A multi-agent approach for autonomous digital preservation

A MULTI-AGENT APPROACH FOR AUTONOMOUS DIGITAL PRESERVATION

Jacopo Pellegrino[1]*, Marco Maggiora*[1] *and Walter Allasia*[2]

[1]Department of Physics, University of Torino, via Giuria, 1, Torino, I-10025, Italy
[2]R& D Department, EURIX, via Carcano, 26, Torino, I-10153, Italy

ABSTRACT

Keeping the content of digital objects accessible is a key challenge for digital archives, especially those with limited resources dedicated to preservation. This paper describes an agent-based model that simulates processes in which a digital object become obsolescent, thus a preservation action, such as the migration to the most appropriate format, is required. Agents monitor and control the local environment and deal with preservation issues individuating obsolescent formats based on global parameters such as their diffusion. They exchange information to find out the most suitable preservation action: suggestions are evaluated and propagated according to their reciprocal level of trust. The level of trust is automatically updated after every interaction through a feedback mechanism based on stigmergy. The paper shows that the framework has a stable behaviour under several use cases. Therefore this approach is suitable for digital archives that may take advantage of a multi-agent system which can either perform an autonomous preservation action or suggest a list of best candidate solutions to the user.

Index Terms— MAS, ABM, Digital Preservation, Autonomous Actions.

1. INTRODUCTION

The digital obsolescence is caused by the ongoing development of new software and new formats, so the risk of obsolescence can be estimated from a global environment. Due to this considerations, digital preservation has to be addressed as a collective and distributed concept[1]. In the agent model being presented two main strategies to cope with digital obsolescence are taken into account: migration and emulation[2]. Migration consists in converting digital objects into a new and more recent format and is the process we are going to focus on in this work. The model will also let the user emulate the installation of an appropriate application under particular conditions. This paper shows a model carrying out autonomous assessment of migration processes by means of a multi-agent system.

1.1. Agent-based Models (ABM) and Multi-agent Systems (MAS)

An agent can be defined as a computer system situated in some environment, and that is capable of autonomous actions in order to meet its design objectives[3][4]. The environment and all the agents within constitute a multi-agent system. In such a system the agents interact with each other and with the environment in order to share information and reach their goals. This interaction may cause a dependence relation between agents which could be unilateral, mutual or reciprocal if an agent depends on another one with respect to different objectives. In order to meet their objectives agents then have to communicate and cooperate to find agreements about which action to perform on the environment. The key idea of this methodology is that a collective and intelligent behaviour emerges from simple design rules inside the agents[5]. An agent-based model could be appropriate to implement the distributed intelligence needed to deal with digital preservation issues. The agents will acquire, evaluate and share a certain set of information in order to understand how serious the risk of obsolescence is and which is the best preservation action to perform according to their internal state.

1.2. State of the Art

One of the first efforts to contrast digital obsolescence has been made by the Research Libraries Group[6] and the Commission of Preservation and Access. They formed the Task Force on Archiving of Digital Information that published a key document about digital preservation: "Preserving digital information. Report of the task force on archiving of digital information"[7]. Based on this report, a reference model named OAIS (Open Archival Information System)[8] was developed. It discusses the concept of long-term digital preservation and aims to point out the various stages of the life cycle of a digital object and of the related preservation process[9]. It is a theoretical reference model for the organization of both conventional and digital archives. The OAIS model has been a guideline for another document: "Trusted Digital Repositories: Attributes and Responsibilities"[6]. In this, the fundamental concept of Trusted Digital Repository (TDR) is defined. As we shall see the concept of trust between archive entities will be a fundamental aspect in the presented model.

It is possible to provide some examples of obsolescence identification and metadata extraction tools such as AONS (Automated Obsolescence Notification System)[10] and AONS II[11]. These systems are capable of analysing digital repositories and identifying objects in danger of becoming obsolescent[10][11]. They take advantage of information about the format which is one of the most relevant metadata. The information is recovered by means of DROID (Digital Record and Object Identification)[12] and JHOVE (JSTOR/Harvard Object Validation Environment)[13], a couple of tools for the format extraction and validation.

The next step in a preservation process is to define a preservation plan. A four-year project named PLANETS (Preservation and Long-term Access through Networked Services)[14] started with the aim to help preservation planning. As explained in [15], Planets involves several preservation functions such as: preservation planning, characterization, preservation action and an interoperability framework.

