Simulation Modelling Practice and Theory 19 (2011) 110–120

Contents lists available at ScienceDirect

Simulation Modelling Practice and Theory

journal homepage: www.elsevier.com/locate/simpat

Social models in open learning object repositories: A simulation approach for sustainable collections

Salvador Sánchez-Alonso, Miguel-Angel Sicilia *, Elena García-Barriocanal, Carmen Pagés-Arévalo, Leonardo Lezcano

Computer Science Department, University of Alcalá, Polytechnic Building, Ctra. Barcelona, km. 33.6, 28871 Alcalá de Henares, Madrid, Spain

ARTICLE INFO

Article history:
Available online 13 July 2010

Keywords:
Learning object repositories
Social networks
Social filtering
Simulation
RePast

ABSTRACT

Learning object repositories (LOR) are digital collections of educational resources and/or metadata aimed at facilitating reuse of materials worldwide. In open repositories, resources are made available at no cost, representing a case of information sharing with an implicit and diffuse social context. In such settings, quality control is in many cases based in some form of community filtering that provides a reliable basis for ranking resources when repositories reach a critical mass of users. However, there have been numerous repository initiatives and projects and many of them did not reached a significant degree of actual usage and growth that made them sustainable in the long term. In consequence, finding models for sustainable collections is a key issue in repository research, and the main problem behind that is understanding the evolution of successful repositories. This in turn requires analyzing experimental models of the behavior of their users that are coherent with the available evidence on their structure and growth patters. This paper provides a partial model for such behavior based on existing reported evidence and on the examination of patterns in a large and mature repository. Agent-based simulation was chosen to allow for contrasting configurations with different parameters. Simulations were devised with the RePast framework and the resulting model implementation constitutes an initial baseline for future studies aimed at contrasting empirical data on repository usage with their community setting. The model described accounts for known user contribution patterns and it is coherent with the implicit social network structure found in an existing large LOR.

1. Introduction

The widespread use of the Web has resulted in a proliferation of systems (often called digital collections or digital repositories) that store resources and/or metadata for some given purpose or domain. For example, MERLOT collects metadata about educational resources on the Web, while Europeana[1] focuses on museums, archives and libraries. The sustainability of open digital collections is a major concern for policy makers and research and innovation funding agencies, as many repositories funded by national or international R&D programmes disappear soon after project end or become outdated as they are not able to gather contributions regularly beyond the initial deployment.

Sustainability in digital collections requires devising an effective socio-technical infrastructure (comprising technology, human factors and communication design) that we are still far from understanding in all of its aspects. A recent report of

* Corresponding author. Tel.: +34 918856640; fax: +34 918856646.
E-mail address: msicilia@uah.es (M.-A. Sicilia).

[1] http://www.europeana.eu/.

1569-190X/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.simpat.2010.06.022

the Australian Partnership for Sustainable Repositories[2] Bradley [4] defined sustainable repositories as "an infrastructure, both social and technical, which is economically viable for maintaining valuable data without significant loss or degradation". It is important to note that this definition is two-sided. On the one hand, it emphasizes economical viability but on the other hand, it is implicitly referring to a second aspect: sustained quality of use or utility. While the sustainability of digital collections has often been discussed in terms of its economic sustainability, having and continually making available valuable resources surpasses financial concerns and relates also to sustained update and evolution. Sustained use is in turn dependant upon the configuration of the social system built around the technical infrastructure [3]. This latter one is the aspect of sustainability that will be addressed in this paper, complementing both financial concerns and purely technical scalability issues (relating to hardware-software infrastructure) that have been discussed elsewhere.

There are a number of proposed models for the long-term economical viability of repositories of open digital contents, that have been tested and refined especially for open access (OA) to scholarly documents [31] and open educational resources (OER) [9]. Financial models for OA are diverse, included those based on author fees [26] but also others that rely solely on the institutional support of a University or non-profit society. Irrespective of concrete financial models, it is widely acknowledged that sustainability can be improved most effectively through a shift from a provider/user paradigm to a community model of collaborative development [9]. Indeed, the idea of the "Web 2.0" emphasizes social aspects, considering services and activities that foster a new kind of media consumer who is more engaged and active in creating and adding value to the content which is the basis for using the Internet [1]. The idea of user-generated content (UGC) is the key element of the new approaches to the social contribution of contents in the Web, which has been found to follow "long-tail" distributions and which is marked with a strong "participation inequality" [22]. "Participation inequality", also known as the 90-9-1 rule (meaning that 90% of the users do not contribute, 9% contribute a few elements and that 1% contribute many) has been considered a typical pattern to measure UGC production.

Learning object repositories (LOR) provide a platform for the sharing of digital educational resources, and currently most of them provide some mechanisms for building community dynamics around their resource base. The community dimension and its social dynamics have been found to be an important aspect for the success of these repositories. For example, Brosnan [5] provided a conceptualization for that importance based on social capital theory, and Monge et al. [19] analyzed the potential impact of Web 2.0 strategies to foster social dynamics and participation in repositories. In a similar direction, Han et al. [14] reported an empirical study on the LON-CAPA repository in which a non-explicit community model was identified on the basis of co-contribution of resources to the same courses by popular authors. Even though our understanding of community dynamics in learning object repositories is still in an inception phase, some scattered evidence and examples can be found to start modeling their evolution.

