Applications, Potential Problems and a Suggested Policy for Institutional E-Print Archives

Christopher Gutteridge and Stevan Harnad

19th August 2002

Abstract

EPrints archives and similar archives promise many benefits for academics and their institutions. This is generally a good thing but there are complications in trying to solve too many problems at once. This article describes the potential uses for an institutional archive running either the GNU EPrints software or software intended to provide similar functionality and how those applications may complement or interfere with each other.

This article then discusses policy decisions which should be made when implementing an archive, and goes on to suggest a possible policy based on our experience.

At Southampton the Electronics and Computer Science Department has been running an archive and database of our publications since 1998 and has provided software and assistance to many other institutions setting up a variety of electronic archives.

1 Possible Applications and Benefits of Institutional Archives

1.1 List of some Applications and Benefits

The benefits attributed to EPrints archives, for academic institutions, are some or all of the following:

- Visibility: More people will read your department's written work as it will be openly accessible on the web.
- Impact: More people will use and cite your publications.
- Preservation: Some or all of your publications will be stored in a central archive with properly managed backups, and URLs which can be maintained for far longer and with less difficulty than those on user home pages.
- Searching: Potential users everywhere will be able to find and use your work much more effectively.
- Integration: Researchers and administrators in your institution as well as external users will be able to find, use and track your research and other written output much more effectively.
- Automating Administrative Data and Analysis: Universities are these days doing more and more compilation and analysis of their research and publication output for funding and assessment exercises such as the Research Assessment Exercise (RAE). A publication database is a valuable resource for this. It can be updated, monitored, reconfigured for different purposes, and analyzed using the scientometric measures of impact and usage that are rapidly developing currently.
- Author Publications Lists: Generating lists of publications for biographies or staff information web pages.

- Sub-Group Publications Lists: Generating lists of all the publications of a single sub-group or project. What a sub-group is would depend on the structure of your organisation.

- Citation Linking: Linking the references in a publication (to the full texts, where possible). Further forms of scientometric analysis are rapidly being developed too as performance indicators.

- Probity: There is some work in systems which can prove that a given document existed on a given date. This functionality may well become an integral part of institutional archives.

- Other: Such as exporting meta-data in formats such as BibTEX or research which uses for the data or the interface.

Some of these benefits will be used to promote the uses and benefits of the archive to the institution management and academics.

1.2 Requirements for various Applications

Exposure, Impact, Preservation, Probity and Citation Linking are applicable only to records with full texts attached. Authors can't cite a paper they can't read.

Automating Administrative Data and Analysis, Searching, Author Publication Lists and Integration only require the meta-data *but,* except for searching, these require meta-data for all relevant records to be deposited.

Author Publication List Generation proved to be a very effective carrot in getting our academics to fill in their records, but it has resulted in our database also receiving records for many papers which had been written by our staff members well before they came to the our institution.

2 Our Experience of Implementing an Archive

Since summer 2002 we have been running GNU EPrints 2, but the meta-data and texts have been accumulating in our previous "Jerome" database/archive since 1998. [1, 2]

There was very little advice or precedent on how to set up such a system so we took the approach of allowing any kind of deposit with the intention of pruning later when we had a clearer policy. Also we took the initial approach of not making any of the meta-data fields "required". We were concerned that if an author became frustrated when depositing items then they might give up entirely.

From an initial investigation of what people were already doing we discovered that two of our research groups maintained BibTEX files of all their publications which could be imported quite easily but without any full texts: meta-data only. To gain the support of these research groups we added a BibTEX export feature to the system which meant that these groups could continue to have exactly the same functionality they had had in their own database.

An unexpected side effect of this was to quickly make the data base look "busy," which in turn inclined other members of staff to become more willing to spend the time entering their own meta-data. Looking at other similar archives, one cannot help noticing that many are almost empty. This may be because busy people are unwilling to be early adopters. Once an archive has a critical mass of records it becomes visible and used, thereby demonstrating its usefulness. Other are then willing to deposit their own work in it too.

