"On the way to a worldwide visible open access repository"

Prof. Dr. P. Schirmbacher
Humboldt-Universität zu Berlin
Computer- und Medienservice
Institut für Bibliotheks- und Informationswissenschaft

Abstract:
The different kinds of repositories offer an opportunity for Open Access publications. We are at a point where the technical requirements have generally been met and repositories exist worldwide for the various disciplines at nearly every scholarly institution. It took a lot of effort to reach this innovative condition. In order to take repositories to higher quality and greater level of acceptance, we need to enhance the networking and provide additional services inside the networks.

Introduction

The present is an exciting time, not only because of global problems such as climate change and the relationship between the rich and the poor, but also in the fields of research, communication and publication, where much is changing. From my viewpoint we face one of the biggest transitions in research communication since Gutenberg created his "movable type" in 1552.

A lot of barriers have been broken:

- digitisation solves the problem of publishing research.
- the World Wide Web answers the question of dissemination.
- economic barriers have in principle broken down, because the cost to publish and disseminate are controllable.

All in all we could say that the essential barriers to free access to knowledge have been broken. The only the people are missing – or rather the attitudes and behaviours of scholars toward this new culture of publication.

It is complicated to change a publishing culture. The reputations and careers of researchers and the productivity of a university or a scientific institute depends on the number of publications and especially on the number in high-ranked journals. There are a lot of additional criteria that characterize a publishing culture, but this paper is too brief to discuss them all in detail.

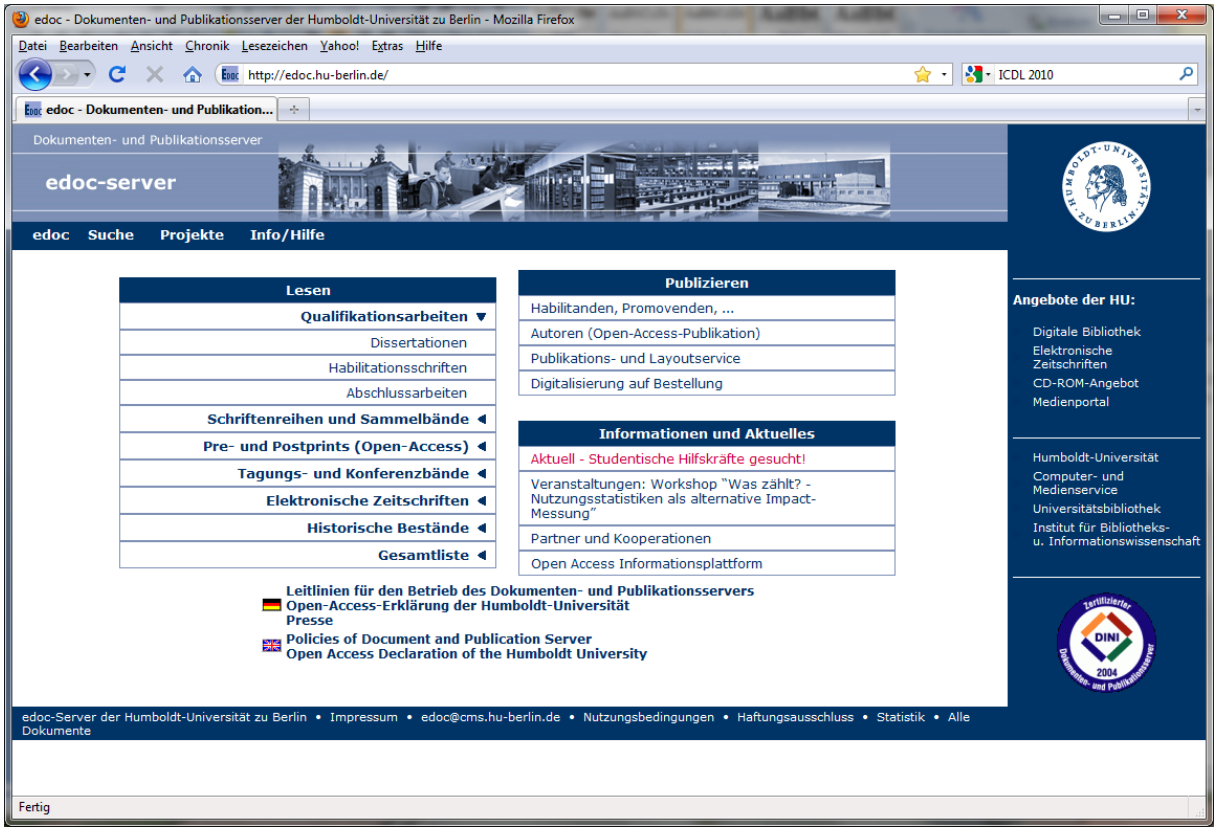
That is why in this article I will concentrate on the role of repositories. Repository-based publication is one road to Open Access, the so called "green road". Worldwide a lot of repositories exist. You can find an overview in the directory of Open Access Repositories (DOAR - http://www.opendoar.org) or in the Registry of Open Access Repositories (ROAR - http://roar.eprints.org/).

