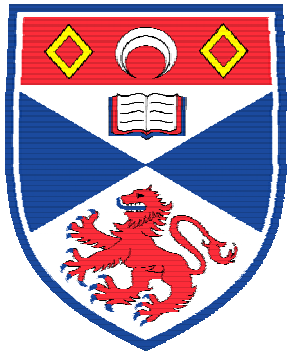


University
of
St Andrews

# HaIRST Project Report

Note: This report has had some of the text dealing with security scripts removed. The full text is available on application to the eprints administrator.

May 2003 – November 2004

**http://eprints.st-andrews.ac.uk**

# HaIRST :: Implementation & assessment of pilot project

*St Andrews is a member of a consortium investigating Harvesting Institutional Resources in Scotland. The project is directed by CDLR, University of Strathclyde.*

# HaIRST Project Report

*University of St Andrews: May 2003 – November 2004*

## The HaIRST Project

The HaIRST (Harvesting Institutional Resources in Scotland Testbed) project commenced at Strathclyde University in 2002 and aims to 'investigate the deposit, disclosure and discovery of institutional resources in the JISC information environment'.

St Andrews University is one of the partners in the project, whose overall management and direction is controlled by CDLR at Strathclyde University. Other partners in the project are Napier University and a consortium of ten Glasgow Colleges of Further Education and the John Wheatley College.

One of the key areas of the project, and the one which St Andrews is primarily involved in, is the creation of metadata which will be harvested and disclosed by Strathclyde. In order for this to be accomplished successfully, *standardisation* and *interoperability* must be issues which receive close attention. However, this will be done by CLDR at Strathclyde; consequently it does not receive attention in this report, but we have borne in mind that part of St Andrews' remit is to:

- create a suitable archive which will deliver metadata in an approved format, simultaneously addressing the issues of standardization and interoperability;
- gather institutional material for the archive. 'Institutional material' means any material generated at St Andrews either in the present or in the past. The focus will be both on research work (at any level) and administrative/informative documents. If the material is from the present, then eprints would act to disseminate current material – in the case of research, it would serve to increase the impact of any work, in the case of informative material it could be used both by prospective and by current staff and students to learn more about the University, its facilities and its regulations. From this point of view an ancillary function might be to fulfil the requirements of the Freedom of Information Act. If the material is from the past, then the eprints repository will function as an archive. Consequently this is an *Institutional* eprints archive which focuses on exposing the resources of a specific academic community unlike the majority of *subject-specific* archives which accept data from a variety of institutions e.g. ArXive, CogPrints;
- report on the problems encountered in the above two actions.

The above has focused on the rôle of St Andrews, as this is the main thrust of the current report. However, if you require more information about the entire project's objectives, especially on the reconciliation of heterogeneous metadata, this can be found at **http://hairst.cdlr.strath.ac.uk/**

### eprints Concept

At St Andrews the HaIRST project is closely tied to the eprints concept (this is not the case at most of the other institutions participating in HaIRST). So it would be useful to give a very brief introduction here. More detailed information on self-archiving and related matters is readily available on websites and email lists such as

**american-scientist-open-access-forum@listserver.sigmaxi.org**; **oai-general@oiasrv.nsdl.cornell.edu**; **http://www.oaforum.org/index.php**; **http://www.erpanet.org**; **http://www.epublishingtrust.org/,** and, of course, **http://www.eprints.org/.** There are many more, and the eprints site give a plethora of links to related sites.

Briefly, eprints attempts to increase the impact of research through self-archiving of preprints and of final papers. In some disciplines, the time between having a paper accepted and its publication can be in excess of five years – during this period the researcher's work has no impact on the academic community, and conversely it reflects neither on him/her self nor on the qualities of the department. This problem is recognized even by many mainstream publishers; for example Elsevier have an Open Archive compliant preprint server on which anyone can place their refereed or final articles (**www.compscipreprints.com**). Articles are not limited to those which are due to appear in Elsevier journals.

This line of though is backed up by recent research, e.g. in Antelman, Kristen. *Do Open-Access Articles Have a Greater Research Impact*? College and Research Libraries, 65(5), 372-382. September 2004 where the abstract reads:

> Although many authors believe that their work has a greater research impact if it is freely available, studies to demonstrate that impact are few. This study looks at articles in four disciplines at varying stages of adoption of open access-philosophy, political science, electrical and electronic engineering and mathematics-to see whether they have a greater impact as measured by citations in the ISI Web of Science database when their authors make them freely available on the Internet. The finding is that, across all four disciplines, freely available articles do have a greater research impact. Shedding light on this category of open access reveals that scholars in diverse disciplines are adopting open-access practices and being rewarded for it.

A large number of similar findings are also reported in Appendix II.

Another related feature of self-archiving is that as any work placed on the server is by default freely available to anyone who has www access, so people in institutions where the cost of subscription to learned journals is not affordable are no longer excluded from current research.

The following two diagrams (figures 1 & 2, on next page), from Stevan Harnad (Professor of Cognitive Science Southampton University) help to explain this further.

Many lecturers self-archive on their own web pages and you may think that this is adequate and less trouble than maintaining an institutional site like eprints. But when they leave, these pages are not maintained and are eventually deleted. This leaves St Andrews with no on-line readily accessible way to demonstrate that quality and quantity of its work over and extended period of time.

Furthermore, even when leturers' pages are on-line they may remain hidden to the casual Internet searcher. You need to remember that indexing by search engines is through keywords extracted from the text, and although it is possible to add metadata to a web page most people do not bother, and some search engines ignore them, anyway. *Structured* metadata implies consistent and accurate information retrieval even if the searcher is unaware of the title of the paper or the name of the author.

An example might make this clearer: if the word *Maastricht* appears in a document is this about the history of the town? Its geography? Or is it about British

foreign policy? Maybe the EU? … . Using Google, I found that *Maastricht* gave about 212,000 hits.

This is not helpful.

This can be avoided to a large extent by centralising web-published papers and adding metadata in a consistent and internationally-recognised format..

This merely needs the will to do it.

The final plank is that the cost to library acquisitions of journals is extremely steep and has been rising very much faster than either library budgets or inflation (see figure 3). Further examples are the journal *Tetrahedron* which was $11624 in 2000 and had risen to $15126 in 2003 for one year's subscription. This is not the most highly-priced journal, either, though if bought as a package with other Elsevier journals its price would be substantially lower. Although the primary purpose of eprints is about dissemination of work, nevertheless the free availability of research papers must also make an impact on the prices charged for journals.

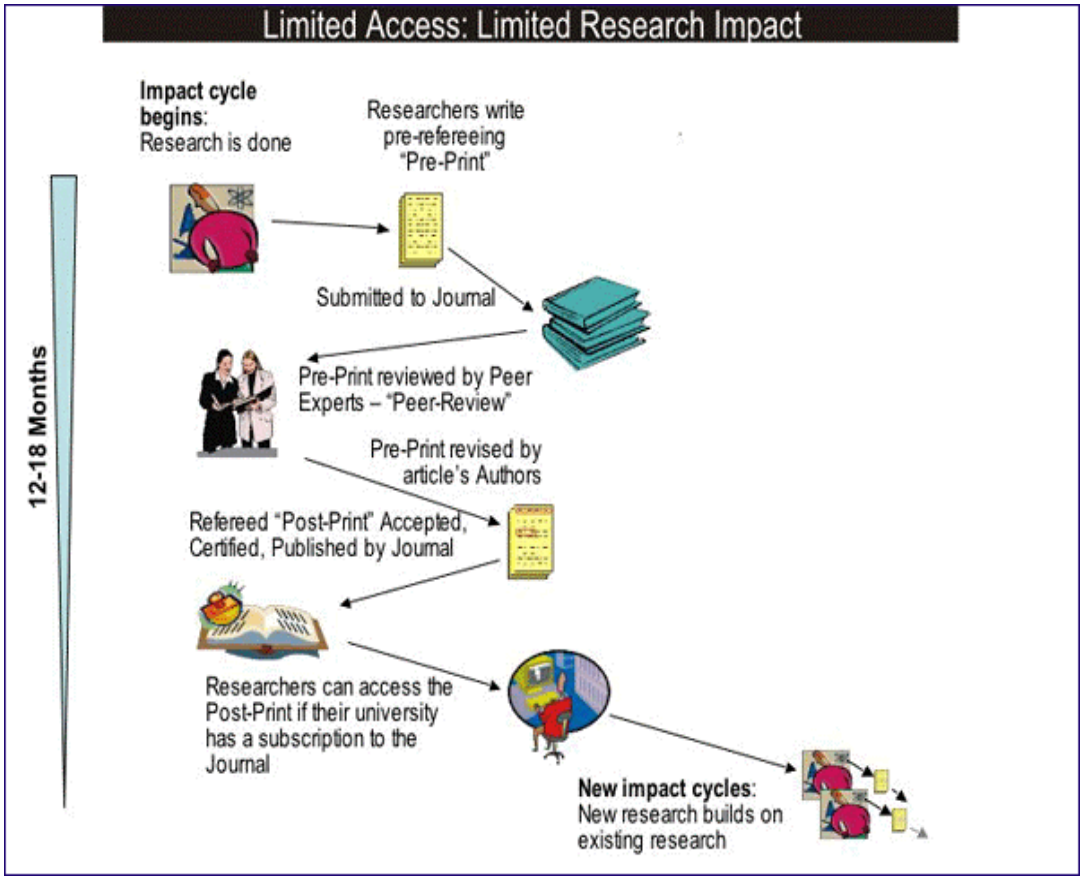



*Figure 1*

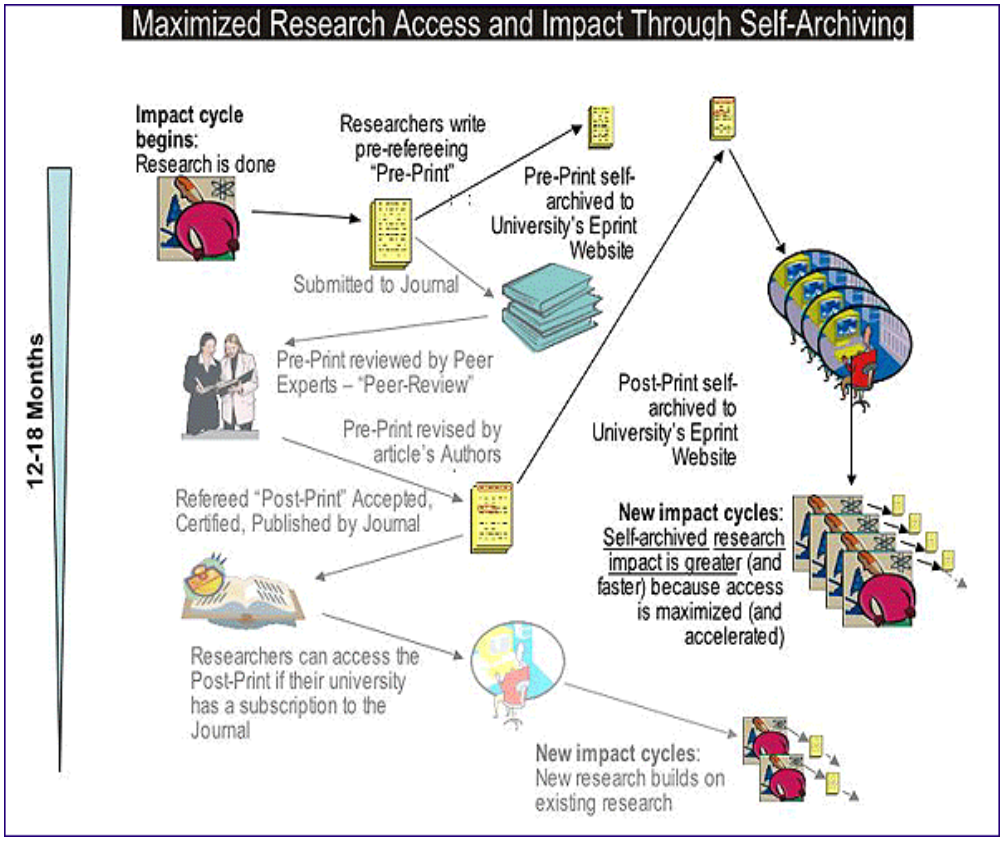



*Figure 2*

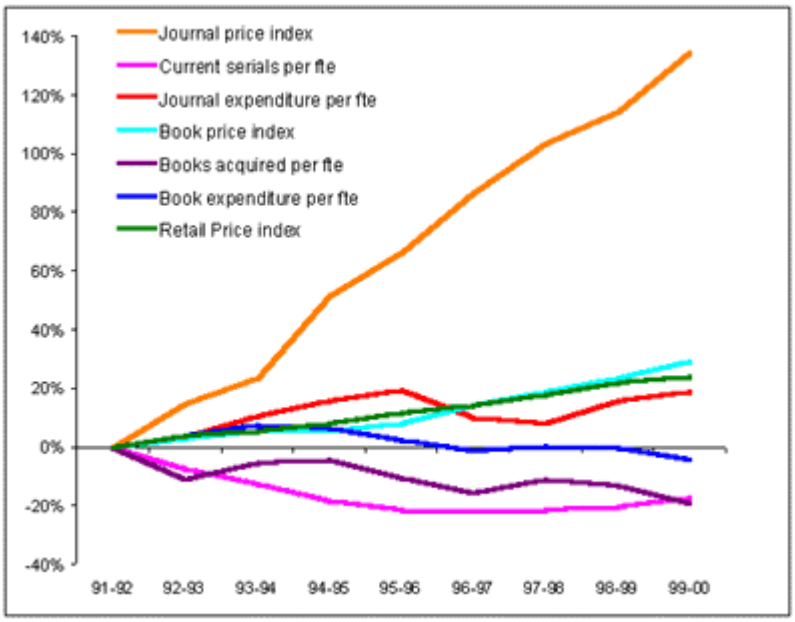



*Figure 3*

## Institutional Repositories

The position of an institutional repository is slightly different from a subject-specific one, as it is designed to serve a specific institution rather than a discipline. Consequently it has two primary aims, viz.

- to increase the impact of research (exactly as for a subject-specific repository)
- to act as a showcase for the University.

In the latter case we took it to mean that any material produced by the University would be considered for entry into the archive. As well as research material, this could be ephemera such as the postgraduate handbook (only available on the web, and regularly updated) or the Computing Laboratory Newsletters, historical material out of copyright but not worth reprinting and theses, whether undergraduate or postgraduate. Much of this material could well disappear or be filed away only to be forgotten. By putting it on the eprints server it can remain as a record of what has happened and what is happening at the University.

This is more than a 'nice-to-have' concept. The reality is that information from the University is hard, if not impossible to find. It is well-known that the internal search engine used here can, at best, be described as awful. An example of this is when I tried to find out about car-parking here. The first few results, presumably the best, are given in Appendix IX. Notice that only one result has 'car-parking' and the most popular one is in fact dealing with car hire. (Google's results are not brilliant, either, but at least they do deal with car parking and the information is readable.)

This is an issue which needs addressing and an institutional repository is one straightforward way of doing this.

We also need to remember that the need to have current *non-academic* material exposed like this is hardly an option at the moment as the Freedom of Information Act will mean that the public has mandatory access to it, and the easiest way to ensure this is to have it freely available on an institutional server. Such as eprints.

# Hardware

At one time, hardware was a major issue in the implementation of any computer-based project. Times have changed: most 'off-the-shelf' hardware will do the job well enough and the short time interval between hardware being cutting-edge and being obsolete is so small – usually in the region of 3-4 years – that no special effort is required to assess whether the project needs hardware updates or not – the standard time for upgrades to servers is almost certain to be adequate; all that is needed would be to insert the eprints server into this schedule. If the server remains shared, as it is now, then upgrades/maintenance are already in place and nothing more needs to be done.

Aside from this, two issues need to be clarified when discussing hardware requirements:

- the requirements for the present project;
- requirements for university-wide operation.

Issue (2) is speculative until the time when the University comes to a decision on this matter, hence it is realistic to focus on the former. However, there are a number of factors which overlap the two domains, as we shall see.

The software chosen for this project (see next section) is designed to run on linux/unix and is open source and makes efficient use of resources. Consequently any computer platform which can run either linux or unix is suitable; in reality the most common system is an IBM or compatible with Intel processor(s), but both Sun and Apple Mac systems have also been used as shown by the correspondence on the eprints technical list (see archive at: **http://software.eprints.org/tech.php/**). It is almost certain that any up-to-date hardware platform which is suitable for the current testbed project would be equally suitable for a full-scale system. Thus the only hardware technical issues we need to consider are:

- the availability of a server running linux;
- disk space.

## Requirements for the present project

Space on a shared server (Kingfisher) was purchased through funds made available through HaIRST. The demands on this are not excessively high so that no more than replacement in the normal upgrade cycle is likely to be needed - say every 3-4 years as already discussed. Disk space has not proved to be a problem, with only 24% of the 30G disk space used by all the data and software (this includes a physics database and other material which is unconnected with eprints). Access time and search time were insignificant – the software reports these values and they are usually in the range of 0.1 of a second or less.

