How and why is the NORA project adding value to the institutional repositories established in Norway?

Elin Stangeland, Marianne Moe
The University of Bergen Libraries, NTNU Library

Summary

In this paper we study the Norwegian NORA project and discuss how and why the services rendered are of value to the participants. For background purposes we begin by looking into topics such as open access and institutional repositories. We then describe the NORA project, its objectives and its progress halfway through the project. Here we also take a look at alternative services to NORA before we discuss how the project adds value for project members and users. Finally we discuss future development of the NORA service.

1 Introduction

The NORA project is a collaborative project between universities and university colleges in Norway, aiming to facilitate national search services for self-archived research material. The project is also concerned with advocacy issues regarding open access and the establishment of institutional repositories (IRs) in Norway. Our aim in this paper is to consider the value added by the project, and discuss the advantages participation gives compared to using other available services.

2 Open access in a short historic framework

The traditional channels of scholarly communication have in recent years been challenged by the potential of the internet. The internet makes it possible to distribute peer-reviewed journal literature world-wide free of charge. At the same time the costs of traditional, print-based scholarly journals have risen at a rate above inflation for a number of years. The budgets of the academic libraries have on the other hand been static, or even decreased. The consequence of this is decreased access to research results.

Both prior to this and as a response to this, a movement of organisations and individuals working in favour of free and unrestricted access to research materials started to emerge. This movement was soon known as the open access movement (Suber 2004). In late 2001 the Budapest Open Access Initiative was launched by the Open Society Institute. In the initiative "open access" is defined as: *free and unrestricted online availability of scientific material* (Budapest Open Access Initiative 2002). The initiative recommends two strategies for open access to scholarly journal

literature. The first is self-archiving in institutional repositories and the second is open-access journals. The focus in this paper will be on self-archiving in institutional repositories.

2.1 Institutional repositories

An institutional repository is defined by Wikipedia as: *an online locus for collecting and preserving – in digital form – the intellectual output of an institution, particularly a research institution* (2006). The idea of posting journal articles, gray literature, theses, and dissertations etc. on the internet either on web pages or in archives/repositories is not new, but the introduction of the OAI-PMH harvesting protocol made interoperability possible. Instead of having separate repositories living their own lives, the OAI-PMH protocol made it easy to harvest metadata from a number of OAI-compliant sources.

There are several reasons why institutional repositories are growing in popularity. The open access advocates usually emphasise the need for free and unrestricted access, but there are also a number of more pragmatic reasons why one should archive such material. A publicly accessible repository can be seen as a way for an institution to display its intellectual output and contribute to the status of that particular institution. It will also be in the authors own interest to publish in an IR. Recent research shows an increase in citations by 45-90% (Antelman 2004), compared to articles that are not posted in IRs. Institutional repositories also make research material available to readers who would otherwise not have access. The most obvious is researchers in developing countries, who in many cases only have access to a very limited number of commercial services. However, institutional repositories are also of interest to the public in the developed parts of the world. For example it may be in our interest that our local physician can update herself on current research and treatment of various diseases.

There are a number of available software packages for institutional repositories. The most common applications are GNU Eprints, DSpace and Fedora (Jones et al 2006).

2.2 Institutional repositories in Norway

There are currently three institutional repositories in Norway, and others are in the process of being established. The largest repository is the DUO service at the University of Oslo which mainly contains masters' theses and doctoral theses. The repository software for DUO was developed locally by the University IT department, USIT. DiVA was chosen as the software at Norwegian University of Science and Technology, this repository mainly consists of doctoral theses and masters theses. The University of Bergen and the University of Tromsø have both chosen DSpace for their repositories. The Bergen Open Research Archive (BORA) mainly contains journal articles and doctoral theses, but a large collection of masters' theses will be added shortly. The Munin service in Tromsø is still under development, and will probably be launched in 2006.

A group of university college libraries have initiated a common development project together with BIBSYS. The project objective is to develop a common solution for institutional repositories. It is named Pepia, and this joined repository is planned to go live in the autumn of 2006 (Kofoed 2005).

