Title
Measuring and Comparing Participation Patterns In Digital Repositories, Repositories by the Numbers Part I

Permalink
https://escholarship.org/uc/item/9js456d5

Journal
D-lib Magazine, 13(9/10)

ISSN
1082-9873

Authors
Thomas, Chuck
McDonald, Robert H.

Publication Date
2007-09-01

DOI
10.1045/dlib.magazine

Peer reviewed
Measuring and Comparing Participation Patterns In Digital Repositories, Repositories by the Numbers Part I

Chuck Thomas,
Florida Center for Library Automation
cthomas@ufl.edu

Robert H. McDonald
San Diego Supercomputer Center
mcdonald@sdsc.edu

Introduction

This article summarizes findings from a study of author/depositor distribution patterns within scholarly digital repositories. At the moment, evaluative frameworks are in short supply for institutional and disciplinary repositories (Kim & Kim, 2006). After a review of issues of scholars participating in digital repositories, author/depositor distribution is analyzed as one possible technique that might be used to judge the success of a repository. This statistical technique was used to evaluate participation patterns among more than 30,000 author/depositors whose works were found in various categories of digital repositories. Findings from this analysis, including comparisons of participation patterns across three categories of scholarly repositories, are presented along with an explanation of the questions and challenges that arose during the study. The article concludes with an evaluation of the analytical technique and its potential as one metric for judging a repository's success.

Context

Digital repositories may assume many forms. The Andrew W. Mellon Foundation's operating definition of a repository is a good starting point:

"A repository is a networked system that provides services pertaining to a collection of digital objects. Example repositories include: institutional repositories, publisher's repositories, dataset repositories, learning object repositories, cultural heritage repositories, etc." (Mellon, 2006)

The Coalition for Networked Information's Executive Roundtable characterized repository categorization as "very problematic" (CNI, 2003). While content within them varies substantially, repositories can be categorized in terms of who funds and administers them. The two main categories in this classification are "disciplinary" and "institutional" (Ibid). Within the realm of institutional repositories, a further important subdivision is emerging rapidly – those whose sponsoring institutions do not mandate faculty/scholar participation (referred to as "voluntary-deposit institutional" in this article) and those that do (referred to as "mandatory-deposit institutional" in this article). Regardless of a repository's category, securing participation by scholars, meaning their willingness to deposit copies of their research output, cannot be taken for granted. In
fact, achieving significant participation rates, particularly in institutional repositories, is cited repeatedly as the most challenging aspect of establishing scholarly repositories (Lynch & Lippincott, 2005).

Participation by contributors is one of the most important indicators of a scholarly digital repository's success. Some repository sponsors try to recruit contributors by emphasizing the need to preserve and measure use of research output (Day, 2004), but the factors that actually motivate scholars to deposit their work are more complex. Recent studies (Swan, et al., 2005; Foster & Gibbons, 2005; Kennan & Wilson, 2006) confirm earlier suspicions that a desire to enhance an institution's prestige or to enable more systematic and automated assessment of scholarly productivity within an organization is most certainly not a motivating factor for repository participants. Instead, the choice to deposit research output usually stems from desires for personal recognition and impact among one's peers (Ibid). Unfortunately, gauging the impact of a deposited paper, report or similar work is not easy for either an individual scholar or a repository manager. Yeomans (2006) reminds us that even the most mature repositories, such as the physics community's arXiv, may generate impressive statistics, but offer little to help anyone know what kind of "success" those figures measure. Some have suggested utilizing the criticized but widely-employed personal "impact analysis" techniques that grew from citation analysis theory (Day, 2004). However, even the creator of modern citation analysis theory warned such techniques “are easily misinterpreted or inadvertently manipulated for improper purposes,” and should take into account differing publishing and citation customs across disciplines, the practice of self-citation, and reasons why individual works are cited (Garfield, 1983).