Understanding the dynamics of successful LOR is key to the realization of the promises of quality increase and economies of scale that are the basic elements of the learning object paradigm [8]. In consequence, finding models for sustainable collections is a key issue in repository research. As practical realizations had showed, the success of LOR is determined by its capacity to create a social space usually reflected in a community of users that register in the system and contribute actively to its development, either by providing resources or other kinds of meta-information as ratings, evaluations or tags. While empirical findings as those reported by Ochoa and Duval [23] and García-Barriocanal and Sicilia [12] provide a picture of the current configuration of LOR and give some directions on their growth dynamics, they do not provide models that can be used to understand how the behavior of users emerges into patterns that result in a sustained maintenance of the quality of the collections. Having such models will be helpful to test and devise community tools and for grasping an understanding on how different incentives to sharing impact the growth and quality of the repository. This situation calls for simulation approaches that depart from the existing scattered available evidence and allow for experimentation and contrast, incorporating new evidence in the model's parameters as it becomes available. Simulation allows experimenting with alternative models where empirical data is scarce or difficult to obtain, as it happens with current repositories.

This paper reports on the first of that simulation models, devised in an attempt to provide a baseline that could be contrasted in further research. As we are considering fundamentally social and community issues, agent-based simulation has been chosen as a paradigm matching the social structure of current LOR as MERLOT or Connexions, which can be considered dynamic systems consisting of a network of interacting actors and processes that adapt to a constantly changing collection content. Concretely, the *Recursive Porous Agent Simulation Toolkit* (`Repast`) has been used for the implementation of the model. `Repast` is an open source modeling framework that permits researchers to create agent-based simulations [21]. The main objective of the model presented is that of capturing user contribution patterns, which are the basis for sustained update. The approach has been that of devising the model after the few scattered research reports available on repository growth together with an analysis of data extracted from a large and mature LOR.

The model presented here focuses on review and rating mechanisms, which can be found in repositories as MERLOT,[3] Connexions[4] and eLERA[5] among others, and that embody the idea that quality control and prestige should be mediated by *social filtering*, following successful applications of models in electronic commerce [25] and information retrieval [24]. The model has been found to have links with incentive models of knowledge sharing inside organizations that have been subject to

[2] Funded by the Australian Department of Education, Science and Training (DEST).

[3] http://www.merlot.org/.

[4] http://cnx.org.

[5] http://www.elera.net/eLera.

experimentation elsewhere [29], but considers a social model that is mediated by the resources contributed by users, which become subject to evaluation following different modalities [27].

The contributions of the research presented can be described in two directions. First, a simulation model is devised to capture the main known aspects of on-line resource sharing inside repositories with community features. And second, the model is able to generate situations that share known metrics with real-world LOR and are able to demonstrate how variations in the initial parameters lead to different results, thus allowing for contrasting incentive and design policies in digital repositories. The focus of the model presented is on sustainability considering quality of resources, so that it abstracts out economical and IT issues and concentrates on the contribution and assessment behavior of users. This is to our knowledge the first model of sustainable repositories available, so that it would ideally be extended, corrected or improved by further research as our knowledge on repositories expands.

The rest of this paper is structured as follows. Section 2 describes related work, the base simulation model and its main characteristics, using the available evidence as the source for choosing parameters. Then, Section 3 details the implementation of that model in `RePast`. Section 4 reports on verification and experimentation with the implemented model and discusses implications. Finally, conclusions and outlook are provided in Section 5.

2. Base model and methods for social analysis

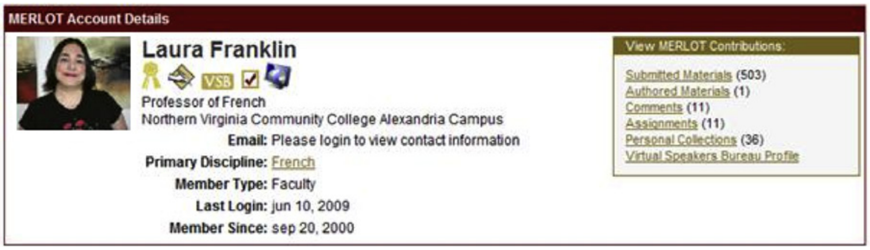
This section first provides an overview of previous research that is related to the approach used in this paper. It then details the foundations of the social model developed and the evidence from which it has been derived, and presents the main modeling elements considered.

2.1. Related work and departure studies

There are several research reports applying simulation to different aspects of social filtering. Notably, reputation has been subject to simulation studies in different domains. For example, Giardini et al. [13] used `RePast` for modeling the effects of transmission of social evaluations in an industrial cluster, and found that evaluations are crucial to selecting trustworthy partners. Social visibility has also been addressed by Malsch et al. [18] in the context of message referencing networks. In a different direction, Wu et al. [32] applied agent-based simulation implemented with `RePast` to study political incentives. While these studies show the viability of agent-based simulation for social sharing contexts, to the best of our knowledge there are no previous models published that address the specific context of digital repositories. Reputation or recognition in repositories is reflected in various forms as is discussed below, but is in all cases related to user contribution. Contributions in digital collections are diverse, and include the resources themselves (authorship) but also describing resources authored by others and shared openly. They also include reviewing, commenting or rating existing resources in several ways.

The point of departure for the simulation model presented in this paper are studies on the size, structure and growth dynamics of popular learning object metadata repositories [12,27,23], other kind of digital repositories [22] and as a complement, studies on community dynamics in LOR [14]. Reusing results described in Sicilia et al. [27], a database from the MERLOT repository [6] was gathered May 2009 by using a crawler that systematically traversed the Web pages of the repository, similar in functionality to the one reported by Biletskiy et al. [2], used to explore patterns that could inform the model. Information of a total of 69,248 users was extracted, of which 1393 were also recognized as resource authors.