Our archive now has over 6500 meta-data records of varying quality. The data which was originally imported from BibTEX has not been as closely checked as might be desirable, but checking thousands of records serially is a significant investment of time. Spending just 3 minutes per record is 50 hours per 1000 records.

An alternative is to distribute the vetting of the deposits in parallel to smaller

groups with designated vettors. This is only necessary for the first wave of legacy material. Daily input can probably be vetted by far fewer designated vettors in the steady state.

Of the 6500+ records, only about 10% have the full text attached. Although more than 30% of the 2002 papers have the full text.

This situation is likely to be common in archives which do not require the full texts and raises some questions about policy.

One possible policy might be to encourage full text for everything but to allow legacy material to consist of meta-data only if necessary, while insisting on full text for current and future material.

3 Questions that should be Answered when Implementing an Archive

There are a number of decisions which should be made before you start expecting more than a test group of staff members to make deposits in your database. Some of these are discussed here.

3.1 Quality Control and Editors

You must decide who are responsible for the data being in the database and accurate: The author of the item, an "editor" for each group or project or an overall editor. Would you rather set up the software and the system in such a way that depositors can enter rough meta-data and then tidy it up when and if they have time or in such a way that depositors must spend their time getting records exactly right the first time (in which case it might be harder to get them to do anything at all).

Who can edit a record once it has appeared in the system? How do you handle requests for deletions? Do you allow records to appear in the database right away or does a designated vettor approve them first?

3.2 Full Text

By full-text I mean an electronic copy of the entire item, as oppose to just the meta-data describing it.

Is the full text required? Do you require that depositors always archive the full text of a document? If you do then people will be unable to list things like books, which will mean you cannot generate listings of all publications of an author. If you do not then some people will just not bother even in cases where they can. (It may be best to have an explicit exceptions policy – for legacy material, books, and other documents for which there is a specific reason why the full text cannot be archived along with its meta-data.)

What full text formats do you require and what formats do you accept? This depends largely on whether your goal is to make your publications visible and openly accessible on the web or merely to provide an internal archive or list of your own publications. It is very possible you will wish to do both.

If you want people to be able to read your texts then the best choices are PDF, ASCII and HTML. PDF requires a browser plug-in and can be harder to cut-and-paste from but it might be the right choice if you wish to have a single required format.

If one of your goals is to preserve copies of the publications produced by your members then you may also wish people to deposit their "source format". The original file from which they were created will usually be a Microsoft Word document or LaTeX although you should be ready to deal with the occasional depositor who insists on using an obscure format which nobody else has ever heard of. In this situation you can still archive the odd format along with a PDF or other format that can be viewed in a web browser.

If you intend to run an archive for the purpose of preserving all forms of institutional digital output then it may be inappropriate to allow members of the public to read all the documents or formats. For example, you may not want people to be able to download the original LaTeX versions. Also, you may want to archive the full text of a book written by one of your members and make it

available internally to members of your group or institution, but it would cost sales of the book if you gave it away free on the web. In this case, only the meta-data would be openly accessible.

You may also wish to allow an author to deposit other formats in addition to the required format or formats. For example, you require PDF or ASCII, but may wish allow an author to optionally deposit a MS-Word or Corel Draw version in addition to one of the required formats.

3.3 Meta-data

Meta-data is the information about your publications which you store. This is used for searching the records and for rendering descriptions of the records. It is also what makes your institutional archive's contents inter-operable with the contents of other OAI-compliant institutional archives.

Does your meta-data have optional and required fields, or just optional fields? If you have required fields then this requires more effort on the part of depositors. If you have very strict rules on what information must be provided, depositors may become frustrated or might find arbitrary ways around your rules. For example, with "abstract" a required field, someone might try to deposit a record which might not happen to have an abstract at all by arbitrarily typing "no abstract" as the text in the abstract field in order to get around the requirement of filling the abstract field. (These improvisations by depositors are not necessarily problems, but it is best to identify as many of them as possible in advance, and perhaps provide specific advice on the "help" page.)

Do you have the concept of different types of record? For example a conference paper may have the required field "Conference Name" but a journal article will not even have that field as optional.

