From Planning to Action: Enabling Preservation in the Digital Repository

David Tarrant

School of Electronics and Computer Science, University of Southampton,
Southampton, UK,
`davetaz@ecs.soton.ac.uk`

Abstract. An increasing amount of born digital materials has led to
the need for environments in which these materials can be managed and
preserved. A Digital Repository provides an example of such a system
which is designed for storing and managing digital objects. Over the
years the digital repository has become a widespread application used
to manage digital objects such as publication, photo and video content
and is increasingly being used for the storage of 'final' works. Digital
Preservation of these works is a very important step in recoding history
as well as enabling access to these valuable materials in the future. Dig-
ital Preservation involves more than the long term storage of a static
object. Moores law sums up the advances in digital technology which
we have seen and are still seeing today and thus digital preservation has
to take the affect that deprecated file formats and software have on the
readability of files within a digital repository. In order to manage these
more complex aspects of Digital Preservation we require a repository
which can utilise services which identify risks and solutions to digital
preservation problems. In this report we outline the proposed techniques
from Joint Information Systems Committee (JISC) Preserv2 project for
interaction between a repository and digital preservation planning ser-
vices. We look at how existing services such as Pronom, DROID and
JHOVE can be utilised effectively alongside the repository architecture
and enable the repository manager to keep their repository future proof.

1 Introduction

Digital Preservation is wi Edinburgh Repository Fringedely misinterpreted to
mean "Long Term Storage" however when looked at in more detail we find that
identification of outdated file types is also a critical problem in digital preserva-
tion. This is also not a simple problem involving the identification based upon
the file extension as common formats such as Portable Document Format (PDF)
and Microsoft Word Document (DOC) have seen many revisions without exten-
sion change. Support for early versions of these formats is now being dropped in
favour of enabling greater capabilities and future proofing. Migration of digital
objects also requires careful consideration when thinking about the properties of
the file format which need to be preserved. In digital libraries the issues of file
format preservation is often simplified to that of submission of a static Portable

Document Format (PDF) file. In a lot of cases however this step has already lead to loss of content such as annotation or presenter notes when the original format was a digital presentation. Loss of annotation notes is just one such example of a files Significant Properties (1), the importance of which can vary based upon the repositories and the authors requirements and policy. From this brief outline it is clear to see that having a single policy or ideal for digital preservation is not going to be suitable for the number of possible use cases, thus providers should look to enable interaction with a broad range of services which can be utilised in ways which suit the individual repositories needs and requirements. In this report we first take a very brief look at a few of the tools which are available to enable preservation planning before outlining the work being done by preserv2 to better integrate and enable integration of these tools with existing repositories. This represents a small section of the work being carried out by the preserv2 project on a web based architecture for digital preservation.

2   Planning Services

The United Kingdoms National Archives (TNA) as part of the Preserv and Preserv2 projects has been developing and enhancing a tool which accurately identifies the format of digital files based upon the content of the file. This tool (DROID) operates by matching file signatures against the content of the file to identify not only the file type but also the version ((2)). The DROID tool operates alongside the PRONOM technical registry which provides an online registry of these file formats and is being expanded to contain information such as format risks ((3)). Beyond simplistic file format identification there are also tools such as JHOVE available which are able to identify a files significant properties for a limited number of file types ((4)). By putting these tools together we can start to build a profile of the files contained within a digital repository and identify any risks which may exist.

3   Unlocking the Repository

With an array of planning tools already available the challenge is to integrate these with the repository in an effective and easy to manage way. In preserv2 we are focusing on a web based architecture and making use of simple protocols such as HyperText Transfer Protocol (HTTP) to utilise services such as those aforementioned. In this way we are focusing on subscription and push services for management of services outside of the repository management environment (EPrints,Fedora etc) which these softwares can then subscribe to in order to receive updates from 3rd party services. Services such as DROID and JHOVE already export their results in the common XML format which was designed as a human and computer readable format for the web for metadata representation. Other services such as OAI-PMH are also able to export a full or customised set of results from a repository in many metadata formats which can thus be interpretted and fed into a tool such as droid for object classification. It is the

linking of these tools which we are considering in this report and which way is best to considering when linking many services together in this way.

