Task	Description	Estimate (days)	Comments
Store indexed PDF text as FULL_TEXT datastream	On ingest, the extracted full text from PDFs will be indexed by Solr and stored as a datastream on the object.	2	This function can be optionally enabled/disabled through an admin interface.
Add option to include text file with PDFs	On PDF ingest, the user may optionally supply a text file to be indexed and stored as a datastream on the object.	1.5	In this case, the system will not attempt to automatically extract and index/store the text from the PDF (instead, the supplied text will be used).
Modify batch ingest to allow text files for PDFs to be included	Appropriately named text files may optionally be included with PDF files when ingesting via batch.	1.5	
Modify book ingest to allow users to supply their own text files with pages images	Allow users to supply their own text files when ingesting book pages instead of generating OCR automatically.	2	We will still need to generate hOCR in order to get word coordinates for highlighting search terms on the page images. The hOCR may not exactly match the supplied text since it will be generated automatically by Tesseract. There is currently no way to manually specify word coordinates (aside from hand editing the hOCR XML)