The International Journal of Digital Curation

Issue 2, Volume 2 | 2007

PRONOM-ROAR: Adding Format Profiles to a Repository Registry to Inform Preservation Services

Tim Brody, Leslie Carr, Jessie M.N. Hey,

School of Electronics and Computer Science,

University of Southampton

Adrian Brown,

Services Manager, Digital Preservation Department,

The National Archives (UK)

Steve Hitchcock,

Project Manager,

Preserv Project

November 2007

Abstract

To date many institutional repository (IR) software suppliers have pushed the IR as a digital preservation solution. We argue that the digital preservation of objects in IRs may better be achieved through the use of light-weight, add-on services. We present such a service – PRONOM-ROAR – that generates file format profiles for IRs. This demonstrates the potential of using third-party services to provide preservation expertise to IR managers by making use of existing machine interfaces to IRs.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

Introduction

PRONOM-ROAR is a file format profiling service that uses the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) standard and the PRONOM-DROID tool from the UK National Archives. PRONOM-ROAR harvests digital objects from OAI-compliant repositories, identifies their file format(s) using DROID[1] and from that provides an exemplar Web interface and email alert service.

The PRONOM-ROAR Web interface generates a profile graph for the repository, a kind of signature of the variety of file formats that the repository is storing. The administrator can then investigate further by listing all objects in a given format to determine, for example, whether a parallel format is also available.

The prototype email alert service provides the user with a periodic email notification of new objects entering the repository. This email contains both the number of new records created (in whatever metadata formats provided by the OAI interface) and, if present and publicly accessible, the formats of any digital objects attached to those records.

PRONOM-ROAR is an exemplar service developed by the Preserv Project[2]. A service-orientated approach to digital preservation has been developed during the Preserv Project, with a focus on providing cross-repository support using flexible tools that can be combined to provide a comprehensive preservation strategy.

Flexible Preservation Services

The Preserv Project is investigating the provision of preservation services for institutional repositories (IRs). Service providers can provide preservation expertise, determined by best practice, enabling repository administrators to focus on supporting and capturing content. By building services that address specific preservation needs, repository software such as EPrints can, apart from offering some elementary preservation support, focus on the primary tasks of user interaction, authorisation, storage, and access.

The choice and provision of preservation services will be informed by repository policies, including policy on preservation. What preservation strategy a repository adopts, hence which preservation services it may require, will depend on the communities it serves. Different communities will have different long-term requirements and use of document types. A key step, however, in developing a preservation policy is to identify the types of material contained in a repository in terms of technical structure, or file formats (e.g. PDF, HTML) - without knowing what they have (and the technical preservation steps required for that material) a repository manager can only guess at the preservation activities they will need to undertake. The National Archives (TNA) curates a database of file formats, PRONOM[3], and this can help to identify repository content by using TNA's Digital Record Object Identification (DROID) open source software, which can be downloaded and applied by any repository.

[1] DROID (file format identification tool): http://droid.sourceforge.net/
[2] Preserv (project home page): http://preserv.eprints.org/
[3] PRONOM (format technical registry): http://www.nationalarchives.gov.uk/pronom/

Format identification is only a first step towards preservation, however. The question is what you do with this information. Format identities need to be verified, and file formats may need to be migrated to other formats in the event of obsolescence. This is where preservation services can help.

Technical Metadata for Preservation: File Format Information

> "Preservation metadata is the information necessary to carry out, document, and evaluate the processes that support the long-term retention and accessibility of digital materials." (PREMIS, 2005)

The Preserv Project has mapped PREMIS preservation metadata fields onto the metadata available from institutional repositories (Hitchcock, 2007). Two PREMIS fields that were identified as required but are not commonly provided by IR software were the format and software elements. These correspond respectively to the format of objects in the IR and the software application environment the IR exists in (i.e. the applications being used to create objects for the IR and applications available to access material from the IR).

The PREMIS format element contains *formatDesignation*, *formatName, formatVersion* and format registry data. The software element describes the software that created the *format* - it is a part of the *environment* description. Together the format and software elements provide the information necessary to, firstly, know how to render an object and, secondly, depending on the availability of the software, whether preservation actions may be necessary to maintain future access to the object.

PRONOM-DROID

The National Archives of the UK (TNA) has developed a technical registry of file formats called *PRONOM*. The first version was developed in 2002 to hold technical data about material being stored in TNA's archive (TNA is responsible for the long-term storage of UK government records, including census data). The genesis of PRONOM is described more fully in Darlington (2003).

In the fourth incarnation of PRONOM, TNA introduced the first of a series of planned preservation tools: *DROID* (Digital Record Object Identification – see Brown (2005a)). DROID performs the automated, batch, identification of the formats of files. Using a database of file format signatures, DROID identifies the format of a file, without relying on a file extension that may be shared by many applications. File format signatures are based on one or more key byte sequences that identify – or at least are intended to identify – uniquely a file format based on its content.

In practice a signature may match more than one format. Formats may be embedded in other formats; for instance OLE embedding allows one document of a Windows-based application to be embedded in another. Other formats may be more generic in nature, for instance a 'text file' may be something that contains (predominantly) characters used in written language on newline-terminated lines, but that could match anything from script files to emails.

The PRONOM database includes unique identifiers for each format record (PUIDs – PRONOM Unique IDentifiers). The current version of DROID provides the PUID for formats it identifies. A PUID resolution service is currently in development that will enable the file format metadata to be extracted by other services e.g. by PRONOM-ROAR which, in time, will enable a *technology watch service* to provide advice based on the PRONOM database. A technology watch service, in addition to containing technical data about a format, includes an assessment of the preservation friendliness of that format. The TNA envisages using such a technology watch service to support preservation planning, that is managing and determining policy to use to provide ongoing access to archived material.