## Requirements for University-wide Service

As already noted, hardware which has proved itself adequate for the pilot project will also be adequate for a full-scale service, and in this case the hardware has shown itself to be reliable and extremely fast. The other issue to consider would be the capacity of the disk which holds the eprints data. Depositing large number of coloured PDF files could use up disk space fairly quickly. For example, at one

extreme a 140-page book scanned in colour occupied 17M; on the other hand the size of a research paper ( PDF, black and white) was only about 100k.

This problem is not as great as first imagined. Many Arts papers are monographs, with few, if any, coloured plates. Similarly in many of the sciences black and white predominates. Unless the eprints are all images (e.g. all scanned PDF files) then the space for each deposit will average at about 0.5 to 1.0 Mb, which is not significant in terms of today's hard disk capacity. If more disk space needs to be purchased, the costs are insignificant; current HD prices can be as little as £75 for a 160Gb disk (Maplin web catalogue, August 2004, **http://www.maplin.co.uk/**) and adding disks to a unix/linux system is usually straightforward.

Another example is that the cost of storing the content of arXive , the physics and computer science eprints archive (contents exceed 250,000 eprints) was less than £300/year (prices for the first half of 2003 and it is reasonable to suppose that they would be lower now). If we add the cost of server and terminal replacement we are still budgeting at under and average of £1500 a year in 2003 figures, probably closer to £1000 when the pilot project ends in November 2004. (The above data is from JISC report: *Feasibility and Requirements Study on Preservation of E-Prints*, October 2003, **http://www.jisc.ac.uk/uploaded_documents/e-prints_report_final.pdf )**

## Scanner

Currently we have access to a fast scanner/printer (HP LaserJet 9000mfp) which not only has a document feeder but will also email the scanned file direct to a mailbox. This scanner was used in the current project to transform archive materials into PDF files, but this could have been done with a small flatbed scanner. The situation might well be very different if the project were to become University-wide. In this case the speed and ability to scan batches of sheets automatically would make this a very desirable purchase. The alternative would be to limit severely the amount of non-electronic material which could be put on eprints and in view of the rich archive materials available this would be a retrograde step.

The situation is complicated in that the scanner has not, apparently, been purchased by the University, but has been given, by HP, for a project to enable students to do distributed printing – they could email scanned materials to their mailboxes and then print them out at another computer. This project appears to have been abandoned, consequently, the ownership of the scanner would need to be clarified before its use was promoted on a wide scale.

The scanner itself can send files in a variety of formats, in black-and-white and in colour. However, the only recommended file format it produces for text is PDF, which is exactly what is required. However, the size of a colour PDF file can be quite large and the University Webmail has severe problems in dealing with files which exceed the pre-defined limit for users – I had to get the helpdesk to move the file to another location on the Sun machine before I could deal with it – prior to that my inbox refused to open and sundry other errors were produced. It is *much* easier to ensure that mailboxes are big enough to start off with.

# Software

The policy of the eprints software project (**http://eprints.ecs.soton.ac.uk/)** is to produce quality software which is freely distributed and available to anyone under the GNU gpl license and to use other software only when it too fulfils this requirement, allowing both the eprints module and its supporting software to be freely distributable. (see **http://www.gnu.org/copyleft/gpl.html** for details of the license agreement and the concept of *copyleft* as opposed to *copyright*).

Unless otherwise stated, the software described below is non-proprietary, stable, scalable and with the source code available. This contrasts strongly with proprietary software (e.g. Windows/Windows-based) which appears to lack most - probably all - of these attributes. If software maintenance is to be considered in the future then issues such as source-code availability are not optional. This is an issue which is important, as the JISC support which eprints software production has had is stopping September 2004. Southampton University are committed to carry on with the eprints project (the software is the most widely used of its type in the world), nevertheless a change in both the support and software creation models is only to be expected.

A final point is that a version 3 of the software is under preparation, though it may be several years before a stable release is offered. Meanwhile the currently distributed version (2.3.x) is actively maintained and will be for a considerable time to come (information from eprints workshop, London, 23-24 June 2004). St Andrews eprints is running on version 2.2.1 which has been upgraded and modified to suit the St Andrews contribution to the HaIRST project and to remove bugs as they were reported.

Current LMS (Library Management Systems) suppliers are also discussing plug-ins which would give eprints-style functionality to current library systems; this included the supplier of this University. Of course, here you are dealing with proprietary software which has to be bought and maintained by an external supplier. How customisable it really is and what its true costs might be have yet to be determined. There is one certainty – you will be locked into the system.

## Front-end

In discussion with CDLR at Strathclyde the GNU-licensed eprints software was chosen to host the St Andrews archive. This had the following advantages:

- relative stability;
- it was customisable;
- runs on linux and the consequently the operating system is stable;
- Perl source code available through the GNU copyleft licence;
- interfaces with a highly-stable and very fast database;
- interfaces with highly-stable web-server.

A Windows machine was used as a terminal onto the linux operating system and also to prepare some of the software customisations which would then be transferred, via a program such as WS-FTP (available though the University), to linux.

Other software for doing a similar job is also available, but many of these systems have aged; consequently they would not be a wise choice. The most popular contemporary alternative to eprints is DSpace (a joint venture between MIT and HP). It would have been useful to investigate this further, and there was a workshop in Glasgow to discuss its implementation. In spite of the fact that this was being run by the DSpace consortium and was the only one in Europe that year, it was impossible to go to Glasgow to investigate this option further. The only local experience is from Glasgow University in 2003, who found that over three months' of part-time work were required to get the basic system running; consequently it was regarded as beta-test quality. During 2004 its stability and ease-of-installation has increase substantially, possibly making it a viable alternative to eprints if its specific features are wanted. For example, Edinburgh University have recently used DSpace for their '*Theses Alive*!' project and Cambridge University is using DSpace 'to develop a digital repository for the University'. See **http://www.dspace.org/** for general information and **http://www.ariadne.ac.uk/issue38/jones/** for an evaluation and comparison of DSpace and ETD-db (from Virginia Tech) in Edinbugh's *Thesis Alive!* It is worth noting that Dspace is also losing large-scale institutional support and, like eprints, is looking for a new funding model to continue support and development.

## Database

Having decided on eprints software, the database which holds the information is effectively pre-determined, as the software is designed to interface with MySQL, a linux relational database. This is a well-proven technology, surpassing all of the commercial databases in speed, stability and scalability, though its speed advantage is at the expense of the non-implementation of some esoteric database commands.

There are times when direct control of the information stored in the database is necessary and the software used for this depends on whether control is to be from a linux system or from a remote machine running another operating system. In the current system a Windows-based program MySQL Control Centre (**http://mysql.com**) was used and proved to be entirely satisfactory. It is distributed under the GNU public license, but unfortunately comes with no documentation whatsoever.

## WebServer

Apache is the software of choice. It is used by the majority of web servers in the world and is stable and well-proven. Eprints interfaces directly with it.

## Web Authoring/Design

Dreamweaver (**http://www.macromedia.com/** and available through the University)was used for some of this work. Note that this is not free software, but is proprietary, though the University has a site license lowing the cost from about £200--£400 to about £30. In fact it was not used a great deal and could be dispensed with, depending on the skills available from technical support or the eprints administrator.

## Connectivity

A variety of software may be needed to connect a Windows or Mac machine to the eprints server for software development or modification. In this case *putty* and

*ws_ftp* were used. Both were free, though proprietary, and available through the University. To make it straightforward to run programs such as Emacs on the linux box, an X-windows connection to linux had to be installed; this was done through the proprietary *Exceed* software, available through the University. Graphic linux programs can now be run on the Windows machine via this connection.

Note that these are not required for a user to access the St Andrews eprints site - this is done through a web interface by using the following address: **http://eprints.st-andrews.ac.uk/**

## Text Preparation Software

Modification or adjustment of the texts was often necessary, so several pieces of software were installed to make this process easier. I list them here for completeness, but their use will be described in greater detail in a later section *Text Preparation*:

- Adobe Acrobat Standard – this is not just the reader, but allows text manipulation and the insertion and removal of pages as well as the conversion of web pages and MicroSoft Word directly to PDF. This was available through the University and is commercial;
- Gimp v.2 – this is the standard image manipulation package distributed under the GNU license;
- Adobe Photoshop – similar to Gimp, but more sophisticated and commercial. This was available through the University;
- WebReaper (**http://www.webreaper.net**) is software which will harvest whole websites, and follow their links to any user-defined depth. It was used to gather websites together as a set of files, which could then be converted to PDFs. This is free software, but not distributed under the GNU license, so no source code is available.

# Maintenance and support

## IT services

For the duration of the project, these are paid for by HaIRST. Longer term costs are affected by the expertise of the eprints administrator. If they have a basic knowledge of Windows (the least necessary skill), linux, Perl and Apache, then little outside help would be required; in most cases problems could be resolved with an average of an hour's work per week or even less from support. This time is likely to be spread unevenly; new versions or upgrades will require more time, whereas a working system will need very little support as it is inherently very stable.

## eprints administrator

The rôle of the administrator in program maintenance will depend on their expertise as discussed above. I would emphasise again that linux skills are of value here - Windows skills are of little use unless reliance is to be placed wholly on IT services (I would assume a basic familiarity with Windows, of course: the discussion here is about maintenance which implies some in-depth knowledge).

I would suggest that passing the maintenance requirements entirely to IT would be a poor policy as even the most trivial glitches or bug fixes would incur delays which would be tiresome in an active system.

Support of users, rather than of computer software, would entail familiarity with areas such as indexing or cataloguing and also being aware of the importance of structured data (such as DC) in order to assist in creating the metadata page required for each deposit.

## General Maintenance

In the current project this has involved the following:

- Installation of the eprints system (eprints software, Perl modules, MySQL, Apache);*
- Modification of the Perl code. The eprints support list (**eprints-tech@ecs.soton.ac.uk** or the wiki connected to this (**http://wiki.eprints.org/**) is quite active, and modifications for bug fixes, improvements and special-purpose modifications are regularly posted. All relevant bug fixes have been implemented and so have some of the improvements. However, this has meant modification of the underlying Perl code of eprints – in this case using *emacs* and *eXceed* (the Windows interface to X on the eprints server). An average modification rate of one change every 2-3 weeks can be expected, though few of these visibly affect the user interface. All these changes have been documented on the relevant pages of the source code printouts and also in the first record of work, both of which accompany this report. The patches are also available on the archive of the eprints software site;

- Some changes to the MySQL database have been necessary, and may need to be done in the future, though this should be very rare. The *MySQL_cc* software* enables viewing and control of the remote database in a very straightforward manner. Equally it allows easy total destruction of the database, so it needs to be used carefully, or only by IT support;
- The main page of the St Andrews eprints site was designed on Dreamweaver and then transferred across to the linux server using ws_ftp. This was the easiest way to design a graphically-oriented page which was meant to be similar in appearance to the University Library home page. Other pages were modified using emacs on linux, but there is no reason why they should not be made in a way similar to the opening page.

Regarding the last point, eprints software has a page-modification facility using XML elements called ‘pins’ which enables text to be inserted or modified. E.g. the majority of the text used is in the file ‘phrases-en.xml’ and have the form:

<ep:phrase ref=”eprint_fieldopt_month_jun”>June</ep:phrase> which allows the word ‘June’ to be automatically inserted at any point in the page layout where ‘eprint_fieldopt_month_jun’ appears.

* indicates that the software was installed entirely or mainly by IT support.

# Content

So far I have looked at the hardware and software requirements for the St Andrews eprints project and also at the maintenance and support issues – all of which need to be addressed before a system like this is launched. It is now time to consider the function of eprints – to present content to the St Andrews community or to the outside world (so long as they have access to a reliable www connection, which will exclude most of the world's population).

It will be best to look at content from several viewpoints:

- The categories of materials deemed suitable;
- Publicizing the archive;
- Depositing
- Self-depositing by users;
- Assisted depositing;
- Transforming the text to a suitable format – this implies a certain regularization of format and issues of file size, whether due to length or to the presence of colour;
- Availability and security;
- Copyright.

… and I will take them in that order.

## Categories

As this is an *institutional* archive then virtually any material released by St Andrews is a potential candidate for inclusion. This would include:

- Research papers at any level; staff, postgraduate and undergraduate. These may be preprints (drafts), material already accepted for publication or material already published. They may also be 'grey literature' – items such as good undergraduate theses which usually disappear onto departmental shelves, never to be consulted again;
- research material not formally refereed, but presented as position papers, conference posters etc.;
- book/monograph chapters (it is not envisaged that whole books would be put onto the server unless they are out of print and of historical interest). These chapters or summaries would act as 'tasters' for the purchase of the whole book;
- material often classed as 'ephemera' such as guides to departments or to libraries. These can be of interest to the future historian, especially as much is now published on the web only, and is changed or deleted as the occasion demands. An example of this is the postgraduate handbook at: **http://www.st-andrews.ac.uk/publications/pghandbook/** whose 2003-2004 version has now been preserved on our eprints server;
- other departmental material;
- reviews of research or of the university featured in popular magazines or newspapers.

## Publicising the Archive

Publicising the archive can be on a number of levels and all need to be addressed, probably over a long period of time, for success to ensue. This is largely due to the need for a change of orientation from staff, both administrative and academic, and a percolation of this change down to the student level. It is unlikely that these changes will come either quickly or without any resistance. Certainly there will be those who would welcome the idea of eprints, and due to the very brief nature of the project, publicity was directed at this group whenever possible, encouraging them to deposit their papers etc. in eprints. On the other hand, other departments had little interest and were very resistant to the whole idea. Some of this may well be a reflection of the culture of St Andrews, on the other hand, some disciplines are known to be resistant to this idea on a world-wide scale; for example there are very few chemistry archives, whereas physics and computer science have been very much at the forefront of eprints internationally, though little enthusiasm was seen at this University.

Apart from approaching individuals with a known interest in eprints, other methods of publicity were tried, including presentations at seminars (e.g. Arts & Humanities computing Workshop, IPR workshop) and specific eprints presentations. In spite of publicity though LIS newsletters etc. the latter had a very poor attendance of non-library staff.

Having contacted individuals and given broad-based presentations, another method for publicity was to target departmental heads and vice-chancellors. These were viewed with interest, but little more came of them with the exception of Art History, which was willing to allow me to write to the students who had obtained a First Class Honours in the summer 2004 exams to see whether they would be willing to place their dissertations on the eprints server after removal of any copyright images. Although eprints could be seen as an institutional resource, therefore making a case for all undergraduate theses to be made available, the feedback from all the schools was that only Firsts should be exposed in this way.

What was strange (and I did not get to the bottom of this) was that a number of schools initially were interested in an eprints presentation, but simply failed to respond when I suggested that we should set a date and time. Nor did they respond to a reminder.

## Depositing

During the publicity stage, one of the matters I was very conscious of was that fact that I was asking people to do extra work. Natural resistance makes it unlikely that they would wish to deposit material if it cost them time and trouble, so we went to some effort to ensure that they all knew that they could either deposit the material themselves, or that if they sent it to me as a PDF, MS-Word or RTF file, then I could do all the work for them. This was also made clear on the website home page with a prominent line labelled '*Let us archive it for you*'.

There is also another factor to consider here and that is of user access to eprints for depositing (there is normally no barrier to access for reading the eprints). After some consideration, it was decided that a working system should allow self-depositing by university staff and postgraduate students, but that undergraduate students should not be able to do this to prevent spurious texts appearing on the website or taking up the eprint administrator's time. This will be considered further under *Security*.

One of the other factors was that the HaIRST project was due to finish in November 2004, giving less than a year of activity (the software had to be customized to suit St Andrews and to become an institutional, rather than a subject-specific, repository, also the last month or two of part-time work would be writing reports etc.) At the time of the various eprints presentations I was asked to make it clear that this was a trial project and that users should not expect this to become a service. Naturally, this would put a damper on even the most enthusiastic advocate of eprints. Why should they bother if everything was to end in less than a year? This is always a problem, but I felt that it was particularly unhelpful in a project which was being promoted as an archive and as a way of displaying work at the University – both having implications of persistent data. One of the lessons learned was that a considerable push 'from the top' is needed for people to take this as a serious option and that the effort to self-archive simply will not take place unless the lecturers have some confidence in the long-term prospects of their work bearing some fruit in terms of greater visibility and a higher likelihood of citations.

I will now deal with the two methods of depositing – *self-depositing* and *assisted depositing.*

## Self-Depositing

This is the ideal of the eprints concept, and has been used successfully for over a decade in some subject-specific repositories such as ArXive, but this depends on a culture of self-deposit either at the University or at a School/Departmental level. I was surprised at the resistance to self-deposit at St Andrews even in departments where lecturers had their papers on their own websites. Unfortunately it might mean that if other universities are pursuing the open access model more vigorously, then much of the research work at St Andrews remains unseen and wasted compared to the impact it would otherwise have (this is probably more so in the sciences than the arts). For more information see *Citations and Open Access* under *Beyond the Project*.

