EPrints Preservation – File Formats and Risk Analysis

Table of Contents

1 Introduction

EPrints 3.2 is able to analyse the files associated with each publication record in order to keep track of risks pertaining to file types.

There are also admin advantages to the new filetype reporting capabilities.

For further background information on this topic:

- Presentation - EPrints and Preservation, In: Tackling the Preservation Challenge: Practical Steps for Repository Managers, 12th December 2008, London
- Publication - Where the Semantic Web and Web 2.0 meet format risk management: P2 registry, In: iPres2009: The Sixth International Conference on Preservation of Digital Objects

2 Tutorial - Aim

The aim of this tutorial is to give some practical experience with some of the features of the forthcoming EPrints 3.2 release. EPrints 3.2 allows the use of DROID at the back-end to classify files in the repository. This allows the assignment of risk analysis scores to the discovered file formats to aid in digital preservation decisions.

Note: At time of writing (September 2009), the National Archives (UK) PRONOM service was not yet providing risk scores and thus a demonstration service is used in part of this tutorial.

3 File Classification

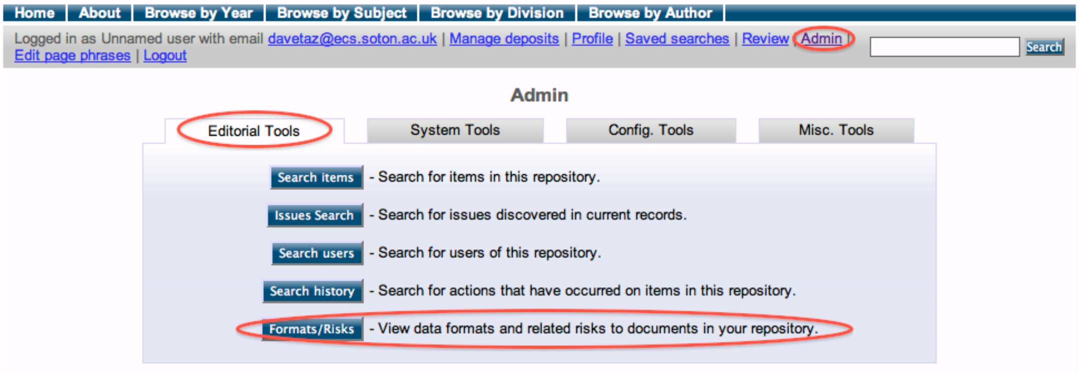
In this section we look at classifying the files in your repository using DROID from http://droid.sourceforge.net/ and the classification add-ons available from files.eprints.org. In the tutorial repositories both of these packages have already been installed.

To classify files in a live repository it is recommended that the process is run using a scheduled job, at most a couple of times a day. For the purposes of the tutorial a button has been provided in the admin interface which invokes it on demand.

The rest of this tutorial is split into exercises applicable to the tutorial and those applicable in all cases.

3.1 The Formats/Risks Screen

This screen will be our main reference point throughout the exercise. Available via the **Admin** interface (shown below) we can view the file types in our repository and any related risk scores.


Home | About | Browse by Year | Browse by Subject | Browse by Division | Browse by Author
Search
Admin
Editorial Tools
System Tools
Config. Tools
Misc. Tools

Logged in as Unnamed user with email davetaz@ecs.soton.ac.uk | Manage deposits | Profile | Saved searches | Review | Admin | Edit page phrases | Logout

Search items - Search for items in this repository.

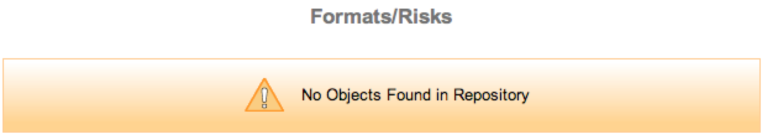
Issues Search - Search for issues discovered in current records.

Search users - Search for users of this repository.

Search history - Search for actions that have occurred on items in this repository.

Formats/Risks - View data formats and related risks to documents in your repository.

