Factors to Assess Self-Archiving in Institutional Repositories

Rutgers University has made this article freely available. Please share how this access benefits you. Your story matters. [https://rucore.libraries.rutgers.edu/rutgers-lib/32566/story/]

This work is an ACCEPTED MANUSCRIPT (AM)

This is the author's manuscript for a work that has been accepted for publication. Changes resulting from the publishing process, such as copyediting, final layout, and pagination, may not be reflected in this document. The publisher takes permanent responsibility for the work. Content and layout follow publisher's submission requirements.

Citation for this version and the definitive version are shown below.

Citation to Publisher

Citation to this Version:
Factors to Assess Self-Archiving in Institutional Repositories
Jingfeng Xia and Li Sun

Abstract – This paper proposes a group of factors that may be used to assess the success of open access self-archiving. It concentrates on self-archiving in institutional repositories. The authors emphasize the importance of examining content materials, particularly the availability of full text versus abstracts and the deposits archived by authors versus by others.

Introduction

After several years of practical implementations, institutional repositories (IRs) have developed into the stage where assessment has become necessary. Good assessment will help IR professionals conclude what has been done in the operation of current IRs and point out future directions. In the past two years, assessment efforts have been scattered in IR literature. Relatively few of them were carried out in a thorough and systematic way.

Self-archiving has been the major part of IR assessments. Evaluations focused primarily on the total number of deposits in IRs, the attitude and behavior of faculty scholars in self-archiving research outcomes, copyrights, and accession policies. Most studies relied on interviews and questionnaires among faculty to collect data. Self-archiving is a complex process that necessitates the involvement of money, time, people, policy, and the like. The assessment of it should be multifarious and take on actual IR contents. Factors applied to the assessment of IRs may not be suitable for the assessment of self-archiving. This paper discusses a set of assessment factors that can cover most, if not all, self-archiving related issues and help measure the success of self-archiving in an IR environment.

Brief History of IR Literature

The invention of the EPrints software in 2000 at the University of Southampton, U.K., enabled the establishment of institutional repositories. In the U.S., DSpace as an IR application was jointly developed in 2002 by Massachusetts Institute of Technology Libraries and Hewlett-Packard Labs. Since then, IRs have experienced tremendous development. By March 2006, the Register of Open Access Repositories listed 637 IRs in the world. This figure did not count those educational institutions that had a plan or desire to implement digital repositories.

Consequently, research outcomes on IRs are abundant. Scholars at the University of Michigan identified a list of more than 200 citations pertaining to the various issues of institutional repositories. Charles Bailey, a well-known open access bibliographer, compiled a more or less similar list of IR literature. Additional articles and reports, although they may not have been published formally, are scattered on the Web in the form of pre-print.

The literature reflects various stages of IR development over time. Suzie Allard et al. carried out a literature review on a group of selected articles over a period of five years from 2000 to 2004. According to their findings, the fundamental issue discussed in almost all the studies was how to define an IR, which indicated the introduction of the new publishing paradigm in the profession. At the same time, “case
“studies” also predominated the IR literature focusing on the planning and implementation of individual IR projects. Although IR management and outcomes have been widely addressed, relatively few studies have discussed these themes in any depth.

Subsequent studies shifted their research topics from how to set up an IR to how to operate an IR. Increasingly, articles are concerned with the way that IR managers solicit content documents. Self-archiving was considered to be the most important way of accumulating content. Other issues also surround self-archiving and often involve faculty. For example, the purpose of understanding the attitudes and behavior of faculty is to work out the strategies of encouraging them to make self-depositing a reality. The purpose of clarifying copyright requirements is to avoid legal conflicts in the self-archiving of published material, and the purpose of promoting a mandate is to improve collecting documents while reducing faculty intervention. IR managers made every effort to encourage the academics to contribute their research results. In order to facilitate such encouragement, many IR managers advocated the concept of marketing their repositories – “marketing, marketing, and marketing.”

Marketing strategies varied from IR to IR. Various outreach plans, including public presentations, personal contacts, and policymaking, brought repositories to the attention of faculty authors in given academic institutions. Studies found that institutional policies have an immense impact on article depositing behavior of scholars. According to Arthur Sale, “a requirement to deposit research output into a repository coupled with effective author support policies works in Australia and results in high deposit rates.” This statement was consistent with studies by Alma Swan and Sheridan Brown, and Stephen Pinfield. To promote self-archiving and facilitate open access, some proposals were made recently by the legislatures in the U.S. and the U.K. that require scholars to deposit their research results sponsored by governmental funding agencies into digital repositories.

