On the Cost and Benefits of Building a High-Quality Institutional Repository

Pieter H. Hartel, University of Twente and Technical University of Malacca

ABSTRACT

Much has been written about the difficulties of getting staff and students to contribute to an Institutional Repository (IR). The most difficult objections of staff to deal with are lack of time, and uncertainty about the benefits. We report on the measures that we have taken when building an IR to address those objections, indicating which were successful and which failed. We then propose a set of parameters to capture the quality, cost, and benefits of maintaining an IR in a quantitative manner, using the parameters to compare three IR. Our proposal is a first step towards a more objective approach towards evaluating IR.

1. INTRODUCTION

For many years all Universities in the Netherlands have been using an information system called Metis[1] to record the performance parameters of the research, including the publications, patents, PhD theses but also the number of projects acquired, the number of staff employed, etc. Metis has been created for, and is used by administrators, solely for the purpose helping the universities to show that they are accountable for the funding received. Full text is of no use to administrators hence until recently Metis did not support full text. At the end of each year the academics at all Dutch universities are asked to enter the meta-data of their publications into Metis. This takes time and effort, and yields no direct benefit to the individual researcher. Until recently, Metis did not produce lists of output in a form that is useable for research, so researchers create their own lists of publications. The main funding organization NWO[2] and its subsidiaries (such as STW[3]) require an annual report for each funded project, again with a list of publications. Summarising researchers have to provide the meta-data of each publication a number of times, in different selections, and in different formats, for which Metis provides little or no support.

The University of Twente (UT[4]) is one of the smaller universities in the Netherlands. The UT has five faculties, one of which is the Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS[5]). EEMCS has 26 research groups with a total of about 150 staff, 450 PhD students and PostDocs, and an annual output of about 50 PhD theses and about 1000 refereed papers, mainly in international journals and conferences. EEMCS staff represent three related disciplines: Applied Mathematics (AM), Electrical Engineering (EL) and Computer Science (CS). The publication cultures in the different disciplines are sufficiently different to pose a challenge to introducing a faculty wide policy about the registration and archival of research output. On the other hand the faculty is small enough for one individual to understand, or at least appreciate the publication cultures of the entire faculty.

Problem. Out of dissatisfaction with use of Metis, EEMCS decided to introduce the Southampton EPrints system (Version 2[6]) with the overall aim of making life easier for the researchers: EEMCS would ask researchers for the meta-data and the full text of their publications only once, and generate all manner of publication list automatically. This led to the following objectives of the EPrints introduction project:

1. to use EPrints instead of Metis as the only tool to enter the meta-data and the full text of a publication.
2. to export all meta-data to Metis in a way that is completely transparent to the users.
3. to improve the quality of the meta-data provided.
4. to create customizable lists of selected publications for web pages, sponsors, the research assessment etc.
5. not to increase the cost to the organisation of the work flow.
6. to provide a basis for long-term preservation of the meta-data and full text.

Contribution. The contribution of this paper is twofold. First, we describe the lessons learned since the introduction of an EPrints Institutional Repository (IR) at EEMCS[7] in October 2005. Second we propose a set of parameters to measure the quality, cost and benefits of building an IR. In particular, we are able to perform a quantitative analysis of the cost of building an IR, at a level of detail which we have not seen in the literature. We then compare our IR to two others on the basis of the partial information found in the literature.

In just over three years, our IR has accumulated about 7360 papers, consisting of the complete EEMCS output of 2006, and 2007,

[1] http://metis.hosting.kun.nl
[2] http://www.nwo.nl
[3] http://www.stw.nl
[4] http://www.utwente.nl

[5] http://www.eemcs.utwente.nl
[6] http://www.eprints.org
[7] http://eprints.eemcs.utwente.nl

XXX YYY, USA

and most of 2008 (the remainder being entered into the IR at the time of submitting this paper), a total of 2930 papers entered on a mandatory basis by users, and 4430 papers provided on a voluntary basis by users for earlier years. The earliest paper dates back to 1972.

We show how and why we have been able to achieve objectives 1, 2 and 3. Objective 4 is partially addressed because the current system is able to provide all relevant selections of output but not always in the requested format. We have a prototype system available to achieve objective 4 fully, which will be launched shortly. We have failed to achieve objective 5. Objective 6 is future work.

The EEMCS EPrints system is an Institutional Repository, but focused on published output. Preprints and postprints are expected to be submitted to our IR by way of a Technical Report. Our IR is the carrier of two such report series: the CTIT technical reports[8] (626 entries, mainly used by CS staff), and the Applied Mathematics reports[9] (415 entries, mainly used by AM staff). Our IR also carries the CTIT PhD Theses[10] (109 entries). All three series have an ISSN and as such represent bibliographically traceable publications. EL staff sometimes makes use of these report series, but not often.

Metis has recently been endowed with capabilities to list the publications of an individual researcher[11], but the presentation is at best as consistent and complete as the data base itself, which in the case of Metis is sub-optimal.

The University of Twente has recently started an IR[12] also based on EPrints software (version 3). Work is in progress to feed the UT IR from the EEMCS IR in a way that is transparent to the users.

The paper is structured as follows. The next section discusses related work. Section 3 explains the basic ideas of the policy for our IR. Section 4 describes how we have adapted the standard EPrints work flow to our needs. Section 5 discusses the main concerns raised by our users. Section 6 discusses where we have been successful in achieving our goals, but more interestingly, Section 7 describes where we have failed. A quantitative analysis of the quality, costs, and benefits of the IR is provided in Section 8, followed by a comparison of thee IR in Section 9. The last section concludes.

2. RELATED WORK

A large number of studies report on the difficulties that may be encountered when setting up an IR [4].

There are three fundamentally different approaches, and depending on the objectives of the IR, some approaches are more natural than others. The first approach is a managed IR, which is typically populated by library staff [2], and as such tends to be relatively well populated. Researchers do not necessarily take much notice of a managed IR [6]. It is hard to put non-standard meta-data into a managed archive as library staff may not know what the appropriate meta-data is (e.g. subject classifications, or funding source). The alternative to a managed IR is self-archiving, where researchers are asked to populate the IR. To make self-archiving a success it appears that a mandated policy is needed [1]. Without a mandate, self-archiving is not something that can be achieved quickly, as it requires a cultural change on the part of the researchers [11]. In this paper we focus on mandated self-archiving [9, 12].