De Roure ((5)) recently did an examination of web2.0 and the number of emerging Application Program Interfaces (APIs) available. A valuable question is raised regarding weather application should be allowed to define their own API or be forced to use standards other than HTTP/XML. Although the majority of web applications export XML with varying schemas the invocation of these applications is often performed in many different ways. De Roure looks at this as the problem of n connection between services Figure 1. If standards are developed then there is only the need for n connections between services and not and n^2 connections as show in Figure 1.
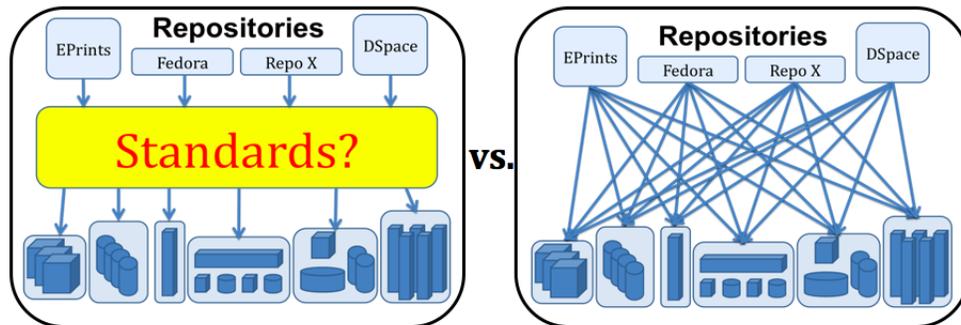

Fig. 1. Standard and the Web

From Figure 1 we can envisage a scenario which fits between the two examples given and here we start entering the realm of having middleware. It is important at this point to consider which version of the diagram the web is closest to and indeed moving towards. We have seen for a while on the web now that the community will tend to elect a new standard through this becoming the most popular but this is not without n^2 iterations coming first.

Within preserv2 we have decided to follow the web2.0 paradigm and construct a series of links between services by writing simplistic wrapper services for classification tools such as DROID which are able to translate and invoke the APIs of many other web based services such as OAI-PMH[1] and Darwin Calendar Server[2].

4 Integrating Droid

In this section we look at the proposed implementation of a basic preservation planning service with a digital library process. By harnessing services such as

---

[1] http://www.openarchives.org/pmh/
[2] http://trac.calendarserver.org/

Darwin Calendar Service for scheduling, results output and history we also provide a method of performing digital provenance, a method also important in digital preservation systems. The use of such a generic calendar services also allows repository administrators to keep track of characterisation events and schedule these from a third party calendar program such as Microsoft's Outlook or Apple's iCal. It is therefor the calendar service which knits the various services together and by defining a simplistic protocol for the contents of a calendar event we can subscribe both the repository and DROID to the calendar server and trigger events such as classifications or updates.

Figure 2 gives a diagram of the overall system implementation showing that all this can be enabled through the use of two wrapper libraries, each of which handles communication between the service it sits on top of and those third party APIs.
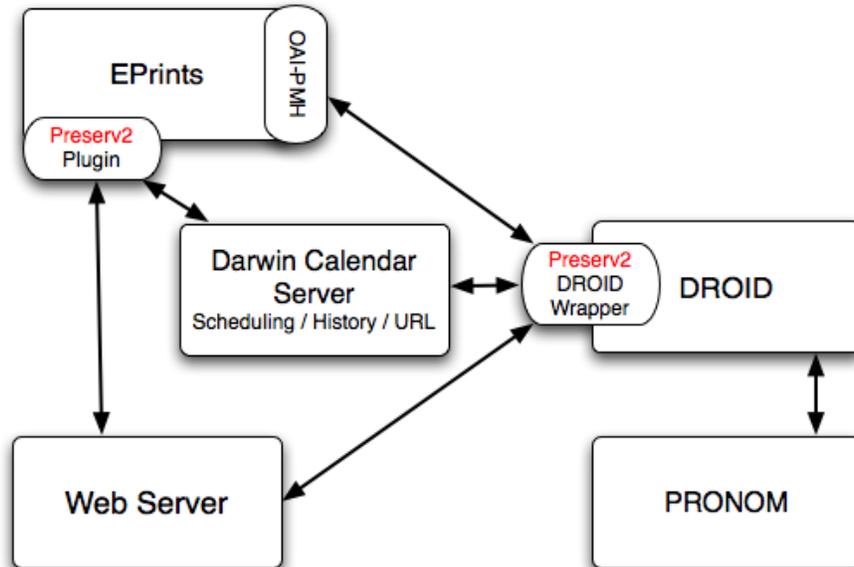