Do you wish to have a subject tree? If so, do you prefer to make it rich and descriptive or simple and easy to use? If you want authors to self archive and your subject list is too long then they may become frustrated or fail to classify

their documents correctly or fully in its terms. The other problem with a very rich subject tree is that the same item could reasonably go in a dozen different places and should really go in all of them yet may only end up going in one.

3.3.1 Names

How do you identify members of staff who are authors or editors of an item? One of the most common searches people will want to do is for "all bob's papers". Names are inadequate. You currently may not have two people with the same family name and initial, but the situation can and will change. You need a way to uniquely identify people. This will depend on what systems you have already. Email address/user-name or staff ID number may work. Remember that the person entering the data must be able to easily find out what the ID is of their co-workers.

Also Identifiers which are not entirely abstract can change. If email addresses are based on name then someone may get married and change their email address. If you used employee ID numbers then someone may leave then come back several years later and be assigned a new ID. You must be ready to handle these situations. This may involve someone doing a search and replace on the database.

Is the name of the author in the meta-data the name as it appears in your staff database or as it appears on the paper/publication?

3.3.2 The Sub-Groups Problem

Another advantage we mentioned was that you could generate a list of all the publications of a sub group. In the case of our database the sub groups in question were research groups of about 50 to 200 people. A field in the meta-data associated each record with one or more research group. This worked fine until research groups got reorganised. A group changing name or joining with another group is relatively easy to handle by running commands on the database, but a group splitting in two will have to be handled by someone

making a decision on a record by record basis.

An alternate solution is to leave the old records belonging to the old group and start the two new groups as having no records. The best solution depends on your organisation.

We have a database which lists which group each person is in. An initial idea was to generate the list of records for a group by listing all the papers of all the people in that group. This proved to be a false start. Occasionally people would write papers which belonged in groups other than their own. If someone moved group then all their papers would move with them. If someone had entered all the papers they have ever written to generate their biography page then all those papers would be listed as belonging to their current research group.

3.3.3 Equations

Another unexpected problem you may encounter is that physics, electronics and maths papers may sometimes have an equation in the title or abstract. Two ways to represent an equation in text are MathML and LaTeX MathMode. You should consider if this problem is likely to occur in your archive, and if so be ready to advise people how to express equations.

3.4 Scope

What kind of records should be deposited? Just academic papers? Other things people may try to add include magazine articles, web-sites, books, chapters in books, patents and even software.

Also do you allow people to deposit items which they wrote before they started at your institution. The advantage of this is that it means you can automate the building of "my publications" pages for staff, which is a great carrot, the disadvantage is that more than half the records in your database may not actually have been written by people while they were at your institution. This prevents you from being able to generate useful statistics like "we wrote 40

papers and 74 conference posters and 8 books in 2001", unless you can clearly identify records which do not belong to your institution. The author-affiliation field on the publication itself might be a way to tag this; or there might be a "document-created-at-institution-X" flag.

3.5 Multiple Publications of a Single Item

A single paper may be published in more than one place. You may wish to require that only one instance is entered in your archive, that all the instances are entered in your archive or that multiple meta-data records are associated with a single document.

3.6 Additional Applications

Do you plan to build lists of staff publications? The pros and cons are discussed in the "scope" section.

Do you want to support OAI or otherwise export your meta-data in standard formats like Dublin Core or BibTEX? If you do then you may need to consider adding additional fields. BibTEX has a "Bib-type" field which is the type of the record and must be one of a limited set. You may or may not be able to map the "type" field from your own records onto this.

If you plan to support OAI then you should consider what rights you are and are not going to grant on your meta-data. Are all the meta-data and full-text data to be open access? Are all harvesters allowed to collect it without your permission and provide searches?

Are they also allowed to charge for that search or sell the meta-data as part of a service?

3.7 What's in a Name?

Whatever working title you pick for the archive may well become what it is known as. We ran into problems as we started calling our archive a "publications

archive" which was not strictly true as it also contains pre-prints. A good choice may be an "e-prints archive" as people will not have any preconceptions about what that is. The same is also true of the meta-data fields. A careful wording of the field names can avoid numerous little confusions and hence save time and improve meta-data quality.