Furthermore, desires for recognition and impact still are not enough to ensure scholars will participate in institutional repositories. Even institutions with mandates requiring faculty deposits face the enduring task of encouragement and mandate enforcement (Sale, 2006). Sale (2007) estimates only 15-20% of faculty will ever choose to participate in voluntary-deposit repositories of any type. Imagine, therefore, the difficulties involved in encouraging or predicting participation in voluntary-deposit institutional and disciplinary repositories!

**Research Questions**

Because scholarly contributions are a vital part of successful repositories, finding a meaningful measure of participation is an important step in developing comprehensive repository evaluation frameworks. One such measurement is to compare the actual number of contributors and their actual numbers of deposits against the total universe of possible depositors and their total research output. Sale (2006) employed this technique to track one university’s successes in requiring faculty to deposit copies of each paper they publish. Ongoing measurement in this manner presents challenges. It requires detailed tracking of individual identities and each scholar's output. Additionally, comparing actual participation rates against the benchmark of possible participation rates is very likely to yield discouraging results, particularly for repositories that cannot leverage an institutional or other mandate to gather content.
Another measurable facet of participation is the distribution of a repository's total content, per contributor. By applying simple analysis of frequencies and distributions of authors and papers contributed, would such a perspective on a digital repository be useful? Because this measurement involves only what is already in a repository, would the data for analysis be simple to obtain from many repositories? The study described in the remaining portion of this article addressed the following three research questions:

- Are contributor distributions a useful analytical metric for repositories?
- Do these distribution patterns differ for institutional and disciplinary repositories?
- Do contributor distribution patterns differ for mandatory-deposit and voluntary-deposit institutions?

A simple research methodology was devised, as described below.

**Methodology**

The study was conducted using the following definitions, processes and guidelines:

**Definitions**

**Repository:** The study focused on repositories containing mainly research papers suitable for publication in serial literature like scholarly journals, professional society newsletters, or related modes of dissemination. Repositories containing significant quantities of learning or instructional objects, institutional records, or other ancillary components of the research and teaching process were disqualified from analysis. Additionally, because Electronic Theses and Dissertations skew author:items ratios toward a 1:1 value, repositories containing ETDs were not considered.

**Depositor:** Though one cannot always assume the author of a paper is the same person who actually deposits it into a digital repository, this study assumes creators of research output generally self-deposit or authorize a proxy to do the task for them. The terms author, depositor and participant are used interchangeably in this article.

**Items:** Discrete manifestations of intellectual creation as described for the term "Repository" above. Typically, an item is equivalent to a research paper, technical report, or similar object.

**Participation:** Allowing one's intellectual creation to be deposited and made available through a disciplinary or institutional repository. This is assumed to be a conscious choice for all depositors.

**Processes and Guidelines**
1. Select a group of repositories for analysis

   a. Open Access repositories were thought to be most transparent and available for inspection, so the OpenDOAR registry of open repositories (http://www.opendoar.org/) provided a starting list of candidate sites for evaluation. OpenDOAR listed 838 registered sites on the date the starting list was compiled.

   b. The study chose to analyze repositories using the Southampton E-Prints software (www.eprints.org/software) because it currently is the most widely deployed digital repository software, and often provides an accessible report of authors and their numbers of deposited items, without having to contact a system manager for such information. Choosing to analyze only E-Prints sites narrowed the starting list to 176 candidate sites for analysis.

   c. Only sites with five hundred or more deposits were considered for analysis. Of the 176 repositories qualified for analysis so far, only 56 contained 500 or more deposited items.

   d. After excluding sites containing ETDs, learning objects and other non-publishable resources, the list of candidates for analysis shrank from 56 repositories of sufficient size to 9 remaining repositories, or slightly more than 1% of the OpenDOAR registry. These 9 repositories provided core groups of data aggregated and analyzed in the study.

   e. Additionally, the study's selection methods were modified to make sure 2 suitably large mandatory-deposit institutional repositories were included. Repositories built on institutional mandates are still rare, so the study's authors accommodated different underlying repository software for both of these cases. The core number of analyzed repositories consequently grew to 11.

   f. In addition to the 11 core analyzed repositories, the study also collected and analyzed separately data from two other large disciplinary repositories. These two extra repositories were the AgEcon Search repository hosted by the University of Minnesota and the arXiv.org repository hosted by Cornell University. Both were selected for their size and their longevity as disciplinary digital repositories. The results of data analysis from these two repositories are discussed in a separate section of this article.

