Found at your repository: how to make the contents of your repository visible, accessible and shareable to users worldwide

Sharon Nuria Mombrú
Senior Product Manager, Scirus (Elsevier), Amsterdam

Keywords: Institutional Repositories, Web search engines, Search technology, Cross-Repository searching

Acknowledgements
Ammy Vogtländer, General Manager, Scirus (Elsevier), Amsterdam

ABSTRACT

The number of repositories set up within and across academic and research institutes worldwide has grown exponentially. These digital archives require critical mass both in terms of content and usage in order to truly benefit the research community. The more visible their content is made to researchers, the more they will be utilised and shared, increasing the volume of submissions to these repositories and benefiting research activities. With the amount of digital content available today, being found is as important, if not more important, as being online. This paper will address how repositories can make their content more visible, including increasing findability in web search engines. With searching overtaking browsing as a means to access information, repositories also need to make their content easily accessible by implementing an optimal search capability on the repository's site, including deep and complete indexing and advanced search features that ensure that relevant results are returned. Finally, offering users the ability to search across a cross-section of repositories increases the visibility of these repositories even further and encourages and supports the sharing of research findings.

1. INTRODUCTION

Institutional Repositories, as a vehicle for holding and preserving an institute's intellectual output[1], have gained increasing attention among the academic and research community in the last three years. This movement reflects the institutes' role in shaping the changes in scholarly communication. Institutional Repositories have grown in number (close to 500 archives) and in size of the content they hold (over 3 million records)[2].

Institutional Repositories have not only increased in volume but also in variety. Traditionally representing a single institute, repositories now host content across institutes (such as NDLTD, the world's largest collection of theses and dissertations) and represent national and regional initiatives (such as DiVA, Digital Scientific Archive, a Nordic initiative including 12 institutes from Sweden, Norway and Denmark). In addition, the type of content has also grown, from conference papers, teaching materials, student projects, theses and dissertations, reports, to images and video recordings, reflecting the value of non-published content for the research community.

This growth has been encouraged by a focus within academic and research circles on policy issues and the technical aspects of setting up and managing repositories. Several open software platforms have been developed to support this movement (the most widely used being MIT's DSpace and Southampton University's GNU EPrints[3]).

However, less progress has been made on the development of a strategy to increase the visibility of the output of the institute, often cited along with preservation of materials as one of the main benefits of establishing a repository. This paper argues the importance of developing a strategy to increase the visibility of Institutional Repositories. Visibility in this context means ensuring findability on the Web, accessibility of content in the repository, and shareability of information across multiple repositories.

2. FINDABILITY

The size of the Web is growing exponentially. At the same time more and more people realise that a large part of the Web, the so-called invisible or deep Web, is not found by the main Web search engines. Estimates of the size of the deep Web vary and some claim that it might be up to 500 times the size of the visible Web[4]. Given the size of the Web, search engines often choose breadth over depth. The seemingly endless supply of content created for commercial purposes (and highly optimised for search engines) will often overshadow the scholarly results in the main search engines. Because scholarly content is not thoroughly indexed by many search engines and the content that is indexed is not necessarily given the appropriate ranking, there are serious issues for institutes using the Web as a means of disseminating information. Being online is simply not enough.

There are multiple options that Institutional Repositories can consider to make themselves more findable by Web search engines. The first, and most popular, is to standardise the technical design of a repository. The OAI-PMH (Open Archives Initiative - Protocol for Metadata Harvesting) is a means of offering structured metadata that not only facilitates interoperability, but also makes Institutional Repositories more findable and available for harvesting. In recognition of this, Web search engines have now begun harvesting OAI-PMH sites. One caveat to this indexing method is that search engines only collect metadata, ignoring the valuable content in the full-text of a document.

An alternative to OAI convergence is to create a highly structured site within the Institutional Repository, where pages guide the Web crawlers through the site. However, even if the site has been optimally structured, one should not expect more than 60% coverage by a crawler given the built-in limitations on the number of pages they will crawl per website[5]. Optimising the site for crawlers is resource-intensive for Institutional Repository managers. This has to be weighed against the expected impact.

To overcome the limitations of the above-mentioned options, Scirus has developed a partnership approach in indexing Institutional Repository content and a special process that allows all-important bibliographical data and the full text to be indexed together. This is achieved by taking the overall structure of Institutional Repository site into consideration and, for each repository, determining the best method of indexing for optimal searching. The unique indexing process developed by Scirus matches the metadata with the full text. For the end user, the results of this indexing process are more complete search results, superior ranking of results based on in-depth classification of the content and a more powerful interface with a more informative display of results.