Another interesting tool is Scout[16] developed within the SCAPE Project[17]. It is a web-based service that helps the users in identifying preservation issues and managing digital repositories. Scout takes advantage of a sort of interaction between users by considering the number of organizations using the same preservation platform.

To the best of our knowledge, none of the tools discussed so far take advantage of the ABM and MAS architecture that is described in the following sections.

2. MODEL DESIGN AND IMPLEMENTATION

2.1. Model Structure

The model aims to emulate a distributed environment where many archive entities share information about their internal state in order to find solutions to their digital preservation issues. In our implementation each archiver is an agent named *institution*, which could be both public, in order to represent large archivers such as libraries, broadcasters, universities or government entities, or private in order to embody small personal archives. These agent species contain other species inside named *pastor*s that manage the digital objects and a *software manager* as regards the applications. The number of micro-species does not represent a constraint since the model can be easily adapted to handle an arbitrary number of digital object categories. The institutions communicate with each other and with their pastors and software manager in order to deal with preservation issues. The pastors and the software manager have the knowledge about the formats and their relation, for instance which format provides the minimum loss of information rather then the maximum compression.

This architecture is one amongst many possible others, it could be possible to add a sort of expert agent which could suggest migration strategies to all the institutions. This choice would not comply to the key concept of multi-agent systems where the intelligence emerges from the interaction of agents and does not reside into a unique entity.

With the aim to be as close as possible to reality, each institution chooses to adopt a particular operating system. This choice guides the possibility to install only a certain list of applications. As far as the digital objects that are managed by the pastors are concerned, they are embodied by agents of the species named *format collection*. We decided to gather all the objects of a given format, which is the key metadata in this model, in order to limit the cpu load and memory requirements and hence to keep the model easily runnable even on a laptop. Implementing each object as a single agent could be another possibility since the model is scalable but it would require much more computational resources. Each collection is named as the format and keeps information about the number of objects, their distribution (there could be either single files or clusters of files) and a hypothetical total size.

The schematic in Fig.1 shows the structure of an *Institution* agent.

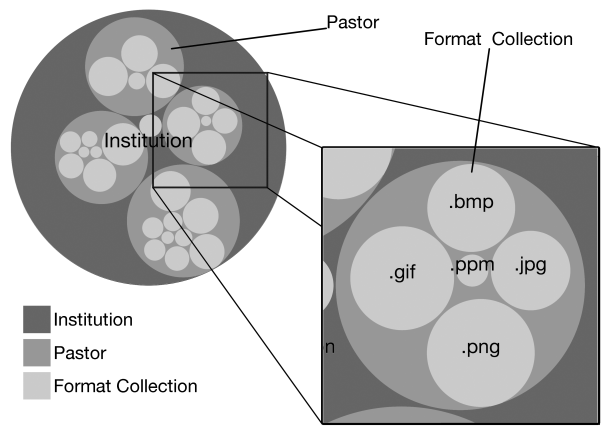


Fig. 1: Architecture of *Institution* agents. The size of the circle representing the *format collection* indicates the numerosity of the related format.

2.2. Model Assessment

The pastor agents check whether they could render one of their format collections using at least two of the installed applications. If this operation can be performed then the format should not be considered to be under risk of obsolescence and no further actions are performed. In the opposite case the pastor informs the institution about the rendering failure. When there is only one application left, which is capable of rendering and eventually migrating the object to another format, the alert process is triggered. In case the trigger number of applications was set to zero the environment would not have been able to handle the obsolete format any more.

The institution receives the warning and estimates how high the risk is for that format to become obsolescent. In this implementation we decided to evaluate the global level of obsolescence of a format by means of three components: the first one is the total number of objects of that format that are contained by all the institutions at the time, this indicates how commonly adopted the format is. The second component is the number of institutions containing at least one object of that format and the third one is the number of installed applications capable of rendering the object.

When the risk level exceeds a defined threshold a preservation actions has to be performed: the institution asks the others about their issues regarding the format under consideration. Each of the other institutions search its own history of actions for any action performed involving the format and, if it has been effective, suggests it as possible solution. In this implementation, the institution that asked for a suggestion chooses the best candidate solution based on three parameters: the level of trust associated with the proposer, the number of files that the other institution has migrated and the number of own files that should be migrated. This two last parameters indicate respectively how significant is the actions performed by the other institution and how potentially dangerous is the action to be performed. As an example, in case the imminent migration regards a huge number of objects, the institution may require the level of trust to exceed a higher threshold than the usual one.

