# Modeling the Macro-Behavior of Learning Object Repositories

*Xavier Ochoa*
*Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador*

xavier@cti.espol.edu.ec

## Abstract

Learning Object Repositories (LOR) are the result of the activity of hundreds or thousands of contributing individuals. It has been shown in a previous work by the author (Ochoa & Duval, 2008) that LORs have an interesting macro-behavior, mostly governed by long-tailed distributions. The shape of these distributions provides valuable information for the management and operation of LORs. However, the reason why these distributions appear is not known. This work proposes a simple model to explain this macro-behavior as the consequence of very simple micro-behavior of individual contributors, more specifically their number, production rate, and lifetime in the repository. This model is formally presented and successfully validated against data from existing LORs. While simple, this model seems to explain most of the large-scale measurements in function of the small-scale interactions. Finally, this work discusses the implications that this model has in the planning and maintenance of new and existing LORs.

**Keywords:** Learning Object Repositories, complex systems, long tailed distributions

## Introduction

The publication of learning materials in online repositories is usually regarded as a simple process. To publish, the contributor provides or uploads the material (or the reference to the material), fills some metadata about the material, and then the material is available in the repository for others to find and reuse. The contributor can repeat this process for more materials as desired, while he or she is still interested in providing content to the repository.

These seemingly simple processes that determine the micro-behavior of contributors and consumers give rise to complex macro-behavior at the repository level once the contribution and preference of hundreds or thousands of individuals is aggregated (Ochoa & Duval, 2008). For example, some learning object repositories grow linearly while others, having a similar number of contributors, grow exponentially. Also, the number of objects published by a given contributor is distributed differently depending on the kind of repository, but always following a long-tailed distribution (Anderson, 2006).

Unfortunately, there is no research available about how the micro-behavior of the individuals is related to the observed macro-behavior of Learning Object Repositories. The fields of Bibliometrics and Scientometrics have been studying a similar problem: the process of paper publication in different venues (journals, conferences, repositories, etc.). In these fields, several models



Editor: Janice Whatley

have been proposed to attempt to explain the observed patterns in the data. For example, De Price Sola (1976) proposed "Cumulative advantage" as a model to explain the inverse-power law distribution, also called Lotka by Coile (1977), observed in the number of papers published by a scientist in a given field. Egghe and Rousseau (1995) and Egghe (2005) refine this notion with the "success breeds success" model. However, the models used for scientific publication cannot be transferred to learning object publication because one of their main characteristics, the increasing rate of production observed in most successful scientific contributors, has not been observed in learning material contributors elsewhere (Ochoa & Duval, 2008). Nonetheless, the methodologies to establish and validate these models will be borrowed and re-used in the present study.

The present work proposes an initial model to explain the macro-behavior of LORs based on the characteristics of their contributor base. This paper is structured as follows: the modeling section presents previous unexplained characteristics of Learning Object Repositories that this work proposes to model. In the next section the model is formally defined and explained. The validation section studies the model, comparing its predictions against empirical data. The paper ends with a discussion of the relevance of this model and further research needed to improve it.

# Modeling the Publication Process

In a previous work (Ochoa & Duval, 2008), several characteristics of the publication of learning objects were measured. That work used data collected from several sources:

- three Learning Object Repositories (LORp): Ariadne, Connexions and Maricopa Exchange;
- three Learning Object Referatories (LORf): Merlot, Intute and Ferl First;
- two Open Courseware sties (OCW): MIT OCW and OpenLearn and
- one Learning Management System (LMS): SIDWeb.

The findings of that work could be summarized as:

- LORp and LORf grow in number of objects linearly in two stages (bi-phase linearly), but OCW and LMS grow exponentially.
- Most LORp and LORf grow bi-phase linearly in the number of contributors. OCW and LMS grow exponentially.
- The number of objects published by a given author follows a Lotka distribution with exponential decay in the case of LORp and LORf. OCW and LMS present a Weibull distribution.
- The rate at which contributors publish materials followed a Log-Normal distribution for all the repositories studied.
- The lifetime of the contributors (time that the contributor remains actively publishing material) is distributed exponentially for LORp and LORf and according to a Weibull distribution in the case of OCW and LMS.