3.8 Copyright Concerns

Authors being asked to deposit their work will often have concerns about the legality of making their work available on the internet. The self-archiving FAQ at eprints.org addresses these concerns.

> "Texts that an author has himself written are his own intellectual property. The author holds the copyright and is free to give away or sell copies, on-paper or on-line (e.g., by self-archiving), as he sees fit. For example, the pre-refereeing preprint can always be legally self-archived." [3]

The exception to this is the case where the author has signed their copyright rights, or exclusive publication rights over to the journal or other publisher. A common example is the text of a book, where the author usually gives exclusive rights to the publisher.

4 A Suggested Policy Based on our Experience

Based on our experiences at Southampton I would recommend the following as a good default set of policies, although these should only be treated as suggestions. Some of these policies we have implemented, some we may implement, some I wish we had implemented when we started four years ago.

4.1 Quality Control and Editors

I would recommend a single editor to make overall decisions about policy. Then a few sub-editors who proof read meta-data before approving it and assist and advise people depositing records. People deposit their own records but once they have been approved for the main archive and become "live" depositors must ask a sub-editor to make any changes in their archived records. The number of sub-editors you require will depend on how many users you have depositing records, the level of support your users require and how many records per month are being deposited.

You will also need a technician who understands the database and can do the "clever" stuff like SQL edits or adding new features. After an initial period of being very time-demanding, the technical staff time to run an archive has been very low, in our experience.

4.2 Full Text

Do not make full-text obligatory in all cases. Simply encourage it very strongly. Especially for things like books. You might perhaps require depositors to specify why they have not deposited full text so as to encourage the sense that full-text is the normal and encouraged option.

You can't expect an academic to figure out how to produce PDF. Some just will not do it. Accept full text in whatever forms it comes in and have an editor produce a PDF copy where necessary before the record is accepted into the main archive.

Access rights are up to you. We allow the depositing staff member to indicate which documents should be password protected, with the option of limiting viewing to members of the department or just the depositing staff member and archive editors.

4.3 Meta-data

This is difficult to advise, but a good approach is to pick an existing meta-data scheme and then add your own extra fields. Four years ago we started with the BibTEX fields, which may not have been a perfect solution but has served adequately. Additional fields we have added include

- Subject
- Research Group
- Status - one of unpublished, in press, published.
- Refereed - Yes/No.
- Performance Indicator - A code used when we are producing our return for the Research Assessment Exercise.
- Full Text Publicly Available - Yes/No. This is an interesting field. It is set automatically by the system. It is only set to Yes if the record has a full text attached which is not password protected.

I would also recommend the possible addition of a "Created-Here" field, to the above list, although we don't have such a field ourselves (yet). This field would indicate whether or not the record was created by someone while in our department or merely included to aid the biography generation.

It may seem that a text field meaning "Created-At" would solve this problem too, but free text fields increase errors. I would recommend using text fields where the data is primarily going to be read and searched by people. When the fields are going to be used to generate statistics they should be as unambigious as possible. An ideal solution would be to have a piece of javascript which enters "University of Foobar" into the "Created-At" field if the depositor sets the "Created-Here" option to "true". This would maximise the accuracy and utility of your data without generating additional work for the person making the deposit.

Searches by the general public should have "Full Text Publicly Available" and "Created Here" both defaulted to "Yes" to avoid people getting unhelpful results.

I would recommend some required fields. Having no required fields seemed a good way to increase the number of records deposited initially, but has lowered the quality of our data.

We do not use a subject tree. If we did we would probably use the library of congress subjects, but only include the detailed levels of the branches relevant to our work. In this way we would have subjects which could be easily fitted into a full Library of Congress listing by an OAI harvester or when we are absorbed by a university-wide database.

4.3.1 Names

We identify members of staff by their user-name. It is not ideal but it has the advantage that people already know each other's user-names. We ask that names should be entered in the way that they appear on the publication and provide separate fields for honorific (Sir), given names/initials (Marvin H.), family name (Fenderson) and Lineage (Junior). Lineage and honorific seemed a good idea but have confused people. Asking for given names and family name instead of first name and surname avoids or at least reduces any confusion when entering names where the family name appears first.