Initial Approach: Integrating Format ID into an EPrints Repository

The initial experiments in Preserv with DROID were to integrate it into the deposit process in *EPrints*, such that, when a user uploaded a file, DROID was run, and the file format metadata were captured and stored within the repository. The original motivation for building DROID into the deposit process was to warn the user or administrator of files that were in a preservation-unfriendly format (thereby prompting some mitigating action, e.g. format conversion). The difficulty is determining what constitutes preservation friendliness or unfriendliness.

The preservation friendliness of a format might be determined by how well it is structured, whether it is proprietary, how well defined it is, the difficulty in producing the format and/or how usable that format is by the object's target community, now and in the future. Without having a method of empirically testing the effectiveness of preservation strategies (we have no parallels with which to compare institutional repositories), its impossible to avoid a degree of guesswork on what constitutes a good preservation format based on these criteria. Regardless of this important question, we investigated the difficulty of integrating DROID into an IR deposit process, in the hope that recommendations for effective preservation formats can be developed in the community.

EPrints has hooks in the deposit process that allow automated processing and validation to occur. One of these hooks is triggered whenever a document is added to an eprint object. (In EPrints' data model, a document is typically a 'format' (e.g. HTML) that may contain many files (e.g. HTML text and inline images in JPEG format). The aim of this initial experiment was simply to determine how difficult it was to integrate DROID into the EPrints deposit process, for instance it was not able to cope with more than one file in an eprint object.)

To store the output from DROID, a metadata field was added to eprint objects called 'PRONOM'. The output from DROID is an XML document that can contain multiple file format matches for any given file, along with descriptive data. In this experiment the XML was simply stored directly in the database, so that, when the eprint object was exported in EPrints XML format, the DROID XML was converted to a string and hence visible in the results (Figure 1).

The intention was that an external service could access the file format metadata, as provided by DROID, and perform some intelligent analysis. In addition, the DROID

tool could eventually be used to provide automatic file format identification to the EPrints deposit process. The drawback with pursuing the integration of these features into EPrints, or any other IR software for that matter, is that it depends on the development life cycle of the software, and after that, on the good will of implementers to adopt our features necessary to test them 'in the field'. This is particularly difficult with live repositories, as we wanted to add the DROID use to the deposit process, something that could possibly interfere with end-users' ability to use the repository.

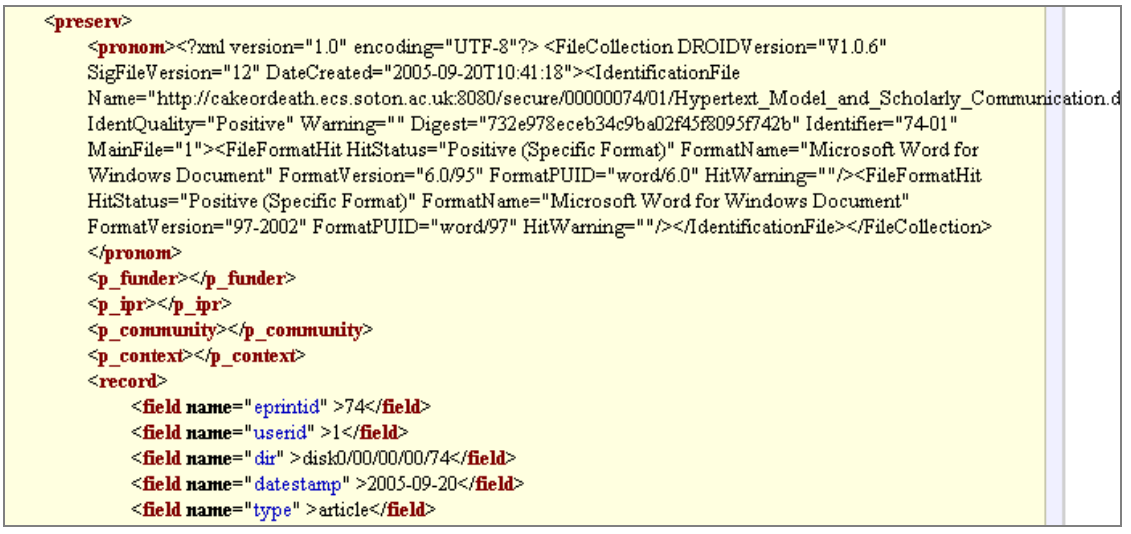


Figure 1. EPrints XML output including DROID XML (in the PRONOM section).

Building a Profile for an Existing IR

Integrating file format identification features into EPrints could provide useful preservation information. However, a real assessment of the value of file format data requires using such data in a live repository. As mentioned in the previous section, integrating file format functionality into an IR is only achievable in the longer term, depending on how quickly a repository wants, and is able, to integrate new features into its live service. In order to get some tangible results in the time required by the Preserv Project, we adapted the DROID tool to be usable with an IR without requiring any modification to that IR.

During the lifetime of the project we have steadily moved away from a model of building functionality into the IR towards a more modular, service-orientated model, where external services can be attached to the IR to provide digital preservation features. However, our first experiments with format profiling were focused more on answering the question of what makes a good preservation format. As mentioned above, two essential criteria for answering this question were identified. Firstly, what formats are being (and can be) produced by the IR's users. Secondly, which formats are available to users, whether of their own choosing or designated by the IR policy. By examining the formats being deposited in IRs, not only can we see the range of formats involved but we can also gain an impression of the complexity involved in managing the differing file formats. Moreover we can begin to determine which formats (by virtue of their popularity) are likely to be most easily accessible to users.

Almost all IRs now provide OAI-PMH support, with at least Dublin Core metadata being exposed. Using our existing expertise with developing OAI-based harvesting tools (Brody, 2003) we coupled a simple OAI harvester with the DROID tool to create an OAI-compliant format profiling tool. This tool retrieved every Dublin Core metadata record from the IR, extracted the URLs of files, downloaded and ran DROID on each of those files. The data was collected as a simple table consisting of the eprint identifier (the OAI record), the URL of the file, and the format as identified by DROID.