Evidence found in these studies has been contrasted with the simulation model for knowledge sharing inside organizations described by Wang et al. [29]. In their model, Wang et al. [29] considered the following variables affected by organizational interventions: *identity transparency*, *benefits* and *costs*. These require a reformulation in our case, as in the open environment of the Web, individuals face different cost-benefit tradeoffs, and identity transparency (defined as "the degree to which organizations are successful in identifying employees who share knowledge or not") has a different consideration. Contributors in open LORs are typically known and contribution information is public, so that identity transparency is not a variable but an intrinsic characteristic. In fact, LORs tend to make individuals and their contributors prominently visible. For example, Fig. 1 shows how MERLOT provides a number of elements describing different aspect of recognition of member behavior. These include the "colored ribbons" quantifying different kind of contributions to MERLOT in the categories: gold,


MERLOT Account Details

Laura Franklin

VSB

Professor of French
Northern Virginia Community College Alexandria Campus

Email: Please login to view contact information
Primary Discipline: French
Member Type: Faculty
Last Login: jun 10, 2009
Member Since: sep 20, 2000

View MERLOT Contributions:
Submitted Materials (503)
Authored Materials (1)
Comments (11)
Assignments (11)
Personal Collections (36)
Virtual Speakers Bureau Profile

Fig. 1. A fragment of a member profile in MERLOT.

Table 1
Overall experimental configuration parameters.

Parameter	Rationale
`Number of initial users`	The "founders community" is assumed to be a small group of initiators. A value of 50 is chosen as an arbitrary small group size
`Early user growth rate`	Growth rate in users per time instant included
`Mature user growth rate`	Growth rate in users per time instant included (decreasing phase)
`Time user growth inflection point`	Inflection point between early and mature phases of user growth

silver, bronze and regular. Similar user profile exposure facilities are provided in the Connexions repository, e.g. through the "Featured content" section in the front page of the portal, that refer to works of particular users considered important.

Benefits can then be considered to be linked to the reputation gained in the portal, that are either recognized by the repository staff (as in MERLOT's ribbons) or implicit but visible in other mechanisms as personal collections or public reviews. These would eventually be linked to rewarding systems, e.g. Namuth et al. [20] mention rewards and evaluations (as in the case of faculty staff assessments) as important sustainability issues that could be implemented through measuring peer evaluations in LOR. The impact of purely altruistic behavior is not discernible from reputation seeking with the available data, so that both elements are not separated in modeling user behavior.

Costs in this case are related to the effort required to contribute materials. In the case of users contributing other's resources in repositories as MERLOT, the costs are low (as they only need to upload basic descriptive metadata, e.g. author, URL, language, primary audience, topical category) but benefits can also be considered reduced, as the author of the resource would be attributed the principal merits. However, the fact that there is a significantly higher number of contributors than authors is significant as it points out that MERLOT is more a community of contributors than of authors. This can be a result of MERLOT being a repository only storing metadata and not the contents themselves, as occurs in other systems as Connexions. In any case, an important departure assumption for our model is that effort to contribute has been found to be largely unrelated to contributor behavior [22] so that it has also been excluded from the model.

2.2. Base model

The main model for the repository will be based on the evolution in time (not necessarily represented as actual time) of the key elements comprising the model: users (of different kinds), resources (learning objects), and assessments (also of different kinds). An example of an assessment may be user *John Doe* that likes very much a learning object in the repository titled *Teaching English Drama* and provides a "five" rating (in a scale from one to five) plus a comment on why he liked the resource. Other elements as the quantity and quality of descriptive associated metadata are not considered here as there is no evidence that it has an impact on the social dynamics of the system.

Repository growth (in terms of number of resources) is modeled according to the comparison between the early and mature growth phases [23], considering as an initial configuration parameter that *Early* growth is lower than the *Mature* growth. It should be noted that this is not the case for all the repositories, so it makes sense to experiment the consequences of having early growth rate higher than the mature one. Also, we have expressed growth in terms of users and not resources because users are the central *agent* concept in our simulation model. Even though there are not studies regarding user growth rates, open repositories typically reach a critical mass at some point in time, and then user growth rate becomes slower. This has been represented by several parameters of the simulation model showed in Table 1. Parameter `number of initial users` represents the initial size of the community, often restricted to the consortium or development team of the repository. From that number, it is expected that the repository will grow in number of users first according to an `early user growth rate` (rates are specified in new users per time unit) and after a given time (`time user growth inflection point`[6]) it changes to a `mature user growth rate`. For example, early rate might be 0.15 and mature 0.4 (users per time unit), and inflection point may be 1100 units (say, for example, days of operation). These parameters account for the existing evidence on user community growth. It is expected that every repository eventually stops operation, but this is not interesting for our analysis that is based on a phase of sustained growth.

Evidence has been found that repositories grow linearly with varied patterns of contributor productivity and popularity [23], but we are still far from fully understanding the social aspects of these systems and the motivations and patterns of interaction of their users. For that reason, instead of modeling resource contribution as a *Information Production Process* (IPP),[7] contribution has been modeled in agent's behavior. This enables experimenting with different agent's parameters and observing its influence in the observed *a posteriori* production functions. Evidence shows that independently of the size of the system, UGC contribution is determined by power laws as Lotka's [22]. This entails that the probability of additional contributions is increased as the amount of previous contribution increases [17]. A similar inequality is modeled in evaluation contribution (rating and bookmarking as will be described in the next subsection) [27].

[6] Here "inflection" is used to refer to change in rate, and not used in the strict differential calculus sense.

[7] Egghe [10], the concept of Information Production Process (IPP). The objective of the IPP is to establish a quantitative relation between the producers and the information items being produced.

