Risk Analysis Overview

Risk Analysis and Preservation

The risk analysis portion of the Preservation module allows libraries to assess current and future risks related to formats of the objects in their repository, whether the risks relate to format obsolescence, application incompatibility, or an attribute that puts one or more formats at risk. Analyzing the current repository holdings from this perspective provides protection against large-scale loss.

During SIP processing, each file that is loaded into the permanent repository is measured for risk when the Risk Identification task is run as part of the validation stack

Types of Risks

Risks are defined as files that cannot be rendered or processed by their institution's software or hardware. Risks are always defined in terms of at-risk formats. There are four types of format-related risks:

- Obsolete formats: Formats that have been deemed obsolete by the global community. All files based or dependent on these formats are at risk.
- Missing applications: Formats that do not have the necessary applications (within the institution that houses them) to render them as originally intended. All files related to these formats are at risk.
- Property-driven risks: Formats with properties (technical metadata) containing certain values that prevent files from being rendered.
- For example, files that belong to the TIFF format may have color encoding with the value CIE_LAB. Only the TIFF files with this value are at risk. The risk is implemented as a query that searches for these values in all the files that belong to this format.
- Tool-driven risks: Formats with properties that are considered to be risky but are not extracted by the existing metadata extractor (JHOVE or NLNZ adapters). These risks are identified by specific risk-extracting plug-in tools and associated with a problematic format through the format's risk record. The extracted properties are stored in the repository database.

Identifying Risks

Rosetta uses following processes to measure the risk status of an institution's repository.

- **Ongoing Risk Identification**
- **Risk Identification Scheduled Process**
- **Ongoing Risk Analysis**

Ongoing Risk Identification

As part of the validation stack phase in SIP processing, Rosetta processes each file and extracts its technical metadata. The information is stored in the HDeStreamRef table.

If there is a risk extractor associated with a file's format, the system also runs the extractor tool and saves the output in the HDeStreamRef table. The extracted technical metadata is stored in a way that allows the risk analysis job to gather the information and summarize it in risk reports.

Risk Identification Scheduled Process

Because a repository contains existing files that have not gone through the ongoing risk identification, the validation stack that runs the risk extractor can be used on this set of files as part of the process automation. To do this, the user must create a set and run this process to identify the risk in these files.

Ongoing Risk Analysis

To generate risk reports, this scheduled process runs on all the risks of each format and checks the database to see how many files match the risk criteria.

For example, the risk analysis process will count how many files exist with obsolete formats. The numbers are stored in a table that is the basis for the format risks reports.

Generating Risk Reports

Rosetta generates reports that allow institutions to determine the degree of risk present in their permanent repositories. Customary, Risk Reports are generated automatically in one step.

Risk Reports

The results of the risk analysis processes are shown in the Global Risk Report. This report summarizes the number of objects for each format and the related risk(s). The reports allow users to understand the degree of risk posed by problematic formats and files in the repository.

In addition, the risk reports allow users to create a preservation set for each format and its related risks. A preservation plan can then be created to handle a preservation set and mitigate a specific risk.

Preservation Sets

A Preservation set is a set of objects specifically designed for Preservation activities. There are two kinds of Preservation sets:

- Format/Risk Set: A set that is based on format and risk and is created from the risk report. Since a Preservation plan must be based on both format and risk, only Format/Risk Sets (or duplications of such sets) can be the basis for Preservation plans.
- Format Set: For research purposes, users can create sets that are based only on format—without a risk code as a parameter. The user can create a set that shows all the risk properties. Format Sets cannot, however, be the basis for Preservation plans because they do not use risk codes.

The difference between both Preservation sets and regular, non-Preservation sets is that regular sets can contain multiple formats and thus cannot show the risk properties in their results. Regular sets show only the common attributes of all formats (mime type, file extension, file size and format id).