Views from others working in this field show that the main way to change this – perhaps the only way to change this – is by a mandated policy to deposit research material, whether this be by self-deposit or by assisted deposit.

As it turned out, only two papers were self-deposited by one author and one paper by another author.

Some other researchers were very interested in depositing older papers which had already been published, but this proved to be difficult as the copyright issues were difficult for me to resolve, and is discussed further in the subsequent section *Copyright.*

Finally, you need to be aware that only those who have registered with eprints can deposit material, though anyone can look at it unless permission to view it outside St Andrews has been revoked. And registration has to be done from a St Andrews domain address; this implies that self-depositing is only possible by those working in the University, making this a truly *institutional* site. You may have questions about what I have just said, so the issues raised in this paragraph are discussed in far greater depth in an entire section *Security*.

## Assisted Depositing

As already mentioned, I thought that assisted depositing would be the preferred method of putting materials into eprints, and this proved to be the case, though taking

the long-term view I think it would change if eprints became an institutional resource. However, there would always be need for assisted depositing as people may not have enough time, or feel familiar enough with computing, to be confident about self-depositing. For those members of staff who did want to deposit their papers, virtually all of them either emailed the PDF or Word file to me or gave me a URL where I could find it. It was then purely a matter of cut-and-paste – the title, abstract, author etc. were normally part of the paper and so could be inserted directly into the eprints metadata page. The paper itself was simply loaded into eprints at the appropriate stage of the depositing process. It was very straightforward, and ideally occupied little more than 10 mins.

The last point to consider in assisted deposits is the creation of metadata. In the case of papers or dissertations, this is not a problem, as items such as keywords, abstract title etc. can simply be transferred by cut-and-paste from PDF files which have been created from text documents (this is not possible with PDFs which are essentially pictures – there is a facility for character recognition in Acrobat but at best this can only be considered as poor. Tests on scanned documents have confirmed this.) The time taken to transfer a document to the eprints archive, including the time taken to create the metadata page, can be in the region of 10 minutes.

If all the metadata is not present (e.g. a keyword list is missing, or there is no abstract) then the time to deposit will reflect the time to create this information. Perhaps 15-20 minutes should be allowed – certainly the time is very unlikely to exceed the latter value. The author could be asked to supply the information, or a list of possible entries created by the administrator could be sent for their approval. Where I have created some of the metadata, I emailed the author(s) and asked them to check it and either correct it themselves (they never did) or email me any corrections (this they did do).

Clearly, assisted depositing is bound to take more time from the administrator's point of view, but it is to be expected that an initial service would be largely of this nature and that self-deposit would creep in over a period of years. I find it difficult to image that self-deposit would replace assisted depositing entirely.

### *Art History Undergraduate Dissertations*

For this pilot project, some students in Art History were approached and asked whether I could put their final-year dissertation on eprints i.e. they were not given the option of self-deposit. This part of the project was accomplished with the help and co-operation of the Professor of Art History, Ian Carradice, and after discussing the matter with him, we agreed that only those who obtained a first class degree should be asked. This is likely to have limited the response to an extent as the students had left University by the time the results were announced, however out of twelve letters sent (14th July 2004), there were five replies by the end of August. They could be classified as follows:

- two still intend to send their work to eprints, but are chasing up copyright issues. I have had recent emails from one and they are still trying to send material to the site. This can show how keen some students would be to have their work publicised in this way;
- two have submitted their work, which is now on eprints. In both cases pictures were removed by the authors;
- one did not want their material on eprints as the pictures were integral to the thrust of the dissertation, and she did not have copyright to any of them.

It is still possible that more material may come to eprints through this, but in essence it shows that the students are prepared to go to considerable effort to put their work on eprints and have all the publicity which that entails (the metadata page is picked up and indexed by all the major search engines within a few days). If a similar eprints strategy were to be implemented earlier in the year (being quite open about the fact that only those getting firsts would be asked to have their material put on eprints) then I think that the take-up across all disciplines is likely to be quite high. Also, copyright issues are usually clearly defined at these levels.

### *Special Collections Materials*

To increase the bulk of the archive, and to have some materials to show during initial presentations, I scanned some archive material of institutional relevance and put it onto eprints. This consisted of the following:

- entire books which were out of print but available in the library. Examples would be some of the Rectorial addresses from the late 19th to the early 20th century;
- publications by the library itself such as a book on the heraldry of St Andrews (this was scanned in colour);
- publications of a genuinely ephemeral nature such as the Computing Laboratory Newsletters, which started in the 1970s and show the evolution of hardware and software over that period, the rules of a Hall of Residence for 1948, floor plans of the original new library building;
- current internet-only material such as the Postgraduate Handbook which would otherwise disappear completely when revised for the next academic year (2004-2005).

Fairly current materials have the advantage of already being in electronic format, hence putting them on eprints is fairly straightforward, taking not much more than about 10-15 minutes, most of this time being taken up with creating PDFs, or downloading websites, then creating and filing the metadata. For the older materials, scanning time has to be taken into account and although no book took longer than about 90 minutes, the work was very tedious.

## Quality Issues

The depositing process in eprints automatically addresses quality issues. The results of the deposit are OIA- and DC-compliant (Open Archives Initiative : **http://www.openarchives.org/** ; Dublin Core: **http://dublincore.org/)** and are harvested by OAI-compliant archives, e.g. OIAster (**http://oaister.umdl.umich.edu/o/oaister/)** whose purpose is to:

> create a collection of freely available, previously difficult-to-access, academically-oriented digital resources that are easily searchable by anyone.

Clearly, it is impossible to expect a member of staff to be able to create the appropriate metadata for their paper or for a report, without some guidelines, or, better still, a template which they could just fill in. Eprints provides the latter on the metadata page, which must be filled in before the user can deposit their paper. This template is dynamic i.e. it changes depending on the type of deposit – a book would require an ISBN, a paper would need the journal etc.

This template is customisable by the eprints administrator and has fairly extensive help associated with it. Customisation can involve the addition or removal of certain fields, making some fields compulsory (such as author, title, abstract, keywords) and giving some fields a default value e.g. the 'Institution' field defaults to 'University of St Andrews'.

When a person self-deposits, the document does not appear on eprints until it has been approved by the administrator, who will check that the fields are completed appropriately. In this way quality is controlled outside the actual depositing process – it is the next stage, which precedes publication on the eprints server.

One of the problems when staff have their own web pages is that metadata will almost certainly be absent, and even if present will not follow OAI/DC recommendations, yet we can easily implement a service which will do all of this automatically – in fact the service exists – it only needs full-scale implementation.

## Copyright

The issues of copyright for depositing new materials are very straightforward in many cases. For example research material produced by the university has copyright vested in the author - exclusively so in the case of undergraduates, but in a slightly more complex way in the case of postgraduates where there may be issues of potential patenting. However, postgraduates have a supervisor whom they would consult in this matter as they would normally do in publishing any of their material no matter what medium or format it is in. Staff are in a similar situation to postgraduates regarding copyright and intellectual property.

A large number of publishers, including the stable of imprints owned by Elsevier and Springer, allow for electronic publication on servers such as eprints, so long as the material does not duplicate the layout of the journal. A full list of these publishers, and the issues surrounding them, are on **http://www.lboro.ac.uk/departments/ls/disresearch/romeo/** or **http://www.sherpa.ac.uk/romeo.php**

For depositing older materials, the situation can be more complex as the author:

- may be a joint author, in which case the other authors would need to be contacted;
- a copyright agreement may already have been signed with the publisher, and it is virtually impossible to find out what this agreement allowed or prohibited.

In the latter case, I tried to obtain a response from six publishers in order to put one lecturer's past papers on eprints. Only one publisher replied, though they were surprisingly supportive considering the rather heavy tone of their website on the subject of copyright; they were happy to have a postprint which was not a copy of the published article i.e. they claimed a copyright on the layout, not on the content. The effort in tracking down the right person to contact, explaining the situation and asking for permission makes this impractical in all but the most important cases unless the publisher has a blanket policy to allow eprints, as for example, Elsevier have (they even have their own preprint server), in which case the paper can be put straight on eprints after ensuring that it is not a copy of the published paper.

And here again, I came across problems.

Unbelievably, some lecturers did not have an electronic copy of their paper – when it was published they accepted offprints and deleted or lost the original files. Consequently even when publishers were happy to have the material on eprints it was

not possible to do this as it was no longer possible to recreate a PDF file from the original text.

If a person self-deposits, then they are responsible for checking the copyright issues. If the depositing is by the eprints administrator, then they will check this as best they can based on the information with which they have, but would also need an assurance from the author. Many publishers ask for copyright to be transferred to them if their journal wants to publish a paper. Some demand this transfer even if they decide not to publish! This transfer of copyright can simply be deleted when the author signs the form, or changed from giving them an 'exclusive right to publish' to a 'non-exclusive' right. The Romeo project has not reported a single publisher objecting to this alteration of terms, and many (like Elsevier mentioned above, and Nature) are building these changes into their conditions for authors.

In view of this, it is recommended that the focus of eprints should be on archiving current content rather than trying to untangle the web of ownerships of past papers. Unfortunately, this does imply that older, but still valuable, papers – especially ones from obscure or obsolete journals – will only be available as photocopies from large central repositories and will be denied wide exposure and straightforward retrieval, but this cannot be helped unless a great deal of extra work is allowed for.

The need for copyright clearance is clearly shown as part of the eprints depositing process, where the following message is shown:

> **For work being deposited by its own author:** In self-archiving this collection of files and associated bibliographic metadata, I grant St Andrews eprint Archive the right to store them and to make them permanently available publicly for free on-line. I declare that this material is my own intellectual property and I understand that St Andrews eprint Archive does not assume any responsibility if there is any breach of copyright in distributing these files or metadata. (All authors are urged to prominently assert their copyright on the title page of their work.)
>
> **For work being deposited by someone other than its author:** I hereby declare that the collection of files and associated bibliographic metadata that I am archiving at St Andrews eprint Archive) is in the public domain. If this is not the case, I accept full responsibility for any breach of copyright that distributing these files or metadata may entail.
>
> Clicking on the deposit button indicates your agreement to these terms

So in both cases it is assumed that no copyright restrictions prevent the dissemination of the paper via the University eprints server. Most publications are happy to allow this (though they would not give permission if this was a 'for profit' company).

In my experience many of any of the problems associated with depositing eprints are perceived to be ones of copyright. Possible solutions to this are:

- increased awareness of copyright issues by university staff. This could be done by:
    - specific training/seminars, such as the one on IPR organised this year by RES;

- explanation of the problem by hooking into other seminars, e.g. computing for the arts and humanities;
- departmental presentations;
- dissemination of information through newsletters.

- Delegating the more complicated areas of copyright to the eprints administrator or some other member of LIS. It would be particularly important for this contact to be made *before* a member of staff signs a copyright agreement with a publisher. This method of dealing with copyright may be particularly appropriate for assisted deposit;

- A change in University policy to encourage awareness of copyright and how to deal with open access issues. This is – at least in theory – simpler than it seems; the University, as employer, owns the copyright of material produced by its employees, but in the case of academic staff has chosen not to exercise this right. It could choose to exercise it if it wanted to (this would probably be a form of suicide), but could also put pressure on staff to become aware of what they are signing away when a paper is accepted for publication.

Ultimately we should remember one important fact – that a publisher or third-party cannot assume any copyright over the author's materials other than that assigned by the author.

# Text Preparation

This naturally falls into several sections:

- choice of acceptable formats;
- preparation of electronic text;
- website downloads;
- preparation of non-electronic text;
- foxing and image manipulation;
- use of colour and pictures;
- keyword extraction for metatdata page;
- converting database entries to eprints format.

I will look at each of these in order.

### *Choice of acceptable formats*

This was governed largely by the recommendations of the AHDS (Arts & Humanities Data Service **http://ads.ahds.ac.uk**) and is presented on the eprints website as well as in Appendix VI of this report. Please note that the recommendations of AHDS may change from time to time, and the information in the Appendix is the same as on the eprints website.

The formats available are *required* i.e. eprints will reject files which are not in these formats. In reality it simply looks at the file extensions, so it would be easy to deposit a file in a different format by changing this extension. However, it would not render (display) correctly as a specific file format means that an appropriate rending algorithm is used to display it.

To quote from the eprints page:

> We have a number of **Required Formats** for your files and you must deposit at least one file in a required format. At present these are:
>
> - PDF
> - ASCII or unicode
> - HTML
> - postscript
> - rich text format
>
> **In general we would like you to save texts either as plain ascii/unicode or as PDF**
>
> Virtually every computer user can read these formats and knows how to open the files.
>
> Multiple files should be zipped (or tar/gzipped) for uploading; eprints automatically unzips them. However, there is an option to add a document after you have uploaded a file, so you can add that file you forgot!
>
> **LaTeX** files can cause a lot of problems with (a) .eps files for illustrations (postscript is proprietory and liable to change), and (b) the host of macros which most LaTeX files need. The advantage of LaTeX is that at core it is an ASCII file so that essential information is normally preserved even in the absence of these items. Comments on the suitability, or otherwise, of LaTeX and its variants would be welcome.
>
> **Microsoft Word** files should be saved as PDF, or, if you really can't do this, then save as rich text format (rtf). We do not accept Word files

unless they are accompanied by another file in one of the required formats.

If people felt unable (or were unwilling) to convert their texts to PDF or any of the other required formats, I was prepared to do this for them i.e. this was now assisted depositing. This fact was clearly presented on the above web page.

It is quite possible to change (add or remove) any of the required formats, but adding a format also implies that some way of rendering the text (to display it correctly) needs to be added to the eprints code; this may, or may not be straightforward, depending on whether there are public domain or copyleft versions of software required to do this. Removing a required format simply means removing it from a list – the rendering code is retained making it a straightforward job to reinstate that format if required.

The principal aim of required formats is simply to allow long-term readability of the submitted paper or document. In general I prefer PDF files as:

- the reader is free for download (contrast this to Word files where you have to have Word, or a Word-compatible program);
- the US government have said that they would maintain this format even if Adobe (the creators of Acrobat) no longer are able to do so;
- most computers have Acrobat Reader installed on them;
- most people recognise what PDF files are and how to open them.

Some of these points may appear to be trivial, but only if we consider computing in the first world. Many other countries do not have the resources to keep up with changes in computing hardware and software, but they, too, deserve consideration if we are looking at word-wide dissemination of the work and history of this university.

### *Preparation of electronic text*

All the texts I received were either already in PDF format or were in Microsoft Word, so little effort was required to make these suitable for eprints.

Using Acrobat Standard it was possible to convert text directly from Word or a Web page to a PDF file and this worked well with no problems with all the texts I processed.

Texts which were already in PDF were either put on eprints ‘as is’ or some very minor changes were made at the request of the author. There are two avenues open for this so long as the original text was in electronic format i.e. not scanned in:

- modify the text directly from Acrobat. As this is line-oriented, no justification or word-wrapping is available unless I do it myself. Hence if words are replaced and the text is right-justified, the modified line is likely to be shorter or longer than the original. Adding or deleting whole paragraphs will create less disruption to the layout than working on word or sentence level. Note that this sort of manipulation is really only effective for the most minor changes. If more substantial alterations are required then it is best to ask the author to do this on the original file and then re-save it as a PDF;
- convert the PDF file to Microsoft Word, make the modifications, then convert the modified file back to PDF. This was attempted with a paper which had the usual scholarly format (footnotes, references etc.) and the conversion was a total shambles. Footnotes appeared in the middle of

text, formatting appeared to be randomised as did the typeface and point size. This option might be worthwhile for short, simple documents, but as formatting gets more complex the usefulness of this process becomes very questionable.

### *Website Downloads*

For some texts it was necessary to download a whole website; an example of this would be the Postgraduate Handbook of this University, which is made up of a number of web pages. To do this I used *Webreaper* (detailed in the previous section *Text Preparation Software*). This was set up to gather all the pages from the site, but not to follow any external links. The output file consisted of the website which could then be converted to a PDF using Acrobat Standard.

All the procedures worked well and were straightforward – *Webreaper* had a 'wizard' to aid the creation of appropriate filters and Acrobat had sections in its help file (very badly indexed, as most help files are) which showed how to convert from website to PDF.

This section, which at first had the potential to be the most troublesome, in fact proved to be the easiest of all.

### *Preparation of Non-Electronic Text*

The only realistic method was scanning straight into a PDF file. We had an appropriate scanner which could email me PDFs of scanned documents (see section on *Hardware* above). One problem which was found is that the default size of mail inbox was far too small to accept these files. Cramming a too-large file into the inbox completely locked up my mail system, and the file had to be deleted or moved by Support. At that point it could be copied by FTP onto the host computer. To avoid this problem in future, the mailbox size was increased to about 80Mb, which is large enough to hold just about any set of scanned files until they are retrieved by the host machine and subsequently deleted from the mailserver. Another problem was the absence of any documentation for the scanner, making its use one of trial and error. There are still facilities which are not available as they are locked by a password.

