A brief literature review on the usability of institutional repositories

Dana McKay, dmckay@swin.edu.au

1. Introduction

Institutional repositories (IRs) are said to be “essential infrastructure for scholarship in the digital age” [40], lessening the access barriers to scholarly communication [15, 26], providing a measure of institutional research prestige and visibility [15], and facilitating formal research assessment [15, 18, 29]. Why, then, with such large potential benefits, have IRs been less frequently implemented [41, 64], harder to fill [9, 17, 21, 36], and less visible [36, 47, 65] than their proponents would have hoped?

While there has been a lot of research time poured into technical platforms such as DSpace\(^1\) [61], ePrints\(^2\) [24], and the VITAL interface to Fedora\(^3\) [62] used by the ARROW\(^4\) project [66], very little is known about the users of IRs, nor how benefits of IRs to users can be maximised. IRs have three main user groups:

1. **Authors**
   Authors are a vital part of any IR, particularly if the approach of the IR in question is that authors self-archive. Even in the case of assisted depositing, however, if authors do not contribute to and do not use an IR it cannot be said to be reducing the barriers to scholarly communication. Authors are a particularly well studied group, and are discussed in depth in Section 2.

2. **Information seekers**
   Information seekers, while well understood in many other contexts, have not been studied as users of IRs. Users of IRs may all be researchers, or may include members of the general public, or may be non-existent, we simply don’t know who they are or what they do, because they have not been studied. Nonetheless, conclusions can be drawn about their information needs from studies of other research information sources, literature on information seeking behaviour, and usability studies. Information seekers’ needs are discussed in depth in Section 3.

3. **Data creators/maintainers**
   IRs create a role not seen before in information systems, that of the data creator and maintainer. This role is often filled by librarians, though if a self-deposit protocol is used authors assist with data creation as well. This group is not well studied, but what research there is on them will be covered in Section 4.

Potentially research assessors could be considered a fourth user group, but their needs are likely to be well-defined by their task, and the usability aspect of their role is likely to be well covered by the usability requirements for information seekers, so they will not be dealt with separately here.

\(^1\) http://www.dspace.org
\(^2\) http://www.eprints.org/software/
\(^3\) http://www.fedora.info/
\(^4\) http://arrow.edu.au
In Section 5 conclusions will be drawn about the usability and usability requirements of IRs, and directions for future work in this field will be suggested.

2. Authors

Authors are a particularly well studied user group, both with respect to their publication preferences, and with respect to institutional repositories. The open access movement has commented extensively on the role of authors, and the potential benefits of IRs to authors, though often they have not taken documented author concerns into account. Here we discuss lack of author participation in IRs, the reasons for it, and ways it can be addressed.

Authors’ primary goal in their work is to do research, with all that implies; they want to read and write, they want to investigate, and they want to share the results of their work [21]. They are very attached to the validation provided by peer-review [57], though often the time-to-press in journals is considered a problem [51]. Authors in some fields circumvent this problem by using subject repositories such as arXiv.org [54], which though they have advantages such as immediate dissemination and “informal peer review” may fall afoul of publisher policies in some cases [8, 11] (and pre-prints are often not cited, meaning work needs to be formally published to gain professional recognition in the form of impact [68]).

The majority of authors claim not to be affected by access issues [17, 58], though many of them have at one time or another used self-archived works [47], and indeed a number of them self-archive on their own websites [4, 42]. Most authors have positive attitudes to both gold (open access journal publishing) and green (author self archiving) [26] open access in principle [57, 60], however they do not wish to engage in the author-pays model unless someone else will pay [3, 60], and often are not even aware whether or not their institution has an IR [2, 47]. Most are prepared to self-archive if required to do so [59, 63], though deposit mandates and author amenability do not always result in deposited documents in an IR [9, 42].