We encountered many of the issues discussed in the literature, such as user concerns over copyright, the lack of interest for self-archiving amongst users, and the fact that that populating an IR does not necessarily fit in the users work flow. See Davis and Connolly [6] for a lucent account of the problems.

A good policy gently persuades users to submit the most relevant part of their scientific output to the IR. A significant part of the output of the users will not be covered by the policy, for example older work. To persuade users to submit also the rest of their work many ideas have been published [5, 7, 8]. We provide some new ideas (such as showing ISI impact factors).

3. POLICY

The essence of the EEMCS policy[13] regarding the use of IR is that all staff and PhD students are required to enter the meta-data and the full text of their publications in the IR before a certain deadline. The basic ideas underlying the policy are:

Bibliographically traceable output. Our IR contains anything that is bibliographically traceable, i.e. any publication with an ISBN, ISSN or patent number. This includes publications that should have been assigned an ISBN or ISSN but due to an oversight or specific policy of the publisher do not have such a number. For example the prestigious IFAC conference in Systems and Control does not have an ISSN or an ISBN. It would be unfair to penalise our authors for the lack of ISBN/ISSN on the proceedings. Therefore we also accept papers that should have been bibliographically traceable. The exception to the rule is formed by the Master thesis, because some groups like to have those online too. For invited talks to be included in the IR, we require that there is at least an abstract that appears in the proceedings of the conference.

Complete meta-data. To be able to satisfy any conceivable present and future demand from the University administration, our sponsors, government or a Research Assessment Exercise (RAE) we ask that all conceivable meta-data and the full text of a paper are put in the IR. Even information that no one is asking for right now will probably be requested eventually. A good example is provided by the DOI, which was not requested when we introduced it, but which is now required for self-archiving by the publishers to accompany the full text on the authors web page.

Flexible interpretation. As with any policy there have to be exceptions. In particular we found it sometimes helpful to allow groups to enter their publications with some of the (meta) data still missing (for example without DOI and full text), so that at least all publications could be entered before University imposed deadlines. It is more work to add missing information later, so in the interest of the overall efficiency of the process exceptions should be granted sparingly.

Quality control by academics. To convince researchers that the IR is created, run and used by academics, rather than administrators, we decided that quality control on the meta-data and full text should be performed by an academic, who acts the editor for a specific research group or set of research groups. Being an academic, the editor can talk to researchers with authority, which cannot be achieved if an administrator acts as editor. The cost to the organisation of appointing an academic as editor is obviously high, which can only be justified if the quality of the meta-data is high too. Our current configuration has one editor for AM, two for EL (one for a specific group and one for the remaining EL groups) and eight editors for CS (i.e. basically one per CS group). The reason that CS has more editors than the other two disciplines is historical: we could not find an academic who was prepared to assume editorial

[8] http://www.ctit.utwente.nl/library/techreports/
[9] http://www.math.utwente.nl/publications/
[10] http://www.ctit.utwente.nl/library/phd/
[11] https://appw2048.utsp.utwente.nl/metis/default.cfm?i=publications
[12] http://doc.utwente.nl
[13] http://eprints.eemcs.utwente.nl/policy.html

responsibilities for all of CS.

4. WORKFLOW

The heart of the IR is a workflow where users (i.e. researchers or their secretaries), the editors, and the library staff use a sequence of linked web forms to enter and process the meta-data and the full text into the IR. The standard steps of the EPrints workflow are extended as follows:

(Optional) Legacy import A consequence of the Complete meta-data idea is that we encourage authors to enter all their publications, regardless of where they worked when they published their papers. For a number of reasons, researchers like to have a complete list of their refereed publications. Such lists are required when applying for jobs, promotion, and personal research grants etc. Therefore we provided a legacy import service for research groups. A number of undergraduate students were hired to collect (from any lists provided by the group, Metis, the digital libraries of the publishers, the Computer Science Bibliography DBLP[14], and Google Scholar) and to enter the meta-data and full text of the legacy of the research groups. In the end 18 out of the 26 research groups of the faculty have made use of this service, indicating a high degree of success. To make sure that the legacy data is imported correctly, the student assistant deposits the legacy in the inbox of the relevant author. The idea being that the author will be able to spot mistakes easily. The author can approve an entry with a single click.

User inbox Users enter the meta-data of a publication and upload the full text (in PDF format). The item is then in the inbox of the user, and may remain there for some time, for instance if the user does not have all the relevant meta-data handy and wishes to revisit the entry later. Once the entry is complete the user deposits the item. The item is transferred from the inbox of the user to the buffer of the appropriate editor.

Editor buffer The editor checks the meta-data for completeness and checks the presence of the full-text, bounces the item back to the inbox of the user in case of problems, or otherwise accepts the item, by transferring the meta-data and PDF from the editor buffer to the archive. An item that has been transferred to the archive will be visible in the IR and it will appear on the various websites produced by the IR.

Archive – copyright check Each item in the archive is checked by the library for copyright compliance and for completeness. Once the check has been completed, the library sets the library check flag.

Archive – consistency check A global consistency check is applied to the whole archive on a regular basis. In case of problems, the editor responsible for the entry is asked to correct the problem. He/she may choose to bounce the item back to the inbox of the user or he/she may fix the problem directly.

Metis A few times per year, the data base administrator exports the recent additions that have been checked by the library from the IR to Metis.

The overall workflow used to manage the IR is shown schematically in Figure 1.

[14] `http://dblp.uni-trier.de/`

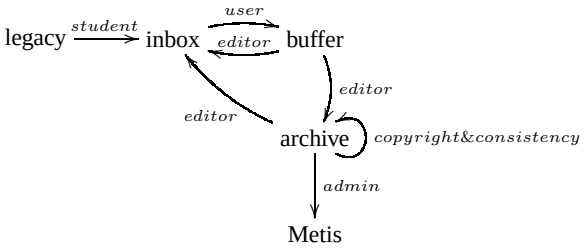

Figure 1: IR workflow. The legacy import is optional.