Simple Dublin Core – as used by IRs – does not explicitly provide the location of all the files that may be attached to that record. Indeed, even if an IR does provide a link to the 'full text' in its Dublin Core output, the mapping between the Dublin Core fields and the location of the files may not be obvious (e.g. historically EPrints has used the *format* Dublin Core field). Initially the mapping used in EPrints was used because EPrints is used by the Preserv partner *e-Prints Soton.* As we used EPrints it was then trivial to apply the tool to two other EPrints-based repositories: the School of Electronics and Computer Science (ECS) at Southampton and the White Rose consortium (Universities of Leeds, Sheffield and York).

In our first tests with DROID we found a number of formats with which DROID encountered difficulty. In particular it could not identify Postcript-format files, of which there are a number in ECS (see Figure 2). There has also been an ongoing development to the Microsoft Office format signatures, as initially (almost) all Microsoft Excel format files were identified as *OLE Compound Document*. This is partly to do with the fact that data from one Office application can be embedded in another (e.g. inserting an Excel chart into a Word document). As problems were identified they were fed back to the TNA which then improved the signatures based on the mis-identified files.

Figure 2 shows the format profile generated for three EPrints-based IRs. We were surprised to find out how homogenous the collection of material was – with virtually all files being in Adobe PDF. In particular it was surprising that the White Rose repository *only* had PDFs and, after querying the repository manager, found that they had a policy of only accepting PDF-based deposits (despite PDF being a *presentation* format, hence potentially losing a lot of useful information from the original document).

That the majority of material is in PDF (as a result of repository policy or community practice) leads us to think that on the one hand digital preservation may be easier, due to not having to monitor many formats, but also that – because PDF often does not preserve the original digital content – the scope for preserving the original authors' work is limited to the presentation of that work only. It should be kept in mind however that this is early days for repositories, so the sample shown here may only be a partial view of what a complete repository may look like.

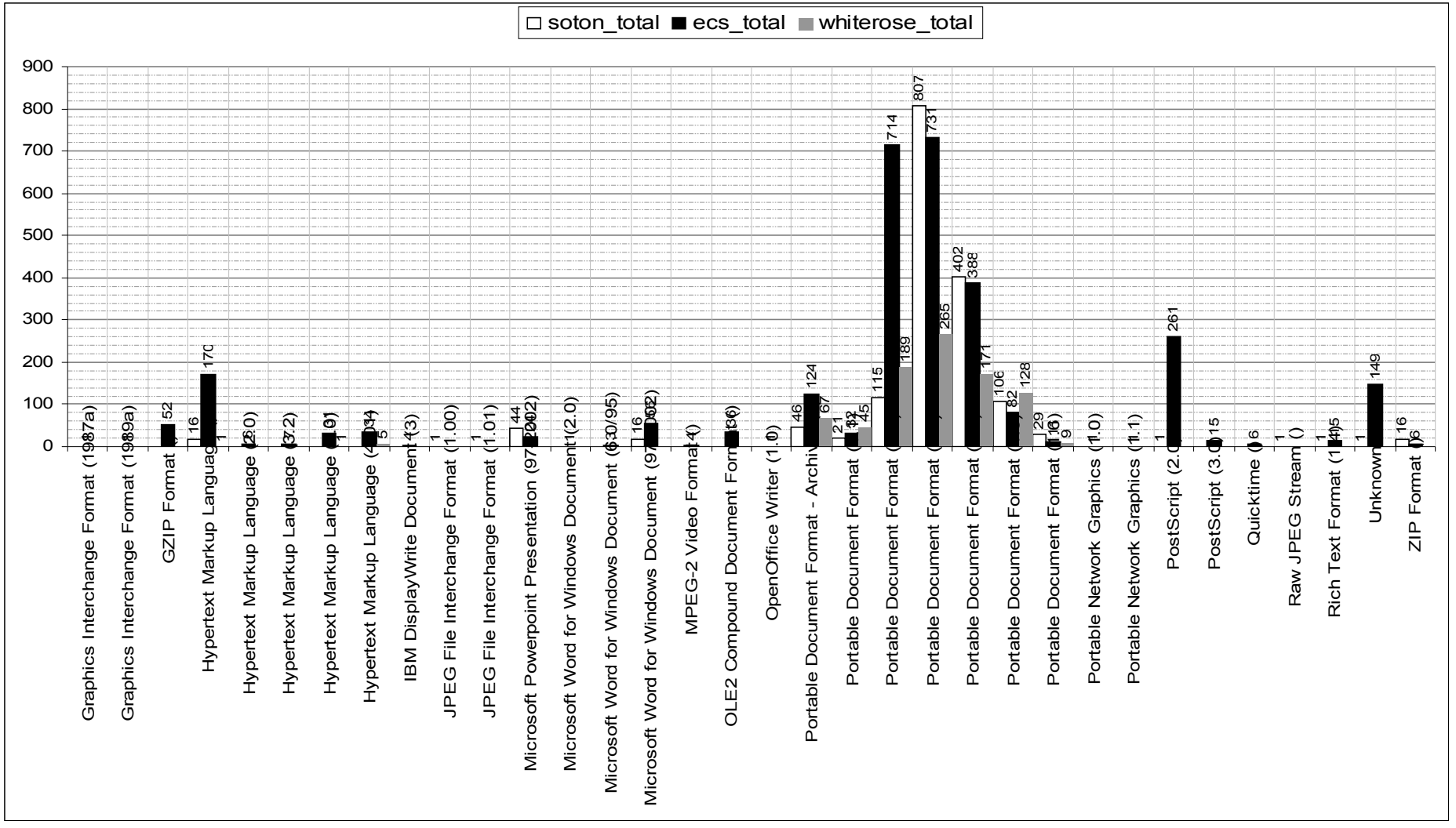


Figure 2. File Format Profile for University of Southampton Repository, ECS (Southampton) and White Rose. June 2006

The Registry of Open Access Repositories

The Registry of Open Access Repositories (ROAR), developed at the University of Southampton, is a listing of open access repositories of research material. ROAR was developed to manage better the existing listing of *EPrints*-based repositories; but it has since been greatly extended to cover all IR softwares as well as providing analytical tools to investigate the rate of growth of IRs and their content.