Authors’ primary goals and above average access to research materials exacerbate authors’ low level of motivation to contribute to IRs. Authors’ barriers to deposit in IRs seen in the literature as being a number of factors:

- Concerns about plagiarism [3, 17] (which are sadly not unfounded [19, 37]). Correct attribution of their work is one of only three restrictions authors would place on the use of their work if setting restrictions was left to their discretion (the other two conditions are that exact copies of the original work must be used, and thirds parties should not sell their work) [22].

- Fears of copyright and publisher policies, notably whether IR deposit breaks copyright agreements authors have with journals [17], and whether publishers consider IR deposit to be “prior publication” [8]. Authors seldom read copyright agreements they sign with journals, and often neither know nor care what they have agreed to [47, 58, 59]. Having said that, authors are prepared to allow more access to their work than copyright law in the UK [22] and Australia [20], (and indeed more than readers would expect [23]).

- Ignorance of the IR’s existence [2, 47] (even in cases where the author may have had material deposited by someone else [36])

- Lack of time and inclination in a busy schedule, where IRs are not seen as directly benefiting the author’s primary goal (performing research) [21, 67].
• Lack of knowledge of how to deposit material, and anxiety about the technological process required to deposit material (lack of computing confidence) [9, 21].

• A perception that material in an IR is second-rate material unable to be published elsewhere (this probably stems from the belief that everything published in a journal is subject to copyright and not available for deposit), or that IRs are for digital theses and working papers only [17].

• The feeling that the IR has no benefit for the author, as it reaches the institution, rather than the research community (this is often coupled with the use of a subject repository to reach the author’s own community) [17].

While these concerns may or may not hold true, they must be addressed if IRs are to be successfully filled and made a useful resource.

There are a number of strategies for dealing with the concerns outlined above that have emerged in this literature review (Mark and Shearer reached the same conclusions [46]):

• Education and publicity about the IR tailored to the individual author’s research discipline, concerns, and technical savvy (this can usually be done at a departmental level) [9, 21]. This education could include information about the increased impact of open access papers [5, 27, 39] and any search engine harvesting of the IR (search engines are one of the primary means by which academics look for and find research papers [47, 57], and search engine harvesting has been shown to greatly increase the use of online journals [60]).

• A mandated deposit policy (thus making deposit of authors’ research outputs part of their job) [9]. Research shows that most authors are not averse to a mandate [58, 59, 63].

• Assistance in the deposit process; either technical assistance or indeed doing the work for authors; where possible information about publications should be gathered from aggregate sources (such as open access journals) [9, 42].

• Value-added services for authors including researcher homepages or profile pages, generation of publication lists, and article download statistics [9, 21, 56].

In combination, these strategies have been shown to increase author buy-in to IRs, alleviate authors’ fears about IRs, and increase the rate of growth of IRs.

While there are readily available strategies to fill IRs, it is important to also take note of the things authors require when they use research materials. First and foremost IRs must be harvested by search engines; they are a primary tool used for finding research [47, 58]. Authors prefer to download PDF documents, and ideally they prefer the journal version of the PDF [47]. When referencing self-archived documents, most authors prefer to reference the journal version of the document [47] (and often they are required to do so by journal publishers, many of whom will not accept papers that reference pre-prints [8]).

Authors are the best-studied user group of IRs, and their needs are very clear. Their primary goal is to do research; peer review is very important, as is access to research, and most important is recognition for their own work. To encourage author engagement with IRs, fears about copyright and plagiarism must be addressed, and engagement must be made worthwhile by mandate and value-added services. Further researcher interest in IRs can be generated by directing researchers from outside the institution to IRs with search engines, and providing
them with PDF documents (preferably journal PDFs) and complete journal citation information.

3. Information seekers

While information seekers are well studied in general, and there is even research on how they find things in research-based online collections there is very little information on they ways information seekers interact with IRs. There are a couple of usability reports about IRs, but there is no usage data; we do not know whether IR end users are primarily institutional, or whether they come from outside the institution, we do not know how they find IRs (for example whether IRs are found via the institutional homepage, or via Google search results), and we do not know what types of activities they engage in while visiting the IR. Nonetheless, we can draw some general conclusions about end-user interaction from other research.

