RecommendCustomIndexingWithTika

Introduction

Tika is an Apache software project that is capable of parsing a large number of rich document formats, including PDF, Microsoft Office, and many others.

Solr includes a contrib handler that utilizes Tika, the ExtractingRequestHandler, also known as SolrCell.

Don't use ERH in production

Rich document formats are frequently not well documented, and even in cases where there IS documentation for the format, not everyone who creates documents will follow the specifications faithfully. This creates a situation where software like Tika may encounter something that it is simply not able to handle gracefully. Although the authors put a LOT of effort into making sure the software runs well in unexpected situations, the reality is that sometimes a document will cause the software to malfunction and even crash.

If the Tika software included in Solr for SolrCell crashes, that means that Solr itself is going to crash too. That is why it is not recommended for production use. SolrCell is a proof-of-concept tool that can get you started with parsing rich documents, but for production we strongly recommend writing an external program that incorporates Tika and sends the discovered data to Solr.

If Tika processing is handled in a separate custom program, then any kind of malfunction or crash can be handled gracefully and will not affect the operation of Solr. With a custom program, the full capability of Tika will be available. The ExtractingRequestHandler is a generic implementation that does not provide access to the full capability of Tika. A custom program is also capable of manipulating the index data in myriad ways.

There is some example code on the Lucidworks blog.