Entries to ROAR are either submitted by repository managers (registering their own repository) or bulk-imported from other listings. Each entry contains a number of metadata fields, including the OAI-PMH interface location and repository software used. The OAI-PMH interface is used to track the size of IRs, by harvesting every record and displaying the results as a graph of records-over-time. The total records in each repository can then be used to rank or filter IRs according to their size.

The metadata exported by a repository allows the location of digital objects to be found, for example by containing a URL from which the object can be downloaded. To be OAI-compliant, a repository needs at least to support *Dublin Core* format metadata; however OAI-PMH allows any XML-based metadata to be exported through the use of parallel metadata records (one or more records that, together with a unique identifier, represent a digital object).

Extending Profiling to Many Installations

Effective preservation services will ultimately be customised to individual repositories, but simpler, initial services such as format identification can be provided to many repositories without much difficulty. To achieve this with PRONOM-DROID a database of repositories and a means to interact with those repositories is required. This is provided by the Registry of Open Access Repositories (ROAR)[4].

Using ROAR and PRONOM-DROID, we have built a service that examines the content of repositories and has produced file format profiles (*Preserv profiles*) for 200+ repositories. The goal of this service is to provide repository managers with a summary of their content and, when the PRONOM database supports it, to provide a 'technology watch' service that can warn them of file formats that are at risk of becoming obsolete and hence inaccessible.

Preserv profiles have been integrated into the ROAR service - PRONOM-ROAR. With this demonstration preservation service Preserv has begun to redefine the role and nature of preservation for repositories: by enabling digital preservation through flexible, external services, IRs (and IR managers) can focus on collecting and serving their users' needs. The IR can 'buy in' preservation expertise through the use of external services, depending on local requirements.

The Preserv profile works by harvesting metadata from ROAR-registered IRs using the OAI-PMH. The metadata is harvested using a tool called *Celestial,* which downloads all the metadata contained in the repository and stores it in a local cache. When a repository is periodically harvested, if it uses EPrints or DSpace repository softwares (the most commonly used softwares), an attempt is made to retrieve the full

[4] Registry of Open Access Repositories (ROAR): http://roar.eprints.org/

text, identify its format using PRONOM-DROID, and store that additional information. For EPrints-based repositories this involves extracting URLs from the Dublin Core metadata (ignoring the URL of the abstract 'jump-off' page). DSpace repositories do not typically include the URL of the full text in its Dublin Core metadata, so instead the abstract jump-off page is downloaded, and the linked full-text URLs located within that. Regardless of the repository software, if the repository supports the Metadata Encoding and Transmission Standard metadata format (METS) the full-text URLs are located and used, because METS provides explicit support for complex objects (i.e. METS does not suffer from the abstract jump-off page/full-text URLs ambiguity). 33 OAI providers support METS according to the OAI Registry at University of Illinois at Urbana-Champaign (UIUC) (November 2007). This is a small minority of the live repositories (1586 in UIUC), but may grow as METS matures and support widens.

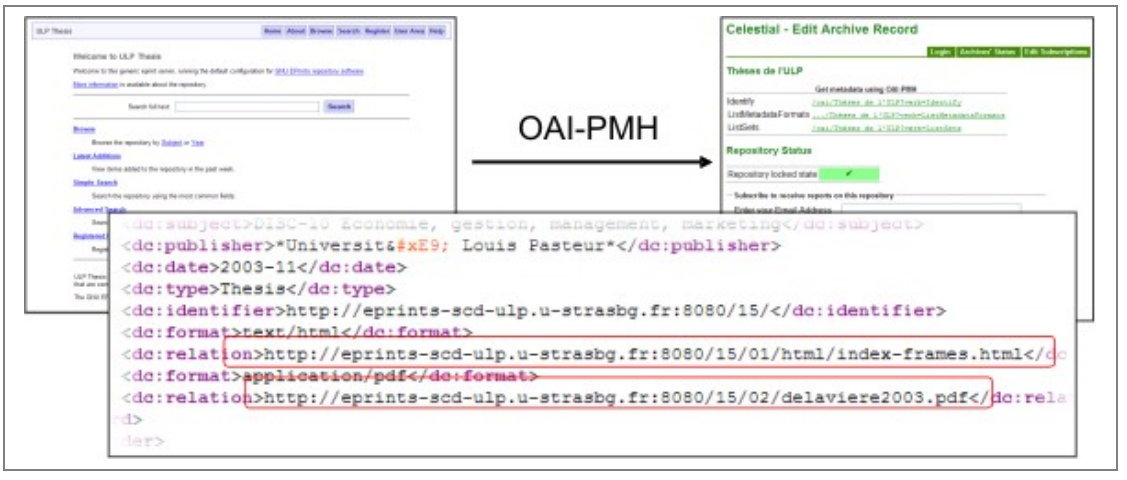


Figure 3. Harvesting Dublin Core metadata from a repository allows the URLs of full texts to be located, harvested, run through DROID and then become part of the Preserv profile.

If a full-text link is located in a record, the file is retrieved and stored temporarily (files larger than 2MB are truncated, so only the first 2MB of the file gets tested). The PRONOM-DROID tool is run on the downloaded file, which uses heuristics to determine the file format and, for some formats, the file format version (e.g. Adobe PDF 1.5). Some files may match more than one format heuristic, in which case only the first format is registered. Other files may not be identifiable at all, in which case they are flagged as 'Unknown Format'. In addition to the file format - as identified by DROID - the MIME-type and last modification time as returned by the Web server are recorded.

