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 file type 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
- Publication - Connecting preservation planning and Plato with digital repository interfaces. In: 7th International Conference on Preservation of Digital Objects (iPres2010).

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 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 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

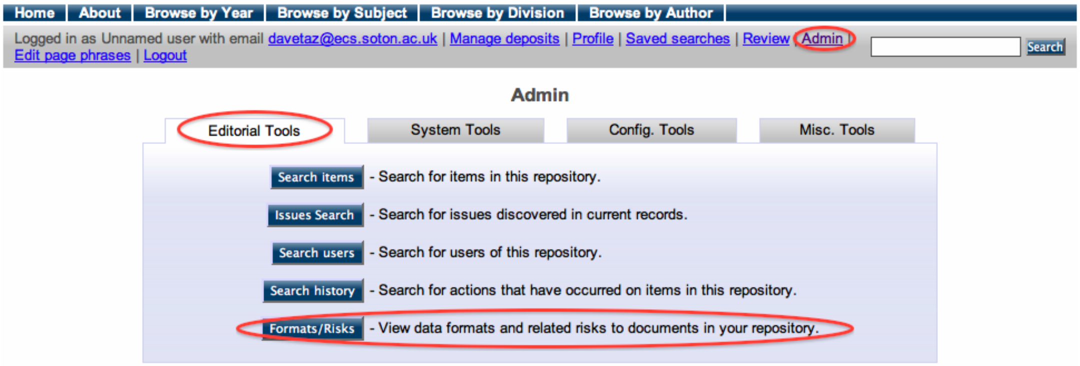
In this section we look at classifying the files in your repository using DROID from http://droid.sourceforge.net/ and the classification add-ons available from files.eprints.org. In the tutorial repositories both of these packages have already been installed.

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.



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

- Search for items in this repository.

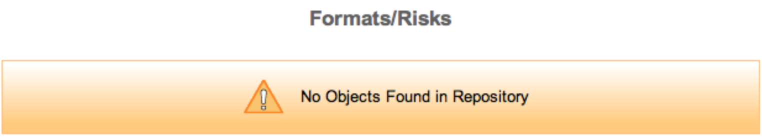
- Search for issues discovered in current records.

- Search for users of this repository.

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

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

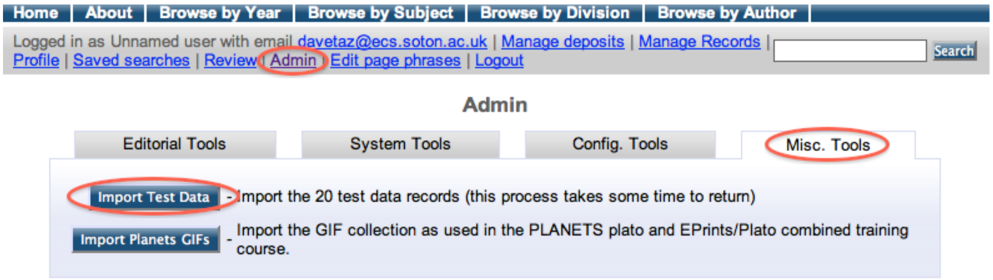
Viewing this page from an empty repository should result in the following screen.



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.

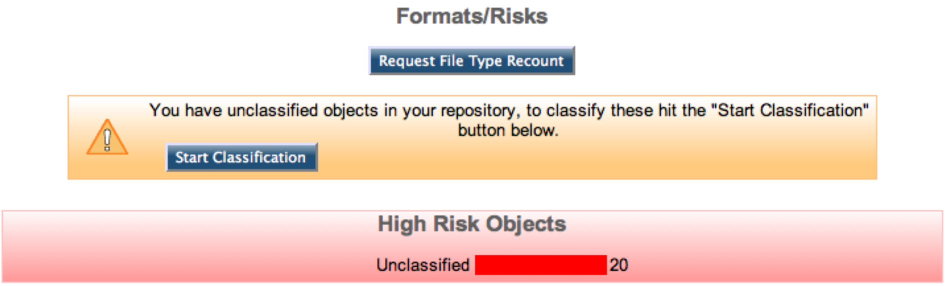


Logged in as Unnamed user with email davetaz@ecs.soton.ac.uk | Manage deposits | Manage Records | Profile | Saved searches | Review | Admin | Edit page phrases | Logout

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

- Import the GIF collection as used in the PLANETS plato and EPrints/Plato combined training course.