In theory, the scanned PDF files could be put through Adobe's word-recognition software and converted to ASCII or Word-compatible format. The results of this were amusing, but of no use whatsoever. Words were misread or omitted, and formatted text (e.g. in columns, or running around a picture) caused nothing but more confusion. I doubt that more than 20% of the words were recognised at all, especially in older books.

Consequently scanned PDFs should be regarded as unchangeable except by image manipulation software.

### *Foxing and Image Manipulation*

Foxing is a problem with older books, or more recent pamphlets which have been stapled together. It consists of brownish staining on the pages of the text: in both cases the brownish stains appear to be due to chemical reactions of iron. As well as this, books which are several centuries old are often printed in brownish inks or inks which have turned this colour with time, and here the problem of removing foxing is difficult as the colour of the ink is close to the colour of the foxing which is to be removed.

In the case of the current archive, the oldest books were from the mid 19th century and all the text was black so the problem of differentiating ink from foxing did not arise, but this may be a problem if we mine the St Andrews collection from earlier centuries. A very slow and tedious method to do this removal stain by stain is possible, by sampling the background colour at that point and then covering it with that chosen colour, but it is likely that the time and effort would only be justified if creating an electronic version of a text was part of a large project to disclose manuscripts or incunabula from the University archives.

Two possibilities remain for dealing with foxing on the documents which I used:

**1. scan in colour**. Here we are relying on the ability of the human eye to distinguish foxing from text on the basis of both colour and context (stains are random, printed text is not). Scans in black and white are completely unsatisfactory as they usually darken the foxing to the point that it is the same shade of black (if black *has* shades) as the text. Consequently the text is completely obscured by the now-black foxing. Scanning in b/w does have the advantage of producing files which are approximately 4x – 6x smaller than their colour equivalents, which could be an important issue if file downloads are over a slow or unreliable connection.

**2. remove the foxing by some form of thresholding** based on colour or brightness. Once this has been done, then the de-foxed page can be saved as a b/w document. Using Adobe's Photoshop or The Gimp it is possible to do this not only for one page, but to create a macro which enables straightforward processing of the whole document, assuming that the staining is of the same colour and intensity throughout. This was tried with Photoshop for the simple reason that there were manuals and help files. At the time of these tests, Gimp v2 had only just been released and although macros were possible there was no information on how to access them – a common problem with GNU software is lack of documentation or documentation that is virtually incomprehensible unless you understand the program in the first place. Full details of the Photoshop procedure suggested by TASI is given in Appendix IV, but my own tests have shown that other methods of thresholding using the Gimp can also be used. An example follows:

- load image into Gimp;
- right click on image;
- choose *layer*;
- choose *colour*;
- increase *contrast*;
- from *colour* choose *levels*;
- pick a point which is heavily foxed;
- This will set the white level to be the same as the foxed point; i.e. the foxing and all colours lighter than this will be assumed to be white, thus removing foxing whilst still leaving the text readable. It can then be saved as a b/w image.

It would be quite easy to build this into a macro and process a whole document – if only there were information on how to create macros in Gimp. At the time of writing (October 04) this documentation is starting to appear.

No doubt there are other, similar methods which are just as effective; the point is that this sort of staining *is* removable without a great deal of effort, making b/w scans of older texts a realistic proposition.

### *Use of Colour and Pictures*

This has been mentioned in passing in the above section, but will be repeated here. The problem is that coloured file images take up about 4x – 6x as much space as their b/w equivalents. I do not think that this should be seen as a problem in terms of disk storage; under £75 can buy up to 160Gb storage (Maplin catalogue, September 04) so adding extra storage is trivial in terms of money, and also in terms of installation in a linux system. The real issue is of download times.

We should not assume that the whole world is connected through a very fast gigabyte network such as JANET. On the contrary, many Universities and most Colleges in the world struggle with slow and unreliable connections. In this case the difference between downloading a b/w and coloured version of a file is the difference between downloading and not downloading. Hence if we are to use eprints as a St Andrews portal, showing the work and achievements of staff and students, then we should be aware of this issue.

Of course, there are files which become virtually meaningless if they are not in colour or if coloured pictures are transformed to greyscale and for these other strategies could be used. For example:

- we could keep b/w and coloured versions; if the b/w downloaded version looks very useful the researcher could ask for a reprint;
- the coloured sections, such as pictures, could be separated from the b/w text and could be downloaded each as a separate file.

Ultimately, this is a management decision which depends on deciding on how broadly we want to disseminate information.

### *Metadata Page*

Before an eprint can be deposited, a *metadata page* needs to be completed. This page is always open to view and can be harvested no matter what permissions are set for the eprint itself (see later section on Security for information about this). It is a template which can be customised to suit the needs of an institution and which is also inherently dynamic. This means that the information requested varies according to the type of eprint being deposited. Thus a book will call for an ISBN, a paper will call for the name of the journal, a thesis will ask for the type of thesis (Mphil, PhD etc.), and so on. Furthermore, certain field can be made compulsory, such as author, title, abstract, keywords etc., other fields, such as the name of the University, can be pre-filled.

The purpose of this is to produce *structured* data, in this case a basic form of DC (Dublin Core). And here we see a huge advantage over the *ad hoc* method of letting staff put items on their own web pages. It is extremely unlikely that they would add this form of metadata to their documents, nor is it likely that it would just happen to have the structure of DC. But this structure is needed for consistency in searching and for interoperability with OAI (open archives initiative) sites, which harvest the metadata rather than the paper itself.

### *Keyword Extraction for Metadata Page*

One of the tasks for assisted depositing is to create a keyword list if the original document does not have one.

Clearly, if the document is current, then the author could be asked to supply one, or more realistically, would be asked to approve one which the eprints administrator has created from the original text – especially the abstract.

In older texts this is not possible, so a short experiment was done on a medium-sized monograph which was scanned in (a rectorial address) to see whether a simple way of extracting keywords had any chance of success using an image as source.

Full details of the algorithm are given in Appendix V, but briefly, it consists of allowing Acrobat to try to recognise as many words as it could and then save a file of these words.

This file would then be processed by reducing all instances of words to singletons, removing any stopwords (e.g. function words) and also removing any words not in a dictionary (hence removing badly-recognised 'words'). The dictionary chosen was the Oxford psycholinguistic dictionary with everything except the head words removed. This was created in about 1980 and has not been updated since, but this is hardly a problem with older texts. The stopwords were based on van Rijsbergen (*Information Retrieval: Butterworth, 1979*) and are adequate for a simple trial such as this one, though they are largely based on business correspondence.

After processing, what is left should be a list of real words, with no duplications, and with stopwords removed. These could be put into the 'keywords' section of the metadata page, or, more realistically due to the length of the list, be put as an invisible, but searchable, layer of the PDF file.

Unsurprisingly, the results were far from useful. As the dictionary was general-purpose and, in contrast, the words which described the contents of the document accurately were specialised, then the keywords garnered by this method were too general to be of use. However, if there is a corpus of texts dealing with a specific subject, e.g. a series of papers from the 19th century about a single topic, then a specialised dictionary could be constructed and used in combination with the more general-purpose one. This approach might well lead to a useful set of keywords which could be incorporated in a hidden layer, but is likely to be intolerably long for metadata purposes.

This remains a topic which could be investigated.

### *Converting Database Entries to eprints Format*

One other problem I came across was the possibility of re-using some database information of researchers' publications which had already been gathered by Research and Enterprise Services (RES). This information was very similar to the contents of the metadata page, but lacked the abstract, keywords and the paper itself.

Although this was not really satisfactory, I thought that if I was to present the core of the information on eprints, I could either:

- inform the member of staff and ask them to add any missing information and approve the publication of the information; or
- ask them for the URL of the paper and use that; or
- ask them to email us a PDF of the paper and I would do the rest.

Although the concept was potentially very good as there were over 1500 entries in the database, in reality the consistency/quality of the data was rather poor, with many missing fields where, presumably, it had been impossible to elicit or discover

the information. It would have taken considerable effort to clean up the data, and even then there may be very few actual papers suitable to put on eprints, that I decided not to pursue this any further. There were, after all, a large number of resources readily available in the library itself which could be used without any of these problems.

On the other hand, I did think that it was worth looking at how we could transfer the database information into eprints if we needed to, as it is quite possible that other, more appropriate, databases may be found in the future, so I did a preliminary investigation into this:

- the database entries were held on a Microsoft spreadsheet (Excel), and we needed to transfer them to XML format as (a) XML is the native format of eprints; (b) Excel is a proprietary format and Microsoft have consistently refused to reveal what it is – the best anyone has done is to reverse-engineer a format which appears to work identically as Microsoft's own;
- the XML list of records had to be pruned (automatically, if possible) to leave ones which fulfil the minimum criteria for eprints metadata, whilst at the same time skipping over irrelevant fields. This is much easier in XML either by writing Perl scripts, using the CPAN modules or pre-existing tools for doing just this, such as the XML toolset from Edinburgh University;
- eprints had to be modified to accept metadata only (no files);
- eprints had to be modified to allow batch loading of metadata.

Apart from the last item on the list, all the other points were resolved as follows:

- *Conversion to XML*: there are two issues here – (a) conversion from Excel, (b) conversion from a csv (comma-separated values) list. As I wanted the solution to be generic if at all possible, I opted for conversion from a csv list. This is quite easy to do as there is a CPAN Perl module *XML::SAXdriver::CSV* which does exactly this job (see **http://www.cpan.org**). There are some help files with this module, but also fuller documentation and sample scripts in the book: *XML and Perl: M Riehl & I Sterin.* If for some reason it was felt necessary to convert Excel files directly, there is also a module *XML::SAXdriver::Excel* which you could use. However, as Excel format may change without notice, this may be a less reliable way of tackling the problem, especially as all spreadsheets and databases can save in csv. The current version of Excel is supposed to save data in XML, but I did not have access to this software and also MicroSoft's track record in following any standards at all is poor.
- Getting the csv-to-XML module to run is not much more complicated than typing in the sample script given in the book and modifying the Writer object to write the output file in the format you want. It works.
- *Pruning* the output list is equally free of problems. A simple Perl script – effectively a set of rewrite rules – would process the XML file a line at a time. As each field now has a start tag and end tag (e.g. <first_name>xxx</first_name> or <published>Y</published> and each field starts on a new line, all that is needed is to test for the presence or

absence of alphanumerics between the start and end tags. So …>Y</… would pass and …></…. would not. Of course the system could be made more sophisticated by testing for the presence of acceptable values on a field by field basis; the principle remains the same.

- *To accept metadata only* eprints has to be modified to accept no papers is explained in the eprints documentation. What needs to be done is to have no 'required formats' in the file ArchiveConfigure.pm i.e. you set
  $c->{required_formats} =
  [
  ];
  After the batch loading, it would be necessary to reset this value back to the original list.

*Batch loading* of the files was not investigated as correspondence on the technical email list for eprints seemed to suggest that although it was possible to do this fairly readily, it was also possible to create major database access problems during the process. As we had a working eprints system with quite a number of large documents deposited, I decided not to experiment any further with this, especially as there was no real data to go into eprints at the end of the exercise.

It is curious that no-one at LIS knew of this database which was duplicating parts of the eprint project, though not giving access to papers and having inconsistent data which was not DC. This form of wasted effort should be avoided by the University giving both publicity (to and consistent, long-term funding

# Security

It is necessary to consider security issues when discussing an institutional archive which is open to the public such as eprints. Problems may arise in two ways:

## 1. Security and access to the University system as a whole

These are issues which are external to eprints and are dealt with by LIS as part of university-wide policy for the implementation of computer systems. Unauthorised access or hacking is the same whether it is done to the eprints server or to any other server on the University network. Consequently, these issues will not be considered any further.

## 2. Internal security of the eprints system

This, in turn, can be broken down to:

- Security controlled by configuration files;
- Registration;
- Permissions for depositing eprint documents;
- Assigning permissions to deposited eprint documents and so controlling access;
- Access to metadata;
- Updating or changing eprint documents or metadata;
- Removal of eprints;
- De-registration.

As before, I will deal with these in order:

### *Security controlled by configuration files*

There are a number of internal features in eprints which can be used for security purposes. These are controlled by modification of configuration scripts such as ArchiveConfig.pm or even more usefully, metadata_types.xml. These scripts can govern whether certain fields are ‘required’ i.e. values for them are obligatory, or whether they can be left blank and also whether they appear in all deposit forms or only those seen by an editor. In this way the archive can be set up to allow only certain classes of people to view selected documents and portions of forms may be revealed or concealed depending on the status of the user.

There is a very large number of these options in eprints, and they are detailed in Chapter 6 of the documentation, Configuring an Archive. This chapter occupies about 30 pages in version 2.2.x, and you are referred to this for further information. The two configuration scripts given above should be given special attention.

### *Registration*

The way eprints software operates is that those who are registered have certain privileges, which will be described later. The software as standard has no control over who can register, nor is the administrator informed of new registrations. This

means that there is no real difference between a casual and a registered user as the former can become the latter with no impediment.

As this is an institutional archive it is necessary to exclude non-institutional registrations, and this could be done by using an *.htaccess* file in the root directory of where the eprints documents are stored.

### *Future enhancements*

#### FURTHER LIMITATIONS ON REGISTRATION

The above is fine for a testbed archive, but is not likely to be satisfactory for a University-wide implementation, as it allows anyone to register from a St Andrews domain. A moment's thought would show that allowing students to register and put material on eprints is likely to cause problems, not only because each document has to be approved by the eprints administrator (a registered user's document is held in a submission buffer until given approval) but there is also the issue of maintenance of a potential total of about 6000 students, of which 1500 may leave and arrive each year. Eprints has no facilities for maintenance of a list such a this.

Some text removed from this page

In summary, I would suggest that:

- Undergraduates should not be allowed to register on the eprints system (they can see eprints but they cannot deposit them);
- Academic staff should be allowed to register;
- Postgraduates should be allowed to register, but their submissions should have the approval of their supervisor;
- Non-academic staff should be discouraged from registering, but unless the LDAP server can distinguish between non-academic and academic it is not possible to do this automatically. (The value of *accountstatus* on the server is currently under revision).

ALERT SERVICE

A useful facility would be to have an alert service which would automate messages to new members of staff or new postgraduates to tell them that they could register for eprints. At the same time new undergraduates could be informed of the papers which are available on the system – probably *via* SIPs

### *Permissions for Depositing eprint Documents*

Anyone who is a registered user can deposit documents in the archive. This is the reason for the complexity of the previous section where control of registration was discussed. If anyone can register (the default) then anyone can deposit. The sense of this is that eprints started out as a subject-specific archive, not an institutional one; hence depositing had to be made straightforward for a wide range of people from all over the world.

The path I have chosen is to control registration, thereby controlling deposits.

Note that deposits do *not* go straight into the archive – they go into a *submission buffer* where they need to be approved by the eprints administrator.

### *Assigning permissions to deposited eprint documents and so controlling access*

This is at the discretion of the user. Three options are available:

- Anyone can view them (the default);
- Only those who are registered can view them. This in effect restricts them to this university – note that even if you are registered you cannot view them unless you do so from a St Andrews domain. The metadata is always visible;
- Only administrative staff can view them; this amounts to about four people and effectively makes the eprints *document* private (the metadata is always visible). This may be useful if someone wants to embargo the publication of a document for a given time. At present there is no automatic way of removing the embargo, it has to be done manually by the administrator, who has to remember when to do it, but this may well change in future versions of eprints, as it has been discussed as an enhancement on the eprints technical list.

### *Access to metadata*

Note that the above section applies to *documents* only. The metadata is always freely available to everyone and will be harvested by search engines and by OAI-compliant (open archives initiative) sites such as *OAIster* as eprints always produces OAI-compliant metadata, ensuring that the information is harvested correctly.

Consequently, no matter what permissions are set, the author, title, abstract, keywords etc. will always be open.

### *Updating or changing eprint documents or metadata*

The only way to change any of the details of the deposited eprint is to ask the eprints administrator to do this (by cloning the eprint and resaving it with changed data, then deleting the original one). This applies whether the change is to metadata, the permissions or to the document itself.

This type of change is *strongly* discouraged as the eprint may have already been used (cited). About the only reasonable reasons would be to fix metadata mistakes or to change permissions.

If the document and its metadata are being *updated* then eprints has a mechanism to do just this. The new document is deposited and a reference is placed to the old copy/copies (you have to know the ID of the document to do this. It is given on the metadata page and there is no quick way to access it from the update page). If anyone looks at an old version this is automatically flagged and they are told that there is a newer version available. This bypasses the most common irritation of using web-based citation – that the document cited has disappeared or has been changed, making a nonsense of the citation.

### *Removal of eprints*

Is very, very strongly discouraged. Plagiarism, major hassle from publishers or similar catastrophes would be OK, otherwise the eprint should be left on the server as it may already have been cited. The only person who can remove eprints is an administrator.