The file format data are stored in Celestial's database, linked to the metadata record. To generate the *Preserv profile* ROAR accesses Celestial's database directly to retrieve the file format data and present it to the end user. In addition to the Web interface described in the next section, an email alerting service is provided as part of Celestial's harvesting tool - because currently email alerts are generated by an OAI interface, and not from repository records in ROAR (which may have more than one OAI interface).

PRONOM-ROAR: The Preserv Profile

The Preserv profile tool is integrated into ROAR's Web interface. Accessing the profile for a known repository requires locating the repository in the ROAR and, if a profile is available, clicking the **Preserv Profile** link in the record for the repository. (If a profile is not available no link is provided). Alternatively a profile can be generated for an aggregation of repositories by selecting the **View repository contents as File Format Graph** link in the result set bar. Figure 3 shows the result set (top) and individual repository (bottom) links highlighted in the ROAR interface.

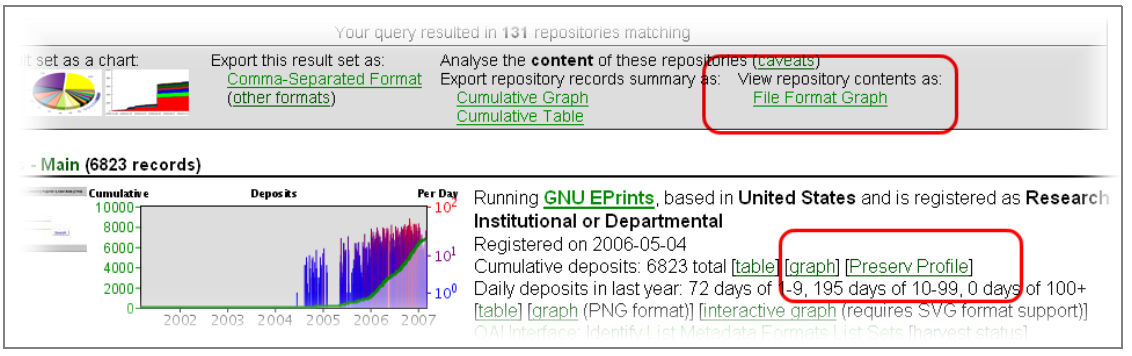


Figure 3 Preserv profiling tool integrated into the ROAR interface.

Following the **Preserv Profile** link displays the profile page, which consists of some brief notes describing what the profile is, a link to the alerting service, the OAI-PMH URL that the profile is generated from, and a histogram of the number of files identified per (PRONOM-defined) file format.

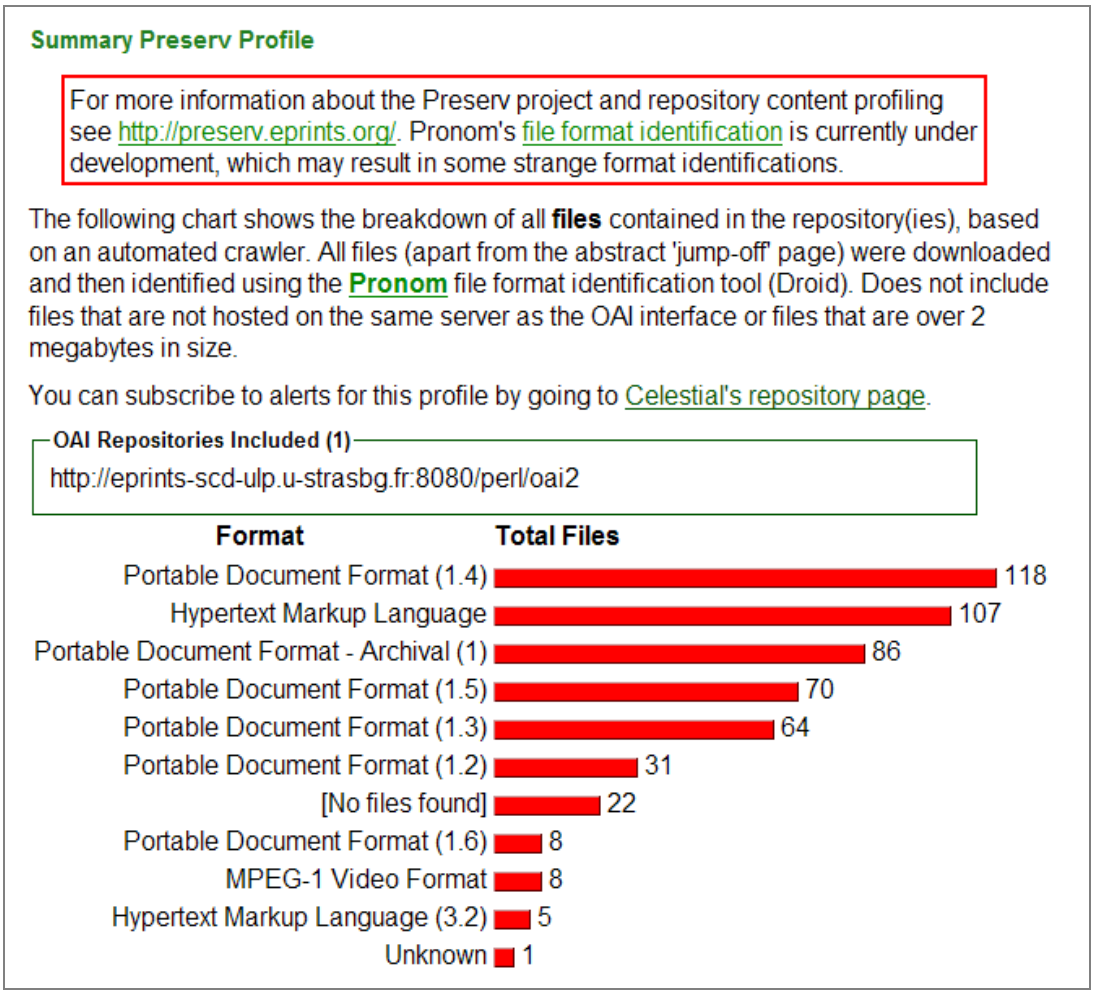


Figure 4. Example Preserv Profile. Each horizontal bar represents a different file format, with the total number of files found shown at the end.