When IRs advanced into a new stage, research concentrations also changed. Now, not only have IR numbers increased rapidly, but their deployments and operations have also matured. Accordingly, there has been an increase of research projects on the assessment of current IRs in the literature lately. In particular, many assessments have focused on the achievements, challenges, and attitudes of open access self-archiving.

One of the assessments was the PALS report that analyzed forty-five IRs on many aspects of IR management including the cultural issues that affected faculty take-up of self-archiving services and intellectual property rights. Another assessment project was supported by the Joint Information Systems Committee (JISC) in the U.K. as an author study. The study collected data from an investigation of 1,296 respondents with the purpose of exploring the current state of play with respect to authors’ self-archiving behavior and outcomes. Similar assessments are either scattered in individual articles or currently in process.

**Assessment Research**

IR research lacks an accepted set of assessment factors. The PALS report measured IR development from the following areas: software technologies, issues affecting the practice of self-archiving, copyrights, organization, funding and business models, accession policies, metadata, long-term preservation, and
access. In another research, Mary Westell proposed a group of indicators of success, including mandate, integration with planning, funding model, relationship with digitization centers, interoperability, measurement, promotion, and preservation strategy.\(^\text{20}\) To M. Kathlenn Shearer, the input activity and use of the IR were the two major factors in determining the success of IRs.\(^\text{21}\) Similarly, Rachel Heery and Sheila Anderson developed a typology that viewed an IR from such aspects as content, coverage, user group, and functionality.\(^\text{22}\)

It is content size that has been prevalently examined in most existing projects of IR assessment as the prime indicator of the success of IRs.\(^\text{23}\) Content size is also core in the evaluation of the success of self-archiving. Findings reveal that current IRs have a rather small content size. As early as 2004, DSpace managers at MIT had noticed a slow rate of self-archiving IR documents in its repository.\(^\text{24}\) Later, the PALS report found an average of 1,250 documents per institution.\(^\text{25}\) By the end of 2005, the statistics for about 250 IRs revealed an average of 3,200 documents per repository.\(^\text{26}\)

At first the sluggish accumulation of content documents was considered to be the lack of awareness of self-archiving by scholars.\(^\text{27}\) Some researchers believed that the lack of a preprint culture in certain academic fields also slowed down the efforts of scholars to contribute to an open access repository.\(^\text{28}\) However, subsequent studies found that faculty were not enthusiastic about self-archiving no matter if they were aware of the existence of an IR or not.\(^\text{29}\) Faculty scholars simply did not have the time or interest to perform the task. Some of them would rather give the rights to IR professionals to help deposit research results on their behalf.\(^\text{30}\)

Current assessment research on self-archiving has some obvious weaknesses. First, few studies, if any, have made efforts to distinguish documents self-archived by authors from documents archived by others before conducting analysis. “Self-archiving” has been a vague term in the IR literature with regard to “self.” Researchers either take for granted to treat all IR content documents as self-archiving deposits or do not care about the distinctions. If it is acceptable that others can help authors to deposit, the different archiving modes should at least become a measurable element in the assessment of self-archiving.

Second, statistics upon which assessments were taken were mostly gathered through questionnaires, interviews, and other personal contacts.\(^\text{31}\) Seldom did studies look into content documents; as a result, the status of individual deposits was largely ignored. Analysts cared only about the total number of deposits in an IR, but not number of deposits by subject, author, or date. There also lacked analysis on the difference of deposit numbers between pre-prints and post-prints, full text and abstract, and local file and harvested metadata. Such dichotomies are, however, important to the understanding of self-archiving.

Ideally, a universally accepted methodology can be developed to standardize the assessments. If this is still an unrealistic expectation, at least it is becoming important to gear the areas in the assessments of self-archiving. Researchers should start paying attention to the status of IR content. They should take into consideration both the quantity of IR documents and the quality of them, rather than only one factor. The following section introduces some useful factors that may help researchers to re-think what new aspects of a self-archiving assessment should be and how to work on them.

