OpenStax Data Collection Procedure

1. List the names of the textbooks we will include in our dataset
   1. Chemistry 2e
   2. Biology 2e
   3. University Physics 1-3 (let’s start with 2)
   4. Other names here…
2. For each textbook in our dataset, create a spreadsheet for the “textbook information”, meaning the names of the chapters, the names of the sub-chapters, the sub-chapter learning objectives. This avoids copying the same information into data spreadsheets over and over again.
   1. Name this spreadsheet “textbook\_name”+”textbook\_info”.xlsx
   2. Column names are: Chapter Name (Topic), Chapter Number, Subchapter Name (Subtopic), Subchapter Number, Subchapter Learning Objectives
3. For each textbook in our dataset, create a spreadsheet for the questions selected from the textbook. This will include the text of the question, any associated picture, the question number, which chapter and sub-chapter the question is from, question type (multiple-choice, free response, etc probably want to use terms from education literature for question types)
   1. Name this spreadsheet “textbook\_name”+”questions”.xlsx
   2. Column names are: Chapter Number, Subchapter Name (Subtopic), Subchapter Number, Question Text, Question Type, (maybe photo)
4. For each textbook in our dataset, select questions from every sub-chapter of every chapter to include in the dataset using the following procedure:
   1. Practice problems for each sub-chapter can be found in the “Exercises” (or other name denoting practice problems) section at the end of each chapter. The practice problems are grouped by sub-chapter. Appendices are not chapters, and they can be distinguished from chapters because they are labeled with letters rather than numbers, and because they don’t have practice problems.
   2. For each sub-chapter, randomly select 5 questions, using a [random number generator](https://www.random.org/) to supply 5 unique question numbers to choose.
   3. Paste each selected question into the questions spreadsheet, along with the chapter, sub-chapter, question number, and question type. If there is a picture included, paste this as well, although it’s unlikely that the pictures will be crucial
   4. If a question has especially difficult formatting that is hard to paste into an Excel cell, randomly choose another question using the random number generator

Notes:

* We can likely automate the question selection and processing for OpenStax.
* For example, if we pasted the text of the entire “Exercises” page for a chapter into a .txt file, we could write a relatively simple program that finds the boundaries between questions, does required processing, and randomly selects a sample of the questions to be included in the dataset. If we choose to include pictures (probably unnecessary) we can do this by hand afterwards since the number would not be that large.
* This would be much more efficient / effective than hand copying questions.