While these findings provided information about how to manage repositories, the quantitative study did not explain the connection between those measurements and the reason why they are found in the first place. For example, Connexions, a LORp, has a linear growth in the number of objects, but an exponential growth in the number of contributors. Also, it does not explain how the behavior of the contributors (publications rate and lifetime) is related to the behavior of the repository (repository growth and distribution of contribution)

This work tries to formulate a model that could simulate the observed results with the lowest amount of initial parameters. The objective of this model is to understand the relation between the micro-behavior (contributors publishing learning objects at a given rate during a given time) and the macro-behavior (repositories growing linearly, publication distribution having a heavy tail). Finally, this model could help us to adjust the initial parameter and simulate the macro-behavior that a hypothetical repository would have. For example, the model will help us to know what type of initial factors give rise to exponential growth.

This model is inspired by the ideas of Huber (2002). Huber modeled the distribution of the amount of patents published among inventors using four variables: the Frequency (publication rate), the Career Duration (lifetime), the Poissoness (the degree to which they conform to a Poisson distribution), and the Randomness. While we use some of his ideas, the methodology used in this paper expands that model in two principal ways: 1) our model is capable of generating non-Lotka distributions, and 2) the predictive scope of our model is larger, including the growth function and total size.

# Definitions

The model is based on three factors. Two of them are directly related to the micro-behavior of the contributor: the rate of publication and the lifetime. The third factor is related to the number of contributors that a repository has at a given time. These factors are defined as follows:

- **Publication Rate Distribution (PRD):** This specifies how talent or capability is distributed among the contributor population. Mathematically, PRD(x) is a random variable that represents the probability of a contributor to publish one object each x days on average. In the case of all the repositories studied in Ochoa and Duval (2008) the Log-Normal is a good approximation of this distribution, although any distribution can be set to test "what-if" scenarios.
- **Lifetime Distribution (LTD):** This specifies the amount of time that different contributors will be active in the repository. Mathematically, LTD(x) is a random variable that represents the probability that a contributor will stay active in the repository for x days. Ochoa and Duval (2008) found that Exponential, Log-Normal, and Weibull distributions seem to represent different types of contributor engagement.
- **Contributor Growth Function (CGF):** This is a repository related factor that, for now, cannot be predicted. Mathematically, CGF(x) is a function that represent the number of contributors that the repository has after x days. Ochoa and Duval (2008) found that Biphase linear and Exponential are a good approximation for the contributor growth.

While the initial factors can be formally defined (distribution functions or growth functions), the process to derive a formal model involves non-linear calculations (Huber, 2002) that make it unfeasible to derive an exact mathematical solution (resulting distribution) that can be easily interpreted. Therefore, a numerical computation is used to run the model. This approach, while less formal, is very flexible to accommodate a greater range of initial factors.

The construction of the model can be described as follows:

1. The period of time, measured in days over which the model is run, is selected.
2. The Contributor Growth Function (CGF) is used to calculate the size of the contributor population at the end of that period.
3. A virtual population of contributors of the calculated size is created.
4. For each contributor, the two basic characteristics, publication rate and lifetime are assigned:

4.1. First, a publication rate value, generated randomly from the Publication Rate Distribution (PRD), is assigned to each contributor.

4.2. Second, a lifetime value, generated randomly from the Lifetime Distribution (LTD) is also assigned to each contributor.

5. Once the virtual contributors' parameters have been set, each contributor is assigned a starting date. The number of contributor slots for each day is extracted from a discrete version of the CGF. Each contributor is assigned randomly to each one of those slots. If the start date plus the lifetime of a contributor exceed the final date of the simulation, the lifetime is truncated to fit inside the simulation period.

Once the simulated population has been created, the model is run. A Poisson process is used to simulate the discrete publication of learning materials. The lambda variable required by the Poisson process is replaced by the contributor's publication rate. The process is run for each day of the contributor's lifetime. The result of the model is a list containing the contributors, the number of objects that those contributors have published, and the dates in which those publications took place. From this data, the macro-behavior of the simulated repository can be extracted in a similar way as for real repositories.