**Assessment Factors**
Information of Depositor

“The self-archiving definition is not quite as cut and dried as we think and this has implications for building research communities and improving individual and field impact through economics of scale,” observed Anita Coleman and Joseph Roback. According to EPrints, self-archiving can be done by authors, by “proxy,” or by digital archivers in the researcher’s institution or its library. In current practices, many IRs have employed administrative staff to deposit articles on behalf of faculty authors. Different archiving approaches address dissimilar operational styles in the operation of IRs. It is necessary to identify deposits on their archiving methods as one of the major factors in assessing self-archiving.

Some repository systems provide information about the depositor of an article, which may or may not be the author of the article. This presentation of the depositor’s name is not broadly available in all repositories, however. The EPrints application has such a feature although not all known EPrints repositories display the information to end-users. Other repository software may or may not require similar information at deposition.

Figure 1 demonstrates how the depositor’s information is displayed. Please note that this depositor is not one of the authors of this article. His name is linked to another page where the number of documents deposited by him in this particular repository is shown (in this case, 355; see Figure 2). In some cases,
the “Deposited By” field is filled in with the name (or abbreviation) of a school/department or something that may indicate the status of the deposition in bulk load by an automated computer application.

Figure 2. Depositor with total number of documents deposited by that person

Also, the depositor’s information is an indirect indication of the attitude of faculty scholars toward self-archiving. An examination of several Australian IRs reveals that in certain subjects only a few of the faculty are active in depositing their research results although the total number of deposits for these subjects appears to be large. Furthermore, the depositor’s information has relevance to the analysis of depositing numbers of an IR. Sometimes, the depositor is a person who co-authored with others for a given article, which is particularly true for research activities in sciences, engineering, and medical sciences. This indicates a not mutually exclusive pattern of self-archiving across authors (if across subject pattern is not considered here). Hence, in any IR, the total number of deposits summarized at the IR level is always not the same as the total number of deposits summarized at the “by subject,” “by author,” or “by year” levels.

**Number of Deposits**

Number of deposits, namely the total number of deposits in an IR, has been commonly used as one of the factors for assessing the success of the IR and self-archiving. It is reasonable that content size of an IR has an important impact on the usability, and thus the value, of the IR. The more content documents it contains, the more researchers rely on it to find reference, and the more likely authors deposit their research results in it. As a result, this IR will become bigger and bigger, attracting the attention of more readers.
Number of deposits by subcategories has been largely disregarded in assessment studies. However, such data are informative in interpreting authors’ attitude, IR development, repository infrastructure, as well as many other important issues. For example, the difference of deposit numbers across departments may point to diverse disciplinary cultures of electronic publishing or to the success rate of a self-archiving campaign in an institution. The information should be valuable in future IR assessments. Similarly, the difference of deposits by document types may help evaluate the usability of an IR content because too many document types in an IR may make it too diverse to be used.\textsuperscript{38} Furthermore, the dates of depositions may help examine the operational style of an IR and the practice of mandate policy by the IR.\textsuperscript{39} Below is a list of subcategories with which deposits can be summarized and analyzed:

- Numbers by class (e.g., education, engineering, science, and social sciences)
- Numbers by subclass (e.g., biology, civil engineering, and sociology)
- Numbers by department/faculty (in some cases, this category is not corresponding to a subclass, e.g., an interdisciplinary unit)
- Numbers by version (pre-prints, post-prints, etc.)
- Numbers by type (report, journal article, conference paper, working paper, etc.)
- Numbers by date (depositing date, mostly by year)
- Numbers by depositor (e.g., author versus non-author)
- Numbers by availability (e.g., full text, abstract, harvested)
- Numbers by location (e.g., local file, external link)
- Numbers by any other categories classified by a particular repository.

\textit{Availability of Full Text}

The availability of full-text documents is also a determinant of the success of an IR. This has become an issue along with the increase in abstracts obtained by harvesting metadata from outside data providers. Metadata harvesting will expand the content size of an IR, making it look useful; however, it is the availability of full text that makes an IR really useful.\textsuperscript{40}

First, if a document is an abstract, regardless if it is available through a harvester or local deposition, its usefulness will be reduced. End-users care much more about reading a full-text article than knowing briefly what it talks about. They will have to look for the full text in other online locations if they find an article interesting by reading its abstract. That extra job may leave them frustrated and cause them to lose patience at this IR. Any reference librarian knows how hard it is for an inexperienced reader to discover a full-text article online.