### *De-registration*

This can only be done by the administrator. As this is an institutional archive, de-registration would naturally happen if the person leaves the University – ways of doing this semi-automatically have been discussed under the previous section *Registration – Future Enhancements*. Other reasons for de-registration might be abuse of the system, plagiarism and similar problems. Note that de-registering does *not* remove eprints deposited by that user, it merely stops them from depositing any more documents. If documents need to be removed then each has to be removed as a separate task.

# Beyond the Project

## Current Trends; or Why we might miss the boat

The following really speak for themselves – all are quite recent initiatives:

### *German government funds OA initiative*

FIZ Karlsruhe and Max Planck Society get £4.2m to develop a collaborative scientific research and funding platform. The German government has awarded Euro 6.1m (£4.2m) to STM publisher FIZ Karlsruhe and the Max Planck Society (MPS) to develop a platform for web-based collaborative scientific work and self-publishing. [...] (In: *Information World Review (1.x.04)* ): (**http://www.iwr.co.uk/IWR/1158510)**

### *Journals & Institutions*

This is a very small sample of the recent trends in adopting Open Access and self-deposit policies:

- *Oxford University* is undertaking a 3-year eprints project and OUP is supporting this initiative (see **http://www3.oup.co.uk/jnls/librarians/OUP%20SHERPA%20PR%20Oct%202003.pdf )**
- Cambridge University is adopting the self-deposit software Dspace in the Dspace@Cambridge Project (**http://www.lib.cam.ac.uk/dspace/**)
- The *Company of Biologist*s have an OA (open access) option on the Journal of Cell Science;
- *Elsevier* allow preprint archiving, and in fact have their own preprint server;
- *Springer* also allow preprint archiving;
- *Royal Society of Chemistry* (RSC) -- **http://www.rsc.org/** -- publisher of 28 journals, announced in August 2004 that as of 6 weeks ago the RSC is happy with author self-archiving for articles they publish in RSC journals. (**http://oasys2.confex.com/acs/228nm/techprogram/S14185.HTM)**
- *Scottish Science Information Strategy Working Group*'s [Open Access] Declaration (Draft) (see Appendix VII for full text)
- The *Canadian Association of Research Libraries (CARL*) is very actively engaged in setting up institutional repositories at all of Canada's research universities. For more information: **http://www.carl-abrc.ca/frames_index.htm**

Note: the policy of journals is constantly changing and in all cases I have heard of this change has been to the advantage of depositing pre- and post-prints. Common sense would suggest checking the policies of the journals/publishers listed here. Try their website, or, better still, the Romeo website (**http://www.lboro.ac.uk/departments/ls/disresearch/romeo/** or **http://www.sherpa.ac.uk/romeo.php** ) A list of the current journals (Sept. 2004) willing to support self-archiving are listed in Appendix I.

The **majority of journals now support some type of self-deposit**, whether this is preprint only or preprint and postprint.

### *Citations and Open Access*

- Perneger's (2004) findings show that download counts (sometimes called "usage impact") of British Medical Journal articles predict citation counts ("citation impact") for those articles in subsequent years (Perneger, T.V. (2004) *Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ.* BMJ 2004;329:546-547 (4 September), doi:10.1136/bmj.329.7465.546 (**http://bmj.bmjjournals.com/cgi/content/full/329/7465/546)**
- Work by Brody & Harnad (2004, in prep) demonstrate the same correlation. Brody, T. & Harnad, S. (2004, in prep.) Using Web Statistics as a predictor of Citation Impact. (**http://www.ecs.soton.ac.uk/~harnad/Temp/timcorr.doc)**
- Brody's online usage/citation correlator (**http://citebase.eprints.org/analysis/correlation.php)** has also been demonstrating this for a number of years in the fields of physics and mathematics;
- The OA impact enhancement effect already reported in physics, mathematics and computer science is also present in biology. (from: Chawki Hajjem, Informatique Cognitive, Université du Québec : October 04 (**http://www.crsc.uqam.ca/lab/chawki/OA_NOA_biologie.gif)**
- Publishing both electronic and print versions of journals (overlay journals), where the electronic ones are open access has been implemented by the University of Warwick with its *Algebraic & Geometric Topology* (**http://www.maths.warwick.ac.uk/agt)**
- Finally, Appendix II list a host of citations which back up this position.

### *Open Access and Citation Impact in non-Scientific disciplines*

Below is the latest evidence that the Open Access Impact Advantage is neither unique to Physical Sciences nor Mathematics, which is the usual claim:

**http://citebase.eprints.org/isi_study/**

nor to the Biological Sciences:

**http://www.crsc.uqam.ca/lab/chawki/OA_NOA_biologie.gif** (see *Figure 4* below)

The Impact advantage is there in the Social Sciences too:

**http://www.crsc.uqam.ca/lab/chawki/sociologie.htm**

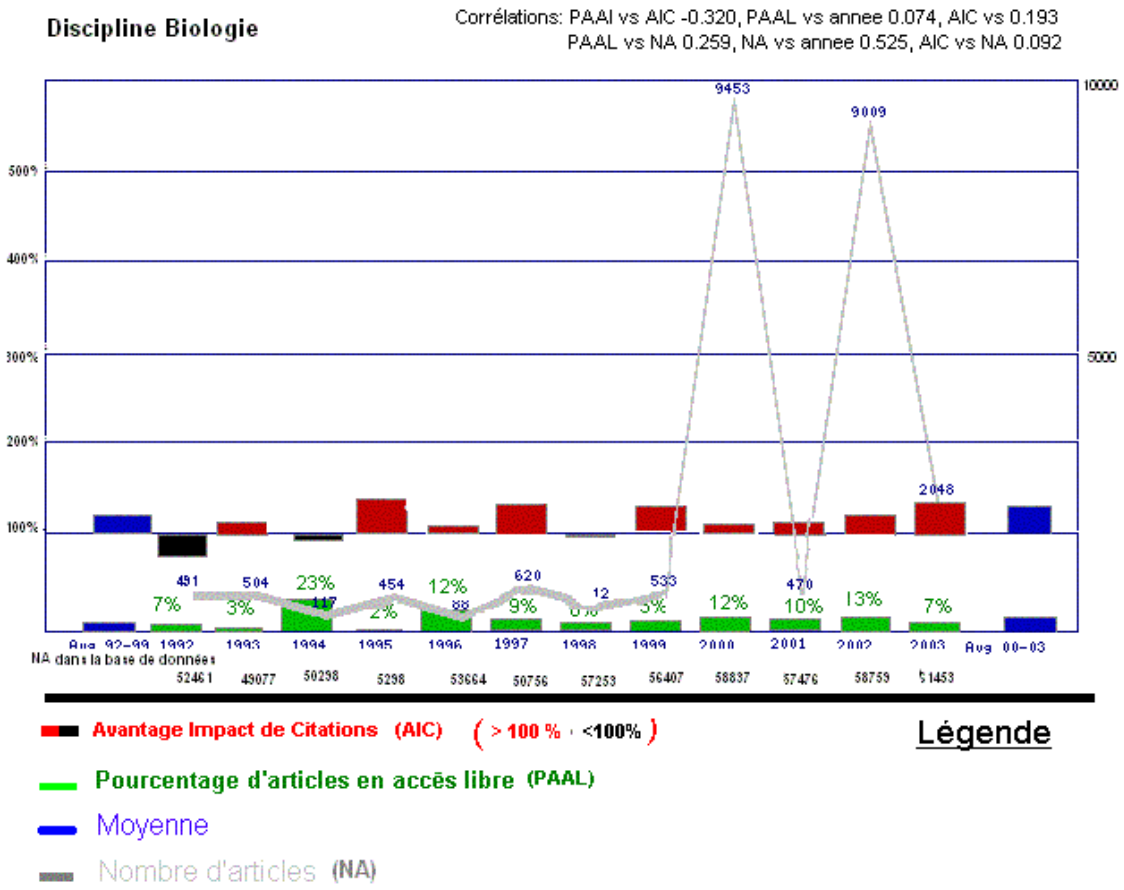



*Figure 4*

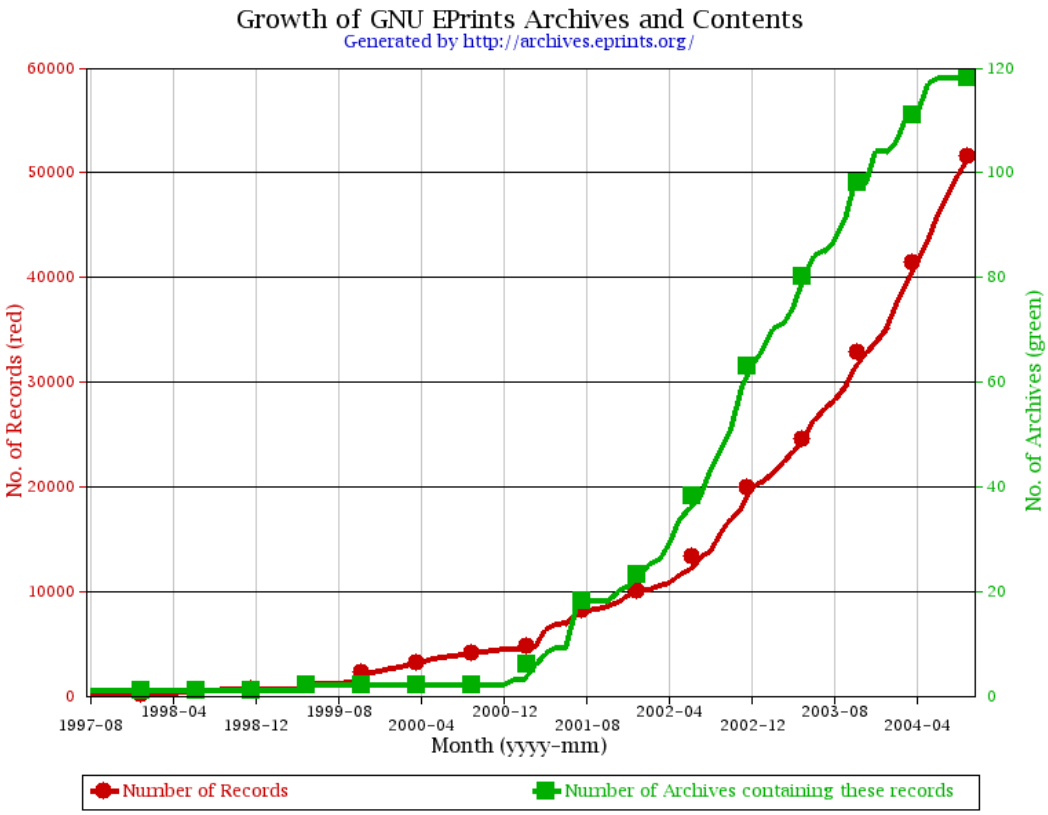
**Open access increases both citations and access to research materials**

*JISC report: central vs. distributed Archives*

"…recent study, carried out in partnership with the Universities of Loughborough and Cranfield on behalf of JISC, produced a recommended model for the delivery, management and access of eprints (both pre- and post-prints)in UK further and higher education communities. We deliberated on the relative merits of central versus institutional archiving and came down firmly on the side of the latter. The reasons for this were several - both technical and cultural - and are set out in detail in our full report, which will be published by JISC within the few days" [Swan,A., Needham, P., Probets, S., Muir, A., O'Brien, A., Oppenheim, C., Hardy, R., and Rowland, F (2004) Delivery, Management and Access Model for E-prints and Open Access Journals within Further and Higher Education]. (**www.keyperspectives.co.uk/OpenAccessArchive/E-prints_delivery_model.pdf**) (from :*ASOA Forum email 3.x.04)*

It would be hard to find a better model for distributed archives than eprints!

*Eprints Archive Growth*




*Data from http://celestial.eprints.org/cgi-bin/eprints.org/graph 11.x.04*

*Action Suggested by Scottish Science Information Strategy Working Group* (see Appendix VII for full report)

I have added my own comments to each point in a typeface like this:

- Set up institutional repositories, and/or liaise with other organisations to establish a joint repository.
  We are already liasing with CDLR at Strathclyde University and continuation of eprints at St Andrews would only strengthen these links. Also it is likely the Mr Bagnall at Dundee University Library would be interested in collaboration as he, along with some colleagues, attended eprint seminars and also came to see the eprints project here;
- Encourage, and as soon as practical mandate, researchers to deposit copies of their outputs (articles, reports, conference papers, etc) in an institutional or co-operative repository;
- Encourage, and as soon as practical mandate, the deposit of PhD theses in an institutional repository;
  A pilot scheme at the School of Art History showed that students were prepared to go to some lengths to deposit their dissertations with eprints. I would suggest that not only PhD theses, but MSc/MPhil and similar theses should be deposited along with Honours dissertations from students who achieved Firsts;
- Review intellectual property policies, to ensure that researchers have the right and duty to provide an open access version of their research.

This is a matter of University policy which would probably flow from the success of depositing theses. After all, the University, as employer, *does* have copyright over all work produced by academics, but it chooses not to exercise this right. The bulk of the research work financed by the US government is by law in the public domain. This has never caused any problems either to the researchers or to the publishers.

### *UNESCO Electronic Theses Project*

This project has been underway for a considerable number of years and intends to disseminate, through Open Access, thesis work completed by students.

To quote:

> Access to information is crucial for education, sciences, and culture and for fostering democracy in the information society. The principle of free and universal access to information as well as freedom in the creation, treatment and dissemination of knowledge, is a fundamental element of the global common good of humanity.
>
> UNESCO is mandated by its Constitutions to ensure "free exchange of ideas and knowledge". One of the mains goals of the Organization consists, therefore, in redefining universal access to information and the minimum level of service to be provided to information users by the public sector. The fair allocation of public resources to public information providers must also be promoted.
>
> New information and communication technologies have the potential to better ensure free and universal access to information and to reduce inequalities in favour of social justice and economic well-being.
>
> This is particularly important to the scientific community which is concerned about topics such as the free flow of scientific information and its universal access; public versus private knowledge; intellectual property and copyright issues; changing practices in scientific communication, etc.
>
> In industrialized countries, the growth of science in terms of the number and complexity of research programmes makes it difficult for researchers to keep abreast of current developments in their subject. In the developing countries the situation is even worse because of poor communications, the comparatively smaller research community and the lack of resources.
>
> New technologies must be used by universities and research institutes in developing countries to keep up with developments and discoveries elsewhere in the world. However, the flow of scientific information between developing countries and from the South to the North is often hampered by poor communication and publication channels and by limited knowledge of application of new technologies in scientific information exchange.
>
> The scientific community, therefore, expects that the principle of free flow of scientific information be respected and that initiatives be taken to facilitate access to scientific information sources by scientists from developing countries. Countries with high scientific expertise must share and transfer knowledge through the support of specific programmes set up for scientists and the users of scientific information worldwide.

(From: **http://portal.unesco.org/ci/en/ev.php-URL_ID=3515&URL_DO=DO_TOPIC&URL_SECTION=201.html)**

### *eprints, Download Statistics and the Humanities*

An email received on 21st October reads:

> … In disciplines where citation rates tend to be quite low (e.g. humanities) , it can be very difficult for researchers to know if one of

their research strands is generating more interest than the others - and thus deserves more attention. Differential download statistics offer some insight.
Just yesterday, an academic called to thank me - he was delighted that his self-archived papers had been downloaded over 1000 times in the past 12 months. He went on to say that the download statistics had made him realise that he should be focusing his research efforts in one particular direction. The download figures for these papers were ten times greater than for his other two research interests. Previously, he had no way of knowing how many times his articles were being read. The data from ISI citation counts for his publications was so sparse that it gave no useful indications. He is now determined to self-archive all his future publications and has been encouraging his peers to do likewise. (see: **http://eprints.qut.edu.au/**)

### *Current Trends - Conclusion*

It seems that *not* to follow up this pilot project with a full-scale implementation would be to put St Andrews at a substantial disadvantage *vis-à-vis* other universities and institutions of higher learning. This implementation necessarily demands that moves to electronic publishing should be encouraged or mandated. A good start would be to make it University policy that electronic versions of theses (at all levels) must be supplied and deposited with eprints in addition to the usual paper copies.

This approach has a number of advantages:

- The copyright situation is clear from the beginning;
- Students' work is invariably in electronic format;
- Students will be familiar with computers in a way that some members of staff may not be – or they could find a 'computer guru' in the class who could help;
- They could refer potential employers to view their work via the URL;
- Results from Art History dissertations show that students are often prepared to go to considerable efforts to get their work displayed on eprints;
- Helping students to prepare and submit their work for eprints would give lecturers a kick-start to try the same with their own work;
- Ultimately it has to be recognized that full-scale implementation is very unlikely to succeed unless the use of eprints is mandated, and that this mandate has to be introduced in an incremental manner.

## Other uses for eprints

Even though eprints software was designed primarily for use in subject-specific repositories, it has matured with time into being more general purpose. Because of this we have been able to use it to create an institutional, rather than subject-specific, repository without having to re-write large sections of the software. Of course 're-branding' would be needed no matter what use it is going to be put to. A subject-specific repository would need to have pages customised to suit its particular needs and to indicate clearly which university or institution is supporting it. The amount of work which this entails merely reflects the level of customisation required.