5. USER CONCERNS

During the implementation of the IR we had to deal with the concerns of our 600 users (150 academic staff + 450 PhD students and PostDocs). Some of the issues took a long time to resolve, but we felt that they were important enough to invest the time and the effort to get it right. We discuss the most prominent user concerns in the remainder of this section.

Does uploading the full text violate copyright? Researchers are not normally experts in copyright law; hence many were concerned that uploading the full text to the IR would get them into trouble with the publishers. To address this problem we reached an agreement with the university library, to check each item for agreement with the policy of the publisher involved. This is a job for experts, as the scientific publishers use wide range of incompatible policies. For example the IEEE allows authors to make only the PDF available on line on the author's web site if the PDF is formatted by the IEEE (with the IEEE copyright notice). Other publishers expressly forbid this. In Dutch law copyright on electronic documents did not exist before 1997, so copyright on older documents can never haven been transferred to the publisher, which allows the author to publish such documents on line in any way he/she sees fit.

To make the handling of the various policies on copyright manageable, our IR supports three classes of protection for the full text: Editors only, University only, and Unrestricted. The default is University only. The library staff will upgrade a University only item to Unrestricted if this is permitted by the specific agreement with the publisher. Meta-data is not copyrighted; hence it is always available with unrestricted access.

How should I classify my paper? The classification of output in the Metis system is a little baroque, probably because the system has grown over many years to incorporate the wishes of 15 universities, covering hundreds of research areas. We seized this as an opportunity to present our EEMC researchers with a manageable set of 12 categories of publications:

1. Article: An article or editorial in a scientific or non-scientific journal, magazine or news paper.

2. Conference or Workshop Paper: A paper, (extended) abstract or preface in conference or workshop proceedings. This includes drafts to the IETF.

3. Internal Report: A memorandum or technical report (to be) published by a school or institution, such as CTIT.

4. External Report: A report commissioned by an external organisation, such as a Ministry, or the European Commission.

5. Book: A book with an explicit publisher, which has UT staff as authors.

6. Book Section: A chapter, section or preface of a book with an explicit publisher.

7. PhD Thesis: A PhD thesis defended at UT and/or written by UT staff or students.

8. Master's Thesis: A Master's thesis written by UT staff or students.

9. Patent: An issued patent or pending patent application.

10. Review: A review of a book, monograph or article in a scientific journal or magazine.

11. Inaugural lecture: The written text of an inaugural lecture.

12. Manual: Technical documentation.

The most important categories are 1, 2, 6, 7, and 9 because these categories usually represent refereed, international work. The definition of the categories and especially the descriptions used to disambiguate the categories has taken some effort. For example item 4 is easily mistaken for a technical report issued by another university, which is less prestigious than what is intended by the description given.

In addition to the 12 categories, we added two important binary flags: Refereed, and International. Both raised discussions as exemplified by the formulations on the IR web forms.

> **Refereed?** Please indicate whether this version of the work has been refereed (you should be able to show the reviews to prove this). EPrints cannot determine this automatically since even scientific journals sometimes accept work that has not been refereed, for example invited papers.

The requirement that authors should be able to show the reviews (even though the editors never actually asked to see a review) caused some debate. We believe that the message is clear though: a one line review (as sometimes produced by workshops) does not make a paper reviewed.

> **International?** Is the journal/event/target audience international in nature? If the majority of the authors of an event work in the Netherlands, International should be set to false. This applies for example to the STW PROGRESS, ProRISC and SAFE workshops.

Again some discussion arose because of this requirement, for example how does one count the authors of a given conference proceedings?

Some publications may still be hard to categorise, for example proposals for new standards to the Internet Engineering Task Force (IETF). These proposals are typically heavily reviewed and as such do qualify as refereed, international output. As a result of a discussion between the authors involved, the relevant editors and the library staff, we now accept IETF proposals as category 2 (Conference item), since the IETF publishes the drafts in electronic proceedings[15]. The result of the discussion is a simple example (see item 2 above) but such examples prove to be effective in preventing the issue from having to be discussed repeatedly. The web forms that we use contain many more examples, to provide as much guidance to our users as possible.

[15] `http://www.ietf.org/proceedings_directory.html`

It is our experience that with the given $12 \times 2 \times 2$ types of publication all output generated by our AM, EL and CS staff can be covered.

Why can't I use the names I want? It is hard to imagine the variety of names that authors choose for themselves (or their co-authors), either by choice, or by mistake. The problem is further compounded by the insistence of the University on spelling certain foreign names in a peculiar format. For example if the current president of the USA had been a researcher at the University of Twente, his name would have appeared as B.O. Barack Obama, whereas his predecessor would have been known as G.W. Bush, as expected.

We have contemplated several solutions to this problem. The first and probably best approach is to maintain a mapping from the various spellings to a single official spelling as chosen by the author, as practiced for example by the DBLP. The necessary support for this is not provided by EPrints so we could not adopt this approach. The second approach, which has been adopted by the University IR is to use the staff number as the official spelling, and provide each author with a list of publications based on the staff number. We found this not user friendly. The solution that we have adopted is to unify the various spellings in order to be able to provide a single list named after the author (using his/her favourite spelling). For University staff we maintain both the unified author name and the staff number, for the purpose of associating the staff number with the favourite name, so that when importing EPrints IR entries into Metis we are able to map author names consistently. In the process of unifying different spellings many emails have been exchanged with the person concerned, group secretaries and administrators, basically to discover whether two apparently different authors are or are not the same person (i.e. is B. Obama the same author as B.O. Barack Obama?).

Why do I have to provide all this information? Consistent with the Complete meta-data idea of Section 3 we always ask users to provide the DOI, and we started doing so at a time when this was not yet standard (for example until recently Springer used Open Url but now seems to have standardized on the DOI). A DOI can be used to identify a document in an unambiguous fashion which is a great tool for the de-duplication process that an IR (including ours) has to face. Interestingly we found publications with more than one DOI issued by Springer. The DOI also provides a permanent link to the publication at the web site of the publisher. Simply prefixing the DOI with *http://dx.doi.org/* gives a permanent URL to the page with the meta-data and the link to the PDF provided by the publisher. At present the DOI is used widely by the scientific publishers, and as a result of our insistence the university has now started to issue DOI for PhD theses. The down side of insisting on authors to provide the DOI is that sometimes it can be hard to find. Our users complained bitterly about having to provide al this seemingly useless information, which is also hard to find. It would help if the scientific publishers were to increase their effort to print the DOI on the first page of each paper, and to make sure that it appears on the appropriate web page.