Second, if a document only has an abstract in an IR but provides a link to an external location to the full-text version, this remote availability may or may not increase the value to the IR. In the short term, such a link can help end-users quickly find materials. In the long run, the link may become a dead link because the destination Web site no longer offers free access, the URL is no longer valid, or the like. Persistent URL can only partly solve the problem. When such a situation becomes frequent, end-users may detour their frustration to the departure databases—the original IR site. The authors took a preliminary assessment of several EPrints repositories in Australia and the U.K. and found that in some IRs more than half of the deposits have an abstract only and provide an external link (see Figure 1). Unfortunately, a certain percentage of these links has already become unavailable. The findings will be published in another report.
Locally deposited full text can be in a variety of formats. A sample list of the formats includes the following file types: .pdf, .doc, .htm, .html, .ppt, .ps, .ASCII, .zip, .gz, and .bz2. The default is .pdf. “Most authors who choose to self-archive a post-print, self-archive the publisher PDF.” In practice, a Microsoft Word or PowerPoint file may not be displayed on a computer if the computer does not have the Microsoft Office suite. Likewise, an HTM/HTML file may not be nicely viewed when Internet Explorer rather than Fire-Fox is the browser, or vice versa.

Authors’ Attitudes

In the existing projects of IR assessment, authors’ attitudes toward depositing articles are a common topic with much information obtained from interviews and questionnaires. For example, Swan and Brown carried out an investigation among 1,296 respondents and then analyzed the data to reach an understanding on how authors reacted to IR self-archiving. Many articles discussing individual IR practice also introduced investigations on faculty attitude in their own institution.

As discussed above, quantitative data such as the numbers of deposits by subject, author, and depositor can help divulge authors’ attitude. In addition, the existence of a mandate policy in an institution is also critical to the practice of self-archiving by faculty authors. A comparison of the total numbers of deposits among several Australian IRs revealed that the Queensland University of Technology (QUT) has the most content documents. QUT set an official requirement for its faculty to self-deposit their research outcomes into its institutional repository. The authors’ assessment also found that QUT has a high rate of full-text availability in its IR database.

An on-going task exists for IR professionals to promote self-archiving among faculty, although studies have discovered a negative attitude of faculty toward participating in the adventure of the new electronic publishing. For a long time, how to secure the participation of faculty will still be an important topic in research. Future assessment projects should try to work out different ways of collecting such data.

Cost per Deposit

The calculation of cost per deposit needs to consider both the total number of content documents and the total cost of an IR development. The cost of IR development may comprise investments in IR start-up, maintenance, activities, and long-term preservation. It may also be measured by looking at expenses in machines and personnel. Either way, IR development involves a very complicated calculation with a very inaccurate figure.

Susan Gibbons notes that “an IR system with attractive services and a strong preservation commitment is not a cheap investment.” Some reports have provided estimates of expenses for individual repositories. For example, it was estimated that the DSpace repository at MIT costs about $285,000 per year covering personnel and systems. The operation of the IR at the University of Oregon costs anywhere from 2,280 to 3,190 staff hours during the fiscal year of 2003–04. And, the University of Rochester spent an estimated $200,000 in its repository from October 2003 through September 2004.
One way of measuring cost per deposit is to divide the total number of an IR content documents by the total number of the IR expenses. Nonetheless, this measure may remain dubious because not only is the total expense a rough estimate, but also that monetary value is not a universally accepted factor for everything. Digital repository as a revolutionary yet experimental way of scholarly publishing has the potential to alter the tradition of information acquisitions and dissemination. Its significance goes beyond monetary numbers. Cost per deposit may not be a valuable assessment in this regard.

Usage Assessment

One of the major concerns for an IR is its usability. Use of a repository’s content “is likely to have an affect [sic] on input activity.” The more its content documents are retrieved, the better usability the IR has. Any effort made by IR managers to attract deposits is with the hope that bigger content can translate into bigger usage.