The Role of Repositories

At Humboldt University in Berlin we started with Electronic Theses and Dissertations (ETD) in 1997 and have accomplished a variety of projects since then. During the initial stages our interests focused mainly on doctoral theses, conference proceedings and speeches by university representatives. In 1998 we published our first electronic journal on Stochastic Programming.

We worked with and are still working with colleagues in Germany and Europe as well as with US universities like Virginia Tech, the University of Utah, Johns Hopkins University and others.

During the first two or three years we had a lot of difficulty in getting sufficient repository content, only about 300 documents, but it provided an opportunity to learn and to develop templates or Document Type Definitions (DTD's) for well-structured documents, like theses.

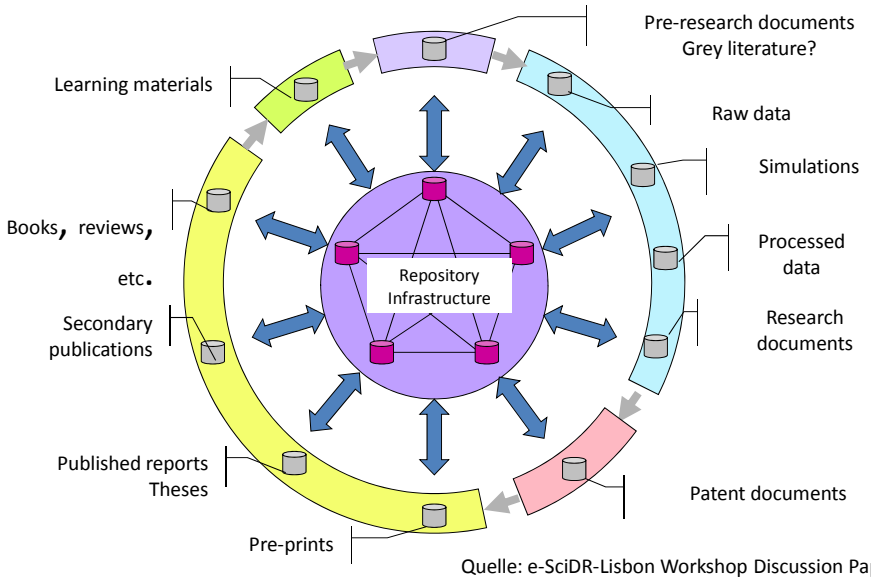


The picture above shows a screenshot of our present homepage. We have currently about 10,000 documents and have changed from being a small ETD-collection to a major Open Access repository.

On the left hand side of the screenshot you can see a column for readers and right hand side you will find a column for authors. The 3,500 doctoral theses are the largest collection. We are presently hosting 8 electronic journals, many special publications from departments across the university, conference proceedings and post-publication Open Access works by the researchers at Humboldt-University.

What is special about this as an institutional repository? From my point of view the content of this repository is extremely broad and comes from a wide range of disciplines so that its character is very heterogeneous. The following picture from a workshop in Lisbon demonstrates this in greater detail.

