amount of knowledge exchange.

Our vision is to introduce social networking into DP, which collectively maps users' needs to solutions, and to use agents to enhance social networks. No particular ontology is strictly needed (though we recommend having one) because in a massive social network, there will always be people who will understand the DP needs of others, and will provide appropriate links to solutions. The links will be part of the heuristics that map needs to solutions, procedures and techniques that satisfy users' needs. These heuristics shape the unique point of view that every single user can contribute to the social network of DP. We expect that new links will be added to solutions for anticipated needs. The anticipated solutions should be acknowledged or rewarded [13]. These links, coming from newly created NS (Need-Solution) pages, could be called "referrals". Unlike links, which are confined to Web pages, *referrals* are addressed to people and have context. The context of a referral is a user's guess of the preconditions under which this referral could apply. In a sense, the referral is like a pre-conditional link to people and Web pages, and could be implemented as a link with semantic tags as preconditions. Because of the diversity of possible ontologies, the preconditions might make sense only in a referring Web page but not in the destination Web page or resource.

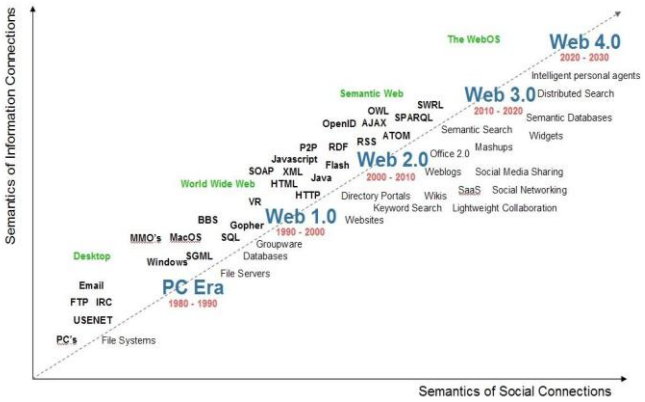


**Figure 4. Semantic map illustrating the ongoing dynamics in relation to technology, applications and architectures (Source: [18])**

In a narrower sense, DP questions and answers (QA) provide ways to describe how needs are defined, how people understand them, and how questions are answered [6] [17]. Thus, our aim is to expand social networks through the use of agents that reduce the burden of answering repetitive questions, motivated by the complex casuistic required by SPADOs. Agents as well as people should link data, agents, and people to find answers. Agents should encapsulate such linking information as well as content, and they should avoid spamming. Thus, DP knowledge will be accessible to crawlers and other agents by explicit queries. Moreover, knowledge itself will be proactively seeking use, unlike current approaches where information either waits to be accessed by links or is poured onto blogs by people. People, in the same way in which they create web pages, blogs, chats, emails and link knowledge, will develop agents to be their personal assistants that will contain their web pages and blogs, and that will be authorized to answer on their behalf and be rewarded on their behalf as appropriate [13]. The concept of *growing* an agent means to add the tools, information, skills, and autonomy for the agent to work on behalf of one user. Users will do the initial processing of information, selection, classification, and tagging, to convert the information into an understandable format for agents. The first stages of developing agents will consist of making agents answer on the users' behalf, then making them contribute on their behalf. Privacy issues should also be taken into account. This is our aim in this approach, to provide tools to help people developing agents and let them create a network of agents that handle DP QAs for SPADOS. Former results are published in [4].

## 5. AGENTIFICATION OF DP SERVICES

According to the W3C Web Services Architecture note [21], a Web Service is an abstract notion that can be implemented by a concrete agent. The agent is the concrete piece of software or hardware that sends and receives messages, while the service is the resource characterized by the abstract functionality that is provided. In this agentification approach, software agents are not used for service communication front-ends or as proxies; rather, they are treated as basic entities that encapsulate Web services. And the definition of an agent by W3C in the Web Services Architecture context is a specialization of the definition in the

architecture of the Web: “an agent is a program acting on behalf of a person or an organization”.

Shen [25] envisions a Web-service-based environment as a collection of economically motivated agent-based Web services. Thus, in this approach, DP web services are agentified so that they look for DOs to be preserved, a try to attract their DP budget.

Our approach [4] named "Shout and Act", a type of swarm intelligence for communication and coordination of agents is inspired by rescue robots: the files, all DOs, that need preservation are called the “victims”. The teams of preservation agents comprise agents of type A, whose main goal is to detect files as potential victims that need migration actions. They spend most of their time exploring the file system looking for victims. Whenever they find one, they call for help, called a *shout*. Their appraisal methods are unsophisticated and their preservation skills very limited, consisting of a very few common image formats (for example, from JPEG to TIFF). On the other hand, agents of type B are fewer and slower in detecting victim than the type A agents, though they have superior abilities to detect, appraise and rescue victims. They follow the shouts that type A agents emit. The shouts are of a magnitude that could be proportional to the severity of the digital injuries of the victim, announced in a blackboard situated in *n* higher levels in the file system tree. Shouts disappear time after being emitted, and disperse with distance in a metric created from file systems.