One way to assess usage is to carry out article citation analysis. Studies found that moving an article online and making it freely available will increase its opportunity to be read and cited. This finding may help convince faculty authors to contribute their research results to digital repositories because scholars have a desire to make their research widely visible to peers.

Citation analysis on self-archived articles in IRs can be a difficult task. Because digital repository is new, the publishing communities still need time to pattern its citation style. Citation analysis is also not yet ready to examine different styles of citations for IR documents. This situation is further complicated by different self-archiving practices, namely the availability of full text versus abstract and local deposit versus remote link. In fact, this can be a totally different area of research.

Another way to assess usage is to rely on the server’s log files of an IR. Most IRs provide usage statistics online that may have been analyzed and formatted by a commercial Web analyzer or its own programming software. Be aware that most analyzers have their technical limitations, and the log data may skew an analysis and make the results unreliable. For example, the IR managers may contribute significantly more to logs than ordinary users due to their maintenance routines on the repository.

Many research projects have been launched to assess the usage of digital collections. One of the projects particularly designed to investigate and develop a joint approach to Open Access Initiatives (OAI) statistics gathering and sharing was Interoperable Repository Statistics, which was funded by the Joint Information Systems Committee. This project attempted to standardize analysis on Web log usage data by interpreting, comparing, and aggregating data across different servers. It aimed to cover all OAI repositories and measure usage of all types of content documents in these repositories. Future research on usage assessment may grab information and collect statistical data from the Web site of this project.

Interoperability

Interoperability as an important factor for assessing the significance of IR development has been broadly emphasized. Its importance in the assessment of self-archiving is also obvious. Most of the current repository applications are OAI-Compliant so that IRs implementing these application systems can both act as data providers and support federated search by adopting a data harvester. Such a harvesting
mechanism can bring in either abstracts or full-text documents from other OAI-Compliant repositories, although metadata harvesting predominates the practice.

Other Factors
Any comprehensive assessment of self-archiving can be complicated. In addition to factors mentioned above, some other issues are also important, such as copyright, quality control, staff support, and software management. Also, it is not always easy to distinguish the assessment of an IR from the assessment of self-archiving for that IR. Therefore, it is essential that we begin discussing some set of criteria by which to measure the effectiveness and challenges of self-archiving.

Conclusion
A thorough and well-organized assessment can provide valuable suggestions to keep self-archiving healthy. This paper compiles a group of factors useful in the assessment of self-archiving. Such factors include those that have been incorporated into previous assessment projects as well as those that are new to studies of this type. For example, this paper recommends the evaluation of the availability of full-text documents and the examination of deposit numbers by different subcategories. It also emphasizes the importance of distinguishing deposits contributed by authors from deposits made by other people.

No one expects that all the proposed factors will be used in a single assessment project, although it is not impossible. The most important point is that the assessment efforts can become more valuable if they are undertaken from different perspectives.

Notes

15. In May 2006, U.S. Senators John Cornyn (R-TX) and Joe Lieberman (D-CT) introduced legislation to require research results, sponsored by U.S. government agencies with research portfolios of over $100 million, to be deposited in a proper digital repository for public availability within six months of their original publication. The U.K. reaction can be found in The United Kingdom Parliament’s Science and Technology – Tenth Report, Session 2003-04, http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/39902.htm.
17. Mark Ware, Publisher and Library Learning Solutions (PALS): Pathfinder Research on Web-based Repositories—Final Report (Bristol, UK: Publisher and Library Learning Solutions, 2004).
25. Ware, Pathfinder Research on Web-based Repositories—Final Report; See also Ware, “Institutional Repositories and Scholarly Publishing.”
30. Ibid.
31. Swan and Brown, Open Access Self-archiving: An Author Report; Foster and Gibbons, “Understanding Faculty to Improve Content Recruitment for Institutional Repositories.”
38. Ibid., 99.
39. Ibid.
43. Foster and Gibbons, “Understanding Faculty to Improve Content Recruitment for Institutional Repositories;” Mackie, “Filling Institutional Repositories: Practical Strategies from the DAEDALUS Project.”
46. Gibbons, Establishing an Institutional Repository, 56.
47. Barton and Walker, MIT Libraries’ DSpace Business Plan Project.
49. Ibid.
52. Foster and Gibbons, “Understanding Faculty to Improve Content Recruitment for Institutional Repositories.”