We are currently introducing Digital Author Identifiers (DAI), to replace the staff numbers.

Why should I bother, my paper is already in a data base! One of the most outspoken critics of the IR wondered why he should bother with it: I only publish in IEEE journals and proceedings so all my publications are available in the IEEE digital library. Others complained that: All my papers are in the Scopus data base. Indeed, if a well maintained data base such as the IEEE DL or Scopus maintains all publications of an author, and provides online access (even at a fee!), there is no reason to create an IR. However, no author in our IR publishes only in IEEE journals and proceedings (a

little pedantic perhaps, but even the critic cited earlier also produces patents which are not indexed by the IEEE). More importantly, the vast majority of our researchers publish in a variety of journals and proceedings without a single data base covering everything. This is mostly an issue for CS, and the AM and EL groups close to CS and less an issue for the EL and AM groups working in an area closer to physics

The most convincing reason we could give to our critics to contribute to the IR is that it contains a significant amount of local meta-data that is not contained in the data base of any publisher, such as the research groups, research programs, research projects, research schools, and a variety of classifications associated with the publication. With this local meta-data, it is easy to produce a list of the output of the research group (which 19 out of our 26 groups do), to produce a nice list of the output of externally funded research projects etc.[16].

6. SUCCESSES

We have achieved some but not all of our objectives. The most important successes will be discussed below.

Metis export operational. The design and implementation of an export facility from EPrints to Metis has been costly and time consuming but not interesting from a technical or usability point of view. This facility is an essential requirement because the performance parameters of the research groups are accessed through the Metis data base by the central University administration. Some parameters determine the level of funding of the groups, hence the relevant data of each research group must be present in the Metis system. In designing the export facility we had to map EPrint meta-data onto the relevant Metis meta-data. For a few fields like the title this was no problem, but for most fields this was a problem, which took a lot of discussion between the Metis Expertise Centre in Nijmegen, the Metis administrators in Twente and our own data base administrators. The Metis import module that was custom-built for us is available upon request.

Auto-complete facilitates meta-data entry. We provide an auto-complete facility for data that is needed often. For example entering a small part of the title of a journal would complete not only the name but it would also fill in the ISSN. To maintain the auto-complete facility we have to enter a variety of data in the IR that is available in other university data bases (such as staff names and numbers, the names of externally funded research projects, titles and ISSN of journals etc.). Thus far this has been done manually, while the question of how to automate this has been pondered for some time. The problem of gathering the auto-complete data from other data bases requires not only trans-coding, but often human intervention. For example the data base of externally funded research projects as used by the University stores project titles in upper case only, and in relatively short fields. Our auto-complete lists carry the official title as created by the researchers instead. Again foreign author names are an issue. We see no other solution than to continue to feed the auto-complete facility manually.

Global consistency check effective. No matter how diligent the users and editors are when they process the meta-data, errors will be made. Therefore, all meta-data of the IR is regularly checked for consistency, such as inconsistent spelling of names, missing information (particularly ISBN, and DOI), and duplicate entries (on average 2 per week). Generally speaking inconsistent spellings are unified and the editors are asked to deal with duplicate entries and to provide any missing information; the editor can decide to fix the problem himself, or ask the user to do so.

University IR easily supplied. In February 2006 the Rector decided that all academics should contribute their publications to the University full text repository to improve the visibility of our research. This of course is a worthy cause, and it would have been easy to implement if Metis had contained the full text of a publication, but unfortunately this was not the case. So researchers from all faculties except EEMCS were asked to enter their publication data twice: in Metis and in the University repository. This was not met by great enthusiasm. However, EEMCS could easily implement the decision of the Rector, without even bothering the researchers.

ISI impact factor appreciated. True to our Complete meta-data idea, each year in the summer we gather the ISSN from our IR and collect for each ISSN the ISI impact factor, loading this information back into the IR. Just a few months after we first loaded the ISI impact factors into the IR one of the institutes (Mesa+[17]) decided that groups should provide a listing of their research output with an ISI impact factor greater than 1.5. We were lucky to have been collecting print ISSN and (not the ISSN for the electronic versions of the journals) as the ISI data base uses print ISSN rather than the electronic ISSN.

Good rapport with the community. While there is no evidence to prove this we believe that the faculty was just large enough to amortise the investments, while being just small enough to be able to oversee the consequences of most (but not all) decisions. For instance it is just about possible to meet with 26 heads of group within a reasonable time span (of about a year). It would be hard to do so with about 100 heads of group for a university wide system. The publication cultures of EL, CS and AM were found to be substantially different (as exemplified for example by the 85% coverage of the ISI for the more physically oriented EL groups versus only 10% coverage of a typical CS group). It would have been substantially more difficult to build a system that caters for the needs of a more disparate user group. In the end local control makes it easier to get buy in from the locals.

IR staff works as a team. The IR is run by a team consisting of the 10 domain editors, and an editor in chief (all academics), a data base administrator, a library specialist, an administrator and a number of student assistants. The domain editors interface directly with the users and as such are the most visible and important to the successful operation of the repository. The library specialist interfaces to the university library, taking care of copyright issues, but also finding important meta-data such as ISBN and ISSN. The data base administrator ensures that the system and the data base are available 24/7. The editor in chief is responsible for the overall functioning of the system. We believe that the composition of the team with well defined areas of expertise is to a large extent responsible for the success (and the cost) of the operation.

Significant quality improvement. The quality of the data in the EEMCS EPrints IR is superior to that of the Metis data base. For example a mandated item in our IR either has an ISSN, an ISBN, or a patent number, or we have verified (with the support of the library) that no such number exists. To compare this to Metis we have counted the number of conference items in Metis from the year 2005 (the last year when Metis was the official IR for the EEMCS faculty). We found that 40% of all 542 conference items were entered without an ISSN or ISBN. Yet for every RAE this data is required, which basically means that the administrator responsible for the publications lists in the RAE documentation has to add all these ISBN at some later stage. We also found that about 100 publications from 2005 that were entered into the IR as legacy had not been entered at all in Metis. Interestingly, we also found several

[16] http://eprints.eemcs.utwente.nl/view/

[17] http://www.mesaplus.utwente.nl/

instances of an incorrect ISBN (i.e. with an incorrect check letter) issued by publishers (such as IFAC).