3 The NORA project

The Norwegian Open Research Archives (NORA) project was started in 2004 by the University of Oslo Library, the University of Bergen Library, Norwegian University of Science and Technology Library and the University Library of Tromsø. The project objectives are to coordinate and promote the development of institutional repositories (IRs) in Norway and to develop a central OAI harvesting service called the Norwegian Open Research Archives (NORA). This service will be open to all Norwegian research institutions that have online material in full text, and metadata in a harvestable format. The project is planned as a three-year project, and received funding from the Norwegian Archive, Library and Museum Authority both in 2005 and 2006. When funding was secured, the project group was extended with five University College libraries.

The goals of the project are the following:
Part I - 2005

- Establish a national search service for open institutional repositories (IRs) in Norway; the service will be part of the Norwegian Digital Library (NDB)
    - This includes the development of an OAI-PMH harvesting service
- Stimulate Norwegian universities, university colleges and other research institutions to establish local institutional repositories
- Create a harmonised metadata model for Norwegian IRs
- Make the service available through a standard search protocol (for example SRU/W) for use in local search systems (such as library portals)
- Establish the service as part of NDB's common search system through a standard search protocol (for example SRU/W)
- Supervise area development and inform the Norwegian research community on open repositories and open access in general

Part II – 2006

- Continue development of the national search service for open IRs in Norway.
    - This also includes implementation of Open URL support
- Assist local IRs to facilitate for metadata harvesting
- Harvest and ascertain quality of metadata in local IRs
- Make NORA available internationally through global vendors of search services
- Harmonise use of indexing schema
- Develop an URN:NBN-service together with the national library, and establish an internationally adapted resolution service
- Create support for submission of articles to the IRs through Frida, a CRIS used by the largest Norwegian universities.
- Establish an information web-site about Open Access in a Norwegian context
- Supervise area development and inform the Norwegian research community on open repositories and open access in general

3.1 Status update

By November 2005 the project had fulfilled most of the years planned activities and most importantly a simple version of the planned search service was implemented (Jakobsson 2005). The implementation of search algorithms and the metadata harvester were conducted by staff from The University of Oslo. Norwegian University of Science and Technology (NTNU) was responsible for user interface design. The designs were thoroughly discussed in the NORA-group and revised several times. For the development of the advanced search interface, the organisation of the interface search elements generated a lot of discussion. The question was whether it should be a traditional bibliographical search, as for example BIBSYS ASK, or whether to use an approach similar to Google, where the search is general, allowing the user to combine terms in different ways. The latter was chosen mainly because it is believed to be more user friendly than a traditional bibliographical search interface.

The developed harvesting service has functions for registration of new archives, uploads and validations of data. It is based on the OAI-PMH which is a protocol that allows metadata from IRs to be harvested by OAI service providers and then made searchable on the internet through services such as NORA. So far 2803 items have been harvested in 2005; these come from five institutions and are made searchable through the NORA service. Data that differs from the metadata standard are either normalised, or the data suppliers are notified and allowed to correct their metadata.

When it comes to area supervision and information to the Norwegian research community, this has been done through several initiatives. Interested parties can access the NORA web-site and find information regarding the project. Also the project members have presented the project and affiliated issues at conferences, seminars etc. both nationally and internationally.

3.2 Standardisation efforts

As mentioned above the participants of NORA has developed a common metadata model for institutional repositories in Norway. This model is made to support the development of national harvesting and searching services. It has two important requirements (Jakobsson & Erlandsen 2005):

1. Repositories that follow the standard have to register data in the Dublin Core (DC) metadata format.
2. In addition the OAI-PMH must be implemented in every repository to allow for harvesting.

The base of the model is the Dublin Core metadata model (2004). This is used by the OAI-PMH protocol and covers the most important types of metadata that is needed for bibliographical descriptions of digital text documents.

The Nora project has selected eleven out of the fifteen original elements in the Dublin Core Element Set as part of the Norwegian metadata model. These have been chosen as vital in any biblio-

graphical description of scientific documents and therefore many of the elements are mandatory to register. So far the project group has standardised the following elements in the metadata model: language (ISO 639-2) and date formats (MMDDYYYY). In addition there is also a field for resource type which adheres to the Dublin Core type vocabulary.

Finally a national effort should be made in order to standardise person names, and potentially publisher names. The library system BIBSYS has a module for authority control which is of very high quality. If this could also be utilised by the IRs, the result would be a comprehensive overview of authors within science.