The above described elements of the model account for the main structural elements of the model: user growth, resource contribution and patterns of rating and bookmarking. The last important element to consider are the requirements for "sustainability as sustained quality of use". The following are the *requirements* considered in the sustainability model considered here:

(1) Repositories require a considerable user base growing at a stable rate in the maturity phase. New users are supposed to replace inactive ones in contributing tasks.
(2) Repositories should have an stable resource growth. Without that stable growth, in general the resource base becomes outdated and no value is added to users.
(3) Contributor productivity is unequally distributed, following a distribution close to a Lotka distribution. This is derived from existing evidence on the existence of a few very productive users and a majority of low activity ones.
(4) Contributor (and thus resource) quality is unequally distributed, following a distribution close to a Zipf distribution. We assume repositories in which there is not an *ex ante* strict quality control, so it can be expected that only a few resources are of high quality.
(5) Highly rated/bookmarked resources appear in the repository with an stable rate in time. Users will recognize high-quality materials progressively on time.

Requirements (1) and (2) are base departure assumptions without which the repository can be considered to stall and quickly approach obsolescence. However, requirement (3) is an empirical law observed in many Web systems, and requirement (4) a similar observation from MERLOT, which entails that there are high-quality resources differentiated from the rest of the collection. The fifth requirement is where the long-term quality approach enters into the model, as some repositories (typically, those built during a project of a duration of 3–5 years) may have all the above requirements except that quality resources were built or acquired at inception, but they are no longer evolving. This will lead to reduced usefulness of the repository in the long term, as high-quality resources may eventually become obsolete, resulting in a progressive degradation of the overall quality of the system.

The consideration of quality in our model is social, in the sense that it is realized by means of the assessments users do about the resources. The social model derived from the user–resource relation materialized in different kinds of evaluations is discussed in the following subsection.

2.3. The dynamics of social reputation

A source of empirical evidence about structural prestige [30] that can be found in some repositories is the availability of references in different forms. These references are often indirect, e.g. some community members store personal "favorite link" collections or provide comments and ratings about some particular resources. Connecting these resources with their authors (or contributors) provides a way to explore indicators of prestige or to seek for potential sub-networks or cluster models. That referencing bears some similarity with citation measures as have been developed extensively in bibliometrics, and they share also the intuition behind models for the ranking of Web pages as the popular PageRank [24], that are based on outgoing and incoming links between Web pages. However, the assumptions under which citation measures used in the context of scholarly literature have been developed need not be directly transferable to the domain of learning resources, as scientific findings and educational material are created, maintained and evaluated very differently. The same occurs with models as the PageRank, which were developed for the open, non-specific context of the Web, which is fairly different from the smaller and more focused context of a learning resource repository. More concretely, learning materials are organized around disciplinary domains as is the case of MERLOT. Further, more elaborated and rich forms of referencing can be found in reviews and ratings provided by community members in learning object repositories beyond those used in these other models, e.g. the peer reviews in MERLOT are evaluations referencing materials that consider specific educational dimensions (e.g. *content quality* or *effectiveness*).

Peer reviews (done only by domain experts) and general user ratings (done by every user of the system) in MERLOT exhibit positive but small correlations in general below 0.2, which suggests that they are capturing different kinds of quality assessment. However, peer reviews have not been included in the current model due to two principal reasons. The first one is that they depend more of the policy of the owners of the repository than of spontaneous social dynamics, which make them less relevant for the current study. The second reason is practical, as the amount of peer reviews available is much less than that of general ratings, so that it is more difficult to find and evaluate a simulation model for them.

Several repositories have started to provide services by which members can share their personal collections of favorite resources and comment or review other's resources. This information is in those sites openly available in some cases, and represents an objective account of usage and expression of preference, at least as much as links in the Web do. Even though it does not directly reflect personal ties between community members, it is a reliable empirical material to analyze the patterns of shared interest between peers, which represents a starting point to gain insights on the communities behind repositories. The social network model for the repository studied here was based on two different relations that account for the two abovementioned kinds of data: personal collections and reviews. As the base data for the two models is of a different nature, we take the assumption that it is representing two different kinds of referential ties. One kind of referential tie would be that of including a resource in oneself's personal collection (a kind of bookmark functionality provided as a feature of the repository), and the other kind includes a potentially critical review of the resource in the form of a review (including a text with the comment and a rating or evaluation, in a numeric scale). It should be noted that a review can be positive or negative, wether including a resource in a personal collection can be interpreted as that resource being useful or interesting for the user. In both cases, the set of actors $M = m_1, \ldots, m_n$ is defined as the user community of the repository considered. Also, the set of learning objects in the repository is denoted as *LO*.

2.3.1. Model based on personal collections

The first model is based in a directed, non-valued network built from gathering the personal collections compiled by repository users. A personal collection is simply a set of bookmarks to resources available in the repository that is openly exposed by a given repository user. The relation *P* of personal collection references represents ties in the form of ordered pairs $P\langle m_i, m_j\rangle$, with m_i and m_j being individuals (and at least $m_i \in M$), and coming from the following transformation:

$$P\langle m_i, m_j\rangle \rightarrow \exists PC\langle m_i, lo\rangle \wedge author(lo, m_j) \qquad (1)$$

where $author(lo,x)$ is true if x is (one of) the authors of the resource *lo* (a learning object), and $PC\langle m, lo\rangle$ is a bimodal, directed relation between *M* and *lo* representing the learning objects bookmarked by users inside the repository. This kind of indirect referencing has been proposed for modeling prestige in the Web, e.g. in the case of PeopleRank [11].