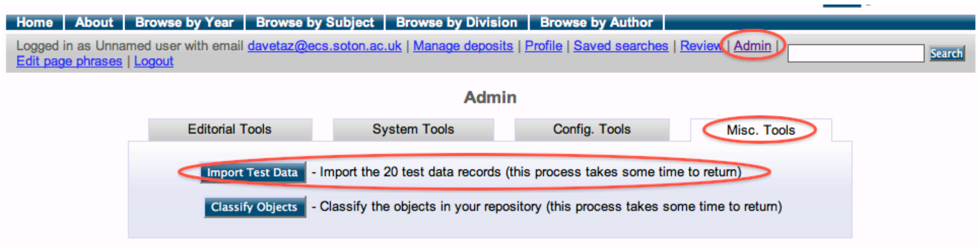
Viewing this page from an empty repository should result in the following screen.



No Objects Found in Repository

3.2 Exercise 1 - Populating the repository

For the purposes of this tutorial, we have provided a set of 20 records from the EPrints test dataset. To import these an **Import Test Data** button is available from the **Misc Tools** section of the **Admin** interface. This process takes some time to return, please be patient.

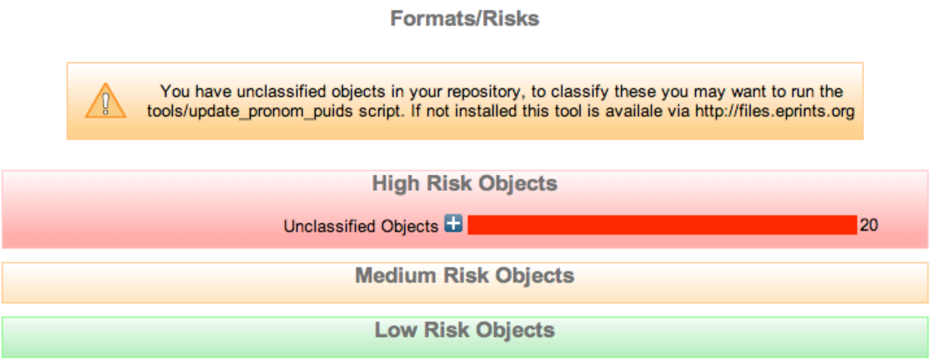

Home | About | Browse by Year | Browse by Subject | Browse by Division | Browse by Author
Search
Admin
Editorial Tools
System Tools
Config. Tools
Misc. Tools

Logged in as Unnamed user with email davetaz@ecs.soton.ac.uk | Manage deposits | Profile | Saved searches | Review | Admin | Edit page phrases | Logout

Import Test Data - Import the 20 test data records (this process takes some time to return)

Classify Objects - Classify the objects in your repository (this process takes some time to return)

After this process is finished, the **Format/Risks** Screen should show that there are 20 unclassified objects.

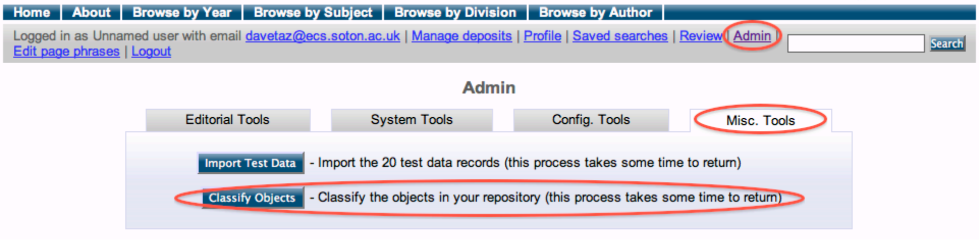


Formats/Risks

You have unclassified objects in your repository, to classify these you may want to run the tools/update_pronom_puids script. If not installed this tool is availale via http://files.eprints.org

3.3 Exercise 2 - Classifying the objects in your repository

The classification process can be performed through the **Classify Objects** button available via the **Misc Tools** tab in the **Admin** interface.



Home | About | Browse by Year | Browse by Subject | Browse by Division | Browse by Author
Search

Logged in as Unnamed user with email davetaz@ecs.soton.ac.uk | Manage deposits | Profile | Saved searches | Review | Admin | Edit page phrases | Logout

Admin

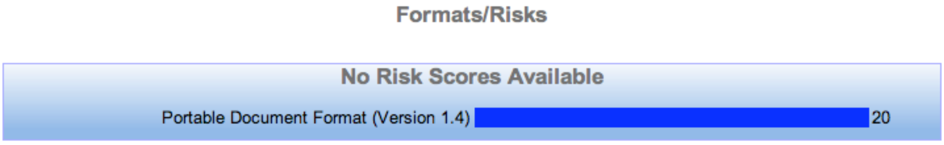
Editorial Tools
System Tools
Config. Tools
Misc. Tools

Import Test Data - Import the 20 test data records (this process takes some time to return)

Classify Objects - Classify the objects in your repository (this process takes some time to return)

In practice, classifying the objects would be run on the server as a background task, but we have provided a button for this tutorial. In a fully populated repository, this task could take considerable time.