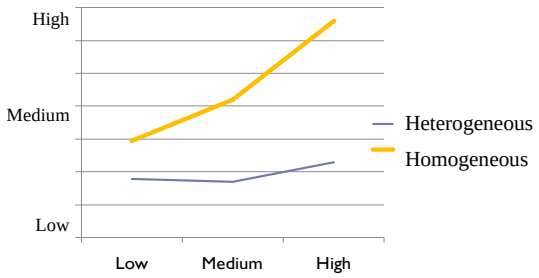


**Figure 5. Performance of homogeneous vs. heterogeneous agents vs. an exponential growth of the number of digital objects. Y-axis is the qualitative average processing effort per agent and x-axis is a qualitative order magnitude of the number of digital objects (high = 10 times medium = 100 times low)**

The result is a number of agents that search a user's file system, a site, for DOs. The agents cooperate with each other by exchanging pieces of information, for example, about the locations of DOs that might need preservation assistance or about what type of assistance they might need. Diverse agents cope with the exponential growth in the needs of DP, as the number of DOs grows exponentially while the computational cost of this agent approach remains flatter. This is shown in Figure 5, where well-designed teams of heterogeneous agents scaled well with the exponentially-growing demand of DOs (x-axis). The implication of this result is that DP web services should be designed in a way such that they do not offer identical services, but rather, should have different properties and capabilities, and should be designed for sociability, calling for other Web services. The DP services then should be designed atomically because having services with different methods of appraisal and migration, to name but a few, is better than having them all in an integrated, powerful, large preservation Web Service integrated. As well, they should be of different costs, for being affordable to all DP budgets of the SPADOs.

# 6. CONCLUSIONS

DP should be taken seriously as a killer application of agents. If all the three agentification approaches were combined, there will result the smart DP environment necessary to support the SPADOs activities, all agents interacting with several objectives and missions. The contribution of the approaches to the problems of DP is described in Table 2 regarding scalability requirements, their degree of openness in future open DP environments, their expected contribution to reduced DP costs while maximizing preservation chances, and an estimate of whether this approach could be already in use as early as 2015.

We have shown the different agentification approaches of DOs, users, and services. This work tells where all the agents are, what they are expected to do, and it settles their level of agency at the DP users’ expectations. Some of the approaches will scale well with the exponentially increasing amount of DOs to be preserved (we have the proof of concept of the agentification of DP services, Figure 5), while others are better at increasing DP awareness and knowledge by taking advantage of newly created social networks on DP (we have the prototype of the agentification of users in DP social networks). The design of DP services inherits a number of requirements from the agents' approaches that enhance their performance, only achievable as a truly distributed system through MAS.

In the future, a bottom up architecture of agents that are agentified DP Web Services will cooperate and compete to preserve agentified digital objects (SPADOs). The objects (SPADOs) will not be passive, but instead, proactive in searching and recommending the best agentified DP web Services. Some infomediaries (those agents in the automated social networks) will assist the other types of agents by providing them with the knowledge they receive from expert DP users. The proofs of concept in this paper give preliminary results that show that resilience under tight DP budgets and scalability are achievable.

**Table 2. Comparison of the AOUS approaches of DP**

| Comparison of the AOUS appr. | **Objects** | **Users** | **Services** |
|---|---|---|---|
| **Scalability** | Expected to be good | Improve the social networks | Good |
| **Resilience** | Good | ? | Expected to be Good |
| **Optimization of the DP budget** | Proved | ? | ? |
| **Openness Improves Digital Preservation** | Very Good | Good | Still a challenge |
| **Awareness** | Good | Very Good | ? |
| **Synergy with** | ? | Web 3.0 | Antivirus and backup services |

# 7. ACKNOWLEDGMENTS

Our thanks to the Spanish MCINN (*Ministerio de Ciencia e Innovación*) project TIN2010-17903 Comparative approaches to the implementation of intelligent agents in digital preservation from a perspective of the automation of social networks, and the AGAUR *grup de recerca consolidat* CSI-ref.2009SGR-1202.

## 8. REFERENCES


[1] Gantz, J. F. The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011: International Data Corporation (IDC), 2008

[2] Daniel, R. & Lagoze, C. 1997. Distributed active relationships in the Warwick framework. Proceedings of the second IEEE metadata workshop, Silver Spring, MD, pp: 16-17

[3] del Acebo, E. and de la Rosa, J.L., Bar Systems, 2006. A Class of Optimization Algorithms for Reactive Multi-Agent Systems in Real Time Environments, 17th European Conf. on AI. (ECAI2006) Intl Workshop on New Trends in Real Time AI, pp: 128-133, Riva de Garda, Italy, Aug 28-29, 2006.