7. FAILURES

We have made many mistakes over the past 3.5 years. We hope by discussing them here that the reader may avoid making the same mistakes.

Accepted papers are not welcome. Some researchers like to enter a publication in the IR when it has been accepted for publication but not yet published. The workflow does not support this, which is a problem. Our policy is to discourage users to enter accepted publications; however, we do put them in the IR when the user insists. Later, when the publication has appeared, the editor has to remember to update the relevant entry in the IR (for example with the volume, issue and page numbers of the paper). This is unsatisfactory, because users and editors tend to forget to update the IR.

Security breaches. The IR provides three levels of protection to the full text: Repository staff only, University only, and Unrestricted. For this to be effective all configurations have to be set appropriately, which for a number of weeks was not the case due to maintenance. We were alerted by a Master student who, while doing an internship at a company was surprised that he could access his Masters thesis from the company network. In his thesis he had been somewhat critical of certain products of the company so he was anxious for us to repair the configuration. Since then we have an automatic check running on a machine from outside the university to test that access control is effective.

Information lost by unifying names. One of the worst ideas we have had is to throw information away by unifying author names. Instead we should have kept the data as input by the users for future reference.

Even relatively small changes take long to implement. It has taken three years to introduce a relatively small change (the migration from Metis to EPrints) into our organization. Some groups have resisted fiercely our demands to provide more meta-data than before, such as the DOI, but also which research project and which research program is associated with the paper. After more than three years the resistance has weakened and now all groups provide complete meta-data.

Editors are not always sufficiently responsive. Editors are not always sufficiently responsive. As a rule of thumb an item should have been processed within a week for users not to become impatient.

University administrators decide stick to Metis. For over a year the university has been debating whether all faculties should adopt the EEMCS model, or whether there should be a university wide database. In the end it was decided to do neither but to stick to Metis. There were two main reasons for this: The cost as measured by EEMCS was considered too high, and the other faculties would lack the expertise to set up and maintain the content of their Eprints based IRs. We suspect that if the decision had been made by the academic deans of the faculties, rather than the administrative directors of the faculties, then a different decision might have been reached.

Publications are entered too late. We have tried to make publicity for the IR in a variety of ways, in order to persuade researchers to enter their meta-data and full text as soon as a paper has appeared. We tried the following:

- The IR editor in chief has visited most of 26 heads of group (three could not spare the time) to explain the advantages and hear the wishes of the group. At this meeting we offered to import the legacy of the group free of charge.
- All group secretaries have been briefed in an hour long session on how the IR affects their work.
- In one of our three buildings we have put up a display with several alternating views showing the latest additions to the IR and, the most frequently downloaded papers.
- We have produced a glossy leaflet[18] summarizing the advantages of the IR and the main steps to be taken when entering meta-data. The leaflet has been put in the 600 pigeon hole the EEMCS researchers, and is handed out to all new staff.
- The home page of the faculty lists the three most recently added items.
- There is an RSS feed of the recent additions to the IR.
- Several editions of the faculty news letter carried some publicity about the IR, new features or show cases of groups, projects and researchers.
- The IR download statistics are a popular feature, people notice heavily downloaded papers and talk about them.
- In 2006 we asked Google Scholar to index the IR.

Every year in January the university requires that groups enter all their publications of the previous year in Metis. For EEMCS this means that all publication data should have been entered in the IR, checked and imported into Metis. The yearly production of about 1000 publications is too large to be processed in a few weeks. However, the campaigns to persuade research groups to enter a publication as soon as it has appeared have consistently failed. (See Figure 3 for details).

High impact publication award too hard to implement. Faculty management has debated the possibility to award on a quarterly basis a prize to the group with a paper in a journal with the highest ISI impact factor. The prize winner would be selected from the IR, thus encouraging groups to enter their publications as soon as they appear. In the end we could not agree on a mechanism to compensate for the large variety in ISI impact factors between different disciplines. For example one of the research groups publishes in biomedical journals which typically have a significantly higher ISI impact factor than other disciplines. Without a compensation mechanism the biomedical group would have won the best paper award every quarter.

Browse lists too limited. The flexibility in producing customised outputs offered by our EPrints version (2) is too limited. Our users wish to be able to configure precisely per category of publication, which fields should be shown, what the order should be, and what formatting to use for the fields, the entry as a whole, etc. To achieve this we are building an output generating system that supports clickable views, i.e. a mechanism whereby users can select from sophisticated menus what the output should look like.

Persistent mistakes The url shown in the browser window is rarely the official url. In particular DOI based urls are resolved to a url that at this particular moment leads to the page with the meta-data and the link to the PDF. The resolvent is likely to change often, and is thus not appropriate to be used as meta-data. We have not been able to find any measures to prevent users from entering the resolvent rather than the permanent link.

[18] `http://eprints.eemcs.utwente.nl/brochure.pdf`

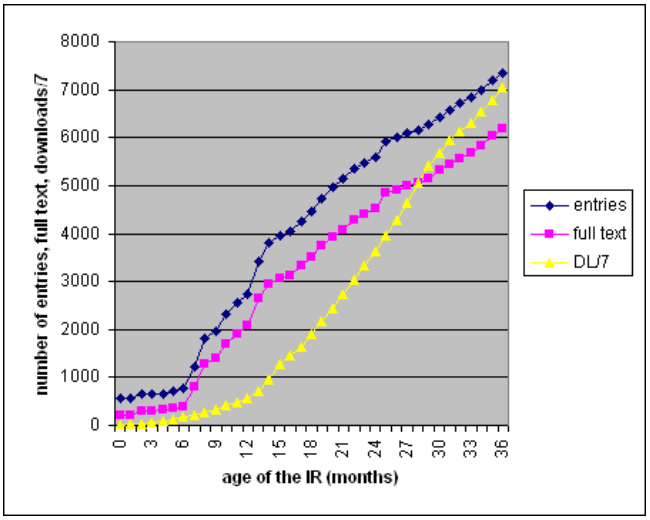

Figure 2: Growth and full text of the IR.