The software can also be used for other purposes, as repositories are not necessarily archives, nor do we need to be wedded to the concept of one general-purpose archive – in fact this may be a bad idea as it would limit the amount of

customisation possible. Rather we should be thinking of a family of archives (maybe an *album* is a better collective name) which can contain other items such as:

- *learning objects* – enabling the software to be a simple and straightforward way of disseminating papers, course work, timetables, etc. to students;
- *data* – experimental data is not usually published in papers as this would make them intolerably long. However, this could easily be published on an eprints site, ensuring that the data does not disappear if the researcher moves to other employment and their personal institutional website is closed;
- items which need to be available under the Freedom of Information Act. The university is currently running a project to comply with these regulations, but eprints, with their division of documents into metadata and the document itself, together with an electronic trail of revisions could also have served the purpose;
- *images*. This would probably need substantial customisation, but little re-writing of the software. All that is necessary is to hide or remove the metadata fields which are inappropriate and generate thumbnails in place of PDF icons (the icons are not available in our version of software (2.2) but are in the current upgrade. These images could be ones which are collected for teaching purposes and could act as a teaching resource (permissions could be set to make them available to St Andrews only). The issue of thumbnails is important as it would prevent downloading a potentially huge image file which turns out to be unwanted. Again not a problem with the University (except for the issue of disk space) but could create difficulties if the eprint (epicture?) was widely available;
- *publishing*. Eprints has been tried in the field of Open Access publishing and this was an issue discussed with Prof. Chris Smith where preliminary discussions were about the revival of the St Andrews University Press, but in an electronic rather than a paper-based format.

Staffing & other costs

## Start-up Costs

These have already been covered in the initial grant given to St Andrews by the HaIRST project. They are included here for completeness but were not part of St Andrews expenditure:

- Cost of computer equipment for eprints Administrator;
- Part cost of server (this equipment is shared);
- Costs associated with initial setting up and continuing maintaining the software by IT services;
- Maintenance and customisation of eprints by the Administrator;
- Initial promotion of the system.

For this project, lasting 18 months, *total* computing costs (administrator's computing equipment, shared server and IT support were £2500; staffing £16000; sundries £200.

## Staffing

### *Role of eprints Administrator*

This should have three foci:

- assisted deposit, as outlined above;
- software maintenance, unless this is to be devolved to IT in its entirety, which is not recommended;
- promotion of eprints to staff and students.

I also suggest that an important skill of the administrator would be familiarity with areas such as indexing, cataloguing and the concept and importance of structured data (such as DC). This would be needed in creating or assessing the metadata page, which is a requirement of each deposit.

#### ***Assisted Deposit***

An option offered on eprints is that of *assisted deposit*, whereby the eprints administrator will deposit a document on behalf of another member of staff, usually the author. The costs involved are:

- determining rights;
- format conversion (only certain formats are accepted by eprints. This is a deliberate policy to aid the creation of archival material rather than material which becomes unreadable with the demise of specialised software);
- creation of the eprints metadata page.

Note that these costs are incurred by anyone depositing material into eprints, but in the case of assisted deposit they all fall on the Administrator, hence need to be reckoned as a cost. Item (3) in particular is going to be higher for assisted deposit as the administrator is probably not familiar with the material and will need to rely on the supplier for advice or for modification of the draft metadata page if the information provided is incomplete.

The information required to make a speedy assisted deposit is on the eprints web page *Let us Archive it for you* and the relevant part is reproduced here:

- Make sure you have gathered all the files you require, including any picture files. Click here to see a range of suggested formats suitable for longer-term preservation;
- if you know how, please compress them all into one file (tar or zip). eprints will automatically unpack them when necessary. If you are not sure how to do this, email all the files as attachments and we will do the rest;
- email me the files with a brief covering note. This should state whether the item is in print (in which case we need to know the journal/publisher), whether it is a preprint, or whether it is an internal document such as a technical report, thesis, etc.;
- if you would like us to archive it for you in other, international, archives, please say which ones.

If you want us to choose the archives, just ask us to put it in any suitable archive.

- We can usually get the other information we need (such as the authors, title, abstract) from the document itself. If you do not have a list of keywords, we will add one. We aim to please.

I suggest that a deposit should take no more than about 15 mins for an unfamiliar subject; 5-10 mins. for a familiar subject. The more helpful the author is the speedier the deposit.

It is unreasonable to assume that there would be a sudden rush to eprints; this is a project which will grow slowly so that deposits should not take a person more than about one day a week (I am assuming that undergraduates are barred from depositing material on eprints). In the case of undergraduate theses, a substantial portion of recent graduates are likely to be able to prepare the material for rapid proxy deposit as basic IT skills are now no longer optional. Also the highest volume of work would occur during the summer when other pressures are likely to be lower.

***Software Maintenance***

Costs of continued staffing will also depend on the balance of skills which would be shared between the Administrator and computer support from IT services (now amalgamated into LIS). The more the Administrator can do in terms of day-to-day maintenance of eprints software, the less the extra expenses involved in asking for external help, though it seems bizarre that payment rather than co-operation is at the core of essential services provided by different sections of the University.

The fewer the computing skills of the Administrator, the more they will need external help, though to counterbalance this the wages paid can be lower. This balance is an administrative decision which will have to be weighed up.

As mentioned earlier, Windows skills are not a requirement (though basic familiarity to use and find software on the Web and use standard Windows software would be a basic requirement to approve deposits and deal with similar matters.

More importantly, they need to be tolerably familiar with linux unless *all* software change in eprints is to be left to IT as the page layouts and software reside on a linux machine. At least a nodding familiarity with Perl would also be useful as eprints is written in this. A scenario such as this would mean that the Administrator would deal with small bugs and be able to repair the software, whilst major changes would be left up to IT.

### *Promotion*

The last duty of the Administrator would be to promote eprints by means of talks and visits to specific people or departments and to more general talks (e.g. on copyright or for Arts and Humanities). This is unlikely to take up more than about 2-4 working weeks of the year, depending on how heavily the University backs the project – if it is promoted at a high level then more visits and meetings would be required, but this would also mean that the project would be very successful and free many lecturers from maintaining links to their own papers. It would also mean that the University would have a record of work accomplished even when the lecturer leaves and takes their own website down – the usual scenario, which leaves no trace of previous work done here.

### *Summary*

Taking all these into account, I currently envisage a total *average* of approximately two days' work per week to fulfil this role, and this is the costing which should be budgeted for initially. Their should be some level of flexibility in this, so that the Administrator can attend extra days some weeks if the timetable demands this e.g. for promotion purposes.

Assuming reasonable computing familiarity by the Administrator a total of about one week's IT support per year would probably be sufficient to keep eprints up-to-date.

Basic knowledge of indexing or cataloguing and also of metadata would be highly desirable for creating or assisting metadata page creation.

## Computing Equipment

This has already been covered in a previous section *Requirements for a University-Wide Service* under the section *Hardware q.v.*

# Acknowledgements

This pilot project would not have been possible without the help of a number of people:

Janet Aucock, the project manager here at St Andrews who was both very helpful and very enthusiastic about the project. It would have been a non-starter had it not been for her.

Duncan Brannen and other members of the IT staff who dealt with problems quickly and were instrumental in getting the software up and running, and then in keeping it running and up-to-date.

Fabio Simeoni, the Senior Research Fellow of the HaIRST project at Strathclyde. We exchanged numerous emails, telephone calls and personal visits in an effort to help me understand eprints. And he showed us a good pub in Glasgow where they do delicious hamburgers. This alone deserves praise.

Numerous members of the eprints technical list whose advice and help were always so useful not only in helping to implement a particular feature, but also in knowing when the effort was probably not worth it.

And, of course, the staff and students who helped by putting material on the eprints server, or gave me access to material which I could use.

# Appendices

# Appendix I :: Publishers willing to have pre/postprints on an eprint server (1.ix.2004)

GRAY = No green light yet from publisher

PALE-GREEN = Publisher's green light to self-archive pre-refeeeing preprint

GREEN = Publisher's green light to self-archive refereed postprint

The number in square brackets is the id number assigned to this publisher by SHERPA

[2] Academy of Management (4 journals) (GRAY)

[3] American Association for the Advancement of Science (1 journal) (GREEN)

[4] American Chemical Society (35 journals) (GRAY)

[5] American Economics Association (3 journals) (GRAY)

[6] American Geophysical Union (19 journals) (GREEN)

[7] American Institute of Physics (11 journals) (GREEN)

[8] American Medical Association (1 journal) (GRAY)

[9] American Meteorological Society (11 journals) (GREEN)

[10] American Physical Society (8 journals) (GREEN)

[11] American Physiological Society (16 journals) (GREEN)

[12] American Psychological Association (48 journals) (GREEN)

[13] American Public Health Association (1 journal) (GRAY)

[14] American Society of Civil Engineers (30 journals) (GRAY)

[15] American Society for Clinical Investigation (1 journal) (GRAY)

[16] American Society for Microbiology (11 journals) (GREEN)

[17] American Society for Pharmacology and Experimental Therapeutics (6 journals) (GRAY)

[18] American Sociological Association (10 journals) (GRAY)

[19] Arnold Publishers (35 journals) (GREEN)

[20] Association of Applied Biologists (1 journal) (GREEN)

[21] Association for Computing Machinery (40 journals) (GREEN)

[22] BioMed Central (163 journals) (GREEN)

[23] Blackwell Publishing (698 journals) (GREEN)

[24] British Institute of Non-destructive Testing (1 journal) (GRAY)

[25] BMJ Publishing Group (23 journals) (GREEN)

[26] CAB International Publishing (17 journals) (GREEN)

[27] Cambridge University Press (186 journals) (GREEN)

[28] Clinical Laboratory Science (1 journal) (GRAY)

[29] Company of Biologists (3 journals) (GREEN)

[30] Elsevier (1882 journals) (GREEN)

[31] Emerald (206 journals) (GREEN)

[32] Endocrine Society (5 journals) (GRAY)

[33] Geological Society (10 journals) (GREEN)

[34] Georgetown University Law Center (10 journals) (GRAY)

[35] Haworth Press (254 journals) (GREEN)

[36] Imperial College Press (9 journals) (GRAY)

[37] Institute of Biology (2 journals) (GREEN)

[38] Institute of Electrical and Electronics Engineers (IEEE) (90 journals) (GREEN)

[39] Institute of Electrical, Information and Communication Engineers (26 journals) (PALE-GREEN)

[40] Institute of Physics (42 journals) (GREEN)

[41] Institution of Chemical Engineers (3 journals) (GREEN)

[42] Institution of Electrical Engineers (IEE) (4 journals) (GRAY)

[43] Internet Journal of Chemistry (1 journal) (GREEN)

[44] IOS Press (62 journals) (GREEN)

[45] John Wiley & Sons, Inc. (378 journals) (GREEN)

[46] Kluwer (837 journals) (PALE-GREEN)

[47] Lawrence Erlbaum Associates, Inc. (89 journals) (GREEN)

[48] Lippincott, Williams & Wilkins (287 journals) (GRAY)

[49] Marcel Dekker (83 journals) (GRAY)

[50] Mary Ann Liebert (56 journals) (GRAY)

[51] Massachusetts Institute of Technology Press (56 journals) (GRAY)

[52] Massachusetts Medical Society (1 journal) (GREEN)

[53] Michigan Law Review (1 journal) (GREEN)

[54] Nature Publishing Group (47 journals) (GREEN)

[55] Oxford University Press (188 journals) (PALE-GREEN)

[56] Physicians Postgraduate Press (1 journal) (GRAY)

[57] Portland Press (50 journals) (GREEN)

[58] Resilience Alliance (1 journal) (GREEN)

[59] Rockefeller University Press (3 journals) (GREEN)

[60] Royal College of General Practitioners (1 journal) (GRAY)

[61] Royal Meteorological Society (529 journals) (GREEN)

[62] Royal Society (7 journals) (GREEN)

[63] Royal Society of Chemistry (28 journals) (GRAY)

[64] Royal Society of Medicine (23 journals) (GRAY)

[65] SAGE Publications (UK and US) (366 journals) (GREEN)

[66] School of Management, University of Bath (1 journal) (GRAY)

[67] Sheffield Academic Press (17 journals) (GRAY)

[68] Society for Endocrinology (3 journals) (GRAY)

[69] Society for General Microbiology (4 journals) (GRAY)

[70] Society of Dyers and Colourists (2 journals) (GRAY)

[71] Society for Industrial and Applied Mathematics (13 journals) (GREEN)

[72] Society for In-Vitro Biology (2 journals) (GREEN)

[73] Society of Photo-optical Instrumentation Engineers (4 journals) (GREEN)

[74] Springer Verlag (Germany) (502 journals) (GREEN)

[75] Stanford University Law School (7 journals) (PALE-GREEN)

[76] Taylor & Francis (917 journals) (PALE-GREEN)

[77] University of Chicago Press (50 journals) (GRAY)

[79] Wiley-VCH Verlag Berlin (122 journals) (GREEN)

[80] Yale Law School (9 journals) (GREEN)

[81] American Society of Limnology and Oceanography (ASLO) (3 journals) (GREEN)

[82] American Institute of Aeronautics and Astronautics (7 journals) (GRAY)

[83] American Mathematical Society (18 journals) (GREEN)

[84] International Union of Pure and Applied Chemistry (1 journal) (GRAY)

[85] Professional Engineering Publishing (Institutional of Mechanical Engineers) (14 journals) (GREEN)

[86] Institute of Mathematical Statistics (7 journals) (GREEN)

[87] American Society of Hematology (1 journal) (GRAY)

[88] Nordic Ecological Society (1 journal) (GRAY)

[89] Medknow Publications (9 journals) (GREEN)

[90] Electrochemical Society (1 journal) (PALE-GREEN)

[91] Annual Reviews (38 journals) (GREEN)

[92] National Research Council Canada (15 journals) (GREEN)

[93] Ecological Society of America (4 journals) (GREEN)

[94] National Academy of Science (1 journal) (GREEN)

[95] American Society of Plant Biologists (2 journals) (GREEN)

[96] Johns Hopkins University Press (58 journals) (GREEN)

[97] Australian Computer Society Inc (1 journal) (GREEN)

[98] Australian Academic Press (10 journals) (GREEN)

[99] Association for the Advancement of Computing in Education (9 journals) (GRAY)

[100] Hindawi Publishing Corporation (15 journals) (GREEN)

[101] International Press (11 journals) (GREEN)

[103] Berkeley Electronic Press (14 journals) (GREEN)

[104] Materials Research Society (****) (GREEN)

[105] Geological Society of America (****) (GRAY)

Data from http://romeo.eprints.org/ on 1.ix.2004

## Appendix II :: Open-Access & Research Impact


Brody, T. & Harnad, S. (2004, in prep.) Earlier Web Usage Statistics as Predictors of Later Citation Impact. (**http://www.ecs.soton.ac.uk/~harnad/Temp/timcorr.doc**)

Harnad, S. & Brody, T. (2004) Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals, D-Lib Magazine 10 (6) June (**http://www.dlib.org/dlib/june04/harnad/06harnad.html**)

Harnad, S. and Brody, T. (2004) Prior evidence that downloads predict citations BMJ Rapid Responses, 6 September 2004 (**http://bmj.bmjjournals.com/cgi/eletters/329/7465/546#73000**)

Harnad, S., Brody, T., Vallieres, F., Carr, L., Hitchcock, S., Gingras, Y, Oppenheim, C., Stamerjohanns, H., & Hilf, E. (2004) The Access/Impact Problem and the Green and Gold Roads to Open Access. Serials Review 30. (**http://www.ecs.soton.ac.uk/~harnad/Temp/impact.htm**l)

Harnad, S., Carr, L., Brody, T. & Oppenheim, C. (2003) Mandated online RAECVs Linked to University Eprint Archives: Improving the UK Research Assessment Exercise whilst making it cheaper and easier. Ariadne 35 (April2003). **(http://www.ariadne.ac.uk/issue35/harnad/**)

Hitchcock, S., Woukeu, A., Brody, T., Carr, L., Hall, W., and Harnad, S.(2003) Evaluating Citebase, an open access Web-based citation-ranked search and impact discovery service.(**http://opcit.eprints.org/evaluation/Citebase-evaluation/evaluation-report.html**)

Kurtz, Michael J.; Eichhorn, Guenther; Accomazzi, Alberto; Grant, CarolynS.; Demleitner, Markus; Murray, Stephen S.; Martimbeau, Nathalie; Elwell, Barbara. (2004a) Worldwide Use and Impact of the NASA Astrophysics DataSystem Digital Library. Journal of the American Society for Information Science and Technology 55. (**http://cfa-www.harvard.edu/~kurtz/jasist1.pdf** : **http://cfa-www.harvard.edu/~kurtz/jasist1.pdf**)