We assume that the level of trust between two institutions I_i and I_j could be calculated as follows:

$$T_{ij} = \frac{1}{A} \cdot \sum_{a=1}^{A} w_a T_{ij}^a \quad (1)$$

Where A is the number of components, T_{ij}^a is each component and w_a is the corresponding weight included in the matrix variable containing the weights of each component. We decided to take into account four component that seemed to be significant indexes of trust: the first one is related to the number of components contained by the institutions, the second and third are the geographical and "cultural" distance (different language or alphabet) and the last is the digital preservation expertise in terms of staff dedicated to this aspect.

This key variable is therefore distributed and it is updated after every interaction, as explained in the next lines.

2.3. Model Update

The model is self-updated according to the preservation actions performed by the distributed agents. We have adopted a reinforcement learning mechanism that rewards those agents that have suggested an effective solution to the requester. It is important to underline that the institution classifies the action as effective with respect to its local environment.

The institution has the responsibility to verify whether this action has been effective or not. If so, then the preservation objectives of the two institutions are similar (in this specific implementation they both needed to migrate the same format) and that the same action has been effective for them. In this case the level of trust between these two institutions is increased. Obviously, in the opposite case it is decreased. Only the trust component related to the object type migrated is updated: in this way the same institution may result trustworthy as regards, for example, the audio files but could be not reliable for text files. Thanks to this feedback mechanism, after every interaction, the institution agents learn which among the others should be trusted more for a given type of digital object.

2.4. Development and Simulation Environment

GAMA (Gis & Agent-based Modeling Architecture) version 1.6[18], [19] is the tool employed for the development of this model. It is a modeling and simulation development environment for building spatially explicit agent-based simulations. It has been developed by several research teams under the umbrella of the IRD/UPMC International Research Unit UMMISCO[18]. GAMA takes advantage of the Gis & Agent-based Modeling Language (GAML)[18]. In general terms a GAML model is made up of a certain number of actions which consist of a sequence of statements.

Each model is made up of three main parts: the first one is named *global* and it includes all those variables accessible to every agent. The second part, named *entities*, contains the declarations of all the species of agents that take part in the model. The last part, named *experiment* is dedicated to the experimental setup.

3. TESTING THE FRAMEWORK

3.1. Experiments

Several models with different sets of initial conditions can be tested in order to evaluate the behaviour of the framework. In particular, the following features have been investigated: first of all the stability of the communication process which is the backbone of the model. The frequency of migrations (defined as migrations over time) performed by the agents has been monitored. The linearity of the asymptotic frequency with respect to the probability for an agent to encounter a new format at every time step has also been monitored both with no time constraint on the migration time and with a temporal dependence. We tested the stability of the feedback mechanism by monitoring the frequency of variation of the level of trust between agents and then we considered again the frequency of migrations in case that each migration required a finite number of time step to be performed. The other features analysed are the trend of the number of migration in progress and the evaluation of the agents' decisions according to the global decision trend. As far as this last feature is concerned, the main idea is that a preservation action can be classified as a good action only if it is performed by the most part of the institutions within the network. If an agent performs a migration that is not globally approved, it may waste the chance to perform a more useful action (in this case the action is classified as false positive). In case an agent decides not to perform a migration that was considered useful it may have lost the possibility to solve its preservation issue (the action is classified as false negative). If the preservation action is considered neither good nor useless it is classified as indifferent action.

3.2. Statistical Analysis of the Results

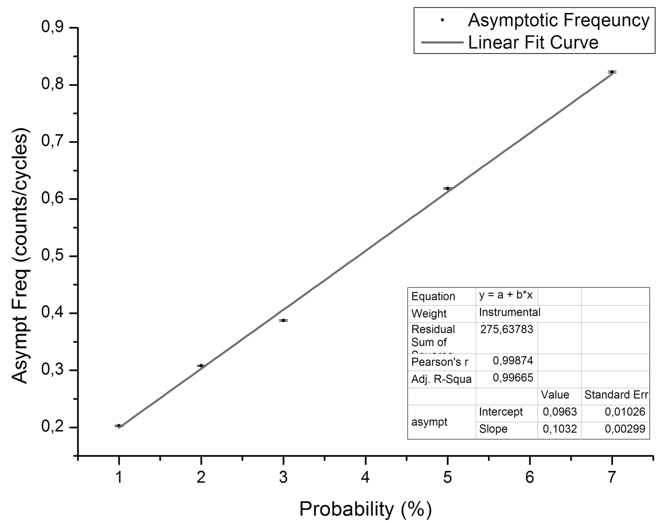
The stability of the communication process, certainly is a key requirement for the framework. We demonstrated with twenty different use cases[20] that the frequency of migrations, without any time constraints, faces an exponential decay until an asymptotic value is approached. This means that the information exchange among agents occurs properly. The trend is confirmed by the very low reduced chi square values obtained from the fit with an exponential function reported in the following expression where the square root of the time may indicate the binary nature of the agents' interaction.