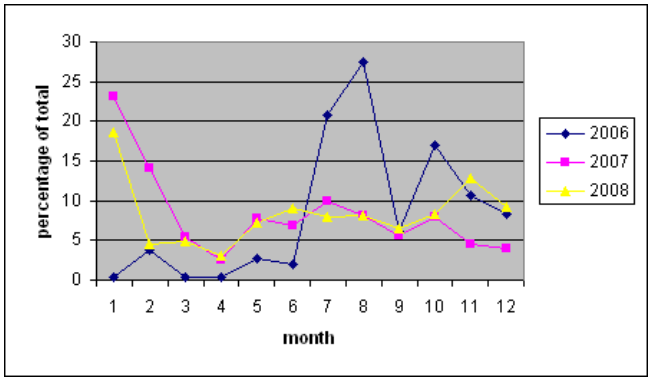

Figure 3: Montly submissions to the IR.

	2005	2006	2007	2008	Total
Legacy	693	1044	1733	1091	4426
Mandate	0	1147	1116	671	2934
Total	693	2191	2849	1762	7360

Table 1: Number of submissions to the IR in each year.

8. QUANTITATIVE ANALYSIS

To quantify the cost and benefits of an IR we need a small set of representative parameters. We propose three sets of parameters. The first set is an indication of the quality of the IR, such as the coverage of the IR. The second set is related to cost, particularly related to the amount of time users and IR staff spend maintaining the IR. The third set quantifies the benefits, such as how many users use lists produced by the IR on their own web sites, and the number of times users did not have to bother with yet another demand for information. We present only an initial set of parameters; we realise that for a comprehensive set more work must be done, particularly in collaboration with the community.

Some parameters are readily provided by the IR but effort is somewhat more difficult to measure, because the only tool we have to measure the amount of time spent by users and editors entering and processing meta-data and full text is the log file of the web server, which shows the date and time accesses to the IR made by users. Users and editors have to be logged in to update the IR so we know who is working on the IR. For example, given the fragment of the Apache log below, we assume that user Pieter has worked for 8 seconds on the IR.

```
130.89.148.32 - pieter [04/Dec/2005:20:42:58 +0100]
"GET /perl/users/staff/edit_EPrint?
            EPrintid=43&dataset=IR HTTP/1.1" 200 15895
130.89.148.32 - pieter [04/Dec/2005:20:43:06 +0100]
"POST /perl/users/staff/edit_EPrint HTTP/1.1" 200 14042
```

We assume that if more than 10 minutes have elapsed between successive log entries for the same user, then he/she has been doing something unrelated to the IR. We have been able to verify that this assumption is indeed true for the editors, and we believe that the same assumption is reasonable for other users as well. We have repeated the measurements reported below with an elapsed time of setting of 1 minute, and 100 minutes. Neither setting alters the essence of our conclusion that the effort expended on the IR is growing.

The IR grows steadily with a significant percentage of full text. Figure 2 shows how the IR has grown over the first 36 months. The top curve gives the total number of records and the middle curve shows the number of records with full text in PDF (about 85%). The graph is typical for a system with a mandatory policy, as it shows a more or less constant growth (of on average 200 entries per month), which is consistent with the idea that researchers never stop publishing papers. The bottom curve shows that the number of full text downloads (divided by 7 to scale the graph) grows more rapidly than the number of items in the archive. We have taken care to exclude downloads by robots and from our university; hence our figures are lower than those often published in related work.

Submissions are not evenly spread over the year. Table 1 shows the number of submissions to the IR during each year, and Figure 3 zooms in on the monthly submission statistics. The peaks in January (month 1) of 2007 and 2008 are due to users (mainly secretaries from the EL groups) frantically trying to enter their publications from last year before the university imposed deadline. In 2005, publications were still entered into Metis; hence there is no peak in January 2006. The peaks in months 7 and 8 of 2006 are due to a campaign to get legacy into the archive, supported by our student assistants. The submissions are spread a little better over the year for 2007 and 2008 than over 2006, but there will still be a peak in January 2009, which at the time of writing (5 Jan 2009) has not materialised yet. The peak in January and February puts significant strain on the editors, which is undesirable. In spite of all our efforts to make publicity for the IR we have thus far failed to level the load.

The effort needed to populate the IR grows. Figure 4 shows how the average time per entry needed to populate the IR has evolved over time. If the number of new entries in a certain month is relatively low (which is the case in months 1, 3 and 4 of 2006 with respectively 9, 6, and 9 entries) it does not make sense to divide the total effort spent on the IR by this relatively small number of entries, as there is always background activity due to the copyright and consistency checks. The year 2007 shows the least amount of effort per entry, because this is the year where the largest amount of legacy has been imported by our student assistants (See Table 1). Please bear in mind that, as explained in Section 4, the student assistants deposit the meta-data directly in the inbox of the relevant user, so that he/she does not have to spend as much time on a legacy entry as for a mandated entry. The time spent by the student asistent is not accounted for in the statistics of Figure 4. The graph

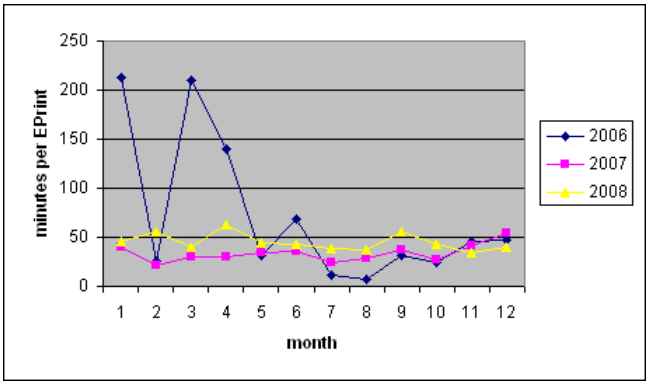

Figure 4: The amount of time spent interacting with the IR web application.