Kurtz, Michael J.; Eichhorn, Guenther; Accomazzi, Alberto; Grant, Carolyn S.; Demleitner, Markus; Murray, Stephen S.; Martimbeau, Nathalie; Elwell, Barbara (2004b) The Bibliometric Properties of Article Readership Information. Journal of the American Society for Information Science and Technology 55. (**http://cfa-www.harvard.edu/~kurtz/jasist2.pdf**)

Lawrence, S. (2001) Online or Invisible? Nature 411 (6837): 521. (**http://www.neci.nec.com/~lawrence/papers/online-nature01/**)

Odlyzko, A.M. (2002) The rapid evolution of scholarly communication. Learned Publishing 15: 7-19. (**http://www.catchword.com/alpsp/09531513/v15n1/contp1-1.htm**)

Perneger, T.V. (2004) Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ. BMJ 2004;329:546-547 (4 September), doi:10.1136/bmj.329.7465.546 (**http://bmj.bmjjournals.com/cgi/content/full/329/7465/54**)

Smith, A. & Eysenck, M. (2002) The correlation between RAE ratings and citation counts in psychology. Technical Report, Psychology, University of London, Royal Holloway. (**http://psyserver.pc.rhbnc.ac.uk/citations.pdf**)

Swan, A. & Brown, S.N. (2004a) JISC/OSI Journal Authors Survey Report. (**http://www.jisc.ac.uk/uploaded_documents/JISCOAreport1.pdf** : **http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/3628.html**)

Swan, A. & Brown, S.N. (2004b) Authors and open access publishing. Learned Publishing 2004:17(3) 219-224. (**http://www.ingentaselect.com/rpsv/cw/alpsp/09531513/v17n3/s7**/)

Below is the latest (October '04) evidence that the Open Access impact advantage is neither unique to the Physical Sciences and Mathematics:

**http://citebase.eprints.org/isi_study/**

nor to the Biological Sciences:

**http://www.crsc.uqam.ca/lab/chawki/OA_NOA_biologie.gif**

The Impact advantage is there in the Social Sciences too:

**http://www.crsc.uqam.ca/lab/chawki/sociologie.htm**

# Appendix III :: Common Image File Formats

| Name and Current Version | TIFF 6.0 (Tagged Image File Format) | GIF 89a (Graphics Interchange Format) | JPEG (Joint Photographic Expert Group)/JFIF (JPEG File Interchange Format) | JP2-JPX/ JPEG 2000 | Flashpix 1.0.2 | ImagePac, Photo CD | PNG 1.2 (Portable Network Graphics) | PDF 1.4 (Portable Document Format) |
|---|---|---|---|---|---|---|---|---|
| **Extension(s)** | .tif, .tiff | .gif | .jpeg, jpg, .jif, .jfif | .jp2, .jpx, .j2k, .j2c | .fpx | .pcd | .png | .pdf |
| **Bit-depth(s)** | 1-bit bitonal; 4- or 8-bit grayscale or palette color; up to 64-bit color[1] | 1-8 bit bitonal, grayscale, or color | 8-bit grayscale; 24-bit color | supports up to 214 channels, each with 1-38 bits; gray or color | 8-bit grayscale; 24 bit color | 24-bit color | 1-48-bit; 1/2/4/8-bit palette color or grayscale, 16-bit grayscale, 24/48-bit truecolor | 4-bit grayscale; 8-bit color; up to 64-bit color support |
| **Compression** | Uncompressed Lossless: ITU-T.6, LZW, etc.<br><br>Lossy: JPEG | Lossless: LZW[2] | Lossy: JPEG<br><br>Lossless:[3] | Uncompressed<br><br>Lossless/Lossy: Wavelet | Uncompressed<br><br>Lossy: JPEG | Lossy: “Visually lossless” Kodak proprietary format[4] | Lossless: Deflate, an LZ77 derivative | Uncompressed Lossless: ITU-T.6, LZW. JBIG<br><br>Lossy: JPEG |
| **Standard/ Proprietary** | De facto standard | De facto standard | JPEG: ISO 10918-1/2<br><br>JFIF: de facto standard[5] | ISO/IEC 15444 parts 1-6, 8-11 | Publicly available specification | Proprietary | ISO 15948 (anticipated)[6] | De facto standard[7] |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Color Mgmt.** | RGB, Palette, $YC_bC_r$,[8] CMYK, CIE L*a*b* | Palette | $YC_bC_r$ | Palette, YCbCr, RGB, sRGB, some ICC[9] | PhotoYCC and NIF RGB,[10] ICC (optional) | PhotoYCC | Palette, sRGB, ICC | RGB, $YC_bC_r$, CMYK |
| **Web Support** | Plug-in or external application | Native since Microsoft® Internet Explorer 3, Netscape Navigator® 2 | Native since Microsoft® Internet Explorer 2, Netscape Navigator® 2 | Plug-in | Plug-in | Java™ applet or external application | Native since Microsoft® Internet Explorer 4, Netscape® Navigator 4.04, (but still incomplete) | Plug-in or external application |
| **Metadata Support** | Basic set of labeled tags | Free-text comment field | Free-text comment field | Basic set of labeled tags[11] | Extensive set of labeled tags | Through external databases; no inherent metadata | Basic set of labeled tags plus user-defined tags. | Basic set of labeled tags |
| **Comments** | Supports multiple images/file[12] | May be replaced by PNG; interlacing and transparency support by most Web browsers | Progressive JPEG widely supported by Web browsers[13] | Multiple resolutions, progressive display, tiling, region of interest coding and many other advanced features | Provides multiple resolutions of each image; wide industry support, but limited current applications | Provides 5 or 6 different resolutions of each image; unclear future | May replace GIF, though market penetration has been spotty | Preferred for printing and viewing multipage documents; strong government use |
| **Home Page** | Unofficial TIFF home page | GIF specification | JPEG home page | JPEG 2000 home page | FlashPix home page | Photo CD home page | PNG home page | PDF home page specs |

[1] Though the TIFF 6.0 specification provides for 64-bit color, many TIFF readers support a maximum of 24-bit color.

[2] LZW is patented and its use in software development may require licensing and royalty payments: Unisys, “License Information on GIF and Other LZW-based Technologies,” LZW Patent and Software Information.

[3] The original JPEG specification included a lossless mode, but most JPEG applications never supported it. Some files referred to as lossless JPEGs are really non-JPEG compressed files in a JFIF wrapper. There is a new specification for lossless JPEG (JPEG-LS) but it has not been finalized. ISO SC29/WG1, “JPEG - Information Links.”

[4] Visually lossless refers to compression techniques that are themselves lossy, but that take advantage of characteristics of human sight to create an image that is virtually indistinguishable from its uncompressed form.

[5] JFIF was released into the public domain by C-Cube Microsystems. The “official” file format for JPEG files is SPIFF (Still Picture Interchange File Format), but by the time it was released, JFIF had already achieved wide acceptance. SPIFF, which has the ISO designation 10918-3, offers more versatile compression, color management, and metadata capacity than JPEG/JFIF, but it has little support. It may be superseded by JPEG 2000/DIG 2000: ISO SC29/WG1, JPEG - Information Links. Digital Imaging Group, “JPEG 2000 and the DIG: The Picture of Compatibility.”

[6] Approved by W3C to replace GIF for Web use.

[7] Adobe has released enough information to allow developers to write applications that read and modify PDF files. However, pdf files are most commonly created and accessed using Adobe's own Acrobat software.

[8] Similar to CIE Lab, YCbCr is composed of three channels: one for luminance (Y) and two for chrominance (CC).

[9] Others are supported in the file format extensions defined in ISO/IEC 15444-2 (JPX file format).

[10] NIF RGB is defined identically to sRGB in the Flashpix 1.0.2 specification. The next revision of the Flashpix specification may move to sRGB.

[11] The JP2 file format also specifies a flexible means to add substantial metadata, either as binary data or in XML. However, this data is considered optional, and baseline JP2 readers are not required to read it.

[12] The TIFF 6.0 specification calls for the ability to store multiple TIFF images in a single file, but not all TIFF readers support this feature.

[13] Some early versions of Internet Explorer may not display progressive JPEGs properly.

From
http://www.library.cornell.edu/preservation/tutorial/presentation/_ftnref13
(1.ix.2004)

# Appendix IV :: foxing

If the documents are being scanned simply to OCR the text, we can be a bit more cavalier about how 'gently' we treat the pixels. If however you want to present the documents as 'images' we might need to use more subtle methods which will preserve the character of the document, so automating things might not be a good idea in that case.

From the sounds of things, the relative brightness of your letters is not much different from the brightness of the foxing stains - so we can not use any tools that rely on brightness differences to select the areas we want to change on the digital documents.

The key to solving the problem will hinge on their being a reasonable colour difference between the 'ink/letters' on the page and the colour of the paper or foxing stains. If, for example the ink is black or bluish, and the paper is reddish/brown/yellow we can first isolate the area needing to be lightened by using tools within the image editor which 'select a specified range of colours'. Then once a selection is made (which usually creates a line of 'marching ants around the pixels to be affected) you can simply adjust the brightness of those colours - essentially if you make them the same brightness as the surrounding non-stained areas of the paper they should almost disappear when the document is converted into gray-scale. These kinds of changes can be saved, reloaded, and then played back on a batch of files using action-recording features of image editors like photoshop.

We played in the office for a bit, and worked out one version of this 'select the colours and then lighten them' method. I have to say that this is one of those multi step procedures that is relatively easy for someone who has played with image editors for a while, but involves enough details that someone new to photoshop (or image editing) will likely find the process very confusing.

Below I have listed a record of the process I used to remove foxing from some images taken of an old book in the TASI office. The pages had significant dark brown stains but the text was printed in black ink. The procedure worked pretty well when run as a batch process on a folder of images - as they all had pretty consistent coloration to the foxing stains:

**Step 1** - creating a 'selection with the colour range tool' and a 'hue adjustment' file that isolate and correct the foxing, save those files.

a) scan the book pages in colour;
b) open one 'typical' page with foxing in photoshop;
c) click on eyedropper tool and make sure it is set to 5x5 pixel sampling;
d) select>colour range ;
e) bring fuzzyness down to an amount where you can see the white patches on the preview are just the areas on the document covered by the stains;
f) SAVE THIS SETTING as a colour range selection file - give the adjustments file a memorable name associated with the book you are scanning;
g) hit ok. Now you should have a selection of marching ants around the stains on the pages - anything you do from this point forward will only affect the pixels inside the selection (ie the stains);
h) then go to the image>adjust>hue/saturation slider;

i) pull back the saturation slider a bit [say -10] (foxing stains have more saturated colour than the pages around them) j) bring up the brightness slider until the stains have the same brightness as the paper around them (on my tests this was around +20 to 25 – but you really have to do this by eye - just try to match the brightness of the rest of the paper in the document;
j) when the stains have become much less noticeable SAVE THE SETTINGS for this hue saturation change by hitting the save button- give the file a descriptive name such as 'hue adjust for bookx foxing' so you can find it easily later;
k) hit the ok button to apply the changes to the image.

At this point we have saved two adjustment files: one that selects the stain colours on the page and a second adjustment that brings the stain colours up to the brightness of the unstained paper. Now you need to record an action that loads these adjustment files and batch applies them to a folder of images with similar stains.

**Step 2** - Creating an action in photoshop which uses those adjustment files to remove the foxing. To make this action,

a) begin recording a new action (give it a name like 'foxing removal for book x'),
b) open one sample image in the series go to select colour range AND LOAD THE SELECTION YOU CREATED EARLIER and press ok to apply the selection;
c) Now go to the image>adjust>hue saturation panel, again load the hue/sat adjustment file you saved earlier and hit ok.
d) Now go to file save and file close.
e) Click the square box on the actions palette to end the action recording.

Now you have an action which removes the foxing, using the settings file you created earlier.

**Step 3** - Running that action as a batch process on a folder of images.

You can go to File>automate>batch and configure the options in the dialogue box to play this action on an entire folder of images. Remember to always save the changed images out to a separate destination folder rather than over writing the original files! If you are not sure how to setup an action in photoshop you can always edit the images one at a time - simply loading the saved adjustment files and applying them in the same order as you examine each image.

After the stains are removed you can convert the images to grayscale and they should be much more readable.

How well it works on each image will depend on how close the stains are in colour among the various pages in the book - if the stains are green at the front of the book but red at the back of the book then the action will only work for stains similar to the adjustment file you created in Step 1. And if you have old brown ink (instead of black) and brown foxing stains that almost match the ink colour - this procedure will not help you much. Also keep in mind that you will need to go

through the procedure again for every book that has stains of a significantly different character.

From Edward Mallon, Technical Training Officer
TASI - Technical Advisory Service for Ima 22.iv.2004

## Appendix V :: Keyword Extraction from PDF image files

Unix shell script:

```
cat input.txt\
|  tr ' ' '\012'\
| tr -sc 'A-Za-z' '\012'\
| tr 'A-Z' 'a-z'\
| sort -bdu > st_1.tmp
 comm -12 st_1.tmp
/home/eprints/lang/psycholinguistic_database/headwords/he
adwordss.txt > st_2.tmp
comm -23 st_2.tmp stoplist.txt > keyword_output.txt
 rm st_*.tmp
```

Explanation:

- List input text;
- Replaces spaces with newlines (each word is now on a line);
- Replace non-alphabetics with newlines;
- Change to lower case;
- Sort in dictionary order, remove duplicates and place in temporary file;
- If words also occur in the dictionary, output them to a temporary file (i.e. remove non-words);
- Remove any stopwords which are common to the stopword list and word file: output final wordlist to file ‘keyword_output.txt’;
- Remove all temporary files.

Note: you will need to change file paths to suit your setup. In this case input file and stoplist are in the same directory; dictionary is in path as given.

This is probably not the fastest or most elegant way to do this, but it is easy to follow and is still extremely rapid.

### Stoplist:

| | | | |
|---|---|---|---|
| A | almost | anything | available |
| a's | alone | anyway | Away |
| able | along | Anyways | awfully |
| about | Already | anywhere | b |
| Above | also | apart | be |
| according | although | appear | Became |
| accordingly | always | Appreciate | because |
| across | Am | appropriate | become |
| Actually | among | are | becomes |
| after | amongst | aren't | Becoming |
| afterwards | an | Around | been |
| again | And | as | before |
| Against | another | aside | beforehand |
| ain't | any | ask | Behind |
| all | anybody | Asking | being |
| allow | Anyhow | associated | believe |
| Allows | anyone | at | below |

Beside
besides
best
better
Between
beyond
both
brief
But
by
c
c'mon
c's
came
can
can't
cannot
cant
cause
causes
certain
certainly
changes
clearly
co
com
come
comes
concerning
consequently
consider
considering
contain
containing
contains
correspondin
g
could
couldn't
course
currently
d
definitely
described
despite
did
didn't
different
do
does
doesn't
doing
don't
done
down
downwards
during
e
each
edu
eg
eight
either
else
elsewhere
enough
entirely
especially
et
etc
even
ever
every
everybody
everyone
everything
everywhere
ex
exactly
example
except
f
far
few
fifth
first
five
followed
following
follows
for
former
formerly
forth
four
from
further
furthermore
g
get
gets
getting
given
gives
go
goes
going
gotten
greetings
h
had
hadn't
happens
hardly
has
hasn't
have
haven't
having
he
he's
hello
help
hence
her
here
here's
hereafter
hereby
herein
hereupon
hers
herself
hi
him
himself
his
hither
hopefully
how
howbeit
however
i
i'd
i'll
i'm
i've
ie
if
ignored
immediate
in
inasmuch
inc
indeed
indicate
indicated
indicates
inner
insofar
instead
into
inward
is
isn't
it
it'd
it'll
it's
its
itself
j
just
k
keep
keeps
kept
know
knows
known
l
last
lately
later
latter
latterly
least
less
lest
let
let's
like
liked
likely
little
look
looking
looks
ltd
m
mainly
many
may
maybe
me
mean
meanwhile
merely
might
more
moreover
most
mostly
much
must
my
myself
n
name
namely
nd
near
nearly
necessary
need
needs
neither
never
nevertheless
new
next
nine
no
nobody
non
none
noone
nor
normally
not
nothing
novel
now
nowhere

o
obviously
of
off
often
oh
ok
okay
old
on
once
one
ones
only
onto
or
other
others
otherwise
ought
our
ours
ourselves
out
outside
over
overall
own
p
particular
particularly
per
perhaps
placed
please
plus
possible
presumably
probably
provides
q
que
quite
qv
r
rather
rd
re
really
reasonably
regarding
regardless
regards
relatively
respectively
right
s
said
same
saw
say
saying
says
second
secondly
see
seeing
seem
seemed
seeming
seems
seen
self
selves
sensible
sent
serious
seriously
seven
several
shall
she
should
shouldn't
since
six
so
some
somebody
somehow
someone
something
sometime
sometimes
somewhat
somewhere
soon
sorry
specified
specify
specifying
still
sub
such
sup
sure
t
t's
take
taken
tell
tends
th
than
thank
thanks
thanx
that
that's
that’s
the
their
theirs
them
themselves
then
thence
there
there's
thereafter
thereby
therefore
therein
theres
thereupon
these
they
they'd
they'll
they're
they've
think
third
this
thorough
thoroughly
those
though
three
through
throughout
thru
thus
to
together
too
took
toward
towards
tried
tries
truly
try
trying
twice
two
u
un
under
unfortunatel
y
unless
unlikely
until
unto
up
upon
us
use
used
useful
uses
using
usually
uucp
v
value
various
very
via
viz
vs
w
want
wants
was
wasn't
way
we
we'd
we'll
we're
we've
welcome
well
went
were
weren't
what
what's
whatever
when
whence
whenever
where
where's
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither
who
who's
whoever
whole
whom
whose
why
will
willing
wish
with
within
without
won't
wonder
would
would
wouldn't

```
x               you             you've          yourselves
y               you'd           your            z
yes             you'll          yours           zero
yet             you're          yourself
```