Studies of online journal database usage have defeated a few myths. Contrary to expectations, it has been found that while half of all journals account for most of the use, there are very few journals and articles that remain completely unused [48]. Moreover, while older articles are less used than recent articles (and certainly less cited) searches that include older articles in their results will often result in the download of these older articles [31]. This extensive use of available materials suggests that IRs should include all available material, not just the newest materials or those expected to be most popular. User behaviour in these databases included a mixture of searching and browsing, and users typically viewed between one and three articles. Most users visited only once in a three month period and only a handful visited more than five times; short sharp interactive bursts are clearly the norm [48-50]. While this cannot tell us much about how to design IRs, it does suggest that users may be infrequent visitors who collect what they need and then leave again.

Transaction log studies of user behaviour in open access research collections (see [34, 43, 44] for all the results discussed in this paragraph) showed consistent information seeking behaviour. Users did not change default search settings, nor did they know how to use many search options (notably Boolean logic). Again, visits were short and sharp, as were queries; the vast majority of queries were 1-3 words long. Users did not refine their queries, and chose documents from the first page of results more frequently than anywhere else. Users often looked specifically for given types of metadata, for example author names or institution names. Few spelling errors were noted in query strings, but technical terms posed some problems with spelling (“modula2 vs. modula-2 vs. modula 2” was the example given). This investigation of search behaviour suggests that insofar as possible, IR software should not require users to use complicated search formulas, and that the default settings for search should provide the most relevant results in the first page of results. Fielded search and search string suggestions are also recommended to allow users to maximise their search power without having to refine their query strings.

While we can assist users in their searching, it is not the whole answer to information seeking; there are a number of models (for example [38, 45]) of an information seeking process that includes but is not limited to search. When information seekers have ill-defined information needs, they engage in this process until they find the information they are looking for (or they find something that is close enough [1], or they give up [52]). This process (according to the models) is in approximately six stages:

1. The information seeker determining that they have an information need
2. Investigating how the information need might be met, including assessing the available information sources (this process may involve some searching or browsing in information sources to get preliminary results)

3. Crystallizing their information need based on investigation of the available resources and personal inclination

4. Querying and information sources to meet their need

5. Investigating query results

6. Assimilating results, and refining their search if their need has not been met.

This process is often iterative, and it is well supported by, for example, librarian interactions [14, 52]. While it is not feasible to provide librarian interactions for every query of an IR, it is advisable to provide flexible browsing for both the entire research collection, and search results. As the process is iterative, search and browse should be able to be interleaved (this type of interleaved behaviour is seen in online journal databases [49], to add weight to this argument).

There have been limited reports of user testing of IRs with respect to end users, and only DSpace and ePrints have been tested. One test, which used a number of known-item search tasks and included a heuristic analysis of the software and an in depth analysis of the reasons for results, concluded that overall users were faster and more successful in their searches with DSpace than ePrints [35]. In contrast, another report concluded that DSpace did not support some of the real world tasks of the local users, and that the terminology in the DSpace interface was confusing [53] (terminology is a notorious problem in library software [32]). These contrasting reports highlight two issues: That IR software is not especially usable for the end user, and that made-up tasks often will not reflect real usage of a piece of software (for an even stronger indication that real-world behaviour results in unintended uses of software, see [14], where librarians use OPAC results collaboratively with patrons). It is best to begin study of software (and understand design needs) using real-world studies such as ethnography, and transaction log analysis and drill down to using lab-based study to solve specific problems.

While we know a lot about information seekers, and even their behaviour in online research databases, we would be well advised to study IR-using information seekers in some depth. This study could add to the understanding we already have of general search and browsing behaviour and result in some IR-specific usability and functionality recommendations.