The profile is based on OAI-PMH records harvested from the repository's registered OAI interface, rather than the Web view of the repository. Thus the URLs shown point to the OAI interface and not the URL of the repository's Web interface. Multiple repositories can be aggregated to generate a single profile, in which case the URLs for all aggregated repositories are shown. In the example illustrated, the repository is represented by a single OAI interface (Figure 6).

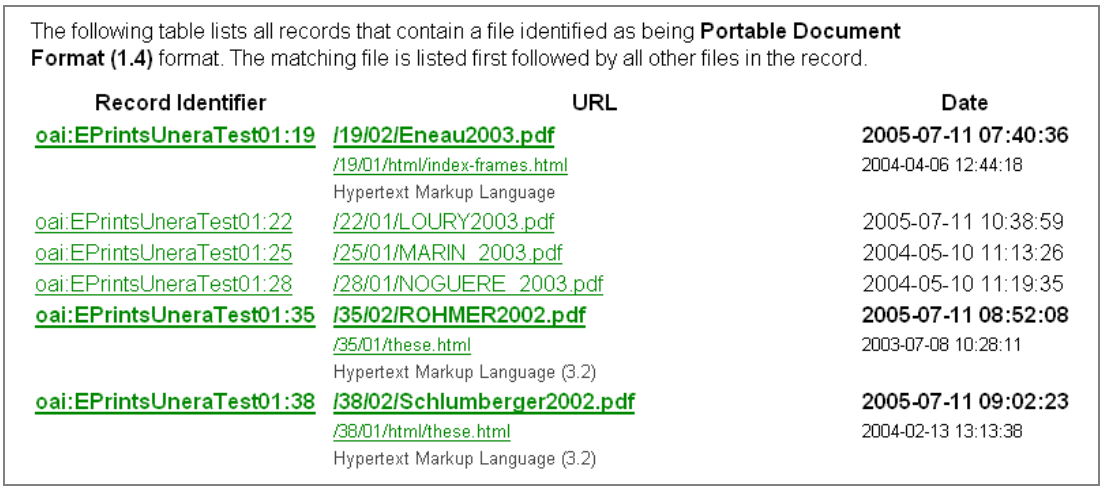
Figure 5. The OAI interfaces from which the data have been harvested are shown at the top of the profile page.

The file format histogram gives an instant overview of the file formats contained in the repository, rank-ordered by the total number of files found for each format. Each record in a repository may contain multiple files in multiple formats, in which case that record will appear multiple times in the histogram (so the number of files shown does not equal the number of records). If a record contained no files it is added to the **[No**

files found] bar. While we are primarily interested in the *files* contained in the repository it is useful to know how many records do not contain any files at all. These records are potentially 'lost' material – without the files there is no way for the repository to preserve access to the authors' work.

The distinction between *records* and *files* is due to repositories using different abstractions of digital objects. The concept of a *record* potentially represents very different things. Typically when a new item is deposited in a repository a (metadata) record is created that describes the item and links to the deposited content files. The record might link to many content files: an HTML file might contain images, or there may be a number of versions of a file. Or, the record may contain multiple formats (or versions) of the same thing.

For preservation purposes we are primarily interested in the formats used for deposited files rather than the more abstract concept of the record format. It is important to show all the files contained in the repository, and not just the formats at the record level.



The following table lists all records that contain a file identified as being **Portable Document Format (1.4)** format. The matching file is listed first followed by all other files in the record.

Record Identifier	URL	Date
oai:EPrintsUneraTest01:19	**/19/02/Eneau2003.pdf**	**2005-07-11 07:40:36**
	/19/01/html/index-frames.html	2004-04-06 12:44:18
	Hypertext Markup Language	
oai:EPrintsUneraTest01:22	/22/01/LOURY2003.pdf	2005-07-11 10:38:59
oai:EPrintsUneraTest01:25	/25/01/MARIN_2003.pdf	2004-05-10 11:13:26
oai:EPrintsUneraTest01:28	/28/01/NOGUERE_2003.pdf	2004-05-10 11:19:35
oai:EPrintsUneraTest01:35	**/35/02/ROHMER2002.pdf**	**2005-07-11 08:52:08**
	/35/01/these.html	2003-07-08 10:28:11
	Hypertext Markup Language (3.2)	
oai:EPrintsUneraTest01:38	**/38/02/Schlumberger2002.pdf**	**2005-07-11 09:02:23**
	/38/01/html/these.html	2004-02-13 13:13:38
	Hypertext Markup Language (3.2)	

Figure 6. Clicking a bar on the Preserv Profile generates a complete listing of all records containing files in that format and all files in those records. The OAI identifier of the record is given on the left; all files contained in the records are shown in the centre; and on the right appears the date the file was last modified, according to the repository Web server. When a record contains more than one file, the file in the current format is emboldened.

Each histogram bar is a clickable and provides a breakdown of all the files identified in that format (Figure 6). This enables users to delve into the content of the repository; for example, if a problem format has been identified they can obtain a list of all records that contain those files.

oai:EPrintsUneraTest01:19	**/19/02/Eneau2003.pdf**	**2005-07-11 07:40:36**
	/19/01/html/index-frames.html	2004-04-06 12:44:18
	Hypertext Markup Language	

Figure 7. A single record containing two files - in Adobe PDF (Eneau2003.pdf) and HTML (index-frames.html) formats respectively.