In formal terms (Equation 1), the random variable N, representing the number of objects published by each contributor, is equal to the PRD, the random variable representing the rate of production of the contributor, multiplied by LTD, the random variable representing the lifetime of the contributor in the repository. Given that solving the multiplication of random variables often involves the use of the Mellin transform (Epstein, 1948) and the result is not always easily interpretable (Huber, 2002), this multiplication is solved through computation methods. Equation 2 shows the resulting distribution of N. The probability of publishing k objects is the combined probability of each contributor publishing k objects. Given that the production of a contributor is considered independent of the production of any other contributor, the combination of probabilities is converted into a product for the N*c* contributors. To calculate the probability with which the *i*th contributor publishes k objects, we use the formula of the Poisson process with production rate R*i* and lifetime L*i* randomly extracted from their correspondent distributions. This formula calculates the probability that the contributor publishes exactly k objects during her lifetime.

$$N = R \otimes L \tag{1}$$

$$P(N = k) = \prod_{i=1}^{Nc} \frac{(R_i \cdot L_i)^k}{k!} e^{-R_i \cdot L_i} \tag{2}$$

# Model Validation

To validate this model the simulated results are compared with the data extracted from real repositories. Three characteristics of the repository are compared: 1) distribution of the number of publications among contributors (N), 2) the shape of the content growth function (GF), and 3) the final size of the repository (S).

The repositories used for this evaluation is a subset of the repositories used in Ochoa and Duval (2008): Ariadne, Connexions, and Maricopa Connexion representing the Learning Object Repositories (LORp); Merlot representing the Learning Object Referatories (LORf); MIT OCW representing the Open Courseware (OCW); and SIDWeb, the Learning Management System used in the Escuela Superior Politécnica del Litoral, representing the LMS. This data was captured between the 5th and the 8th of November 2007. To perform the evaluation, the initial factors were

taken from the data extracted from these repositories and are presented in Table 1. These factors were fed into the model and used to run the simulation. To have a statistically meaningful comparison between the real data and the output of the model, 100 Monte-Carlo simulated runs were generated for each repository.

**Table 1.** Initial factors extracted from the empirical data of the studied repositories

| ***Repository*** | ***PRD(x)*** | ***LTD(x)*** | ***CGF*** |
|---|---|---|---|
| Ariadne (LORp) | Log-Normal (μlog=-3.25, σlog=1.27) | Exponential (λ=0.0010) | Bi-Phase Linear (slope1=0.02, slope2=0.06, breakpoint =1277) |
| Connexions (LORp) | Log-Normal (μlog=-4.11, σlog=1.36) | Exponential (λ=0.0012) | Exponential (λ=1.2x10-3) |
| Maricopa (LORp) | Log-Normal (μlog=-5.18, σlog=0.95) | Exponential (λ=0.0012) | Bi-Phase Linear (slope1=0.06, slope2=0.28, breakpoint =1095) |
| Merlot (LORf) | Log-Normal (μlog=-2.47, σlog=1.11) | Exponential (λ=0.0015) | Bi-Phase Linear (slope1=0.12, slope2=0.54, breakpoint 401) |
| MIT (OCW) | Log-Normal (μlog=-1.68, σlog=1.07) | Weibull (k=1.72, λ=325) | Exponential (λ=3.7x10-3) |
| SIDWeb (LMS) | Log-Normal (μlog=-2.57, σlog=0.96) | Weibull (k=1.21, λ=588) | Exponential (λ=1.8x10-3) |

First, a comparison is made of the distribution of publications (N) between the empirical and simulated data. To have a meaningful comparison, the parameters of the distribution of the simulated data are estimated with the same methodology used to obtain the empirical measures (Ochoa & Duval, 2008). As expected, each simulated data set was assigned slightly different parameters values. However, the values were normally distributed. A simple t-test was applied to establish whether it is reasonable to assume that the parameters assigned to the empirical data set belongs to the same population as the simulated parameters. If all the parameters of the empirical distribution belong to the same population as the simulated ones, it can be concluded that the empirical and simulated data sets have the same distribution. The p-value for the t-test is provided in Table 2 together with the mean values of the simulated parameters.

For LORs, the model is able to accurately simulate the alpha value for all the repositories. The alpha parameter basically determines the general shape of the Lotka distribution. The rate parameter, on the other hand, has a more subtle effect. This parameter determines the slight reduction in probability of finding very productive contributors. The model does not seem able to consistently reproduce this value. The subtle effect that determines the exact value of rate is most probably lost during the simplifications of the model. An example of the simulation of the Connexions repository is presented in Figure 1.