# Appendix VI :: Recommended Data Formats

**Table of suggested formats**

| | |
|---|---|
| **Digital Resource Type** | Database |
| **Preferred Deposit Formats** | Delimited text (tab or pipe delimited, comma delimited with quotes around textual values) with SQL setup |
| **Acceptable Deposit Formats** | Database software formats with full description of database structure (tables, fields, data types, keys and relationships):<br><br>Access95+<br>FoxPro 2.5+<br>Paradox<br>Filemaker Pro<br>Delimited text with full description of database structure (tables, fields, data types, keys and relationships) |
| **Problematic Deposit Formats** | Obsolete database software formats |
| **Problematic Aspects** | User interface forms, queries using custom extensions to SQL.<br>Report templates |

| | |
|---|---|
| **Digital Resource Type** | Plain Text |
| **Preferred Deposit Formats** | ASCII (7 bit)<br>UTF-8 UNICODE<br>UTF-16 UNICODE |
| **Acceptable Deposit Formats** | ISO 8859 character sets<br>MS-DOS codepages<br>MS-Windows codepages<br>Apple codepages<br>Other UNICODE encodings |
| **Problematic Deposit Formats** | EBCDIC |

| | |
|---|---|
| **Problematic Aspects** | - |

| | |
|---|---|
| **Digital Resource Type** | Word Processor Document |
| **Preferred Deposit Formats** | Rich Text Format<br>Word |
| **Acceptable Deposit Formats** | PDF<br>WordPerfect<br>StarOffice / OpenOffice |
| **Problematic Deposit Formats** | Early versions of word processor packages.<br>Word processor packages for platforms other than Windows, Mac, Unix, Linux |
| **Problematic Aspects** | - |

| | |
|---|---|
| **Digital Resource Type** | Mark-up |
| **Preferred Deposit Formats** | XML (including XHTML) with DTD or schema<br>SGML (including HTML) with DTD |
| **Acceptable Deposit Formats** | - |
| **Problematic Deposit Formats** | Custom mark-up without DTD or schema |
| **Problematic Aspects** | - |

| | |
|---|---|
| **Digital Resource Type** | Raster Image |
| **Preferred Deposit Formats** | TIFF v6+,<br>PNG |

| | |
|---|---|
| **Acceptable Deposit Formats** | GIF<br>BMP<br>PCX<br>Photoshop*<br>Paintshop Pro*<br>CGM<br>PhotoCD<br>GeoTIFF |
| **Problematic Deposit Formats** | Any lossy compression (e.g. JPEG)<br>Minority image formats (e.g. .bob)<br>PDF |
| **Problematic Aspects** | - |

| | |
|---|---|
| **Digital Resource Type** | CAD |
| **Preferred Deposit Formats** | - |
| **Acceptable Deposit Formats** | STEP<br>DXF |
| **Problematic Deposit Formats** | - |
| **Problematic Aspects** | - |

| | |
|---|---|
| **Digital Resource Type** | GIS |
| **Preferred Deposit Formats** | None |
| **Acceptable Deposit Formats** | GML (version 2 or above)<br>ESRI Shape Files<br>ESRI Export formats (.e001)<br>MapInfo Formats<br>SDTS<br>DXF<br>DWG |
| **Problematic** | NTF |

| **Deposit Formats** | |
|---|---|
| **Problematic Aspects** | The NTF format is supported by the OSGB but they have announced their intention to move away from NTF in favour of GML2.1.2 |

| **Digital Resource Type** | Spreadsheets |
|---|---|
| **Preferred Deposit Formats** | Delimited text files(tab or pipe delimited, comma delimited with quotes around textual values) |
| **Acceptable Deposit Formats** | Excel<br>Lotus<br>Quattro |
| **Problematic Deposit Formats** | Obsolete spreadsheet software formats |
| **Problematic Aspects** | Functionality of formulas, results of functions, embedded charts, complex visual layout (borders, fonts, colour, column widths etc) |

| **Digital Resource Type** | Executables |
|---|---|
| **Preferred Deposit Formats** | None |
| **Acceptable Deposit Formats** | ANSI C or Java 1.2+ source code |
| **Problematic Deposit Formats** | Compiled code<br>C++, C#, Visual Basic, Pascal, Ada, assembler or other common programming languages source code |
| **Problematic Aspects** | - |

| | |
|---|---|
| **Digital Resource Type** | Audio |
| **Preferred Deposit Formats** | WAV |
| **Acceptable Deposit Formats** | MP3<br>Ogg Vorbis |
| **Problematic Deposit Formats** | Real Audio<br>Modules<br>Other minority audio formats |
| **Problematic Aspects** | Streamed audio |

| | |
|---|---|
| **Digital Resource Type** | Moving Image |
| **Preferred Deposit Formats** | None |
| **Acceptable Deposit Formats** | MPEG-1<br>MPEG-4 with common codec<br>AVI with common codec<br>MJPEG with WAV file |
| **Problematic Deposit Formats** | Streamed audio/video<br>Unknown codecs |
| **Problematic Aspects** | Sync issues |

| | |
|---|---|
| **Digital Resource Type** | Vector Graphics |
| **Preferred Deposit Formats** | SVG<br>DXF |
| **Acceptable Deposit Formats** | Adobe Illustrator |
| **Problematic Deposit Formats** | - |
| **Problematic** | - |

| Aspects | |
|---|---|

| | |
|---|---|
| **Digital Resource Type** | Geophysics Datafile |
| **Preferred Deposit Formats** | AGF |
| **Acceptable Deposit Formats** | Contors<br>Geoplot |
| **Problematic Deposit Formats** | - |
| **Problematic Aspects** | - |

| | |
|---|---|
| **Digital Resource Type** | Statistical Dataset |
| **Preferred Deposit Formats** | SPSS Portable<br>Delimited text files(tab or pipe delimited, comma delimited with quotes around textual values) with data dictionary and codebook |
| **Acceptable Deposit Formats** | Stata<br>SAS |
| **Problematic Deposit Formats** | Fixed width text files without appropriate documentation |
| **Problematic Aspects** | - |

| | |
|---|---|
| **Digital Resource Type** | Virtual Reality |
| **Preferred Deposit Formats** | None |
| **Acceptable Deposit Formats** | VRML |
| **Problematic Deposit Formats** | Proprietary virtual reality modelling software formats |

| | |
|---|---|
| **Problematic Aspects** | - |

This page is adapted from the information held at the Arts and Humanities Data Services (ADS) , University of York. If you are interested in finding out about preservation of digital media in greater depth, the Conservation on Line (COOL) project is also a good starting place.

# Appendix VII :: Scottish Science Information Strategy Working Group Declaration (August 2004)

'We believe that the interests of Scotland will be best served by the rapid adoption of open access to scientific and research literature.'

- ***Preamble***. The timely, universal and organised dissemination of advances in scientific and public policy research is fundamental to the proper operation of a modern society, in terms of community awareness and empowerment, economic advance, and optimal functioning of health, education and other vital services. For Scotland, this means not only gaining access to the fruits of research from throughout the world but also exposing the endeavours of our researchers as widely as possible to the world at large.

***The Present Situation***. Until recently, the current system of scholarly communication and dissemination of research results has worked well for society, learned institutions, universities and individual researchers, given the restrictions of print-based publication. These restrictions are caused not only by the limitations of print as a medium for presenting research, but also by the high annual subscription charges and price increases well above inflation in some disciplines which distort the 'traditional' publication system for research journals by reducing the availability of journals in all disciplines, often at the expense of small learned societies.

This subscription-based system is showing signs of increasing strain, and we believe that it is no longer the most advantageous means of disseminating crucial research results to all those interested, whether in our leading research institutions or in the wider community. By its very nature, it severely restricts access to leading edge research, published only in appropriate scientific journals and subscribed to by at best a handful of institutional libraries. Yet the advent of digital content and the web has the potential to render the current system obsolete, and there are signs now that the full power of networked digital content to change the system for the benefit of research and the diffusion of knowledge generally is beginning to be understood. The research pooling agenda within Scotland depends on wider access to research, if it is to achieve its aim of maximising Scotland's research potential.

New options have been developed in recent years, potentially removing constraints upon access and opening up research literature to be available online to everyone. These developments go under the broad heading of Open Access, which has been defined as 'free availability on the public Internet, permitting all users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal or technical barriers other than those inseparable from gaining access to the Internet itself'[1].

***Open Access.*** This vision of open access to the scientific and research literature has captured the imagination of research funding agencies, international

[1] *Budapest Open Access Initiative*, February 2002 [http://www.soros.org/openaccess/read.shtml]

organisations and other nation states and we seek to follow the example of similarly sized countries in Scandinavia and elsewhere in seizing the opportunities now open to us. Other initiatives are being undertaken in this area worldwide: in particular the House of Commons Select Committee on Science and Technology has recently reported in a favourable way on current developments[2,3,4,5]. We believe that the interests of Scotland – for the economic, social and cultural benefit of the population as a whole, and for the maintenance of the longstanding high reputation of research within Scottish universities and research institutions – will be best served by the rapid adoption of open access.

There are two main routes to achieving open access, and we wish to register our support for both. The number of open access journals has been growing in recent years, with some publishers offering all their journals on an open access basis, and others offering it only for selected titles. There are of course still significant costs associated with publishing online, in particular the cost of organising the essential peer review service, but for open access journals these are covered by publication fees rather than subscriptions (with appropriate exemptions for those who cannot pay for various reasons, including financial constraints in the case of researchers in the developing world).

The second route is usually described as 'self-archiving', where authors deposit the final, post peer review, electronic version of their articles in an institutional, or subject-based, repository: appropriate software adhering to open standards and encouraging interoperability allows these repositories to be searched jointly, and relevant articles retrieved from repositories located worldwide. Some subject-based repositories (for example, for high-energy physics[6]) have been in existence for a number of years. Several universities in Scotland have already established institutional repositories (which include theses, departmental reports, conference papers, etc as well as journal articles), and plans are underway to enable other Scottish research institutions to deposit their own research output appropriately. It should be noted that a growing majority of publishers, but not yet all, expressly permit self-archiving of the final version of an article.

***Conclusion.*** There is mounting evidence to suggest that open access increases the reach and impact of research. More people can and do view and read open access articles, and there are indications that these articles are cited more frequently and earlier than is the case for articles not available in this way.[7]

---

[2] *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities,* October 2003 [http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html] - signed by research organisations from Germany and other European countries

[3] *Bethesda Statement on Open Access Publishing,* April 2003 [http://www.earlham.edu/~peters/fos/bethesda.htm] - signed by a number of US learned societies

[4] *Scientific publishing: a position statement by the Wellcome Trust in support of open access publishing,* September 2003 [http://www.wellcome.ac.uk/en/1/awtvispolpub.html]

[5] House of Commons. Science and Technology Select Committee. *Scientific Publishing:Free for All? Tenth Report of Session 2003-04. Volume 1: Report* [http://www.publications.parliament.uk/pa/cm200304/cmselect/cmsctech/399/399.pdf]

[6] *arXiv.org e-Print archive* [http://www.arxiv.org/]

[7] See, for example, Hitchcock, S. et al 'The impact of OAI-based search on access to research journal papers', *Serials* Vol 16, No 3, November 2003, 255-260 [http://opcit.eprints.org/serials-short/serials11.html]

Open access publishing therefore provides a more cost-efficient means of disseminating the outputs of research funded from the public purse than does the current system which requires that public money be paid over to external bodies in order to gain restricted access to the same research outputs. In the light of these developments, and recognising the huge potential gains to Scotland in terms of impact, comparative advantage, and return on public investment if open access to our research can be established quickly, we will use our best endeavours to ensure that research carried out in Scotland is published in an open access format, recognising that a transition phase may be necessary in some areas.

***Action.*** The signatories to this declaration endorse the general principles of open access, and commit themselves to implementing as and when possible the following actions, thereby ensuring a national commitment to the free-est and fullest access to scholarly information:

RESEARCH FUNDERS

- Require as a condition of grant that publications resulting from funding are available on open access by means of self-archiving in an appropriate repository.
- Allocate funds for fees for publication within research grants, to facilitate publication in an open access journal where appropriate.
- Encourage traditional research publishers to offer open access publishing streams with fair pricing.

UNIVERSITIES/RESEARCH INSTITUTIONS

- Set up institutional repositories, and/or liaise with other organisations to establish a joint repository.
- Encourage, and as soon as practical mandate, researchers to deposit copies of their outputs (articles, reports, conference papers, etc) in an institutional or co-operative repository.
- Encourage, and as soon as practical mandate, the deposit of PhD theses in an institutional repository.
- Review intellectual property policies, to ensure that researchers have the right and duty to provide an open access version of their research.

SHEFC

- Develop sector-wide policies in this area.
- Consider open access issues when taking forward the research pooling agenda.

SCOTTISH EXECUTIVE

- Recognise the benefits to society as a whole of wide access to knowledge.
- Endorse implementation of open access within broader initiatives such as Smart, Successful Scotland and Openscotland.

- Take a leading role by working with other national governments in promoting open access.

# Appendix VIII :: Security scripts

Text removed from this page

Text removed from this page

Appendix IX :: Search Results

These were from putting ‘car parking’ in the St Andrews search engine and are fairly typical of the output you get. And yes, they are exactly as they appear on the results screen, complete with the formatting commands shown.

**Documents 1 - 10 of 173 matches. More ☆'s indicate a better match.**

---

**School of Chemistry, University of St Andrews**☆☆☆☆

Getting to St Andrews Air Nearest airport is Edinburgh. **Car**-hire facilities at airport; otherwise airport bus to centre of Edinburgh (Waverley Station) and by rail to Leuchars. Rail Nearest station is Leuchars (5 miles) on main line from London (King's Cross) to Edinburgh and Aberdeen. St Andrews ...

*http://chemistry.st-andrews.ac.uk/gettinghere.html* 19-10-2004, 11006 bytes

**Weddings**☆☆☆

... * Bells * Pipers * Flowers * Candles * Photography * Alternative photography * Recordings: Video and Audio * Confetti * How to get to St Andrews * **Parking** * Catering and reception * Contacting the Chaplaincy * Final checklist for wedding We are delighted that you are considering being married in ...

*http://www.st-andrews.ac.uk/services/chaplaincy/weddings.shtml* 19-10-2004, 55307 bytes

**Weddings**☆☆☆

... * Bells * Pipers * Flowers * Candles * Photography * Alternative photography * Recordings: Video and Audio * Confetti * How to get to St Andrews * **Parking** * Catering and reception * Contacting the Chaplaincy * Final checklist for wedding We are delighted that you are considering being married in ...

*http://www.st-andrews.ac.uk/chaplaincy/weddings.shtml* 19-10-2004, 55307 bytes

**[minibus.rtf]**☆☆☆

{\rtf1\ansi\ansicpg1252\uc1 \deff0\deflang1033\deflangfe1033{\fonttbl{\f0\froman\fcharset0\fprq2{*\panose 02020603050405020304}Times New Roman;}{\f3\froman\fcharset2\fprq2{*\panose 05050102010706020507}Symbol;}}{\colortbl;\red0\green0\blue0; \red0\green0\blue255;\red0\green255\blue255;\red0\green255 ...

*http://www.st-andrews.ac.uk/services/safety/webpages/minibus/minibus.rtf* 19-07-2002, 83667 bytes

**[minibus.rtf]**☆☆☆

{\rtf1\ansi\ansicpg1252\uc1 \deff0\deflang1033\deflangfe1033{\fonttbl{\f0\froman\fcharset0\fprq2{*\panose 02020603050405020304}Times New Roman;}{\f3\froman\fcharset2\fprq2{*\panose 05050102010706020507}Symbol;}}{\colortbl;\red0\green0\blue0; \red0\green0\blue255;\red0\green255\blue255;\red0\green255 ...

*http://www.st-andrews.ac.uk/safety/webpages/minibus/minibus.rtf* 19-07-2002, 83667 bytes

**8** 

**...** single and 1 shared Meals: 19 Contract length: standard, although the option to stay during Christmas and Spring vacation may be possible. Limited **car parking** Interior views Back to "A Guided Tour of the Residences" *http://www.st-andrews.ac.uk/resbus/SALLYSview.html* 03-07-2003, 1585 bytes

*Results downloaded 14.x.04*