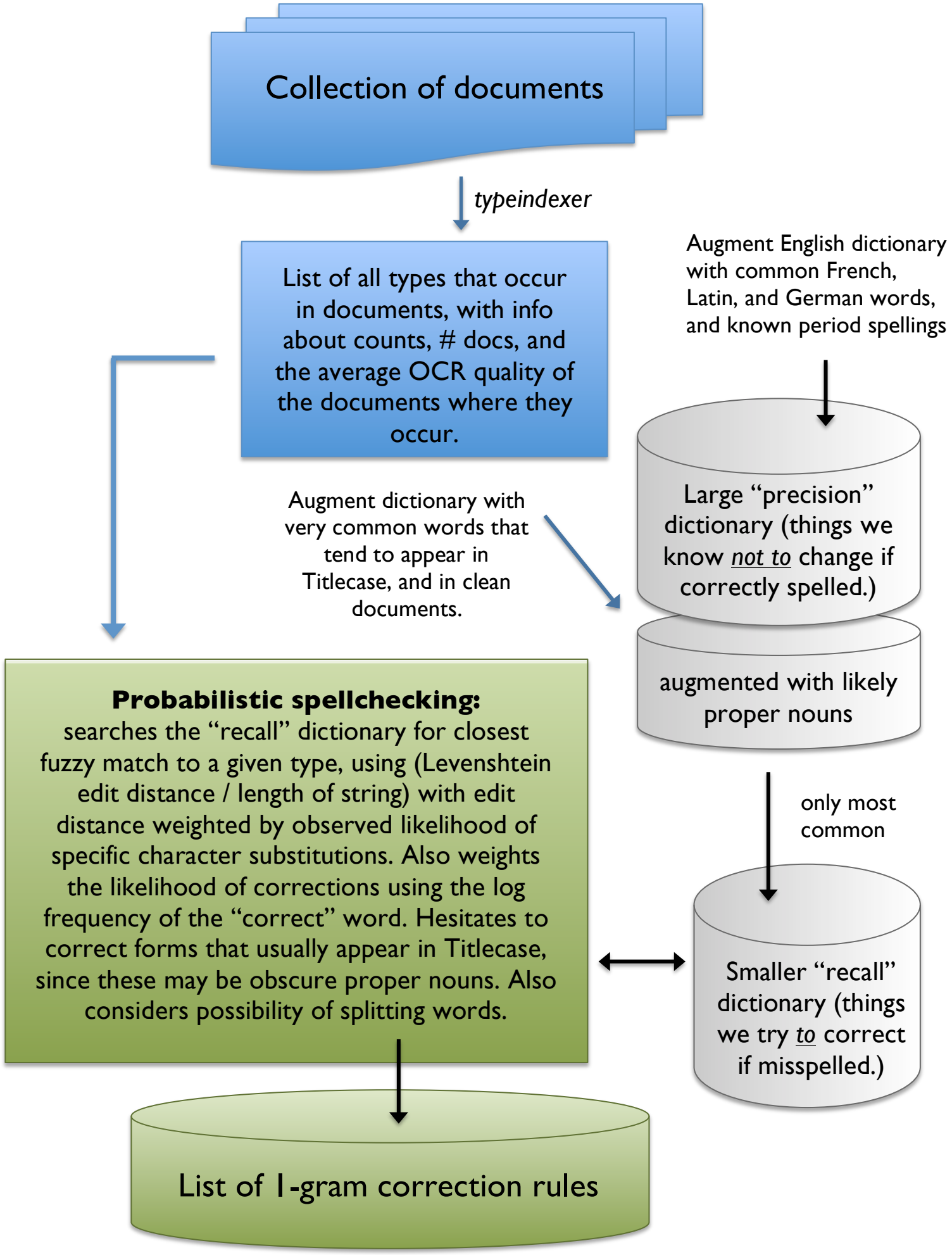
Collection of documents
typeindexer
List of all types that occur in documents, with info about counts, # docs, and the average OCR quality of the documents where they occur.
Augment English dictionary with common French, Latin, and German words, and known period spellings
Large "precision" dictionary (things we know not to change if correctly spelled.)
augmented with likely proper nouns
Augment dictionary with very common words that tend to appear in Titlecase, and in clean documents.
Probabilistic spellchecking: searches the "recall" dictionary for closest fuzzy match to a given type, using (Levenshtein edit distance / length of string) with edit distance weighted by observed likelihood of specific character substitutions. Also weights the likelihood of corrections using the log frequency of the "correct" word. Hesitates to correct forms that usually appear in Titlecase, since these may be obscure proper nouns. Also considers possibility of splitting words.
only most common
Smaller "recall" dictionary (things we try to correct if misspelled.)
List of 1-gram correction rules