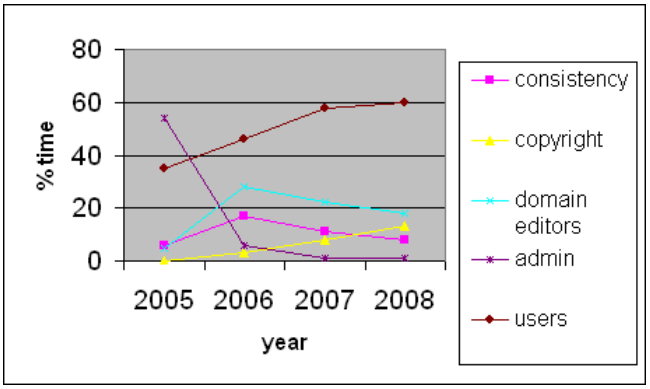

Figure 5: The division of labour.

for 2008 shows an overall increase in the amount of time per entry, whereas we had expected that the cost of populating the IR would fall as users became trained. The year average amount of time per entry in 2006 is 25 minutes, in 2007 it is 33 minutes, and 2008 it is 43 minutes. These times may appear already high but in reality this is only a lower bound because we have not accounted for the time that people need to discuss matters, to meet, and to email each other etc. The inescapable conclusion is that we completely failed to meet objective 5 of Section 1. We do not know what the cause of this increase is.

The work is well distributed over the users and the editors. Figure 5 shows how the relative workload of the users and the IR team has evolved. In 2005 the lion's share of the work was done by the administrator setting up the IR, preloading a first batch of about 600 publications to initiate the system. In 2006 the domain editors were getting used to the system; as a result in 2007 and 2008 their relative workload dropped. The work of the library for the copyright and completeness checks has steadily increased, as we were able to delegate an increasing number of checks to the library. The work involved in the consistency checks has dropped because of the implementation of better online checks (summer 2008) as part of the user interface. The work of the users has increased because over the years because the editorial team has become stricter in adhering to the policy, thus bouncing back more items.

Our IR directly benefits only 1/3 of our users. We believe that a good indication for the popularity of an IR is the fraction of researchers that use the output of the IR to produce publication lists on their personal home page. It would have been too time consuming to count how many of our 600 researchers use lists generated by the IR on their personal or project web pages. Instead we have looked at the personal home page of the first and the last full time academic listed on the web site of each of the 26 groups by way of a sample. From the sample 52 academics (out of 150), 19 present a list generated by the IR on their home page. Of the 15 who do not present a list of publications on their home page, 12 are well covered by the ISI and also by Scopus, hence they do not need a personal publication list. Unfortunately 18 academics maintain their own list, thus duplicating the work of entering meta-data and uploading the PDF. The most plausible reason for this is that the academic already had a list of publications before we introduced the IR, and he/she does not want to spend time entering legacy into the IR. We take these statistics as an indication that only about 1/3 of our users actually use the IR for their personal web site, 1/3 do not need it, and 1/3 do not want to use it. This is a pity, because once an individual uses the IR output himself, he will be motivated to enter publications as soon as they appear.

9. COMPARISON

Two IRs resemble our own, both in terms of size, and in nature (mostly mandated self-archival): QUT eprints [3], and RepositóriUM [10]. We discuss the salient characteristics of both before giving a comparison in terms of our parameters.

QUT eprints[19] started about a year earlier than our IR, and by Feb. 2008 had collected 9673 entries from 1100 academic staff, i.e. approximately 9 per staff. 75% of the QUT entries have full text. There are two important differences between our IR and that of QUT. Firstly, QUT EPrints has reached 50-60% coverage whereas we have reached 100% coverage. Since the academics in both institutions are required to provide a complete list of their publications, we do not understand why QUT does have such a low coverage. Perhaps our IR team has been more insistent to have all output entered in the IR. If this is so, then at least the significant cost of populating our IR can be justified. Secondly, while we insist on the authors providing full meta-data, QUT EPrints has few mandatory fields: *If the library can source it, then it is not mandatory*. This means that some types of output list (i.e. all publications pertaining to a particular project) cannot be created by the QUT IR. The QUT study [3] claims that entering a publication takes a user about 5 minutes. No data is provided to back this up, nor does the QUT study state how much effort the library has to spend completing the meta-data.

RepositóriUM[20] started about two years earlier than our IR, and by Feb 2008 had accumulated 6296 entries from 1087 academic staff, approximately 6 entries per staff. Unlike any other IR that we know of, RepositóriUM has awarded a financial incentive to groups who actively enter their publications into the IR. This has been successful, but it has not lead to 100% coverage. The RepositóriUM study [10] notes a sharp increase in the number of submission towards the deadline in order to qualify for this incentive. We take this as an indication that a financial incentive will not solve our problem of levelling the workload for our editors. The RepositóriUM study [10] claims significantly higher download rates (on average 100 downloads per paper) than we have observed (about 7). The study does not comment on the amount of work required to populate the IR, except to mention that on the library side approximately 2.5 fte work on the repository. Assuming a 1350 hour work year,

[19] `http://eprints.qut.edu.au`
[20] `https://repositorium.sdum.uminho.pt`

	EEMCS	QUT	RepositóriUM
Quality related parameters			
age (years)	3	4	5
entries	7360	9673	6296
acad. staff	150	1100	1100
av. entries/staff	49	9	6
% full text	85%	75%	?
% ISBN/ISSN	100%	?	?
Cost related parameters			
av. cost/entry (min)	>30	>5	>120
Benefit related parameters			
coverage	100%	50-60%	?
av. downloads/entry	7	?	100
staff popularity	33%	?	?
pre-empt demand	1	1	?

Table 2: Quality, cost and benefit related parameters of three similar IR.

this translates into $4 \times 2.5 \times 1350/6296 \approx 2$ hours per entry.

Comparison Table 2 summarises the three groups of performance parameters of the three IR compared.

- *Quality related parameters* The three IRs are comparable in size but ours caters for a faculty, whereas the other two cater for a whole university. We interpret our relatively high number of entries per staff as an indication that it is easier to reach out to a relatively small community. All three IR have a high percentage of full text, which is the main idea of building an IR. Our IR boasts complete meta-data for the mandated period of 2006-present. No data on this parameter is available for the other two IRs.

