

# EPrints Preservation – File Formats and Risk Analysis

## Table of Contents

---

1 Introduction .....	1
2 Tutorial - Aim.....	1
3 File Classification .....	1
3.1 <i>The Formats/Risks Screen</i> .....	2
3.2 <i>Exercise 1 - Populating the repository</i> .....	2
3.3 <i>Exercise 2 - Classifying the objects in your repository</i> .....	3
3.4 <i>Exercise 3 - Adding an "at risk" example file</i> .....	3
4 Risk Analysis (Exercise 4).....	4
4.1 <i>Exercise 5 – Moving Risk Boundaries</i> .....	4
5 Gathering File Samples (Exercise 6).....	5

## 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 EPrints 3.2. 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 (March 2010), 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

---

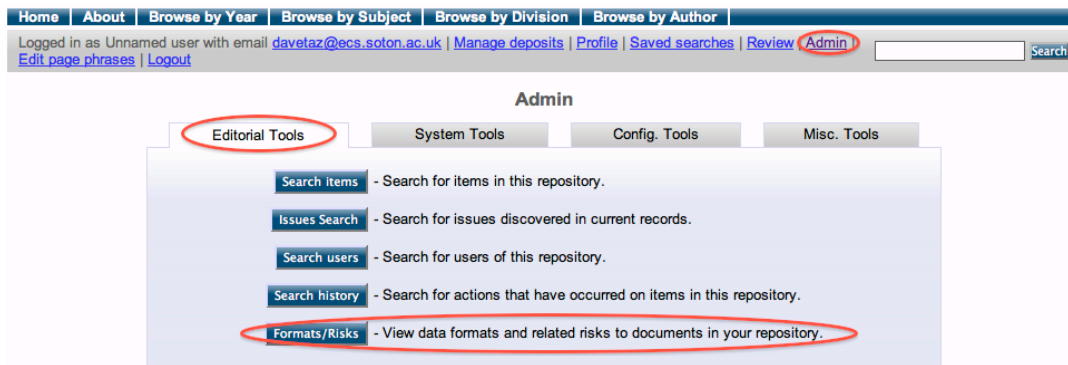
In this section we look at classifying the files in your repository using DROID from <http://droid.sourceforge.net/> and the classification add-ons available from [files.eprints.org](http://files.eprints.org). In the tutorial repositories both of these packages have already been installed.

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.

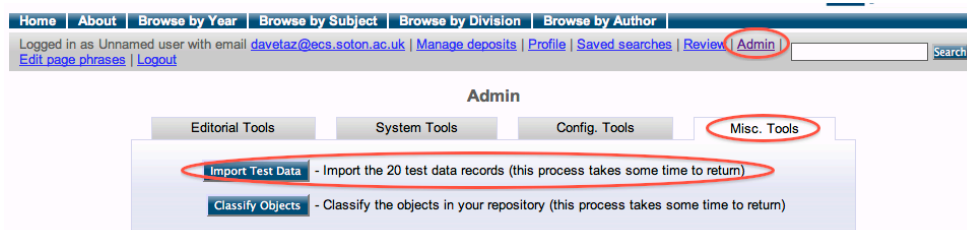


Viewing this page from an empty repository should result in the following screen.



### 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.



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 available 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.

The screenshot shows the Admin interface with the 'Misc. Tools' tab selected. The 'Classify Objects' button is highlighted with a red circle. The button text reads: 'Classify Objects - Classify the objects in your repository (this process takes some time to return)'. Other visible buttons include 'Import Test Data' and 'Editorial Tools'.

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



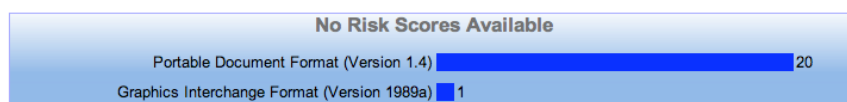
### 3.4 Exercise 3 - Adding an "at risk" example file

For this part of the tutorial we have provided a set of files available at [http://www.eprints.org/software/training/3.2/march\\_2010/](http://www.eprints.org/software/training/3.2/march_2010/). We recommend that you download the zip file containing these onto your computer so you can view and upload them into a new EPrint.

At this point you will need to create a new EPrint, adding each file one at a time from the zip. This will create you a collection of GIFs. Please don't upload the zip file as this won't help with later demonstrations.