2. Classify each repository for analysis as "voluntary-deposit institutional", "mandatory-deposit institutional" or "disciplinary"

3. Obtain reports from each repository on the total number of items in the repository, the total number of contributing authors, and the corresponding number of items created by each author
Based on Steps 2 and 3 of the study methodology, the 11 analyzed repositories were grouped as:

a. 6 voluntary-deposit institutional repositories (18,326 Authors; 14,829 Items)

b. 2 mandatory-deposit institutional repositories (4,167 Authors; 5,920 Items)

c. 3 disciplinary repositories (6,895 Authors; 6,773 Items)

All together, these 11 repositories represented 29,388 authors and 27,522 deposited items.

d. The AgEcon Search disciplinary repository was analyzed separately and contained 24,567 deposited items from 19,700 authors.

e. The arXiv.org disciplinary repository was analyzed separately and contained 406,857 deposited items from 105,131 authors.

4. Calculate author:contribution frequencies by repository

Simple frequency and distribution descriptive statistics were generated for each repository using the SPSS statistical software.

5. Aggregate repository frequencies into voluntary, mandatory, and disciplinary groupings

After examining each individual repository's frequencies and distributions of total number of items contributed by each author, author names were anonymized and cases were combined for aggregate analysis by each category (voluntary-deposit institutional, mandatory-deposit institutional and disciplinary).

6. Analyze distribution patterns within and across categories

The results of each aggregated analysis were then compared across categories looking at results such as the range of contributions per author, the mean, median and mode number of contributed items per author, standard deviations among values in each set, and distribution curves (including skew and kurtosis of each distribution) for each dataset.

Issues

During this study, the following issues had to be considered:

1. Identifying the scope, content and context of individual repositories is an imprecise and subjective exercise. Categorizing the types of content within any repository is difficult. Though the content of both institutional and disciplinary digital repositories
certainly will be an important part of future "distributed libraries" (Brogan, 2006), even the most mature sites have trouble providing clear categories and indicators for measurement (Yeomans, 2006). The initial selection of archives for evaluation was one of the most time-consuming aspects of this study.

2. Most repositories do a poor job of maintaining standard forms of names for contributing authors, so the same author may be listed under multiple name variants and treated as separate people. Responsibility for name consistency in most repositories seemed to rest with the depositors themselves. This study could do little to correct for such variations within the 29,388 contributors listed by the repositories analyzed. However, as the results of this study will show, a large number of authors only have contributed one item to any particular repository, so perhaps the issue is not yet important enough for most repository managers to notice.

3. Multi-author papers are common in all of the repositories analyzed. Most repositories could provide a list of the total number of items deposited, and the total list of authors, but none were able to easily identify or count the total number of multi-author papers. This is a significant shortcoming in the reporting functions for repository software, because knowing such publishing patterns is key to understanding the impact of organizations and individual scholars. Customized reports would have been necessary to determine what percentage of each repository's total deposits are multi-author papers. Fortunately, this issue did not affect the research questions asked in this study, for an author's total number of contributions to a repository is still available, whether as sole author or as part of a group. However, knowing whether a paper has a sole creator or was authored by a group would be a useful metric for evaluation. Sale (2006) also identified this issue as a complication when the creators of a multi-author paper must decide where it should be deposited, especially if authors work for different organizations.