After this process is finished, this screen should list 20 new EPrints that have been imported. The **Format/Risks** screen should show that there are 20 unclassified objects.



Formats/Risks

Request File Type Recount

You have unclassified objects in your repository, to classify these hit the "Start Classification" button below.

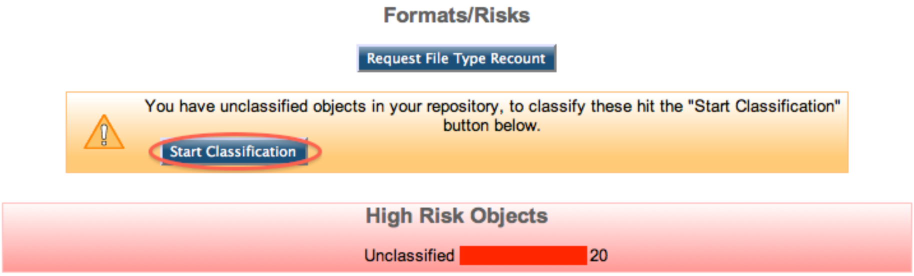
Start Classification

High Risk Objects

Unclassified 20

3.3 Exercise 2 - Classifying the objects in your repository

The classification process can be performed by simply pressing the **Start Classification** button on the **Formats/Risks** screen.



Formats/Risks

Request File Type Recount

You have unclassified objects in your repository, to classify these hit the "Start Classification" button below.

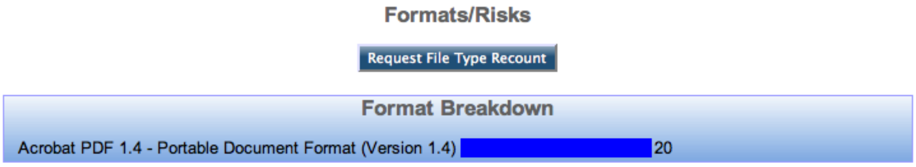
Start Classification

High Risk Objects

Unclassified 20

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. If you refresh the page you should see a ticker which tells you how far the classification is through the process.

Once done another refresh should show the result of the classification process:



Formats/Risks

Request File Type Recount

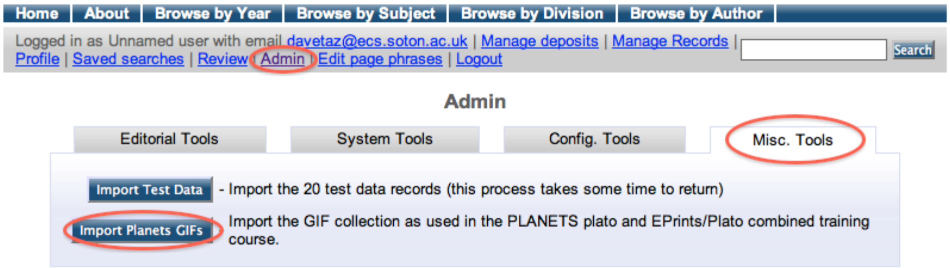
Format Breakdown

Acrobat PDF 1.4 - Portable Document Format (Version 1.4) 20

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

For this part of the tutorial we are going to use the dataset provided and used by the PLANETS Plato training. This dataset consists of a collection of 12 GIF images relating to cars.

Importing this collection can be done in the same way as the previous dataset via the **Misc Tools** tab in the **Admin** interface.