In addition to the metadata model, the project group has decided on using the Norwegian Science Index (Walløe, 2003) for indexing purposes. This is a fairly shallow subject set originally used to statistically document research output. Also it was intended to align with international standards, so that the Norwegian statistics would be comparable to international research statistics. Currently there are a number of indexing schemas in use in Norwegian academic institutions, some are international, such as NLMs Medical Subject Headings, or they can be national schemas such as Humord created at the University of Oslo Library of Humanities and Social Sciences. Each repository can choose to register such subject data, but the project group meant that the resources in NORA ought to be searchable in one comm. on indexing schema, and therefore the Norwegian Science Index was chosen. Another advantage of using the Norwegian Science Index comes from it already being utilised by Norwegian CRIS' such as Frida.

4 Alternatives to NORA

There are several alternatives to services such as NORA available. This can be either subject discipline repositories or general OAI service providers such as OAIster. Also search engines such as Google (Scholar), Scirus and others harvest the metadata and make these available in their general searches.

Subject discipline repositories are repositories where academics within a discipline can self-archive their academic papers. Examples of subject repositories are arXiv.org, which contains e-prints in physics, mathematics, Computer Science and quantitative biology, or EconPapers that contains e-prints and working papers concerning economy.

Internationally, the OAI-PMH is utilised by several service providers which base their existence on harvesting other OA services and making the metadata searchable in central searching facilities. An example of this is OAIster, run by the University of Michigan Digital Library Production Service. Any digital collection that has OAI compatible metadata can request to be harvested by OAIster, the data are then made searchable via the OAIster search interface. Another similar service is the Scirus search engine. This has a strict focus on scientific content, and lets the user combine searches in IRs metadata with searches in e-journal resources such as ScienceDirect and BioMed Central.

Finally there is Google with its numerous searching possibilities. Google collects information in two ways, partly by harvesting metadata with OAI-PMH (currently this only happens in Austra-

lia), and partly with its googlebot. The data is then made available for searching through its general search pages and Google Scholar.

All of these come in addition to IRs, not instead of, but academics may choose subject repositories as their primary publishing place rather than IRs. This is a problem for the IRs who aspires to cover as much as possible of a university's academic output.

5 NORA compared to other services

In the following, we will discuss the issues making the NORA project and its achievements worth its while for the project members, and we will also compare NORA to some of the other services available. To begin with, the financial advantages by such collaboration projects need to be pointed out. By managing to get funding from government bodies, the local repositories can use their resources on other important issues, such as acquiring content. Many of the tasks that were done would also have been difficult to do locally for the various IR projects, partly because they do not have the technical competence required, and also because of lack of personnel. Examples of this are the development of an article submission solution via Frida, and the support for SRU/W.

Seen from a Norwegian point of view, the development of a common search facility for academic publications in full text is a great advantage. Instead of having to access the IR of the individual institution, openly accessible research material can be found at one central service. For institutions that have library portals, the planned implementation of SRU/W[1] support will be an additional advantage. NORA can then be searched via the local portals, allowing employees of a research institution to search for various types of scientific publications in different sources from one search interface.

Another important part of the NORA project, is the standardisation efforts that have been made. By standardising the metadata model and making some metadata fields mandatory we ensure that even if the data come from many different sources they will look consistent to the user. The standardisation of field contents adds to this; for example when describing language, the user knows that if she searches for documents in German she will get everything German in the database, while if the language had not been standardised the encoding in some IRs could be German and in others Tysk or Deutch.

By choosing a standard for indexing (the Norwegian Science Index) it will be possible to search for documents within a specific subject, across all participating institutions. This makes it easier to get a general overview of available works on for instance biophysics at the PhD-level.
Standardisation on both indexing and document types will make it possible for other repositories to harvest selected metadata in NORA. This could be of interest to various subject repositories that want to import a very specific set of metadata to their repository.