Authors are the original creators of the resources. However, as MERLOT does not store the resources themselves but only the metadata describing them, it is frequent that the person that initially described the resource is not (one of) its original authors. Then we make a difference from authors and contributors of the resources, being both sets overlapping only to some extent. All the contributors are users in *M*, however, some authors are not MERLOT users, as their resources have been described in MERLOT by others. If instead of $author(lo,x)$, we use the alternative $contributor(lo,x)$ a variant P' network is obtained. This alternative deserves exploration, as ratings and prestige could be hypothesized to be attached to contributors rather than to authors, since all contributors are MERLOT members and they take over the work of selecting useful resources from the Web.

2.3.2. Model based on reviews

The model based on learning object reviews can be represented as a valued graph in which values represent the ratings given to resources by users evaluating them. The relation *R* of review-based references represents ties in the form of valued ordered pairs $R\langle m_i, m_j\rangle$: v, coming from the following transformation:

$$R\langle m_i, m_j\rangle : v \rightarrow \exists RR\langle m_i, lo\rangle : r \wedge author(lo, m_j) \wedge v = \Phi(m_i, m_j) \qquad (2)$$

where r is the rating the user m_i provided as evaluation of the resource by m_j (it is assumed a user cannot rate a resource more than once). And relation *RR* represents the association between users and the learning objects they have reviewed. The aggregation operator $\Phi(x,y)$ has the purpose of combining the opinion of user x to resources authored by user y. In a simple formulation, it can be implemented as an average of the ratings, in which case v will be an average of several ratings r. The domain of v is determined by the rating scale given by the repository, and it might represent a simple rating (typically in a scale [1..5]) or a vector of ratings. This later case occurs in eLERA, in which the multi-item rating instrument LORI [28] is used for the evaluation of the objects. As in the previous case with personal collections, changing *author* to *contributor* in the expression results in a related but different network R'.

3. Model implementation

RePast is an extensible software framework for agent-based simulation created by Social Science Research Computing at the University of Chicago [7]. It provides an integrated library of classes for creating, running, displaying, and collecting data from an agent-based simulation. Repast has been used to develop simulations for contexts similar to our object of study. *Repast Symphony*[8] has been selected for the implementation since it provides built-in capabilities for dealing with, displaying and exporting social network data.

Contexts represent abstract populations of agents in RePast. In our implementation of the socio-technical model, the repository is considered a single context labeled `learning-object-repository`, and time is maintained in that context. Clock unit is the "tick", which is used to be the index of the order (in a concrete run of the model, ticks can be assimilated to actual time units as days and the parameters can be adjusted to model a concrete setting). A global behavior was made responsible for the addition of users according to the growth rate (early or mature, as described above), with a random variation to avoid a perfectly constant growth.

Physical space is not relevant to our simulation model, but the digital resources and the users (both contributors and consumers of resources) are 'located' in a disciplinary space, basically represented as a tree-like structure of disciplines and sub-disciplines (e.g. *Science and Technology/Biology/Genetics*). This is required to make the model flexible enough to support differentiated behaviors for users with different disciplinary backgrounds. Evidence supporting this impact is in that it is already known that there are significant differences in the statistical profiles of highly rated resources [12], and the

[8] http://repast.sourceforge.net/.

Table 2
Parameters of simulation agents.

Parameter	Explanation
`date registered`	Timestamp of the creation of the agent
`peer reviewer?`	Boolean attribute differentiating users that qualify as peer reviewers
`contributions count`	Count of contributions so far
`contribution propensity`	Accounts for individual differences in motivation to contribute

Table 3
Tasks associated to user agent behavior.

Parameter	Explanation
`perhapsContribute`	Decide on contributing according to Lotka's model
`perhapsComment`	Decide on commenting according to resource's visibility
`perhapsBookmark`	Decide on adding a item to personal collections according to resource's visibility

distribution of resources and evaluations are also significantly different when changing disciplines. However, the disciplinary space is not used in the experiments reported here as they would make the model tied to a particular topical structure.

Projections in `RePast` are used to impose an structure to the meta-population of proto-agents as defined in a Context. The rating relations P and R described above can be modeled through (directed) projections. Concretely, they are defined with the `Projection-Network` type of `RePast`.

Agents in our simulation are the users that contribute and evaluate resources. Table 2 summarizes the main attributes of users. As the network projections in our case involve users, resources and evaluations, we used a generic `LorElement` agent subsuming the three sets of elements so that the networks are defined on that class. However, `LorResource` and `LorEvaluation` (representing respectively the two latter categories) are actually not agents but passive configuration elements with no scheduled behaviors. The social network models described in Section 2 are built from the `RePast` network projections by reducing the graph to a unipartite model containing only instances of `LorUser`.

Table 3 summarizes the agent's behaviors and tasks. Users are modeled as infinite loops in which behaviors trigger according to the parameters described in the table. The contributions count is used to model the contribution behavior making more probable additional contribution with the increase in previous size of contribution history of the user (x). This has been implemented by a dynamic threshold on random numbers on a Lotka with exponential cutoff (Eq. (3)) with the parameters found by Ochoa and Duval [22], i.e. $\alpha = 1.87$ and $\lambda = 6 \cdot 10^{-4}$

$$p = x^{-\alpha} \cdot e^{-\lambda \cdot x} \tag{3}$$

Some modifications were require to account for differences in user's interest to contribute described below in the results section.

The `perhapsContribute` task combines the Lotka distribution described above with the `contribution propensity` attribute to model contribution behavior (details on adjustments required for this initial model are provided in the next section). Contrary to contribution, evaluation tasks (comment and bookmark) depend on the previous history of evaluations related to the resources already in the repository. The `perhapsComment` and `perhapsBookmark` tasks respond to the reviewing and bookmarking (personal collection addition) that conform respectively networks R' and P' as discussed above. The authoring-based variants R and P were discarded in the model as there is less empirical data available [27] and they represent a different kind of relation that surpasses the social community structure of the repository—in Sicilia et al. [27] it is reported that only a 2% of users are also resource authors in MERLOT.