4. Only a small number of institutions require scholars to deposit copies of all their research output like papers and technical reports. Of five mandatory-deposit institutional repositories considered for this study, two were too small and another contained dissertations and other content that would have skewed distributions toward a 1:1 author:items ratio. Only two mandatory-deposit institutional repositories were consequently analyzed. The patterns identified within these two repositories indicate a possible and intuitively logical difference from disciplinary and voluntary-deposit institutional repositories, but the sample size is too small to be reliable.

5. The age of a repository, and the age of items it contains, can significantly confuse any analysis and comparisons of scholarly repositories.

**Findings**

Analysis of the selected repositories yielded the following results:

1. *Cumulative Findings*
The 11 analyzed repositories ranged in size from 505 to 6,244 items
The number of authors in each of the 11 repositories ranged from 828 to 7,516
Authors in these repositories each deposited between 1 item and 215 items
Combined, the 11 repositories contained 27,522 deposited items by 29,388 authors

In the core group of 11 repositories, aggregated data showed:
- 21,236 of 29,388 authors (72.3%) contributed 1 item to an individual repository
- 3,660 of 29,388 authors (12.4%) contributed 2 items to an individual repository
- 1,537 of 29,388 authors (5.2%) contributed 3 items to an individual repository
- 758 of 29,388 authors (2.6%) contributed 4 items to an individual repository
- 2,197 of 29,388 authors (7.5%) contributed 5+ items to an individual repository

The following sections detail the study’s findings by each repository category, with separate analysis devoted to the AgEcon Search and arXiv.org repositories.

2. Voluntary-Deposit Institutional Repository Findings

Data from 6 repositories, collectively containing 14,829 deposited items by 18,326 authors, were analyzed. On an aggregated basis:
- 13,588 of 18,326 authors (74%) contributed 1 item to an individual repository
- 2,362 of 18,326 authors (13%) contributed 2 items to an individual repository
- 944 of 18,326 authors (5%) contributed 3 items to an individual repository
- 480 of 18,326 authors (2.7%) contributed 4 items to an individual repository
- 952 of 18,326 authors (5.3%) contributed 5+ items to an individual repository

Figure 2a. Breakdown of Contributions by 18,326 Authors to Voluntary-Deposit Repositories
Summary Notes:

Nearly three-fourths (74%) of the authors listed in these repositories each had contributed only one paper to an individual digital archive. The remaining 26% of authors were each responsible for from 2 to 156 papers.

3. Mandatory-Deposit Institutional Repository Findings

Data from 2 repositories, collectively containing 5,920 deposited items by 4,167 authors, were analyzed. On an aggregated basis:

- 2,561 of 4,167 authors (61%) contributed 1 item to an individual repository
- 625 of 4,167 authors (15%) contributed 2 items to an individual repository
- 261 of 4,167 authors (6%) contributed 3 items to an individual repository
- 149 of 4,167 authors (4%) contributed 4 items to an individual repository
- 571 of 4,167 authors (14%) contributed 5+ items to an individual repository
Figure 3a. Breakdown of Contributions by 4,167 Authors to Mandatory-Deposit Repositories

Figure 3b. Distribution of 4,167 Authors by Number of Deposits

Summary Notes:

61% of the authors in these 2 repositories were responsible for only one title each. The remaining 39% contributed from 2 to 215 items. Compared to the data from voluntary-deposit institutional repositories, Figures 3a and 3b indicate a significantly greater tendency for authors listed in mandatory-deposit institutional repositories to have
contributed more than one item to the repository. Expressed another way, these authors are much less likely to have only one item in an individual repository, and are more likely to be represented in the positive skew of the distribution. The distribution curve in Figure 3b shows more authors distributed on the positive side of the curve.