Fig. 2. Wrapping and Integrating DROID

4.1 Roles: DROID and The Preserv2 Wrapper

DROID acts purely as an object classification tool and has little or no idea about the rest of the system other than the direct communication with PRONOM which is built into each distribution of DROID. DROID takes an XML file with a listing of resources to classify of which the results are also given in XML format. The Preserv2 wrapper sits in front of DROID and handles all communication

with external services. Upon request either by a manual process call or a scheduled event on the calendar server the wrapper locates new objects to be classified through querying a configured OAI-PHM interface (note that this OAI-PMH interface doesn't necessarily have to exist on top of a repository). Any resources which require classification will then be formated into the DROID XML input format before DROID is then invoked. Upon completion these results are analysed and posted directly onto a web server and a record of the event and urls where results can be found are published on the Calendar Server.

4.2 Roles: Web Server

This part of the system simply holds output documents in XML format which relate to calendar events and invocations of DROID. The web server can either be open or require some sort of authentication to gain access to these documents which can be held for varying amounts of time.

4.3 Roles: Darwin Calendar Server

This is the most crucial part of the system, providing an authenticated means by which events can be scheduled and history recorded about events which have happened. There are two main types of event on the calendar server: a future scheduled event which causes a DROID classification to occur and a history event which gives details of an event which has occurred and urls which will contain infomation or results of the event. Due to the use of generic calendar server we can subscribe many desktop applications to the calendar server as well as the two preserv2 interfaces which act as a layer on top of other services.

4.4 Roles: EPrints and the Preserv2 Plugin

At this stage the Preserv2 plugin is performing very little work other than recognising that an event has occurred and can interface with the relevant services to update EPrints as required. EPrints acts as our repository as it is envisaged that the Preserv2 plugin simply takes the classification information provided by DROID and imports this into the EPrints metadata set. By separating out the bulk of the work into 3rd party services we have reduced the amount of work required to be carried out by the repository to the extent that it would not be hard to write a corresponding plugin for other softwares such as Fedora and DSpace.

5 Future Work - Acting on Repository Policy

With the future planned release of PRONOM also providing format risk scores we can envisage this being one of the most useful pieces of information which can be fed back into the repository. Based on local policy the repository could set thresholds for scores which constitute files in immediate danger and those which

formats which are safe for the foreseeable future. It is envisaged that using this data that we can demonstrate a simple traffic light type scale for files in different risk levels. Following on from this looking at the work which is being done within the European Planets project, a server can attempt to locate any migration tools which suit the local policy on the importance of significant properties for each format. A migration process could then be scheduled via the calendar server and the output later imported back into the repository alongside the current version.

Authors Note: The neatest thing about all of this is that by using an external calendar server to schedule and record history about events enables third parties to interact with resources directly rather than only via the repository. If the resources related to our repository are all stored on an open platform, 3rd parties such as preservation providers would be able to perform operations on these objects independently and post an event to the calendar server. The repository could then either take notice of this event and import any new data or simply ignore it. Thus change control remains with the repository whilst allowing other to manipulate the repositories resources freely and externally.

Bibliography

[1] Wilson, A.: Significant properties report. Proceedings of the Meeting on What to Preserve? Significant Properties of Digital Objects (2008)

[2] Brown, A.: Automatic Format Identification Using PRONOM and DROID. The National Archives, Digital Preservation Technical Paper **1** (2005) 17

[3] Adrian, B.: Automating Preservation: New Developments in the PRONOM Service. RLG DigiNews **9**(2) (2005)

[4] Donnelly, M.: JSTOR/Harvard Object Validation Environment (JHOVE). Digital Curation Centre Case Studies and Interviews (2006)

[5] De Roure, D.: How Repositories can Avoid Failing like the Grid. Proceedings of Repository Fringe Workshop, Edinburgh, UK (2008)