Research cycle & repositories

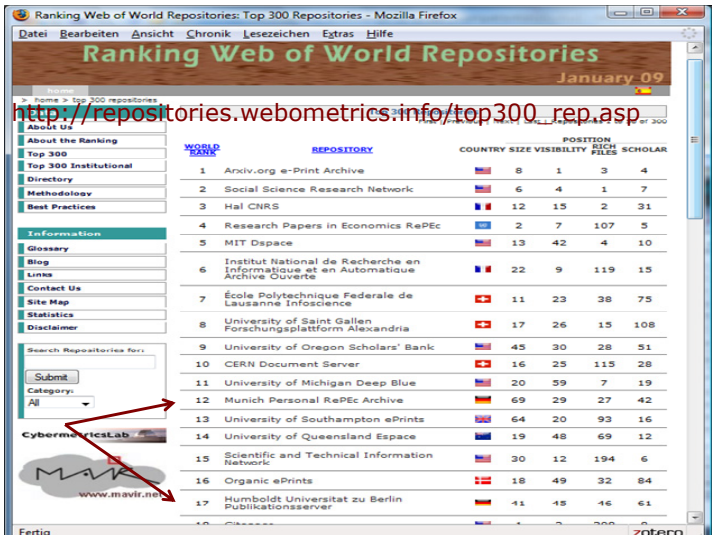


Quelle: e-SciDR-Lisbon Workshop Discussion Paper

Based on the "Research Cycle", the repository could represent a comprehensive collection of the most activities inside a university. The classical output of the research process varies by discipline and includes monographs, preprints, post prints, and journal articles. In additional an institutional repository can simultaneously collect raw data, processed data, and research documents like reports. It is a question of visibility for both the readers and the public. In some cases you could find also learning materials in repositories.

The discussions about content and its variety are intense. A lot of rating and ranking activities have started in order to determine the usefulness of repositories. One of these activities is the ranking table of the Spainish institute of webometrics (http://repositories.webometrics.info/). Three years ago they published their first ranking list for repositories. The image below shows the ranking list as of the beginning of 2009. It is a list from the top 300 repositories. In most cases highest ranking repositories are disciplinary. Of course the oldest and most popular repository, archive.org, is at the first place.

Under the first 20 you will find two German repositories; one of these is the eDoc server of my university.



Another ranking table focuses only on Institutional Repositories. Here the Humboldt-server is listed at place 11.

What are the criteria for such a ranking?

Let me cite from the Web-site of Webometrics:

> "The aim of this Ranking is to support Open Access initiatives and therefore the free access to scientific publications in an electronic form and to other academic material. The web indicators are used here to measure the global visibility and impact of the scientific repositories.
> We encourage the web publication as a way to communicate both formal and informal scholar material, maintaining the high standards of quality of the peer review processes. Web sites reach much larger potential audiences, offering access to scientific knowledge to researchers and institutions located in developing countries and also to third parties (economic, industrial, political or cultural stakeholders) in their own community.
> With the aim to improve visibility of repositories and good practices in their web publication we have extracted the following quantitative **web indicators** from the most important search engines. The methodology is similar, but not exactly the same, to those use in our other Rankings:
> **Size (S)**. Number of pages recovered from the four largest engines: Google, Yahoo, Live Search and Exalead.
> **Visibility (V).** The total number of unique external links received (inlinks) by a site can be only confidently obtained from Yahoo Search and Exalead.
> **Rich Files (R)**. Only the number of text files in Acrobat format (*.pdf*) extracted from Google and Yahoo are considered.
> **Scholar (Sc)**. Using Google Scholar database we calculate the mean of the normalised total number of papers and those (recent papers) published between 2001 and 2008.
> The four ranks were combined according to a formula where each one has a different weight but maintain the ratio 1:1 between activity (size *sensu lato*) and impact (visibility)"

The Visibility of Repositories

Of course every ranking or rating is a matter of discussion and it is the same with the webometrics ranking. From my viewpoint the visibility of repositories is most important in order to improve Open Access publication of this sort. I will give a summary of the different views about repositories. Every player in the research and publication process

has different opinions about functionality, visibility and service. It is important to understand these different views in order to make improvements

Let me start with the users:

When looking at Open DOAR or ROARMap, the most popular directory of repositories, users will find a lot of repositories and will have problems finding the right one. They have difficulties because, in most cases, they lack information about the type of the content.

Some repositories have only scholarly materials, but there are also repositories with e-learning modules or other contents. I already mentioned the research cycle and the range of materials. Rarely will the user find any comments about content quality.

Another problem with repositories is the number of documents. Most of the repositories have no more than from one thousand to between four and five thousand documents. In the last ten years at Humboldt-University we collected about 10,000 documents. While that sounds good, the university library has more than five million documents.