4. Disciplinary Repository Findings

Data from 3 disciplinary repositories, collectively containing 6,773 deposited items by 6,895 authors, were analyzed. On an aggregated basis:

- 5,087 of 6,895 authors (73.8%) contributed 1 item to an individual repository
- 673 of 6,895 authors (9.7%) contributed 2 items to an individual repository
- 332 of 6,895 authors (4.9%) contributed 3 items to an individual repository
- 129 of 6,895 authors (1.9%) contributed 4 items to an individual repository
- 674 of 6,895 authors (9.7%) contributed 5+ items to an individual repository

![DISCIPLINARY REPOSITORIES: DEPOSITS PER AUTHOR AS % OF REPOSITORY](image)

Figure 4a. Breakdown of Contributions by 6,895 Authors to Disciplinary Repositories
Summary Notes:

Data from the three analyzed disciplinary repositories show the vast majority of authors (nearly 74%) are responsible for only one item. In a manner to similar to the other repository categories discussed earlier in this article, the remaining 26% of listed authors contributed from 2 to 96 items, with the number of contributions per author dropping off rapidly after two papers to a "long-tail" distribution (Anderson, 2004).

5. Additional Analyzed Disciplinary Repositories Findings

In addition to the 11 repositories described in sections 1-4 of this study's findings, the authors collected and analyzed data from two other large, well-established disciplinary repositories. Because of each repository's size, longevity and other distinguishing characteristics, data from each archive was analyzed separately using the same techniques applied to the 11 core repositories. Following are the results of each analysis.

AgEcon Search Repository Analysis

The AgEcon Search repository contains research papers and reports in the broad field of agricultural economics. Author and contributions data was obtained with the
cooperation of the repository's managers. On the date this repository was analyzed, it contained 24,569 deposited items by 19,700 authors. According to this data:

- 12,781 of 19,700 authors (64.9%) contributed 1 item to the repository
- 2,917 of 19,700 authors (14.8%) contributed 2 items to the repository
- 1,198 of 19,700 authors (6.1%) contributed 3 items to the repository
- 693 of 19,700 authors (3.5%) contributed 4 items to the repository
- 2,111 of 19,700 authors (10.7%) contributed 5+ items to the repository

![Agecon Search Repository: Deposits per Author as % of Repository](image)

Figure 5a. Breakdown of Contributions by 19,700 Authors to AgEcon Search

**Summary Notes:**

As with other digital archives examined for this study, the majority (nearly 65%) of authors listed in AgEcon Search have only 1 item in the repository. The remaining 35% of authors were responsible for between 2 and 166 items. Compared to the other disciplinary repositories described in section 4, however, AgEcon Search authors were significantly more likely to have 2 or more contributions listed.

**arXiv.org Repository Analysis**

The arXiv.org repository contains current research papers and related materials from the communities of researchers in Physics, Mathematics, Computer Science and Quantitative Biology. Author and contributions data was obtained with the cooperation of the repository's managers. Data from this repository were extensive and represented 406,857 deposited items by 105,131 authors. According to this data:

- 41,869 of 105,131 authors (40%) contributed 1 item to the repository
- 19,183 of 105,131 authors (18%) contributed 2 items to the repository
11,364 of 105,131 authors (11%) contributed 3 items to the repository
7,596 of 105,131 authors (7%) contributed 4 items to the repository
25,119 of 105,131 authors (24%) contributed 5+ items to the repository

**Summary Notes:**

Administrators of arXiv.org were very responsive to requests for data, but warned us that categories such as "author", "depositor" and "item" are less clear in the complex group of research materials and depositors represented within this repository. Comparing arXiv.org data with other repositories introduced new areas of uncertainty into this study, and therefore should be considered with caution.

Keeping this caveat in mind, one can see in the arXiv.org Figure 5b a pattern noticeably different from other analyzed repositories, with only 40% of listed authors having only 1 item in the repository. Nearly one-fourth (24%) of listed authors, in fact, are each responsible for 5 or more titles in arXiv.org, with the number of contributions per author ranging from 1 to 446 items.