$$f(t) = a \cdot \exp^{-b \cdot \sqrt{t}} + c \qquad (2)$$

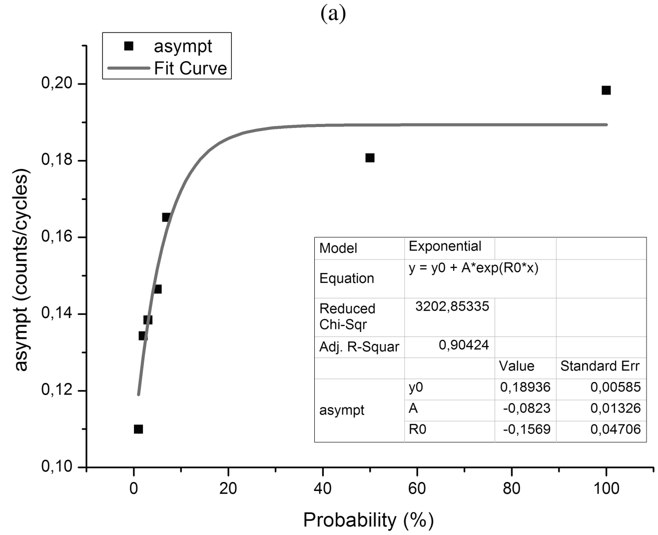
Concerning the linearity of the asymptotic frequency value, which is the parameter c of expression 2, has been proved for low probability values. We refer to "probability" as the chance for an agent to encounter a new format or delete one of its own format collection at each time step. Both when migration time is an issue and when there are no constraint we observed a linear trend followed by a saturation effect as shown in Fig. 2.

Once that the communication process has been proved to be reliable, the feedback mechanism was investigated. We analysed the frequency of variations of the trust weights matrix and observed that the trend of the frequency of both positive and negative variations faces a slow linear decrease when the system is into a stable condition after a certain number of time steps. Moreover we observed a significant difference between the value of positive and negative variation frequency. This result indicates that, even tough they were educated guesses, our assumptions were effective enough to make the system capable of identifying the sources of wrong suggestions and discard them.

Due to its relevance, a temporal dependence has hence been introduced. We decided to test the behaviour of the framework when a certain number of cycles was associated with each migration. Moreover, each institution has the capability of refusing migrations in case the number of cycles required is considered too high. The analysis of the frequency of migrations shows that its trend faces an exponential decay with increasing time as happened in the stability experiments, as depicted in Fig. 3a.



(a)



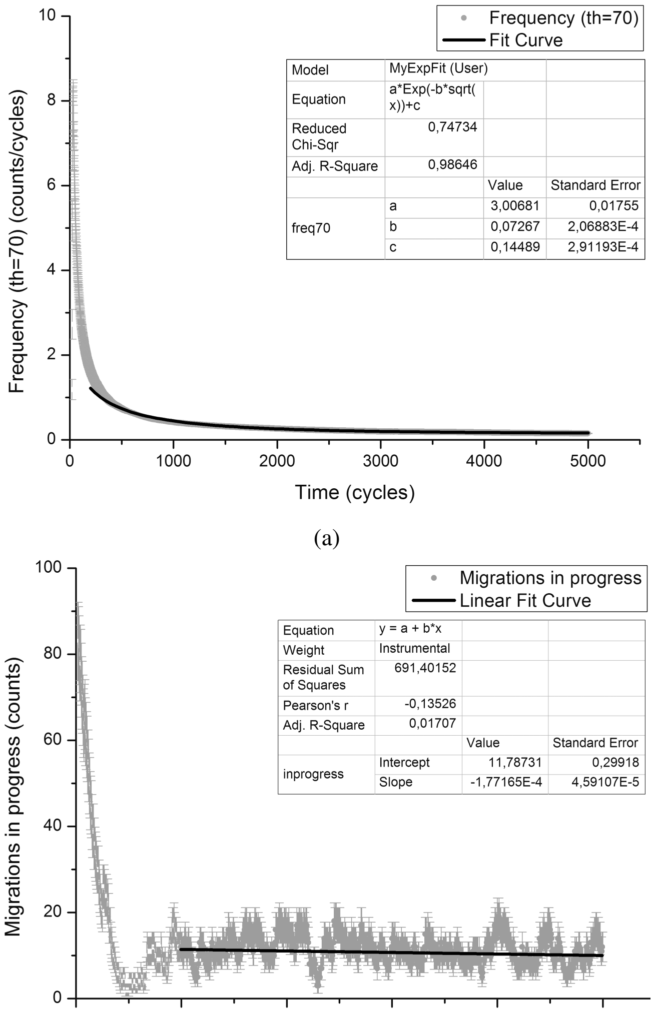
(b)

