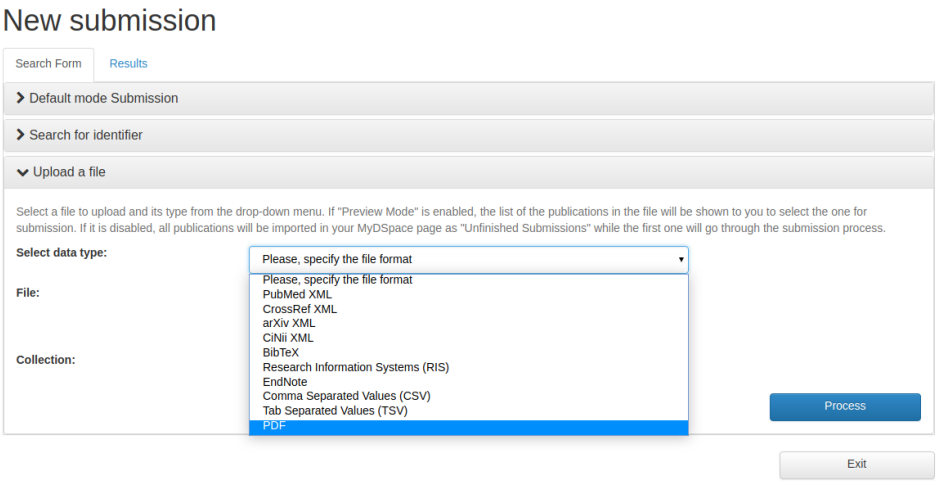
Capture metadata from the PDF file (Grobid Integration)

Through the integration with GROBID Library, DSpace-CRIS since 7th March 2018 (e65e0fa) leverages machine learning technologies for extracting information directly from PDF publications. Metadata such as title, authors, abstract, keywords, identifiers, source, etc. are extracted through parsing of uploaded full-text.

The feature is implemented as a BTE dataloader and accessibile in the import for a file section of the New submission



The PDF file is also automatically attached to the new item in the ORIGINAL bundle if the import is done directly **skipping the preview mode** of the BTE framework.

To enable the feature you need

1. to install grobid as a local accessible webservice. It is not needed to expose it over internet, in fact, it is recommended to keep it visible only to the dspace server to reduce security risks. Please refer to the Grobid documentation for the installation. The integration has been tested against Grobid 0.5.0 but it should work with more recent versions that doesn't break the grobid REST contract
2. to enable the grobid data loader in the bte.xml file, see https://github.com/4Science/DSpace/blob/dspace-5_x_x-cris/dspace/config/spring/api/bte.xml#L136 and https://github.com/4Science/DSpace/blob/dspace-5_x_x-cris/dspace/config/spring/api/bte.xml#L21 (if you want to use it also from the command line import)
3. to configure the grobid server connection in the grobid.cfg file, see https://github.com/4Science/DSpace/blob/dspace-5_x_x-cris/build.properties#L342 (use http://localhost:8070 for a default grobid installation)

The mapping between the extracted metadata and the dspace metadata is managed in the usual way of the BTE framework, where the mapping between the Grobid metadata and the BTE model is defined in the Grobid Data loader field map and the final mapping in the output map