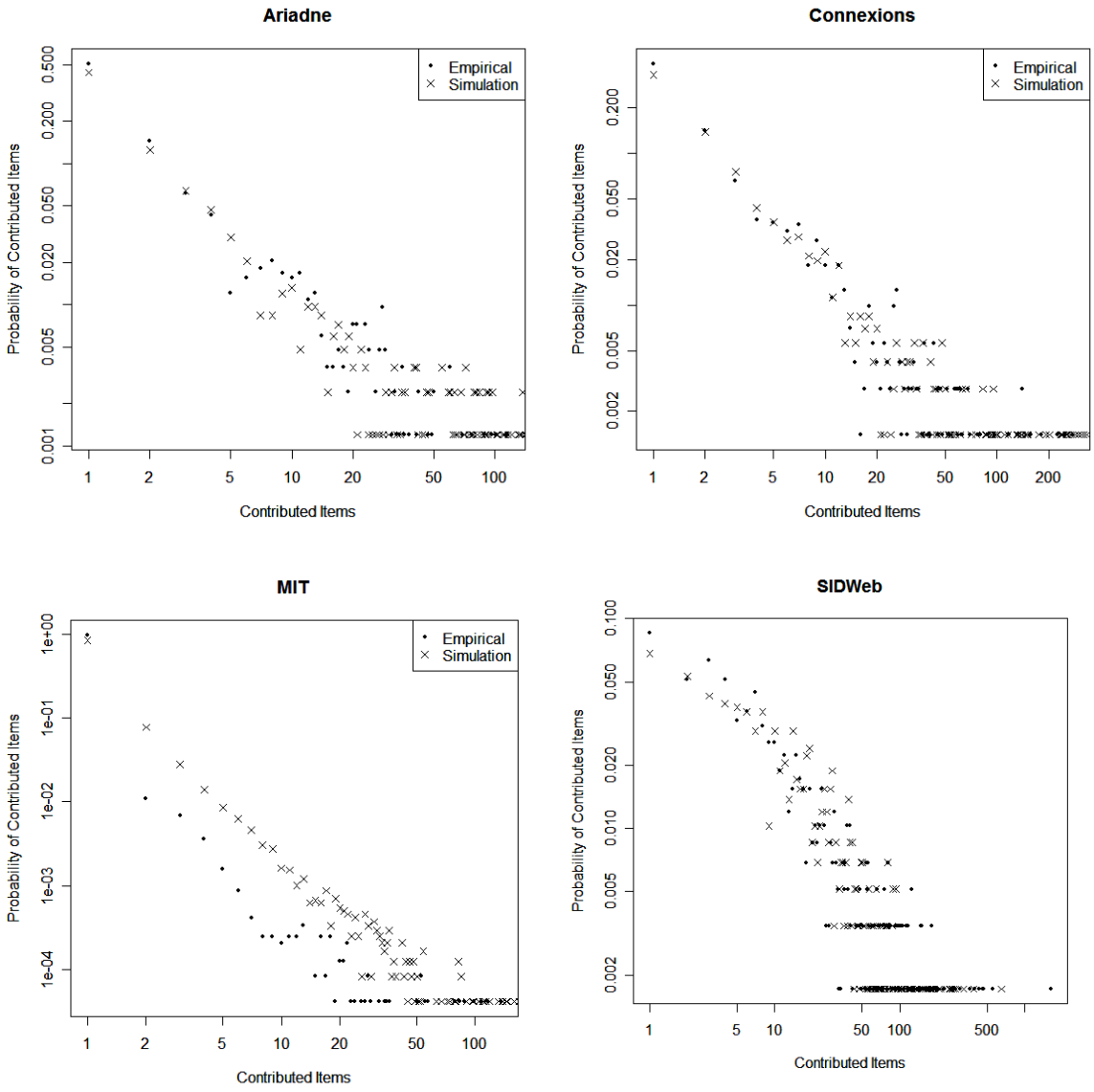


**Figure 1.** Comparison between the Empirical and Simulated Distribution of the Contribution (N)

**Table 2.** Comparison between the empirical and simulated distribution of the number of objects published per contributor (N)