Fig. 2: Low probability values (1-7%) linearity Fig. 2a, Saturation effect for high values (50 and 100%) Fig. 2b.

This is a significant result: it means that users are allowed to adopt specific temporal units to simulate the duration of migrations, without affecting the behaviour of the framework.

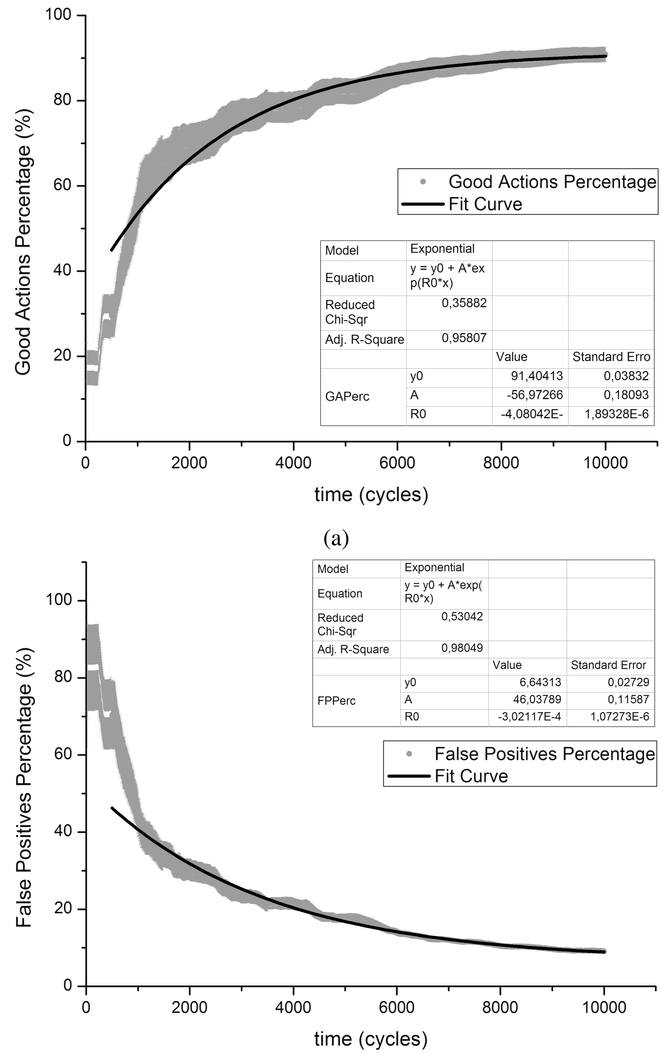
With the same use case considered where the focus was the trend of migration frequency, we also monitored the number of those migrations that were being performed by the institution at time. In particular we observed that this value oscillates due to the duration of the migrations. These oscillations, by the way, occur around a constant value which is further confirmation of the stable behaviour of the framework as shown in Fig. 3b.

The last but most promising results concern the evaluation of the agents' decisions. This experiment shows how at first, agents perform several unnecessary migrations, thus



(b)

Fig. 3: Exponential decay and stability of the frequency of migrations Fig. 3, constant trend of the number of migrations in progress Fig. 3b.



(b)

Fig. 4: Exponential increase of good actions Fig. 4a, exponential decrease of false positives Fig. 4b.

classified as false positives. With increasing time the information exchange mechanism takes place and agents start learning which is the most effective preservation action to perform. Moreover, the percentage of good actions and unnecessary migrations (false positives) is complementary at each time step. That means that either indifferent actions or missed migrations (false negatives) never happened.

