Our sources for our web scraping can be found here:

- USM building/classroom data:
  - https://tdx.maine.edu/TDClient/2624/Portal/KB/?CategoryID=22631
- USM Library study rooms data:
  - https://libguides.usm.maine.edu/guides/group-study-rooms/
- American Colleges data:
  - https://en.wikipedia.org/wiki/Lists_of_American_universities_and_colleges

Cleaning was only needed for our USM building/classroom data. The cleaning process for this is shown below. Note that the validation process was done for all data acquired through scraping, not just the cleaned data.

Sample room data gathered from our scraping process:


```
  "building": "List of Rooms in Bailey Hall on USM's Gorham Campus",
  "rooms": [
    {
      "room": "Room 10",
      "attributes": [
        "Web Conferencing Lecture Style",
        "Customizations"
      ]
    },
```


Sample room data after being cleaned by building_cleaner.py:


```
{
  "institution": "USM",
  "campus": "Gorham",
  "building": "Bailey Hall",
  "room": "Room 10",
  "room_number": "10",
  "attributes": [
    "custom",
    "web_conferencing"
  ],
  "has_web_conf": true,
  "has_pc": true,
  "customized": true
},
```


Data Cleaning & Validation Process:

- The raw scraped entry did not include institutional or campus context.
- Based on the known source of the data (USM classroom listing), the following fields were inferred and added to match with our database schema:
    - “Institution”: “USM”
    - “Campus”: “Gorham”
        - The campus data was attained through parsing the string value of the ‘building’ field in the original scraped data.
    - “Building”: “Bailey Hall”
        - The building data was attained through parsing the string value of the ‘building’ field in the original scraped data.
- With these changes, each room record aligns with the database’s hierarchy:
    - University → campus → building → room
- Extracted room number / name for location schema
    - The ‘room’ field keeps the original name of the room that was scraped, while the new ‘room_number’ field extracts the numeric part of the room name if there is one.
- Miscellaneous fields and boolean flags
    - Some of the data found in the original ‘attributes’ field of the scraped data was converted into flags in case they were ever needed. As of right now, no fields after ‘room_number’ are used or inserted into the database.
- Validation
    - Validation was done by running the script data_validation.py to view the highest occurring lengths of values in each field throughout the JSON file. These values were then manually compared to the limits set in our schema. The data is considered valid for the database as long as those returned values are within the designated limits.