| *Repository* | *N empirical* | *N simulated (average)* | *p-values* |
|---|---|---|---|
| Ariadne (LORp) | Lotka exp. cut-off ($\alpha$=1.57, $\lambda$=0.011) | Lotka exp. cut-off ($\alpha$=1.58, $\lambda$=0.001) | p-$\alpha$: 0.60<br>p-$\lambda$: 0.02 |
| Connexions (LORp) | Lotka exp. cut-off ($\alpha$=1.35, $\lambda$=0.0094) | Lotka exp. cut-off ($\alpha$=1.42, $\lambda$=0.0002) | p-$\alpha$: 0.31<br>p-$\lambda$: 0.07 |
| Maricopa (LORp) | Lotka exp. cut-off ($\alpha$=2.12, $\lambda$=0.0067) | Lotka exp. cut-off ($\alpha$=2.39, $\lambda$=0.04) | p-$\alpha$: 0.60<br>p-$\lambda$: 0.02 |
| Merlot (LORf) | Lotka exp. cut-off ($\alpha$=1.88, $\lambda$=0.0006) | Lotka exp. cut-off ($\alpha$=1.76, $\lambda$=0.002) | p-$\alpha$: 0.28<br>p-$\lambda$: 0.10 |
| MIT (OCW) | Weibull (k=1.07, $\lambda$=40.5) | Weibull (k=0.68, $\lambda$=35) | p-k: 0.00<br>p-$\lambda$: 0.22 |
| SIDWeb (LMS) | Weibull (k=0.52, $\lambda$=17.14) | Weibull (k=0.60, $\lambda$=19) | p-k 0.21<br>p-$\lambda$: 0.55 |

The shape of the OCW MIT Weibull distribution of publications seems to present a major challenge for the model. The almost horizontal head of the distribution cannot be accurately simulated with the current calculations. The shape parameter is vastly underestimated. However, the tail of the distribution is reasonably matched by the simulated values and the scale parameter is correctly estimated. The comparison between one simulation run and the empirical data can be seen in Figure 1. The model, nonetheless, can model less extreme Weibull distributions, as can be seen in the estimation of the SIDWeb parameters.

The next step in validating the model is to compare the shape of the content growth function (GF) and the final size of the repository (S). For the GF evaluation, the daily simulated production of objects was counted across contributors. First, the count was fitted with the same methodology and functions used to obtain the empirical results in Ochoa and Duval (2008). Counting the times that the correct function, Bi-Phase Linear, was selected, provided the best-fitting alternative. For the S evaluation, the total number of objects produced in each simulated data set was counted. The distribution of the final size follows a left-skewed distribution. The Empirical Cumulative Density Distribution (ECDF) was used to calculate the chances that the empirical size came from the same population. The results of these evaluations are presented in Table 3.

The simulated growth functions seem to agree with the empirical measurements most of the time (>50%). When the contributor base growth function is also Bi-Phase Linear (Ariadne, Maricopa, MERLOT), the accuracy of the prediction is high (90% or higher). However, when an exponential contributor rate growth is involved in the calculation (Connexions, MIT OCW, and SIDWeb), the identification rate decreases (60-80%). It is interesting to note that thanks to the variability in the lifetime, an exponential contributor growth does not necessarily means exponential growth in the number of objects. However, as the miss-interpretation rate shows, when there is exponential growth in the number of contributors, exponential growth in the number of objects is a viable

outcome. These effects can be observed in Figure 2. There, a graphical representation of random simulated growth functions is presented.

The actual parameters of the Growth Function (GF) are not analyzed, as they varied widely from simulation to simulation. The implication of this variation is not clear; it could be that natural variation creates several types of growth from the same contributor population, or that this model, in its simplicity, does not take into account some relation between lifetime and production rate that is responsible for the shape of the function. More research is needed to solve this question.