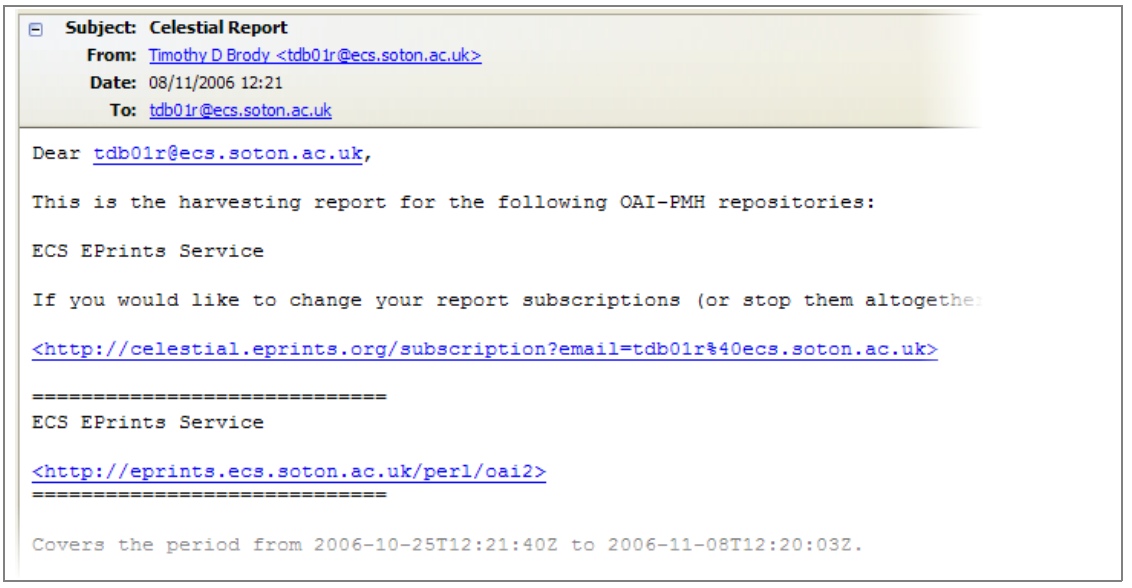
If a record contains more than one file each file is shown on the breakdown page (e.g. see Figure 7). The format selected from the histogram is emboldened to distinguish it from files in other formats from the same OAI record. In the example the record contains two files, one in Adobe PDF 1.4 format and the other in HTML. A possible use of the breakdown is to list file formats identified as a preservation 'risk' to check what other parallel formats may be available. Currently this is difficult to do automatically, because the relationship between multiple files contained in a single record is not captured (e.g. HTML and dependent inline images or subsequent versions).

Email Alerts to Technology Watch

In Figure 4 Brown (2005b) outlines the components of a technology watch service. The function of a technology watch service is two-fold: firstly, to act as an advisor to the archive's ingest process (to notice original material that may be troublesome); and secondly, to watch for existing content that may be at risk. Having identified the content that is at risk, a migration service can convert the object to an up-to-date format (or advise the archive administrator of the issue). This is similar to the AONS (Automated Obsolescence Notification System) Project (Curtis, 2006). In AONS a tool is integrated into the repository that scans the repository's content and, combined with a database providing obsolescence information, provides repository administrators with an email notifying them of problems. This differs to the approach taken by PRONOM-ROAR, which is to use existing public interfaces in the repository (OAI-PMH) to provide the same functionality, but without requiring repository software-specific modifications.

PRONOM-ROAR has taken the first tentative steps towards a technology watch service for open access/institutional repositories by generating email alerts to make repository managers aware of what material is entering their repository. The file format identifications could in future be linked to a technology watch database (i.e. something that can flag up at-risk formats). Consequently warnings of objects at risk can be sent to the repository manager.

The alerting service linked from the profile page sends periodic emails, set by the users, that notify them of new records and files in the repository to which they are subscribed. The first part of the email shows the date period covered by the report, and may show any errors that have occurred during the OAI-PMH harvest (Figure 8). The main part of the email shows the number of new OAI records and files (by file format) that have been identified in the repository during the period covered (Figure 9). In a technology watch service those file formats at risk could be flagged, with a link to the technology watch database explaining why the format may need to be acted upon.



Subject: Celestial Report
From: Timothy D Brody <tdb01r@ecs.soton.ac.uk>
Date: 08/11/2006 12:21
To: tdb01r@ecs.soton.ac.uk

Dear tdb01r@ecs.soton.ac.uk,

This is the harvesting report for the following OAI-PMH repositories:

ECS EPrints Service

If you would like to change your report subscriptions (or stop them altogethe

<http://celestial.eprints.org/subscription?email=tdb01r%40ecs.soton.ac.uk>

==========
ECS EPrints Service

<http://eprints.ecs.soton.ac.uk/perl/oai2>
==========

Covers the period from 2006-10-25T12:21:40Z to 2006-11-08T12:20:03Z.

Figure 8. An email alert sent from Celestial on ECS EPrints Service. The Celestial alert service was developed as part of PRONOM-ROAR but comes from the harvesting tool (Celestial) and not ROAR (which provides the Web interface).

```
New Metadata Records

       20      oai_dc
       20      uketd_dc

New Fulltexts

       1       Fixed Width Values Text File

       2       Hypertext Markup Language (4.01)
       1       IBM DisplayWrite Document (2)
       1       IBM DisplayWrite Document (3)
       1       MS-DOS Text File
       1       MS-DOS Text File with line breaks
       1       Macintosh Text File
       1       OLE2 Compound Document Format
       1       Plain Text File
       3       Portable Document Format (1.3)
       5       Portable Document Format (1.4)
       1       Portable Document Format - Archival (1)
       1       Rich Text Format (1.5)
       1       Rich Text Format (1.6)
       1       Tab-Delimited Text File
       1       Unicode Text File
```

Figure 9. File format summary enclosed in an email alert.