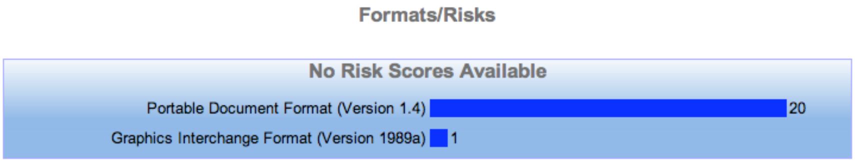
As a result of the above process our Format/Risks Screen should now be showing classified objects.



Formats/Risks

3.4 Exercise 3 - Adding an "at risk" example file

As part of the tutorial we have provided a set of files available at http://www.eprints.org/software/training/3.2/gif_collection/. Uploading one of these to the repository and then repeating the classification process at the end of Exercise 2 should lead us to to a repository profile much like the one shown below.



Formats/Risks

No Risk Scores Available
Portable Document Format (Version 1.4)
20
Graphics Interchange Format (Version 1989a)
1

4 Risk Analysis (Exercise 4)

To enable risk analysis we need edit the config file for PRONOM available via the **View Configuration** button in the **Config Tools** tab of the **Admin** interface. The pronom.pl config file is near the top in the first cfg.d section. Click on this file to view and edit it.

In this file, find the following line:

```
$c->{"pronom_unstable"} = 0;
```

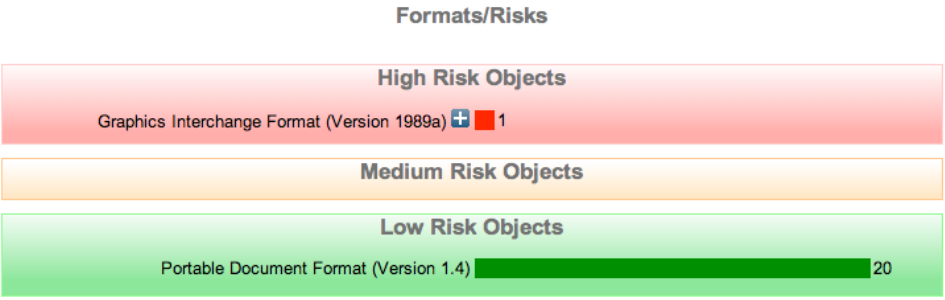
and change it to:

```
$c->{"pronom_unstable"} = 1;
```

Pronom Unstable is a development library that mimics the behavious to the National Archive's PRONOM web service for obtaining file format information.

Finally click on the **Reload Configuration** button in the **Config Tools** tab of the A**dmin** interface to get EPrints to load the changes you have made.

This should now lead to the **Format/Risks** screen displaying the following, showing that the previously added gif format is high risk. Note that this is test data, and no reflection on the riskiness of gif files.


Formats/Risks

High Risk Objects

Graphics Interchange Format (Version 1989a) 1

Medium Risk Objects

Low Risk Objects

Portable Document Format (Version 1.4) 20

The Planets Project part of this tutorial will go further to explain what can be done with potential high risk formats. Also the introductory presentation should have gone some way to help explain the importance of file formats in digital preservation.

4.1 Exercise 5 – Moving Risk Boundaries

The configuration file we edited in the last section can also be used to control the risk score boundaries between high, medium and low risk. The National Archives (UK) schema is to provide a score based on 8 classification categories between 0 and 3000. Thus for simplicity EPrints' default boundaries have been set at 0-1000 for high risk, 1001-2000 for medium risk and 2001-3000 for low risk. By moving these to be 0-100,101-200 and 201-3000 respectively you should be able to change the classification of the gif risk score (which is 183.24 in our test data). In the configuration file these boundaries are listed:

```
$c->{"high_risk_boundary"} = 1000;
```

and

```
$c->{"medium_risk_boundary"} = 2000;
```

Modify the boundaries so that gif are medium risk. Don't forget to reload the configuration. After verifying that this has worked, change it again so that gifs are low risk.