**Table 3.** Validation of the Growth Function (GF) and the final Size (S) of the repositories

| ***Repository*** | ***GF Empirical*** | ***% GF Simulated = GF Empirical*** | ***S Empirical*** | ***Average S Simulated (p-value)*** |
|---|---|---|---|---|
| Ariadne (LORp) | Bi-Phase Linear | 100% | 4,875 | 5,516 (0.48) |
| Connexions (LORp) | Bi-Phase Linear | 73% | 5,134 | 6,052 (0.50) |
| Maricopa (LORp) | Bi-Phase Linear | 100% | 2,221 | 3,105 (0.36) |
| Merlot (LORf) | Bi-Phase Linear | 98% | 18,110 | 20,389 (0.61) |
| MIT (OCW) | Exponential | 65% | 53,880 | 48,320 (0.52) |
| SIDWeb (LMS) | Exponential | 76% | 21,675 | 25,443 (0.20) |

Finally, a comparison is made of the final number of produced objects when the simulations have been run for the same period of time as measured in the empirical data sets. As can be seen in Table 3, the size values for all the repositories were estimated correctly, even in repositories where the simulated and empirical publication distributions does not completely match (OCW). The reason for this resilience is that the tail of the distribution (or the head, in the case of OCW) is responsible for a small fraction of the objects. If the simulation can match the head (or the middle section, in the case of OCW), where most of the objects are published, the total simulated output is similar to the original repository. These results support the use of this model to calculate growth and required capacity.

