IEs failing to index due to a huge significant property

- **Product:** Rosetta
- **Product Version:** 5.0 until 5.2
- **Relevant for Installation Type:** Dedicated-Direct, Direct, Local, Total Care

Description

Please note the issue is solved in SP 5.2 and starting at SP 5.2 there is no limitation on the significant property length.

Solr index engine can handle properties only up to a certain size limit. Larger properties result in errors such as:

> ERROR [org.apache.solr.core.SolrCore] (http-bio-1801-exec-41) [] org.apache.solr.common.SolrException: Exception writing document id FL1311 to the index; possible analysis error.
>
> Caused by: java.lang.IllegalArgumentException: Document contains at least one immense term in field="FILE.significantProperties.significantPropertiesType.nisoImage.stripOffsets.norm_string.single" (whose UTF8 encoding is longer than the max length 32766), all of which were skipped. Please correct the analyzer to not produce such terms. The prefix of the first immense term is: '[56, 32, 49, 50, 55, 56, 50, 32, 50, 53, 53, 53, 54, 32, 51, 56, 51, 51, 48, 32, 53, 49, 49, 48, 52, 32, 54, 51, 56, 55]...',
>
> original message: bytes can be at most 32766 in length; got 35413

If this issue is encountered, the IE will not be indexed, and it will be placed in the Index Exception Queue.

Some of the problematic properties might be removed by the FLWG from a future version of the Format Library for all customers, but until then, it can be temporarily removed locally.

Examples for significant properties that tend to grow to such sizes are

IE.FILE.significantProperties.significantPropertiesType.nisoImage.stripOffsets.norm_string.single

or FILE.significantProperties.significantPropertiesType.nisoImage.stripByteCounts.norm_string.single

Workaround

The following procedure removes the property or properties chosen from the list of extracted properties.

This means that they will not be extracted and indexed for any file regardless of their size.

It should be noted that this is a temporary workaround. The Format Library is being overwritten every time a FL update is being performed, so if the FL update does not include the removal of the relevant property, you will have to repeat the procedure.

1. Change from Local to GLOBAL format Library:

    a. Go to Administration module > General > General Parameters

    b. Choose parameters Module : Format Library

    c. Change the parameter format_library_is_global to be TRUE

    d. Click Update

2. After that, in the Management module,

    a. Go to Home > Preservation > Manage Global Format Library > List of Extractors

    b. Edit the extractor that was used (i.e. PDF-hul-1.10, TIFF-hul-1.10)

    c. Remove the significant property that is causing this issue

    d. Re-run the TechMD task.

    Please note that running the task will remove the problematic metadata/ significant property, however, it will create new version of the IE.

- **Article last edited:** 03-JUL-2016