[1] http://www.loc.gov/standards/sru/

As we have discussed above there are several good services available for those who search for research material on the internet. The Scirus service mentioned above covers more than 200 million science related web pages and have 22 million documents indexed from a variety of sources (http://www.scirus.com/srsapp/aboutus/). Scirus classify documents according to 20 subject fields, but it is not known to us how well these subject fields are mapped against the indexing schemes used in the various repositories harvested by Scirus. Another large service is OAIster.org with more than 6 million records from 600 institutions (http://oaister.umdl.umich.edu/o/oaister/). The subject field is free text, and no indexing scheme is implemented. A service specialising in theses and dissertations is the NDLTD Union Catalog Project (http://oai.dlib.vt.edu/~etdunion/cgi-bin/OCLCUnion/UI/index.pl) with more than 137 000 records from over 50 institutions. The primary quality of these services as we see it is the enormous amount of material available. However, with no common indexing scheme, any search could return a very high number of hits, making it difficult to separate relevant material from less relevant material. In this aspect Scirus holds an advantage against the others because it is possible to limit the search to one or more broad fields of subjects. The advantage of NORA compared to the mentioned services is the quality of the metadata that allow its users to find relevant material in a simple manner.

More than anything else, the NORA project is about dissemination of Norwegian research results openly. It is well known that the number of access points to a document influences its use. Statistics for the repositories in the DiVA consortium shows that the documents are accessed from both OPAC links and from different search engines, and that the users find the material using a number of different search phrases (http://www.diva-portal.org). Therefore it is important to make the content of the IRs accessible from as many search engines and repositories as possible.

6 Future developments

So far the NORA project has come a long way in developing common services for Norwegian IRs. In this section we will discuss some of the opportunities and challenges that the project will have to handle in the near future.

The NORA group should agree on a vocabulary used to describe the type of document, whether it is a book, a journal article or a PhD thesis. Norwegian universities use a common OPAC (BIBSYS), and most of these terms are already agreed upon, but documents such as journal articles or conference papers are not part of the OPAC. The CRIS (in our case FRIDA) will contain data on these documents and most of the content of the OPAC, but not necessarily data on work done by students such as masters theses (which will be part of the IRs). The situation we see ourselves in is that a rather large proportion of what we expect to be the content of the IRs is not a part of the OPAC, and some of the content may not be part of the CRIS either (preprints, masters' theses or other work done by students). We therefore find it necessary for the NORA group to agree upon a uniform vocabulary for document types and a common understanding on how we deal with them.

Another question the authors believe should be addressed is how we deal with various versions of a document. The lifecycle of a document normally starts with a journal article being self-archived as a manuscript, the second version could be a preprint submitted to a specific journal and the

third version is the post-print. The prior versions cannot be deleted from the IRs, but we believe that the user needs to be informed that there is a newer and hopefully better version available. Exactly how we can achieve this is a question we need to discuss. The problem is how we can relate documents to each other and when a document should be regarded as a completely new document (and not be related to prior documents). Although the Norwegian IRs use different software, a common understanding on how we deal with this would be useful. The implementation of the recommended strategy will have to be handled at each individual institution.

As mentioned above the use of the Norwegian Science Index will make it possible to search for documents within a specific subject field. This indexing could also potentially be used to find other subject related documents in the various IRs if you find the first document through a search engine such as Google. This is an idea we will have to discuss further in the NORA group; we will also have to consult technical expertise for possible solutions. It could also be useful to look at the indexing schemes used in neighbouring countries such as Sweden and Denmark. The SVEP-project in Sweden has recommended a scientific index for use in Sweden (http://www.ub.uu.se/epub/categories/), and it would be interesting to see how well this matches the Norwegian Science Index and to explore the possibilities of creating a Nordic search interface for scientific material standardised on subjects.

One important issue for the next year is to make sure that new repositories in Norway are added to NORA. As we have mentioned earlier, the institutions that want to join NORA will have to implement the common metadata model and be OAI-compliant. There will be a very limited opportunity for new IRs to make changes in the metadata model or various definitions used, whereas the existing members already have come with their suggestions. This can be an obstacle for the inclusion of IRs in NORA, but we believe that with other project members assisting the newcomers these issues could be solved.

Another challenge that we mean is relevant for both NORA as well as for the local IRs is the issue of acquiring content to the IRs. This is a known problem (Foster, & Gibbons 2005) which also the Norwegian repositories are familiar with. Advocacy is the responsibility of each individual institution, but the integration between FRIDA and the IRs, and use of common indexing schemes in both systems are attempts to make it easier for authors to self-archive their research.