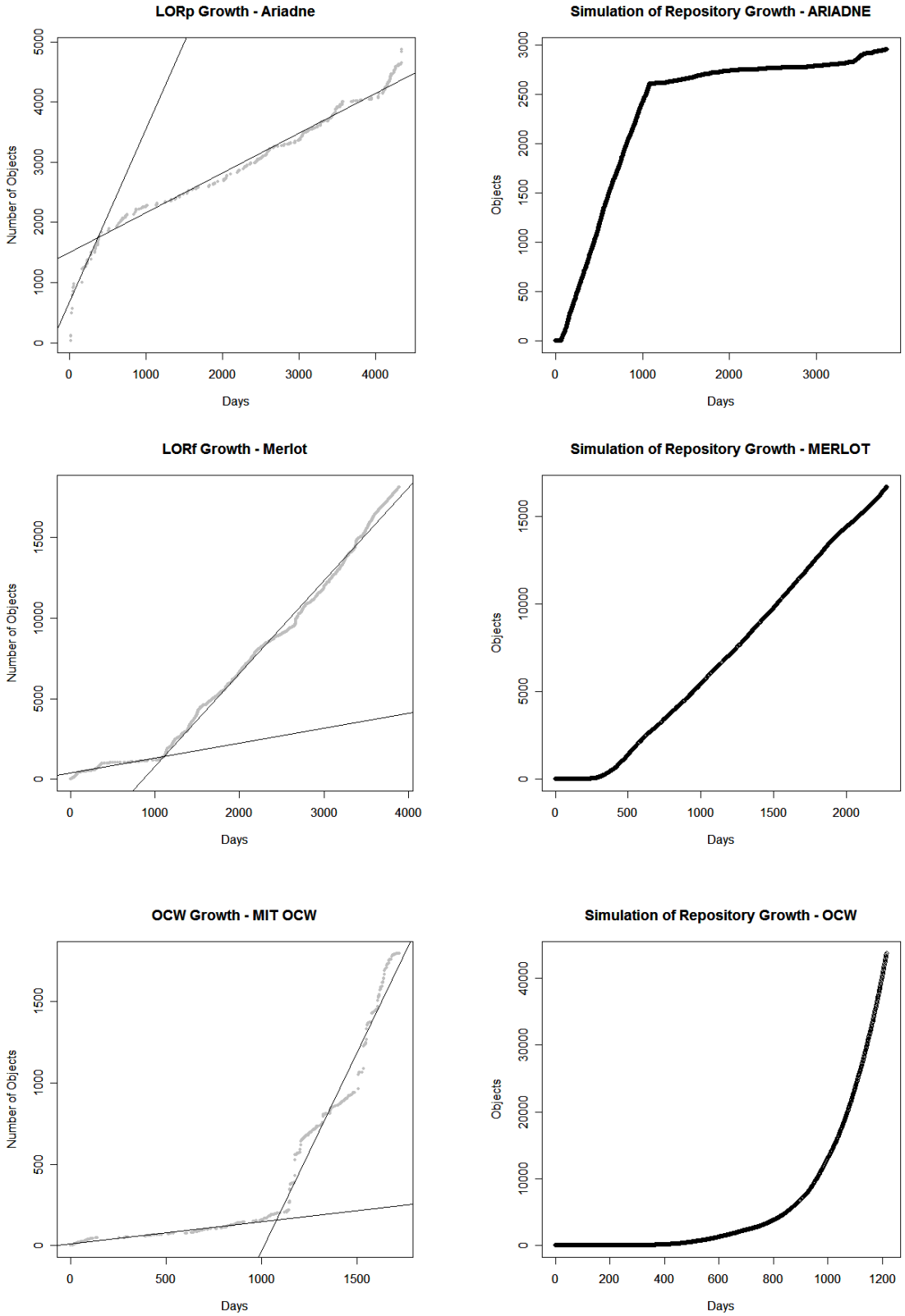


**Figure 2.** Comparison between the Empirical Growth Function (left) and the Simulated Growth Function (right)

# Conclusions and Implications

The model presented makes the simple assumption that the only variables that affect the characteristics of a repository are how frequently the contributors publish material (publication rate), how much time they persist in their publication efforts (lifetime), and at what rate they arrive at the repository (contributor growth function). The model combines those variables through a computational simulation that is capable of predicting other repository characteristics, such as the distribution of publications among contributors, the shape of the content growth function, and the final size of the repository.

The model has been evaluated with the data presented in the analysis sections. From this evaluation, it can be concluded that the simple model is capable of simulating quite well the characteristics observed in real repositories based only on the initial factor. However, the simplicity of the model can be seen when the model tries to simulate repositories with special characteristics, for example, when it tries to simulate repositories like OCW that have a small low-publishing community. Nonetheless, the model can be used at it is to predict future growth of current repositories or to simulate repositories with characteristics not seen naturally. For example, what the publication distributions will be like if the publication rate is uniformly distributed. Improvements of this model to include special cases, as well as interactions between the factors, are an interesting topic for further research.

The most important implication that the development of this model has for LOR administrator-management is to provide a tool that can be used to predict growth and behavior of the repository. For example, based in the observed growth in the number of contributors, their rate of publication and current lifetime, the model can be used to predict the number of objects that the repository will have in the future. Also the model shows that if the lifetime distribution is altered, the growth will be immediately affected. For example, if the repository could retain its contributors for longer periods of time, its growth could change from linear to exponential.

Finally, the most important characteristic of the proposed model is its testability. It would be easy to construct competing models and test if they predict different characteristics of the repositories and can handle special cases that the current model cannot. This testability provides a way to measure progress in efforts to understand the nature and workings of the learning object publication process.

# References

Anderson, C. (2006). *The long tail*. New York: Hyperion.

Coile, R. (1977). Lotka's frequency distribution of scientific productivity. *Journal of the American Society for Information Science, 28*(6), 366-370.

De Price Sola, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5-6), 292-306.

Egghe, L. (2005). The power of power laws and an interpretation of Lotkaian informetric systems as self-similar fractals. *Journal of the American Society for Information Science*, *56*(7), 669–675.

Egghe, L. & Rousseau, R. (1995). Generalized success-breeds-success principle leading to time-dependent informetric distributions. *Journal of the American Society for Information Science*, *46*(6), 426–445.

Epstein, B. (1948). Some applications of the Mellin transform in statistics. *The Annals of Mathematical Statistics*, *19*(3), 370–379.

Huber, J. (2002). A new model that generates Lotka's law. *Journal of the American Society for Information Science and Technology*, *53*(3), 209–219.

Ochoa, X. & Duval, E. (2009). Quantitative analysis of learning object repositories. *IEEE Transactions on Learning Technologies*, *2*(3), 226-238.

# Biography

**Xavier Ochoa** is a professor at the Faculty of Electrical and Computer Engineering at Escuela Superior Politécnica del Litoral (ESPOL) in Guayaquil, Ecuador. He coordinates the research group on Teaching and Learning Technologies at the Information Technology Center (CTI) at ESPOL. He is also involved in the coordination of the Latin American Community on Learning Objects (LACLO), the ARIADNE Foundation, and several regional projects. His main research interests revolve around measuring the Learning Object economy and its impact on learning. More information at http://ariadne.cti.espol.edu.ec/xavier