The following Fig. 4a and Fig. 4b show how the number of good actions faces an exponential growth while the number of false positives decreases exponentially with increasing time.

4. CONCLUSIONS AND FUTURE WORK

The work presented in this paper provides a novel approach to the decision processes concerning one of the most common digital preservation issues such as the migration process. The agents are endowed with the capability of communicating, cooperating and propagating information about the performed actions in order to help each other in finding the best solution to a given preservation issue with respect to their internal state. The goal is to provide a framework that can be used as a flexible test bed in which the user is able to simulate several dynamic and distributed digital preservation scenarios and to probe different approaches in defining the trust rules for the network. Moreover, the stability of the framework has been proved under various use cases.

By means of the several tests performed, a significant dependency from the initial conditions has been observed which allows the users to evaluate the effect of different initial rules on the effect on the environment. Every set of initial conditions can be introduced by the user in order to verify how they affect the stability of the evolution of such a complex system.

Several models with specific use cases have been tested demonstrating how institutions could benefit from an interaction as the one presented in this work. It is possible to think of the design and the implementation of a software application based on the discussed framework. Such an application will be capable of either performing autonomous preservation actions or helping the user in taking decisions about the best preservation strategy to adopt. The communication process will take place through a real network thus allowing the digital archives to share information and knowledge. The decisional processes and also the preservation actions would be no longer simulated, but executed on actual digital objects for real.

5. ACKNOWLEDGEMENTS

This work was partially funded by the European Commission in the context of the FP7 ICT project ForgetIT (under grant no: 600826).

6. REFERENCES

[1] Fabrizio Falchi Francesco Gallo Walter Allasia and Carlo Meghini, "Autonomic preservation of "access copies" of digital contents,," in *Proceedings of The Memory of the World in the Digital Age: Digitalization and Preservation*, 2012.

[2] Henry M. Gladney, *Preserving Digital Information*, Springer, 2007.

[3] Michael Wooldridge, *An introduction to Multiagent Systems*, John Wiley & Sons Inc, second edition, 2009.

[4] J. M. Epstein, *Generative social science: Studies in agent-based computational modeling.*, Princeton University Press, 2006.

[5] P Terna, "The epidemic of innovation–playing around with an agent-based model.," *Economics of Innovation and New Technology*, vol. 18(7), pp. 707–728, 2009.

[6] Robin Dale Marianne Doerr Margaret Hedstrom Maggie Jones Anne Kenney Catherine Lupovici Kelly Russell Colin Webb et al Neil Beagrie, Meg Bellinger, "Trusted digital repositories: Attributes and responsibilities,," Tech. Rep., Research Libraries Group & Online Computer Library Center, 2002.

[7] Donald Waters and John Garrett, "Preserving digital information," Tech. Rep., Task Force on Archiving of Digital Information, 1996.

[8] Alex Ball, "Briefing paper–the oais reference model," 2006.

[9] Lothar Schmitz Uwe M Borghoff, Peter Ro dig and Jan Scheffczyk, *Long-term preservation of digital documents*, Springer, 2006.

[10] Joseph Curtis, "Aons system documentation," Tech. Rep., Australian National University, 2006.

[11] David Pearson, "Aons ii: continuing the trend towards preservation software nirvana," Tech. Rep., Web Archiving and Digital Preservation Branch, National Library of Australia, 2006.

[12] "Droid (digital record object identification)," 2006, `http://www.nationalarchives.gov.uk/information-management/projects-and-work/droid.htm`, Online; accessed 18-February-2015.

[13] "Jhove (jstor/harvard object validation environment)," 2009, `http://jhove.sourceforge.net/`, Online; accessed 18-February-2015.

[14] "Planets (preservation and long-term access through networked services)," 2006, `http://www.planets-project.eu/`, Online; accessed 18-February-2015.

[15] Adam Farquhar and Helen Hockx-Yu, "Planets: Integrated services for digital preservation," *Serials: the Journal for the Serials Community*, vol. 21, no. 2, pp. 140–145, 2008.

[16] ," 2013.

[17] ," .

[18] "gama-platform," 2013, `https://code.google.com/p/gama-platform/`, Online; accessed 25-June-2014.

[19] Andrew Crooks, "Gis and agent-based modelling," 2014, `http://www.gisagents.org/2013/03/gama-gis-agent-based-modelling.html`, Online; accessed 25-June-2014.

[20] Jacopo Pellegrino, "A multi-agent based digital preservation model," *CoRR*, vol. abs/1408.6126, 2014.