**Discussion & Future Work**

The findings presented in the preceding sections would tell a manager several useful facts about a repository, including:

- The range of contributions per author from lowest to highest number
- The general spread or distribution of contributors within this range
- Rates of participation that can be expected from most scholars
- Other data that would be required for more in-depth analysis of a repository
According to data from the 11 core analyzed repositories, contributions to non-mandatory institutional repositories and disciplinary repositories can best be characterized as widespread but shallow. The distributions patterns for both of these datasets were surprisingly similar, and raise more questions. Would other disciplinary repositories show similar results? Unfortunately, obtaining this kind of data for analysis was very difficult; many repositories provide automated reports for online users to browse repository contents by year, topical group, or even author, but seem to have taken extra efforts to not provide correlations between individual authors and their total number of papers deposited. Responses by many repository managers to requests for author:items reports indicated their concern over releasing such information. Nonetheless, examination of a greater range of disciplinary repositories was warranted, and led the authors to acquire and analyze the data described in this article for the AgEcon Search and arXiv repositories. Data from Ag-Econ Search tend to support the validity of the patterns discovered in the 11 core repositories. Data from arXiv.org shows a very different pattern, but the reasons for this variance are likely explained by one of the many issues identified by the authors with counting and comparing figures without the benefit of common categories, terminology, or reporting standards among digital scholarly repositories.

As for mandatory-deposit repositories, the limited available data indicate authors represented in such repositories tend to contribute more of their intellectual output. Sale (2006) predicted institutions establishing deposit mandates were likely to see such results within three years of implementing these policies. Harnad (2006) cited surveys showing 95% of scholars comply if their university mandates depositing in an institutional repository. This study’s findings only reinforce such predictions and arguments favoring institutional mandates. As the data in this article show, a mandate is arguably the "tipping point" described by Gladwell (2000) that can make depositing behavior among scholars not just widespread, but also more of an ingrained and complete behavior.

Mandates for Open Access and deposit are proliferating internationally, and are sure to create a noticeable impact upon institutional and disciplinary repositories in terms of the characteristics analyzed in this study. Intuitively, one would expect the average number of contributions per author to increase in many repositories. However, the overall number of participants is likely to increase also. What effect will this have in overall distributions and relative measures of participation? This question will be a fertile area for evaluation in the future. These studies, along with recent articles like Carr & Brody’s (2007) report on deposit profiles among digital repositories, affirm the need for continued research such as:

1. Extending the methods and research questions of this study to more of the repositories listed in the OpenDOAR registry.

2. Performing more qualitative analysis to particular repositories to learn who is actually depositing items into repositories for authors, especially in the case of multi-author papers.
3. Applying social networking theory methodologies as proposed by Lopez-Fernandez (2004) to look for social factors influencing decisions to participate in repositories.

4. Analyzing correlations within individual and aggregated repository datasets between variables such as total number of deposits, publication dates and deposit dates, and disciplinary affiliations of contributing authors.

5. Conducting overlap analysis for authors' works in institutional and disciplinary repositories.

6. As better categorization and evaluative frameworks emerge for repositories, applying these at the early stages of the processes described in this study.

Conclusion

This study revealed a heterogeneous universe of repositories. Many of these repositories defy easy classification into the groups defined for this study. Of the repositories listed in the OpenDOAR registry, many are less transparent than one might prefer when trying to analyze characteristics such as participant distributions. The large number of complications encountered during this study indicates the Open Repository community might benefit by endorsing some sort of standard set of harvestable reports for all repositories, similar to those emerging for other scholarly databases.

Despite the difficulties in categorizing repositories and their content, and obtaining needed datasets, the participant distribution analytical techniques used in this study were valuable for the new perspective they provided on individual and grouped repositories. At a time when too few measures of success are available for repository implementers, the patterns shown here should be of great interest to local repository managers. One also can imagine the additional precision and certainty that would be gained if these techniques and statistics were not tabulated manually for a limited group of repositories, but instead were part of routine analysis reports run on massive repository metadata harvests using the OAI-PMH protocol. However, for such efforts to be possible, many of the uncertainties involved in comparing repositories need to be addressed first.
Citations