The case of commenting and rating that results in R' was modeled excluding the consideration of ratings, as the distribution of ratings in R and R' (see Table 4) shows that most reviews are positive (i.e. ratings above three). In consequence, comments can be considered in general as supporting statements, with only a few acting as negative filters, following a similar univalent consideration as in backlink models [24].

P' features a distribution of indegrees (one of the major tools to analyze structural prestige) that appear to follow a kind of power law. Concretely, Fig. 2 shows the log–log plot of the indegree in the y-axis and the rank of each vertex in the x-axis. The points appear to fall along a single line segment, which is typical of Zipf law distributions (see Fig. 2), except for the higher rank values. This suggests that there is a high inequality in the indegree of contributors. This evidence is translated into the simulation model via `perhapsComment` and `perhapsBookmark` tasks, that select resources to rate and bookmark respectively implementing a propensity to reference resources that have already attracted previous ratings or bookmarks. The implementation of those behaviors includes a simple Zipf-like formula determining the probability of a resource to be evaluated according to x, the number of times it has been evaluated so far, i.e. $p = \frac{c}{x}$. In this case, there is no previous evidence to give a concrete value to constant c, so that it was heuristically adjusted to produce a coherent distribution matching the other parameters.

Table 4
Distribution of ratings in R and R'.

Rating	R freq.	R% freq.	R' freq.	R'% freq.
1	18	2.4	90	1.7
2	25	3.3	172	3.3
3	114	15.1	733	14.2
4	308	40.7	2204	42.6
5	292	38.5	1971	38.1

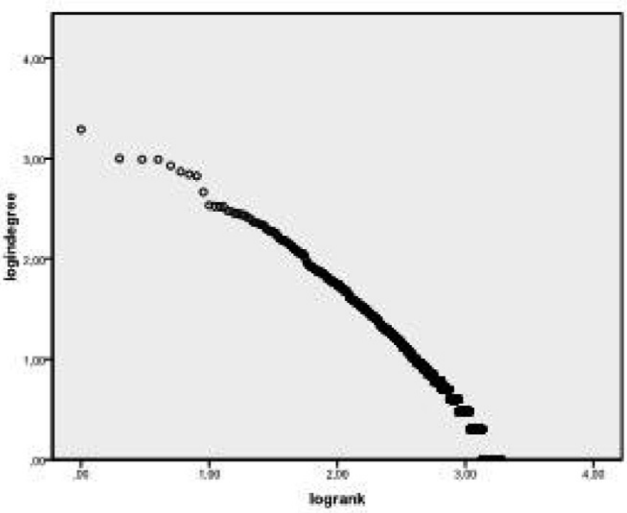


Fig. 2. Log–log plot of indegree and vertex rank in P'.

4. Experimentation results

4.1. Contrast and discussion

A simulation experiment was carried out to explore the long-term impact of the parameters of the model. Before the simulation was devised, a number of verification contrast were carried out to evaluate the coherence of the model implemented with the departure evidence and assumptions (described above). The first contrast was that of checking if the user contribution pattern generated was producing distributions similar to Lotka's. This revealed the need to randomize the `LorUser` agents actually deciding to contribute at each step, and also randomly making a small percentage of them more propense to contribute.

A key contrast carried out was that of the production or approximately constant resource growth rates. This required the inclusion of a small contribution propensity decrease in the constantly growing user base, which is coherent with the fact that there is a significant number of long-lasting inactive user accounts in repositories. With this additional adjustment, the model produces a stable resource growth rate consistent with the available evidence. Fig. 3 shows the histogram of a run of the simulation model in which it is showed that the growth rate is not increasing in time but varies with the modifications in propensities to contribute that compensate the effect of Lotka's based increase in propensity. The peak at the beginning corresponds to the initial launch of the portal, in which all the founders contribute without having the effect of propensity modification. A Kolmogorov–Smirnov test was performed to evaluate differences in 50 runs of the simulation model, showing no significant differences with $\alpha = 0.1$.

4.1.1. Evaluating the generation of social nets based on evaluations

The first main experimentation consisted on comparing the overall distribution of social networks R and P taken from actual MERLOT data and the corresponding ones generated by the simulation model. Comparing network structure is in general a challenging problem, but some general metrics can be used to evaluate a degree of similarity. From the available measures, we have focused in density and connectedness.

The evaluation consisted in the batch execution of 100 runs of the simulation under the same parameters. Table 5 shows the original parameters obtained from the crawling of MERLOT and the average of the same parameters for the simulation runs. None of the simulation measures was subject to a high variance. Densities can be considered similar in magnitude, and

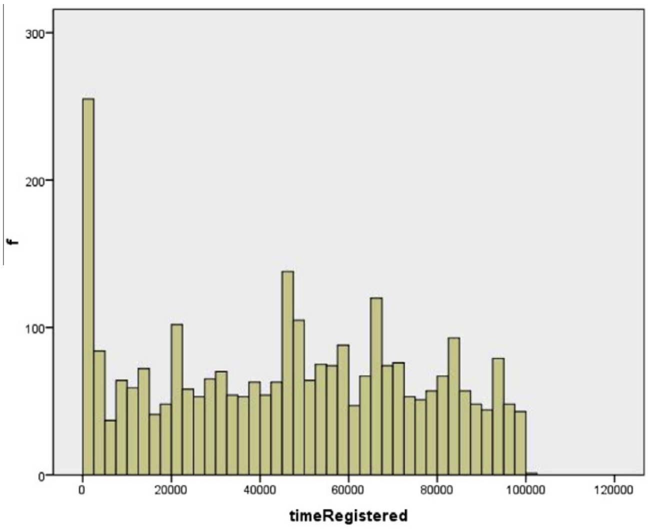


Fig. 3. Frequencies (*f*) of resource addition for time = 100,000 with a correction of 0.0001 in propensities to contribute for non-successful contribution cycles.

