Help collect dataset(s)

1. 【done】Basic learning about quantitative literary research => doc already
   1. 【undone】(search for big data literature)
2. Read their websites and log requirements
   1. Intelligent archive似乎比较image focus啊，比如古物文字识别，我不是特别感兴趣
3. Choose among the following approaches:
   1. The search engine approach: e.g. google dataset, gale dataset etc
   2. The article writing approach: see notes
   3. The question/research approach: what am I interested in?
   4. The helping approach: what are they interested in using, especially testing format?
   5. a simpler / faster ready dataset: <http://www2.informatik.uni-freiburg.de/~cziegler/BX/>
      1. good for building the entire system fast, but no use for me to get information about big book data
4. Collect dataset(s), each will:
   1. Input format
   2. (cleaning or state that a notebook/pipeline is needed for the dataset)
   3. Output format (storage required?)