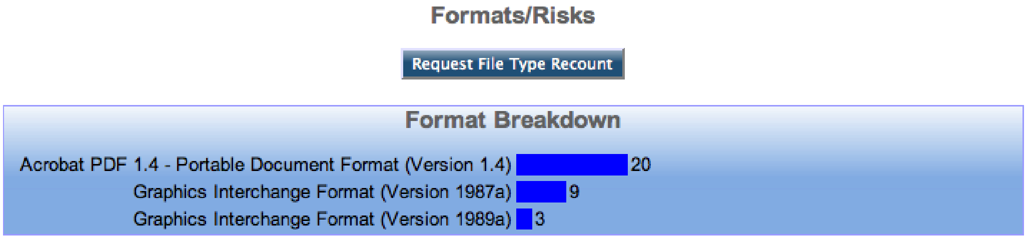
Logged in as Unnamed user with email davetaz@ecs.soton.ac.uk | Manage deposits | Manage Records | Profile | Saved searches | Review | Admin | Edit page phrases | Logout

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

Import the GIF collection as used in the PLANETS plato and EPrints/Plato combined training course.

Once finished this screen will provide a link to the EPrint containing the imported GIF files. It is worth looking at this now and getting familiar with how it looks at this stage, prior to later exercises involving possible migration.

After repeating the steps in 3.3 to update your file classifications the Format/Risks screen should display something similar to the following:



Format Breakdown

Acrobat PDF 1.4 - Portable Document Format (Version 1.4)	20
Graphics Interchange Format (Version 1987a)	9
Graphics Interchange Format (Version 1989a)	3

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 A**dmin** 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;
```

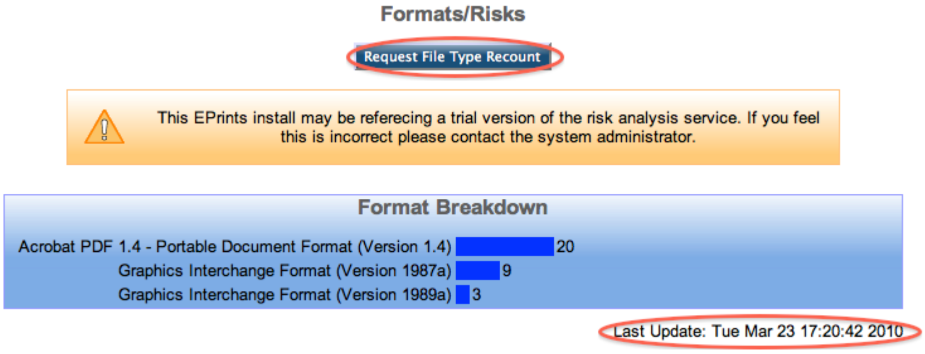
and change it to:

```
$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 A**dmin** interface to get EPrints to load the changes you have made.

In order to update the risks scores we simply need to request a re-process of the objects (not a complete re-classification). This can be done by clicking the **Request File Type Recount** button on the **Formats/Risks** screen. Note that this button queues the event to happen on the indexer and you may need to wait up to a minute for it to complete. If you keep refreshing you will notice the **Last Update** time change, indicating that something has happened.


Formats/Risks

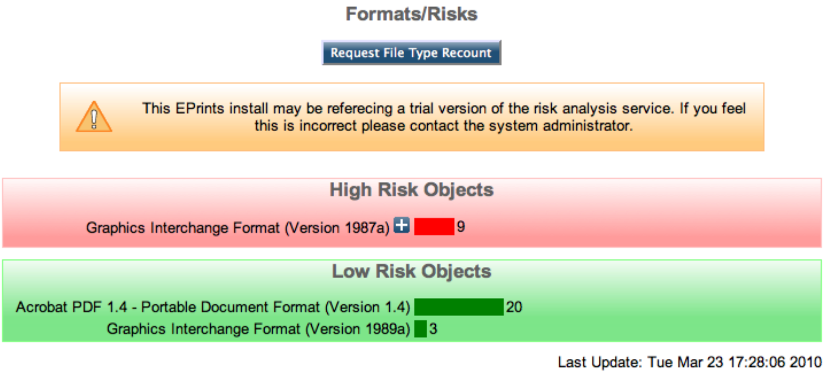
This EPrints install may be referecing a trial version of the risk analysis service. If you feel this is incorrect please contact the system administrator.