4. Data Creators and Maintainers

Data creators and maintainers are a well defined group, usually including librarians and, in the case of self-deposit, including authors as well. While we know who is in this group, there has been very little study of the ways in which they use IR software and their usability needs. There is no comparable role to that of data creator and maintainer in any other system; authors who self-archive or deposit to subject repositories have been highly motivated enough to ignore usability problems [54], and actively disseminating information (as opposed to transferring it) is an entirely new role for librarians [13].

IRs are generally seen as positive developments by libraries, providing scope for research affiliations with other libraries [56], greater visibility [7] and the provision of new and useful services to researchers [33]. Moreover, libraries are well placed to offer services such as the creation of metadata, liaison with authors, and collation and cataloguing of materials [12, 30,
Nonetheless, this new role for libraries may eat into staff time [13], and libraries are not predicting the savings on journal subscriptions that other open access proponents are [25]. Furthermore, authors have consistently displayed a low level of motivation to create IR data [9, 21, 42].

It is of concern, then, that the usability reports of the administrative components of IR software (again, focusing on ePrints and DSpace) are resoundingly and universally negative. They highlight flaws in both interfaces, the most problematic of which are listed here:

- Audience-inappropriate terminology; the words on the interface were often not from the right vocabulary, whether the user group was librarians or authors [6, 16, 53].
- Rigidly linear process; to deposit or edit an item requires click-through of a number of screens, and pages where no edits are needed must still be clicked past [6, 16, 53].
- Low-power searching and browsing (DSpace); in its native form, DSpace did not allow librarians to perform many reporting functions, because they could not search with the required criteria, nor sort properly by date [6].
- Difficulties with suppression (DSpace); it was hard to suppress material not meant to be available to the public [53].

It should be noted that, while DSpace is more criticised, it is also more extensively tested than ePrints (though in the test that compared the two, a slight preference for ePrints was seen among users [16]).

It is also of some concern that although it has been commented that it “should” take authors less than an hour a year to self-archive [10] (and authors have been criticised by open access proponents for not self-archiving more [28]) that there has been no investigation of how data creation and maintenance fits into the work practices of authors and librarians, and what it is reasonable to expect of each of these groups given the other demands on their time.

Libraries show an interest in and the capabilities for the role of IR data creator and maintainer, however there is little study of how this should or does work in practice. Limited study of IR data creation and maintenance interfaces have shown that they are inadequately usable, which is of grave concern when adding something extra to a group’s workload. Further study is needed to understand how IR maintenance fits into the work processes of the people involved, and to engender greater usability and utility of IR software.

5. Conclusions and future work

The goals of IRs include greater dissemination of research and increased institutional prestige, both within the community and under formal assessment. These goals are not being met, however, IRs remain under-deployed under-filled and likely under-used.

An examination of the usability aspects of IRs from the perspectives of three major user groups (authors, information seekers, and data creators and maintainers) shows that IRs are not well investigated from a usability standpoint, and that this is likely to be impacting uptake.

Authors are the best investigated group, and a combined strategy for getting author buy-in to IRs is now well-confirmed. Information seekers are well studied for other information systems, and we can conclude that in many aspects their behaviour will be similar in IRs, meaning we know that we need to support searching and browsing, short queries, infrequent
visits, and “once off” searching (as opposed to searching where interface defaults are changed and queries are refined). Data creators and maintainers are poorly studied at best, and while librarians are well suited to this role and interested in being involved, the usability of the software that is available to them is very poor and may discourage further interest.

Future work falls into two categories: understanding of information seekers, and understanding of data creators and maintainers. We know very little about information seekers, and though we can guess at ways they may behave and draw design guidelines from those conjectures, it would well to have a deeper understanding of who they are and their behaviour, and thus their usability needs. Data creators and maintainers, on the other hand, are a well-defined group; what we don’t understand is how their IR work fits into their usual work processes. We also clearly need to perform further usability analysis on software they are asked to use, with the goal of better understanding their use of the software and providing software better suited to their needs.

6. References