Once you have created a new EPrint containing the GIF dataset, follow the steps in 3.3 for classification. The Format/Risks Screen should display something similar to the following:

Formats/Risks



## 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;
```

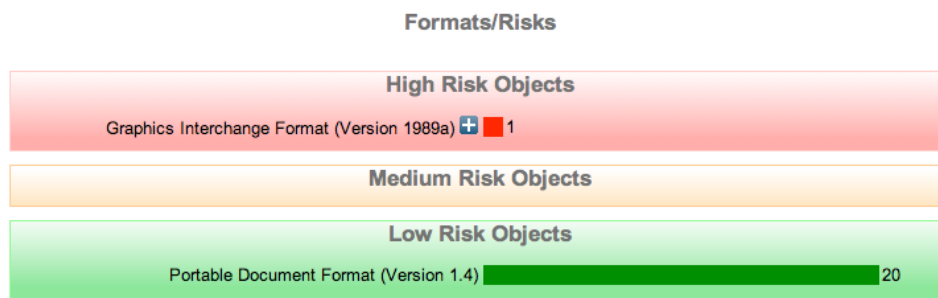
and change it to:

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

Pronom Unstable is a development library that mimics the behavior of 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 **Admin** interface to get EPrints to load the changes you have made.

After performing a re-classification (section 3.3), the **Format/Risks** screen should display 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.



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;
```

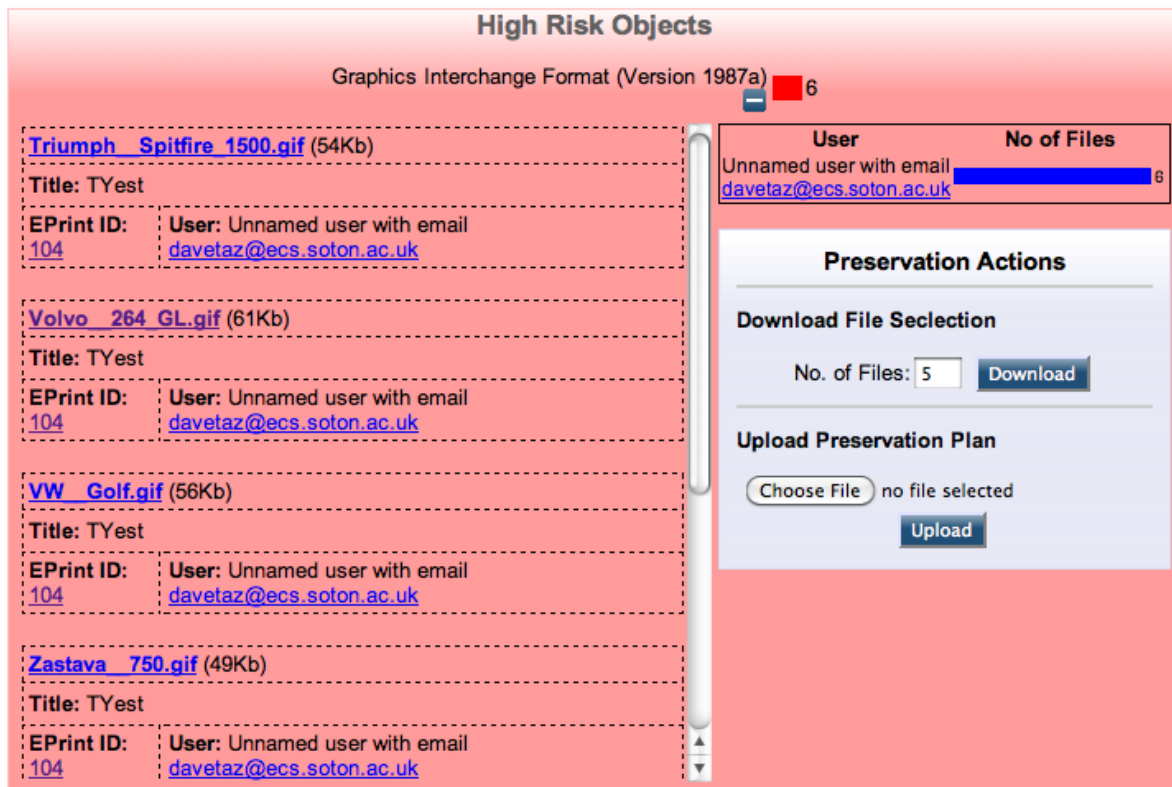
Modify the boundaries so that GIF are medium risk. Don't forget to reload the configuration and re-classify. After verifying that this has worked, change it so that GIFs are high risk again.

## 5 Gathering File Samples (Exercise 6)

While EPrints 3.2 can help identify risks related to the files contained within your repository, an external process is required to help build your preservation plan.

In order to help this process EPrints can provide a selection of files of a particular format.

By clicking the + button relating to the file type which is listed as at risk, we are presented with the following screen.



On the left hand side of this screen you can browse the files of this particular format, including information pertaining to who submitted each file and which EPrint they are part of.

Through the preservation actions panel we can request a selection of files from the repository and later, upload our completed preservation plan.

In order to provide the broadest selection of files EPrints will select them according to the following policy:

```

If more than 1 file requested:
  Provide Newest and Oldest
If more than 3 files requested:
  Also provide Largest and Smallest
Then
  Provide a random selection

```

To finish this exercise, input a number of files and press download.