[4] de la Rosa, J. Ll., Trias, A., Aciar, S., del Acebo, E., and Quisbert, H. 2009. Shout and Act: an Algorithm for Digital Objects Preservation Inspired from Rescue Robots, in Proceedings of the 1st Intl Work. on Innovation in Digital Preservation, Austin, Texas, June 19, 2009

[5] 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

[6] Gosain, S. 2007. Mobilizing software expertise in personal knowledge exchanges, The Journal of Strategic Information Systems, Volume 16, Issue 3, September 2007, pp: 254-277

[7] Hendler, J. 2007. Where Are All the Intelligent Agents?, IEEE Magazine Int. Syst., Vol. 22 (3), pp: 2-3, 2007

[8] Hendler J. 2009. Web 3.0 emerging, IEEE Computer 42 (1) (January 2009) pp.111-113

[9] Lagoze, C., Lynch C. A., & Daniel, R. 1996. The Warwick framework: a container architecture for aggregating sets of metadata. Cornell University Computer Science Technical Report TR-96-1593, 1996, Available at http://ncstrl.cs.cornell.edu/Dienst/UI/1.0/Display/ncstrl.cornell/TR96-1593

[10] Luck, M., McBurney, P., Shehory, O., Willmott, S. 2005, Agent Technology Roadmap: A Roadmap for Agent Based Computing, AgentLink III

[11] 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. Available at http://www.dlib.org/dlib/march99/maly/03maly.html

[12] Marazakis, M., Papadakis, D. & Papadakis, S. A. 1998. A framework for the encapsulation of value-added services in digital objects. In C. Nikolaou & C. Stephanidis (eds.) Research and advanced technology for digital libraries, second European conference, ECDL '98, pp. 75-94, 1998. Berlin: Springer.

[13] Moreno, A., de la Rosa, J. L., Szymanski B. K., and J. M. Bárcenas, J.M. 2009. Reward System for Completing FAQs, ISSN: 0922-6389, Artificial Intelligence Research and Development, Vol: 202, pp: 361-370, Nov 2009, IOS Press.

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

[15] Robles, A., Noriega, P., Luck, M., and Cantú J.J., 2006. Using MAS Technologies for Intelligent Organizations: A Report of Bottom-Up Results, A. Gelbukh and C.A. Reyes-Garcia (Eds.): MICAI 2006, LNAI 4293, pp. 1116–1127, 2006. Springer-Verlag

[16] Tadokoro, S., Takahashi, T., Takahashi, H., et al. 2000. The RoboCup-Rescue Project: A Robotic Approach to the Disaster Mitigation Problem. ICRA 2000

[17] Trias, A., de la Rosa, J. L., Galitsky B., and Dobrocsi G., 2010. Automation of social networks with QA agents, 9th Intl. Conf. AAMAS 2010, Toronto.

[18] Van Oranje, 2008. The future of the Internet Economy: a discussion paper on critical issues. http://www.future-internet.eu/fileadmin/documents/netherlands/Netherlans_Future_Internet.pdf , 2008

[19] Payne, T. 2008., Web Services from an Agent Perspective, IEEE Intelligent Systems, vol. 23, no. 2, pp. 12-14, March/April, 2008, ISSN 1541-1672

[20] Bush, V., Public Health Rep. 1953. February; 68(2): 149–152

[21] Web Services Architecture note, W3C, Feb. 11, 2004, www.w3.org/TR/ws-arch

[22] MacWorth, A. 2009. Agents, bodies, constraints, dynamics and evolution, AI Magazine, 30(1). Association for the Advancement of Artificial Intelligence, Spring 2009, ISSN 0738-4602, Accessible at http://www.aaai.org/Library/President/Mackworth.pdf

[23] 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

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

[25] W. Shen et al., 2007. Robotics and Computer-Integrated Manufacturing 23 (2007) 315–325

[26] Phelps, T. A. and Wilensky, R. 2000. Multivalent documents. Communications of the ACM, 43(6), 83-90, 2000

[27] Howe, J. 2008. Crowdsourcing: Why the power of the crowd is driving the future of business, Crown Business, 2008, ISBN: 978-0-307-39620-4

[28] 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

[29] Berman, F. 2008. "Got Data? A Guide to Data Preservation in the Information Age." Communications of the ACM 51(12):50-56.

[30] Quisbert, H, 2008. On Long-term Digital Information Preservation Systems – a Framework and Characteristics for Development. Ph D Thesis. Luleå University of Technology.