Table 5
Network parameters for the different relations considered.

Parameter	*R*′	*P*′	*R*′ (sim)	*P*′ (sim)
Density	0.00117	0.006	0.0021	0.0048
Number of strong components	12	1	6	2
Max. size of strong components	27	171	30	215
Number of weak components	5	2	6	8
Max. size of weak components	1786	3286	1532	4201

structurally, both the sample and the simulations show comparable values. Component sizes are higher in the case of simulations but still representing a similar proportion with respect to the size of the networks.

From the structural analysis it cannot be disregarded that the referral networks resulting from the model are different from those found in actual empirical data.

4.1.2. Experimenting with quality-based sustainability

A critical element in the sustainability model presented in the list of requirements discussed in Section 2.2 is that of the requirement on the continuous appearance of new high-quality resources. This has been considered problematic in large repositories due to a known effect on rating-based search systems. For example, Jansen and Spink [15] studied user logs in several query engines, concluding that users are viewing less result pages, in some cases no going beyond the first page of results. This phenomenon may lead to a design problem with the requirement of continuous appearance of high-quality resources. Concretely, it is possible that newly added high-quality resources are not made visible in querying and browsing at the LOR due to competition with older resources that have attracted a great deal of (positive) user evaluations.

The procedure for contrasting this was that of observing the pattern of temporal high-quality resource appearance. This was measured as the moment in which a resource accumulated a number of ratings above the average plus two standard deviations. The examinations showed a clear decreasing tendency of this kind of resources in time, which can be attributed to the ranking effect mentioned above, as there are no other reasons of degradation that could be hypothesized as alternate explanations. The way ratings and bookmarks were generated in the implementation of the simulation model produced a similar effect. This required a change in the `perhapsComment` and `perhapsBookmark` tasks to introduce a decrease in tendency to be evaluated with the passing of time, which gave more opportunities to newly created resources to gain a critical mass of evaluations and rank higher globally. In the case of real-world LOR, this kind of effect can be removed in several ways. These include transversal browsing mechanisms in which rank criterion is based on content similarity (e.g. topic or educational characteristics).

4.2. Limitations

The main limitation of the model presented lays in that its underlying assumptions are based on scattered empirical research reports, which may eventually be challenged in forthcoming studies. However, this does not invalidate the model as a first attempt to use simulation for gaining understanding on the sustainability of LOR. An additional limitation is that currently it does not account for content-related factors as the actual quality of the resources, but this element could not be included as the evaluations of the resources in LORs are based in the same social dynamics that are part of the model, so there is not an external criteria to use as parameter.

Also, there is a need to implement simulation models for alternatives to the evaluation mechanisms presented here that were based on MERLOT. For example, the concept of *lenses* in Connections [16] could bring insights and contrast as it represents an endorsement model, qualitatively different from the rating and peer review mechanisms used in the model presented here.

5. Conclusions and outlook

The sustainability of digital repositories is a critical aspect of digital resource preservation that requires a socio-technical approach for its understanding. Indeed, community and social reputation issues are known to play an important role in digital repositories. A model for sustainability based on growth and community activity (complementary to economical models) has been proposed in this paper, using available evidence about resource and user base growth and the distribution of evaluative activity. The model has been implemented using `RePast`, producing repository situations that follow the main assumptions in the model, and model parameters can be used to explore and experiment with different rates and user behavior. Even though the main supporting evidence for the model comes from studies regarding LOR, the model has the potential to be translated to other kinds of repositories in which user behavior can be assumed to be similar, as in scientific literature databases, with the necessary parameter changes. The consideration of the time pattern of appearance of high-quality resources has confirmed the concern for alternatives to browsing materials that allow newer materials to get accessed and eventually gather a high number of positive evaluations.

Further research should develop in the direction of enriching the model as more empirical studies complementing existing knowledge arise. There are a number of variables that are potentially affecting the model of digital repository presented and that has not been included so far due to the current unavailability of relevant studies that result in a lack of understanding of their impact. These include the possible competence between repositories covering similar domains (which is specially relevant in the case of generalistic repositories), which may affect incentives and patterns of usage. That multi-repository analysis would require a multi-context simulation.

Acknowledgements

The results presented in this project have been partially funded by the University of Alcalá and the CAM (Comunidad Autónoma de Madrid), as part of Project MARIA, code CCG08-UAH/TIC-4178, and by the Spanish Ministry of Education through Project MAPSEL, code TIN2009-14164-C04-01.