Format Breakdown

Acrobat PDF 1.4 - Portable Document Format (Version 1.4) 20
Graphics Interchange Format (Version 1987a) 9
Graphics Interchange Format (Version 1989a) 3

Last Update: Tue Mar 23 17:20:42 2010

After performing a re-count, 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.


Formats/Risks

This EPrints install may be referecing a trial version of the risk analysis service. If you feel this is incorrect please contact the system administrator.

High Risk Objects

Graphics Interchange Format (Version 1987a) 9

Low Risk Objects

Acrobat PDF 1.4 - Portable Document Format (Version 1.4) 20
Graphics Interchange Format (Version 1989a) 3

Last Update: Tue Mar 23 17:28:06 2010

The preservation planning 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;
```

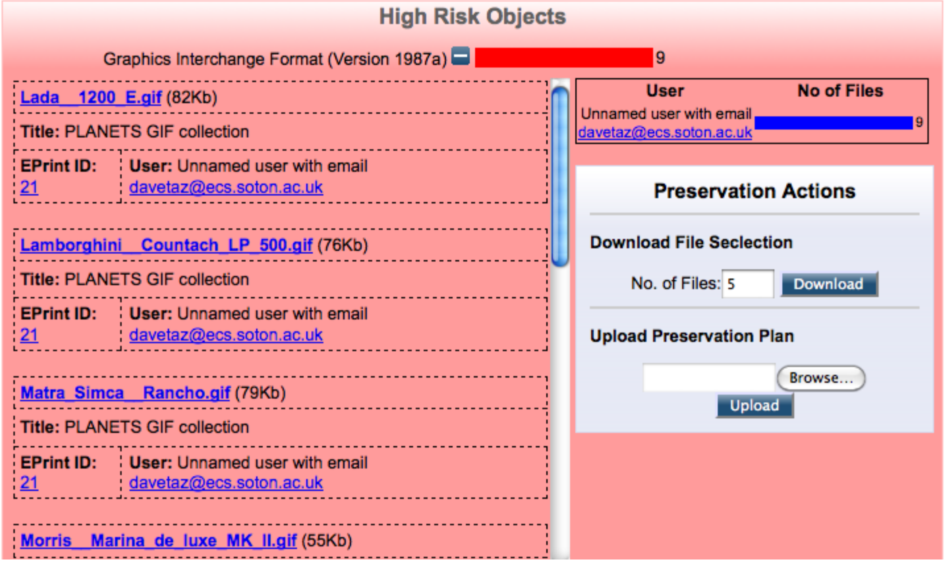
Modify the boundaries so that GIF are medium risk. Don't forget to reload the configuration. After verifying that this has worked, change it so that GIFs are high risk again.

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.



Lada__1200_E.gif (82Kb)
Title: PLANETS GIF collection
EPrint ID: 21 | **User:** Unnamed user with email davetaz@ecs.soton.ac.uk

Lamborghini__Countach_LP_500.gif (76Kb)
Title: PLANETS GIF collection
EPrint ID: 21 | **User:** Unnamed user with email davetaz@ecs.soton.ac.uk

Matra_Simca__Rancho.gif (79Kb)
Title: PLANETS GIF collection
EPrint ID: 21 | **User:** Unnamed user with email davetaz@ecs.soton.ac.uk

Morris__Marina_de_luxe_MK_II.gif (55Kb)

User	No of Files
Unnamed user with email davetaz@ecs.soton.ac.uk	9

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 morn 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.