3. ACCESSIBILITY

The second factor which contributes to the visibility of an Institutional Repository is the degree to which content is made accessible on the repository's website. A good search technology deployed on the site will determine to a great extent how accessible this content is to the repository's patrons. With the influx of information available on the Web, searching has become the primary method for users to access content, including scholarly information, rather than browsing or going through library portals. It is not only the increased volume of available information but also the trend towards cross-faculty collaboration in research that has led users to search and discover across a variety of documents and disciplines.

Implementing a good search functionality on an Institutional Repository is therefore very important. User's expectations have increased significantly. For example, users now expect full-text search capability. Furthermore, since 85% of users only view the first page of a results list, presenting the most relevant results of their search first is paramount[6]. Speed has become a de facto attribute for searching and is often associated with quality. Search queries are expected to take a fraction of a second (many search engines display the speed at which a result has been returned), regardless of the volume of information being processed. Functionality and results need to be presented in a very user-friendly interface. The increasing role of search coupled with higher user expectations makes a good search solution difficult to deliver, especially given the complexity of the content housed in Institutional Repositories

The three key features that can contribute to good search functionality on a repository's site are indexing, ranking and the display of results. The indexing process needs to be optimised for the type of content it covers. In the case of Institutional Repositories this would mean, as discussed above, having an indexing system in place that combines the metadata and full-text of a document.

While an appropriate indexing method will contribute to greater recall and precision of a search query, a sophisticated ranking algorithm is needed to ensure that the most relevant results appear first. A good ranking mechanism must consider the location of the query terms (title, author, abstract, etc.), their frequency of appearance in the document, the proximity of terms and freshness (the date of the document), among other factors. These technical aspects need to be managed as technology evolves and ranking algorithms become more sophisticated. Even in the cases where the appropriate technology might already be offered by some search solutions, specific search expertise is needed to obtain an optimal implementation and tuning of the index.

User expectations are such that the benefits of having the appropriate search technology might be lost to a poor design. The way the results of a search are displayed and the overall aesthetic are just as important to a good search tool. Only the most appropriate details of the content found in repositories, such as the author, date and title, should be displayed, avoiding extraneous details. Combining the all-important bibliographic data with a fragment of the full text document in the teaser gives the user a clear picture of the quality of the results.

Both the Scirus interface and the technology behind it are optimised for researchers and academics. By partnering with an institute to index its repository, Scirus can power the search of the repository's site ensuring comprehensive indexing, relevant ranking of results and an interface which is optimised for the repository's content. As such, Scirus Search is able to offer a solution to Institutional Repositories who want a serious and relevant search functionality to make their content more accessible to patrons.

4. SHAREABILITY

The third factor increasing the visibility of Institutional Repositories is shareability. Research has become more complex and multi-disciplinary. Users are less restrained by departmental boundaries and want the ability to search across institutes. This is compounded by the growing number of communities being formed across disciplines for joint research. More than a place for locating information, the Web has also become a social network where researchers can find each other and exchange ideas.

In response to this trend, some Institutional Repository platforms are being designed to share the contents of multiple Institutional Repositories. The larger the pool of repositories, the more these platforms will be used. Critical mass will be achieved much faster by this sharing process. Through its Institutional Repository program, Scirus can offer customised searching across a selected group of repositories.

5. CONCLUSION

In conclusion, search plays a critical role in increasing the visibility of a repository. Being online doesn't guarantee visibility – ensuring that Institutional Repositories are made visible within Web search engines and implementing a good search capability on the repository's site can be difficult to implement. The latter in particular requires expertise in search technology coupled with knowledge of the structure of scholarly material. In addition, technological developments are evolving and user expectations are rising. Partnering with the right entity will help to make this a successful effort.

6. REFERENCES

[1] Crow, Raym (2002). The Case for Institutional Repositories: A SPARC Position Paper, The Scholarly Publishing & Academic Resources Coalition, 2002
[2] Institute Archives Registry (http://archives.eprints.org)
[3] Institute Archives Registry (http://archives.eprints.org)
[4] Bergman, Michael K (2001). The Deep Web: surfacing hidden value, The Journal of Electronic Publishing, Volume 7, Issue 1
[5] According to research carried out by Scirus
[6] According to research carried out by Scirus