Making Agreements between IRs and Services

It is anticipated that some kind of technical agreement will be needed between repositories and preservation services. At the moment the PRONOM-ROAR service does not harvest more than 2MB of any one file. This is to avoid overloading repositories and wasting bandwidth in an experimental tool. Given repositories are

likely to contain both very large collections and potentially large individual objects (e.g. live performance videos and multimedia), an agreement between the service and repository would allow the service to access the contents of the repository without fear of compromising the repository's performance.

An institutional repository may contain objects that are not publicly accessible either because of copyright or other legal restrictions or because users do not want their material to be made publicly accessible. Regardless of whether an object is publicly accessible or not, a repository will still want to apply digital preservation techniques to that material. The PRONOM-ROAR service is predicated on public access, however it would be (fairly) trivial for a repository to provide a 'back door' to PRONOM-ROAR, based on remote host restrictions. A more secure solution would be to develop either a certificate- or password-based access to the repository's content, e.g. through secured Web services.

Conclusion

PRONOM-ROAR is the most visible output from the Preserv Project, being part of an active resource within the IR/open access community. With PRONOM-ROAR we intended to demonstrate that digital preservation can be achieved through light-weight web-based services. While, in Preserv, we started with an expectation of building preservation functionality into the IR, we no longer see this as an optimum solution. In a second phase, Preserv 2, is investigating a series of such services that build on file format characterisation, including technology watch, preservation planning and preservation actions.

There are a number of drawbacks in basing the preservation of material solely within the IR. Preservation features added to IR software would need to be duplicated across many installations, with the subsequent risk of conflicts with local customisations and errors in implementation. As digital preservation is still in development basing a preservation strategy on IR software will result in a slow process of preservation development, due to the long lead times involved in modifying stable, live repository installations. More fundamentally, we do not consider the IR as the best place to go to for preservation expertise.

Digital preservation awareness is growing, but is of most interest to the major archivists (the British Library, the Library of Congress, et al). Rather than duplicating their effort at the IR level, we see them informing, or even providing themselves, preservation services that can be used by IR managers. The latter can then pick and choose which preservation features their repository may need, e.g. mirroring, format migration, emulation services, etc.

The envisaged role of the IR manager in preservation is therefore to *manage* the preservation of their collection. The IR manager can develop a strategy based on the profile of their contents and in coordination with their communities. That strategy – at the IR level - could range from simply making backups (a 'mirror' service) to a commitment to ensuring long-term access to content through migration and emulation services.

The concept of the community is very important for digital preservation – it is the needs and requirements of a community that determine whether users require long-term access and the nature of that access. Some communities may not care if material is no longer accessible while others may require access to the material exactly as it was created in perpetuity. A repository based within an institution is unlikely to be able to cater to diverse sets of requirements, whereas a service-based approach could leverage centralised, community-specific tools to enable digital preservation at the repository. Alternatively the repository may just be a source of material for centralised, community-specific data archives.

Regardless of which services IR managers may use to implement their preservation strategy, it is essential they have the legal right to preserve that material. That involves not only the right to store a copy, but also to migrate that copy to other file formats and to be able to provide access to external services. That requires not only the agreement of depositing authors but also their provision of a version of the document that is unencumbered by access controls (for instance the PDF read-only attribute which, if set, prevents migration to other formats).

By separating preservation from the IR, the IR can focus on providing efficient object collection and metadata capture. By interrogating the IR external services can provide preservation expertise and later use the IR's dissemination and ingest functions to modify objects within the system. IR software is already moving towards this: for instance, Fedora provides for the ingest and dissemination of complex objects. Two potential preservation services for IRs are file-format identification, which this paper has demonstrated, and migration. Migration could be achieved through the IR dissemination and ingest interfaces: that is, the service harvests the object, migrates it, and adds it into the repository with suitable linking between the original and replacement objects. In the preservation services model envisaged, these two services and others can be developed in a repository-agnostic fashion, thereby maximising flexibility through a 'pick-and-choose' selection of services.

Acknowledgements

The Preserv Project was funded by JISC, within the programme Supporting Digital Preservation and Asset Management in Institutions[5]. The authors would like to thank the reviewers for their detailed comments and suggestions.

References

Brody, T. (2003). Citebase Search: Autonomous Citation Database for e-Print Archives. Unpublished. Retrieved December 7, 2007, from http://eprints.ecs.soton.ac.uk/10677/

Brown, A. (2005a). Automatic Format Identification Using PRONOM and DROID. *RLG DigiNews,* Volume 9, Number 2. Retrieved March 1, 2006, from http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/automatic_format_identification.pdf

[5] http://www.jisc.ac.uk/index.cfm?name=programme_404

Brown, A. (2005b). Automating Preservation: New Developments in the PRONOM Service, *RLG DigiNews*, Volume 9, Number 2, April 15. Retrieved December 7, 2007, from http://www.rlg.org/en/page.php?Page_ID=20571#article1

Curtis, J. (2006). AONS System Documentation, Australian Partnership for Sustainable Repositories, The Australian National University, Revision 169 2006-09-29, September 2006. Retrieved December 7, 2007, from http://www.apsr.edu.au/publications/aons_report.pdf

Darlington, J. (2003). PRONOM—A Practical Online Compendium of File Formats. *RLG DigiNews,* Volume 7, Number 5. Retrieved December 7, 2007, from http://www.rlg.org/preserv/diginews/v7_n5_feature2.html

Hitchcock, S., et al. (2007). Preservation Metadata for Institutional Repositories: applying PREMIS *Pre-print.*

PREMIS (PREservation Metadata: Implementation Strategies) Working Group. (2005). Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group (May 2005). Retrieved December 7, 2007, from http://www.oclc.org/research/projects/pmwg/