References

[1] M. Allen, Web 2.0: an argument against convergence, First Monday 13 (3) (2008).
[2] Y. Biletskiy, M. Wojcenovic, H. Baghi, Focused crawling for downloading learning objects – an architectural perspective, Interdisciplinary Journal of E – Learning and Learning Objects 5 (2009) 169–180.
[3] A.P. Bishop, N.A.V. House, B.P. Buttenfield, Introduction: digital libraries as sociotechnical systems, in: A.P. Bishop, N.A.V. House, B.P. Buttenfield (Eds.), Digital Library Use: Social Practice in Design and Evaluation, MIT Press, 2003, pp. 1–21.
[4] K. Bradley, Digital sustainability and digital repositories, in: Proceedings of the VALA 2006 Conference, 8–10 February 2006, Melbourne, Australia, 2006.
[5] K. Brosnan, Developing and sustaining a national learning-object sharing network: a social capital theory perspective, in: J.B. Williams, M.A. Goldberg (Eds.), Proceedings of The ASCILITE 2005 Conference, 2005, pp. 105–114.
[6] R. Cafolla, Project merlot: bringing peer review to web-based educational resources, in: Proceedings of the USA Society for Information Technology and Teacher Education International Conference, 2002, pp. 614–618.
[7] N. Collier, RePast: an extensible framework for agent simulation, 2000. <http://repast.sourceforge.net/>.
[8] S. Downes, Learning objects: resources for distance education worldwide, International Review of Research in Open and Distance Learning 2 (1) (2001).
[9] S. Downes, Models for sustainable open educational resources, Interdisciplinary Journal of Knowledge and Learning Objects 3 (2007) 29–44.
[10] L. Egghe, The duality of informetric systems with applications to the empirical laws, Journal of Information Science 16 (1) (1990) 17–21.
[11] E. García-Barriocanal, M.A. Sicilia, Filtering information with imprecise social criteria: a FOAF-based backlink model, in: Proceedings of the Fourth Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2005), 2005, pp. 1094–1098.
[12] E. García-Barriocanal, M.A. Sicilia, Preliminary Explorations on the statistical profiles of highly-rated learning objects, in: Proceedings of the Third Metadata and Semantics Research Conference (MTSR 2009), Springer Communications in Computer and Information Science, vol. 46, 2009, pp. 108–117.
[13] F. Giardini, G. Di Tosto, R. Conte, A model for simulating reputation dynamics in industrial districts, Simulation Modelling Practice and Theory 16 (2) (2008) 231–241.
[14] P. Han, G. Kortemeyer, B.J. Kramer, C. von Prummer, Exposure and support of latent social networks among learning object repository users, Journal of Universal Computer Science (JUCS) 14 (2008).
[15] B.J. Jansen, A. Spink, How are we searching the World Wide Web? A comparison of nine search engine transaction logs, Information Processing & Management 42 (1) (2006) 248–263.
[16] C.M. Kelty, C.S. Burrus, R.G. Baraniuk, Peer review anew: three principles and a case study in postpublication quality assurance, in: Proceedings of the IEEE 96(6), 2008, pp. 1000–1011.
[17] A.J. Lotka, The frequency distribution of scientific productivity, Journal of the Washington Academy of Sciences 16 (12) (1926) 317324.

[18] T. Malsch, C. Schlieder, P. Kiefer, M. Lbcke, R. Perschke, M. Schmitt, K. Stein, Communication between process and structure: modelling and simulating message reference networks with COM/TE, Journal of Artificial Societies and Social Simulation 10 (1) (2006). <http://jasss.soc.surrey.ac.uk/10/1/9.html> .

[19] S. Monge, R. Ovelar, I. Azpeitia, Repository 2.0: social dynamics to support community building in learning object repositories, Interdisciplinary Journal of E-Learning and Learning Objects 4 (2008) 191–204.

[20] D. Namuth, S. Fritz, J. King, A. Boren, Principles of sustainable learning object libraries, Interdisciplinary Journal of Knowledge and Learning Objects 1 (2005).

[21] M.J. North, N.T. Collier, J.R. Vos, Experiences creating three implementations of the repast agent modeling toolkit, ACM Transactions on Modeling and Computer Simulation (TOMACS) 16 (1) (2006) 125.

[22] X. Ochoa, E. Duval, Quantitative analysis of user-generated content on the Web, in: Proceedings of the First International Workshop on Understanding Web Evolution (WebEvolve2008), 2008, pp. 19–26.

[23] X. Ochoa, E. Duval, Quantitative analysis of learning object repositories, in: Proceedings of the World Conference on Educational Multimedia, Hypermedia and Telecommunications, AACE, Chesapeake, VA, 2008, pp. 6031–6048.

[24] L. Page, S. Brin, R. Motwani, T. Winograd, The PageRank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford Digital Library Technologies Project, Paper SIDL-WP-1999-0120, 1999 (version of 11.11.99).

[25] J.B. Schafer, J. Konstan, J. Riedl, Electronic commerce recommender applications, Journal of Data Mining and Knowledge Discovery 5 (1/2) (2000) 115–152.

[26] S. Schroter, L. Tite, Open access publishing and author-pays business models: a survey of authors' knowledge and perceptions, Journal of the Royal Society Medicine 99 (2006) 141–148.

[27] M.A. Sicilia, S. Snchez-Alonso, E. Garca-Barriocanal, D. Rodrguez, Exploring structural prestige in learning object repositories: some insights from examining references in MERLOT, in: Proceedings of the International Conference on Intelligent Networking and Collaborative Systems (INCoS 2009), Barcelona, Spain, November 4–6, 2009.

[28] J. Vargo, J.C. Nesbit, K. Belfer, A. Archambault, Learning object evaluation: computer mediated collaboration and inter-rater reliability, International Journal of Computers and Applications 25 (3) (2003) 198–205.

[29] J. Wang, K. Gwebu, M. Shanker, M.D. Troutt, An application of agent-based simulation to knowledge sharing, Decision Support Systems 46 (2) (2009) 532–541.

[30] S. Wasserman, K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge, New York, Melbourne, 1994.

[31] J. Willinsky, Scholarly associations and the economic viability of open access publishing, Open Journal System Demonstration Journal 1 (1) (2005).

[32] J. Wu, B. Hu, J. Zhang, D. Fang, Multi-agent simulation of group behavior in E-Government policy decision, Simulation Modelling Practice and Theory 16 (10) (2008) 1571–1587.