

EPrints Preservation – File Formats and Risk Analysis

Table of Contents

Table of Contents.....	1
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>	2
4 Gathering File Samples (Exercise 3).....	3

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

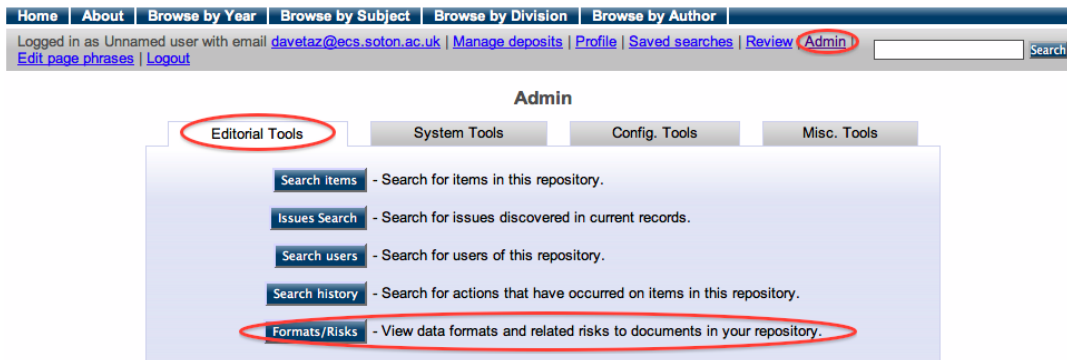
In this section we look at classifying the files in your repository using DROID from <http://droid.sourceforge.net/> and the classification add-ons available from files.eprints.org. In the tutorial repositories both of these packages have already been installed.

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.



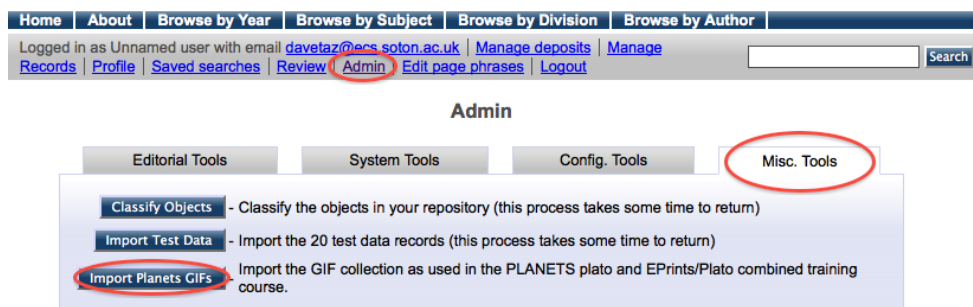
Viewing this page from an empty repository should result in the following screen.



3.2 Exercise 1 - Populating the repository

For 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.

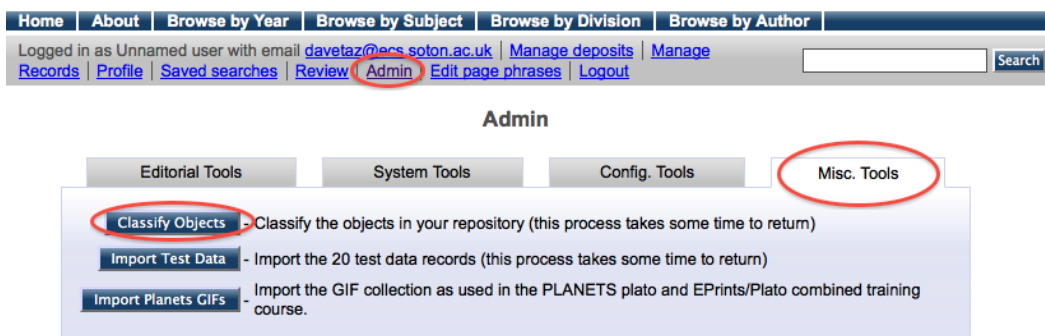
To import these an **Import Planets Gifs** button is available from the **Misc Tools** section of the **Admin** interface. This process takes some time to return, please be patient.



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.

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

Logged in as Unnamed user with email davelaz@ecs.soton.ac.uk | [Manage deposits](#) | [Manage Records](#) | [Profile](#) | [Saved searches](#) | [Review](#) | [Admin](#) | [Edit page phrases](#) | [Logout](#)

Admin

Editorial Tools | System Tools | Config. Tools | **Misc. Tools**

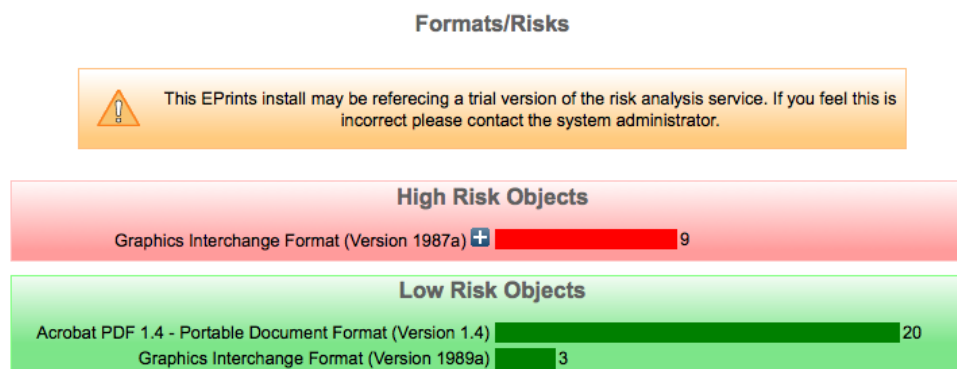
Classify Objects - Classify the objects in your repository (this process takes some time to return)

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


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

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


 This EPrints install may be referencing 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

4 Gathering File Samples (Exercise 3)

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.