- *Cost related parameters* The costs of stocking an IR seems to vary wildly. We believe that the cost cited by QUT is not a tight lower bound. Our 33 minutes figure is a better lower bound; we estimate the total cost, including all hidden costs to be at most double the lower bound (i.e. approximately one hour per entry). Perhaps the estimate of RepositóriUM is on the high side.

- *Benefit related parameters* With an IR that has 100% coverage it is possible to satisfy information demands without bothering the users (shown as pre-empt demand in the table). In the short history of our IR this has happened just once (See *University IR easily supplied* in Section 6). With the next RAE on our doorstep, we anticipate another success. The staff popularity (measured by counting how many staff use IR generated lists on their personal web site) indicates how popular the IR is. We are hoping to be able to improve this in the years to come.

On the whole we believe to have built a better IR than the QUT and RepositóriUM but we have had to deal with far fewer users.

10. CONCLUSIONS

As many before us we can confirm that building an IR, even with a mandated self-archiving policy is hard. We are the first to quantify the costs of building an IR. We have succeeded in building a well stocked (100% coverage over 2006-present, on average 49 papers per academic), high quality (all ISBN/ISSN accounted for, 85% full text) archive. The most alarming find is that the cost of populating the archive seems to be rising rather than falling.

We would like to suggest a series of recommendations for developers of an IR:

- Develop a clear policy and implement it gently. However, be sure that your users will disagree with the policy and with good reasons too. Take your time to listen and to discuss the issues, adapting the policy where possible, interpreting the policy with flexibility where necessary, and taking a firm stance otherwise. With time most issues will be resolved.

- Plan for things to go wrong, because they will. For example the data base will become inconsistent at some point, the access control will be broken, and data will magically disappear. Keep good backups and monitor the health of the system.

- Build a team with all the relevant areas of expertise present. Building a repository is hard enough if you have people on your team who know what they are talking about.

- Try to develop clear and consistent web forms, and don't stop refining them. It is amazing how a few well chosen examples can prevent common mistakes.

- Building a repository with high quality meta-data and a large percentage of full text is expensive. Make sure that the management is aware of this.

- We believe that building a repository with complete information will be cheaper in the long run, than satisfying ad-hoc demands for a variety of information by university management, sponsors, RAE, government etc.

- Quality control for an IR is best done by academics. This is not a matter of push back but a matter of authority and credibility.

- Build in measures to quantify the quality, cost and benefits of the IR.

We believe that in order to be able to compare the quality, cost, and benefits of different IRs in an objective manner, much work remains to be done. It will be necessary to extend and refine the relevant, to develop better measurement techniques for the parameters, and to aggregate the parameters in a more meaningful and compact set of figures of merit. We believe that our approach towards measuring user time is a first step towards an objective framework for evaluating IRs.

As future work we plan to address the following issues:

- Improve the durability of the full text in collaboration with the university library.

- Reduce the effort required to navigate the data entry forms. There are indications that the present user interface, which requires the user to fill in a series of forms, is not optimal. Perhaps a single form is more appropriate.

- Reduce the effort required to gather relevant data. It might be possible to extract some of the meta-data from publisher's web sites, or from the PDF. However, since there is no uniform presentation of information, even on the website of a single publisher, it will always be necessary to check the information manually.

- Increase the intrinsic motivation of the user to contribute to the IR. It may be possible to develop a policy that gently persuades more users to stop maintaining their own lists of publications, and instead to start using the lists generated by the IR.

Acknowledgements

The support of present and past members of the EEMCS IR team is gratefully acknowledged: Peter Asveld, Axel Belinfante, Marc Berenschot, Karel van den Berg, Erik van Doorn, Hans Groenland, Geert Heijenk, Maurice van Keulen, Jan Kuper, Wim Oosterling, Walter Nijhuis, Lynn Packwood, René de Ridder, Klaas Sikkel, Frank van Viegen, and Petri de Willigen.

Les Carr, and Chris Gutteridge of the Southampton Eprints team have provided a great service to the community and to us.

We are grateful to the many student assistants that have helped us.

11. REFERENCES

[1] P. Baudoin and M. Branschofsky. Implementing an institutional repository : The DSpace experience at MIT. *Science & Technology Libraries*, 24(1 & 2):31–45, Jun 2004.

[2] S. J. Bevan. Developing an institutional repository: Cranfield QUEPrints - a case study. *OCLC Systems & Services*, 23(2):170–182, 2007.

[3] P. Callan. QUT repository case history. In *3rd Int. Conf. on Open Repositories*, pages 118–126, Southampton, UK, Apr 2008. University of Southampton, UK.

[4] L. Carr and L. Lyons, editors. *3rd Int. Conf. on Open Repositories*. University of Southampton, UK, Southampton, UK, Apr 2008.

[5] T. Cochrane and P. Callan. Making a difference: implementing the EPrints mandate at QUT. *OCLC Systems & Services*, 23(3):262–268, 2007.

[6] P. M. Davis and M. J. L. Connolly. Institutional repositories – evaluating the reasons for non-use of Cornell university's installation of DSpace. *D-Lib Magazine*, 13(3 & 4), Mar 2007.

[7] M. Ferreira, E. Rodrigues, A. A. Baptista, and R. Saraiva. Carrots and sticks – some ideas on how to create a successful institutional repository. *D-Lib Magazine*, 14(12), Jan 2008.

[8] M. Greig and W. J. Nixon. On the road to enlighten-ment: Establishing an institutional repository service for the university of Glasgow. *OCLC Systems & Services*, 23(3):297–309, 2007.

[9] M. A. Kennan. Academic authors, scholarly publishing, and open access in australia. *Learned Publishing*, 20(2):138–146, Apr 2007.

[10] E. Rodrigues and R. Saraiva. RepositoriUM case history. In *3rd Int. Conf. on Open Repositories*, pages 80–90, Southampton, UK, Apr 2008. University of Southampton, UK.

[11] G. van Westrienen and C. A. Lynch. Academic institutional repositories: Deployment status in 13 nations as of mid 2005. *D-Lib Magazine*, 11(9), Sep 2005.

[12] M. Ware. Institutional repositories and scholarly publishing. *Learned Publishing*, 17(2):115–124, Apr 2004.