From the user's view it is important to know which kinds of services are available within the repository, such as search machines or print-on-demand functions. The quality of search machines is improving, but are the repositories prepared with the right metadata to work with the search machines?

The view of the authors:
There is a big different between the view of the users and the view of the authors about the visibility and usefulness of repositories. It is a pity but it is also reality that the basic goal for most authors is not Open Access publication, even though they are interested in being read. Often they do not know that a freely available publication will be read more than five to eight times more frequently.

In most cases OA-publication are second or post publications, and the author dreads the extra work and effort involved in the uploading.

In addition, the scope of service in most repositories is limited and not comparable with the author-service from a commercial publishing house.

What we need are additional facilities like a metadata front page that offers a persistent identifier, print-on-demand services, metadata export possibilities and especially usage statistics. In the two screenshots below you will find to examples of how to present publications in a repository.

The View of the operators:
In order to discuss the visibility of repositories, it is also necessary to consider the view of the repository operators. In most cases the operator can do a lot to improve the visibility of an institutional repository.

There are, however, some obstacles, the largest of which is the lack of integration with and the lack of support for the repository in the regular operations of the library. At least in Germany you will find many enthusiastic people running repositories but in most cases it is not their main job. Only a few university libraries in Germany have special groups for repository operations. Such tasks are often a kind of "odd man out".

There is no doubt that a key disadvantage of institutional repositories is the absence of the operator's influence on the content of the document. Peer review is rare. Most repositories have only general statements about quality, like the declaration on Humboldt-edoc-Server that we collect all documents from scholars of the university from

doctoral theses to journal articles, but that we collect student papers only when a professor recommends them.
Therefore the main target of the operator's activities is directed at quantity and not the quality of the documents.

Another observation is that the operators are oriented more toward providing a perfect system and less toward service. From my point of view one reason for that may be that many operators studied computer science instead of library and information science.

The perspective of repositories
In Germany these different views were discussed within the "Electronic publishing group" of the German Initiative for Networked Information (DINI), as I mentioned in prior presentations during the ICDL-conference in 2007. Therefore we developed a special DINI certificate for documents and publication services, which I presented in detail 2007. Seven criteria must be fulfilled in order to get certification. It is not my aim to describe the certificate again here, but these criteria can help operators improve the quality of their servers.

Seven criteria for the DINI-Certificate:

1. Visibility & Server Policy
2. Authors Support
3. Legal Issues
4. Authenticity and Data Integrity
5. Indexing (Subject Indexing, Metadata, Interfaces)
6. Visibility / Impact /Access Statistics
7. Long term Availability

For more detail see: http://edoc.hu-berlin.de/series/dini-schriften/2006-3-en/PDF/3-en.pdf

These criteria focus on the technical quality of a repository. To improve the visibility of whole repositories, we have to do more. The most important step is built a network of repositories.

We have different projects that are to improve the visibility of repositories, including the edoc-server of Humboldt-University. First we got the funding from the German Research Foundation (DFG) for the Open Access Network project.

Let me give a short explanation of the main goals of this project. The project should include all repositories in Germany that received a DINI certificate. This is important so that we can work with repositories that have a minimum shared technical standard for the different documents. Currently we have 29 servers that have the certificate. We have built a service provider, organized browsing for all repositories based on DDC and we offer additional services as mentioned above.

OA network is also the German node for the European Project called DRIVER.
DRIVER means "Digital Repository Infrastructure Vision for European Research" and aims at improving the visibility of repositories in Europe.

Additional and closely cooperating projects include:

- Open Access statistics
  This project aims to create usage statistics about all networked repositories

- Open Access Citation Service
  This project aims to analyze the citation rates of the different repositories. Both projects are underway and I hope to present results from them next year.
- Two other projects, which DINI supported, are an OA Information Platform and OA-policies. Both projects target improving knowledge about Open Access publications in Germany.

Conclusion

The various kinds of repositories offer good opportunities for Open Access publication. Currently we are at a point where, in most cases, the technical requirements have been met. We have scholarly repositories all over the world for the different research disciplines and there are repositories at nearly every scholarly institution. These efforts have created new and innovative conditions for the scholarly community. In order to raise the repositories to further levels of acceptance, we need to take steps to network them and to provide additional services inside these networks.