I think that Given Name, Family Name, Identifier (optional) is probably enough but make sure that your interface can handle an unlimited number of authors.

4.3.2 International Characters

It is easy to assume that you will never have to deal with non-ASCII characters but it has been my experience that occasionally our staff will collaborate with staff from countries which use other characters. All modern web browsers understand unicode and so I would encourage you to ensure that your meta-data

is stored as UTF-8 or UTF-16 rather than ASCII or ISO-LATIN-1. [4]

4.3.3 Equations

We found a workable solution to this problem. Which was to ask the author to enter equations in LaTeX mathmode in and store them as such in the meta-data. When rendering the meta-data, a script looks for anything which appears to be latex and replaces it with an image. The URL of the image is a CGI script with the LaTeX string as a parameter . The CGI script returns a rendered image of the equation. Both scripts are part of the GNU EPrints system, but could be easily adapted to work in other archives.

4.3.4 Superscript, Itallics and Other Mark-up

Some authors may wish to include markup in their abstracts and titles. From setting the "th" of "5^{th}" to be super-script to wanting itallics and bulleted lists in their abstracts. This is not an essential feature so may not be worth the effort. A useful minimum might be to support two blank lines indicating a paragraph break in an abstract or other large text field.

4.4 Scope

This is up to the editor, but I would suggest starting with a very open policy and then narrowing it later if you feel you need to. A "created here" field will avoid confusion as you can clearly specify "not created at Southampton University" (insert your institution) next to items which were not so that people do not think you are taking credit (or blame) for them.

You will probably not guess every kind of record people will want to deposit so be ready to add new record types (or an explanation of why you will not).

A good place to draw the line on what to accept would be software, web-sites and other things which cannot be rendered into printed pages. An alternative is to throw the doors wide open in the first instance and see what people want to use the archive for.

4.5 Multiple Publications of a Single Item

If multiple instances of the same item with slightly different meta-data would cause problems for some additional service you are offering then require that people handle these cases in the manner suited to this service. Otherwise let the author select how they wish to deposit their own work.

4.6 Additional Applications

In our experience, generating lists of staff publications encouraged people to enter their data. We generate staff information pages which contain a list of that person's publications. When a staff member discovered that their page said "John Smith has 2 publications in the ECS publications database" it seemed to prove an incentive to add the remaining 150.

The publications database can also be an important resource for generating, maintaining and updating an on-line CV. Meta-data tags can be generated to cover other typical CV items (former institutions, employment, talks in a given year, grants, etc.). These can all be put in the same database and used to generate an up-to-date CV.

I would recommend supporting OAI as this will improve the "discovery" of your records. Unless you have reason to do otherwise I would suggest allowing the meta-data you export to be used in any way people wish on condition that it is not distributed in a modified form without your permission. People should only have the right to download the full text of a record for their personal use unless stated otherwise in the meta-data of the record, or your permission is asked and given (and yours to give).

I would be wary of adding other additional applications. Which is not to say do not do it. But if you create an experimental service it has been my experience that you often discover people now depend on it, no matter how much the documentation explained it was experimental. If you add a new service, be prepared to support it.

References

[1] GNU EPrints Software - http://software.eprints.org/

[2] University of Southampton Department of Electronics and Computer Science EPrints Archive - http://eprints.ecs.soton.ac.uk/

[3] Self-Archiving FAQ for the Budapest Open Access Initiative (BOAI) - http://www.eprints.org/self-faq/#self-archiving-legal

[4] What is Unicode? - http://www.unicode.org/unicode/standard/WhatIsUnicode.html

[5] Example of metadata containing equations: http://eprints.ecs.soton.ac.uk/archive/00006296/

[6] Stephen Pinfield, Mike Gardner and John MacColl "Setting up an institutional e-print archive" 11-April-2002, Ariadne Issue 31, http://www.ariadne.ac.uk /issue31/eprint-archives/intro.html