The work on NORA so far has been very fruitful and has opened up for many interesting discussions. One of the issues we find most interesting is open access. We currently see a paradox in the process of funding research in Norway, where open access publishing (in open access journals) is endorsed by the Norwegian Association of Higher Education Institutions in their comments to a report presented to the Parliament (Det Kongelige Utdannings- og Forskningsdepartement 2005). At the same time the funding bodies focus their attention on the traditional journals and rewards researchers and institutions according to number of publications in such journals. An important upcoming task for the NORA project is therefore to persuade the funding bodies to make self-archiving a prerequisite when funding research.

7 References

Antelman, K. (2004): Do Open Access Articles Have a Greater Research Impact? College & Research Libraries News, Vol. 65, No. 5, p. 372-382.

Budapest Open Access Initiative, http://www.soros.org/openaccess/read.shtml, viewed 10.01.2006

Dublin Core Metadata Element Set, Version 1.1: Reference Description (2004) http://dublincore.org/documents/dces/, viewed18.01.2006.

Diva-portal.org, http://www.diva-portal.org/stats/2005/awstats.ntnu-fulltext.html, viewed 13.01.2006

Foster, N.F.; Gibbons, S. (2005) Understanding faculty to improve content recruitment for institutional repositories. D-Lib Magazine, Vol. 11, No. 1, http://www.dlib.org/dlib/january05/foster/01foster.html, viewed 01.02.2006

Jakobsson, A (2005). Nasjonal søketjeneste for vitenskapelig informasjon i åpne institusjonelle arkiv. Statusrapport 01.11.2005, http://www.ub.uio.no/ubfelles/prosjekter/NORA/NORA-status_01-11-2005_version6.doc, viewed, 13.01.2006

Jakobsson, A; Erlandsen, M (2005) Metadatamodell for åpne institusjonelle arkiv ved universitetene i Norge. http://www.ub.uio.no/ubfelles/prosjekter/NORA/Metadatamodell_for_NORA.htm, viewed 09.01.2006

Jones, R; Andrew, T; MacColl, J (2006) The Institutional Repository. Oxford: Chandos Publishing

Kofoed, JE. (2005): PEPIA-konsortiet er stiftet. Bibsys-nytt 4. desember 2005, http://www.bibsys.no/bibnytt/05-4/05-4.html, viewed 24.01.2006

Det Kongelige Utdannings- og Forskningsdepartmenet (2005): *Vilje til forskning*, http://www.odin.dep.no/filarkiv/241962/PDF-TS020.pdf, viewed 27.01.2006

NDLTD Union Catalog Project, http://oai.dlib.vt.edu/~etdunion/cgi-bin/OCLCUnion/UI/index.pl, viewed 1. February 2006

OAISter.org, http://oaister.umdl.umich.edu/o/oaister/, viewed 1. February 2006

The Open Archives Initiative Protocol for Metadata Harvesting – protocol version 2. http://www.openarchives.org/OAI/openarchivesprotocol.htm, viewed 18.01.2006.

Scirus. com: http://www.scirus.com/srsapp/aboutus/, viewed 1.february 2006
Subject Categories for Scientific Publications, http://www.ub.uu.se/epub/categories/, viewed 20.01.2005

Suber, P. Timeline of the Open Access Movement. http://www.earlham.edu/~peters/fos/timeline.htm, viewed 10.01.2006

Walløe, R. et al (2003): Norsk inndeling av vitenskapsdisipliner, http://www.uhr.no/utvalg/forskning/dokumenter/forskdokNorskvitdisinnst.htm, viewed 10.01.2006

Wikipedia, Institutional Repositories, http://en.wikipedia.org/wiki/Institutiona_repository , viewed 09.01.2006

8 Contact information

Name Elin Stangeland
The University of Bergen Libraries
Nygårdsgt. 5
5015 Bergen

e-mail: elin.stangeland@ub.uib.no

Name Marianne Moe
NTNU Library
Development and Coordination Section
7491 TRONDHEIM

e-mail: marianne.moe@ub.ntnu.no