Website 1:

[Cincinnati Children's Hospital Medical Center (cincinnatichildrens.org)](https://www.cincinnatichildrens.org/)

[cincinnatichildrens.org/robots.txt](https://www.cincinnatichildrens.org/robots.txt)

Cincinnati Children’s Hospital provides information about the hospital and patient resources. The business domain is of healthcare. I chose this website because it’s a good example as to why some websites don’t allow web scraping. Hospitals must protect their patients’ privacy by keeping their medical history confidential. The website holds personal information as it allows users to check messages from their doctors and pay their medical bills. If they allowed web scraping on their website it would be a HIPAA violation. The website has an entire page dedicated to its privacy practices stating, “We are required to protect the privacy of your information, establish Polices and Procedures that do so, provide this Notice about our privacy practices, and to follow the practices derived in this Notice.” Allowing webscaping could also increase the website’s susceptibility to security threats that could lead to data breaches and theft.

Disallow: /404.htm

Disallow: /site-error.html

Disallow: /Workarea/workareaError.aspx

Disallow: /templates/cchmc2010/PageTools/Share.aspx

Disallow: /templates/cchmc2010/Content.aspx

Disallow: /content-gen.aspx

Disallow: /templates/cchmc2010/TemplateFinder.aspx

Disallow: \*?=

Disallow: \*&=

Disallow: \*pagestart=2

Disallow: \*?q=

Disallow: /notfound

Disallow: /bio/photo/

Disallow: /sp20/

Disallow: /contact/email

Website 2:

[List of Friends episodes - Wikipedia](https://en.wikipedia.org/wiki/List_of_Friends_episodes)

[en.wikipedia.org/robots.txt](https://en.wikipedia.org/robots.txt)

Wikipedia is an online encyclopedia that is made up of various topics. Anyone is able to edit the information on the website. The business domain is mostly involved in spreading information. It’s a non-profit organization with information that’s public and relies on volunteer editors. I chose this Wikipedia page specifically because the contents of the webpage could be formatted as a CSV file, since it was in a table. This page specifically pertains to the show “Friends” and contains information about the episode numbers, titles, directors, etc. Text on Wikipedia can be reproduced or modified as long as it abides by the licensing terms. Wikipedia is under the Creative Commons Attribution-ShareAlike license. The website says, “The Wikimedia Foundation does not own copyright on Wikipedia article texts or illustrations. It is therefore pointless to email our contact addresses asking for permission to reproduce articles or images.” It also allows people to extract public information from Wikipedia using Python’s web scraper API.

# robots.txt for http://www.wikipedia.org/ and friends

# Please note: There are a lot of pages on this site, and there are

# some misbehaved spiders out there that go \_way\_ too fast. If you're

# irresponsible, your access to the site may be blocked.

# Observed spamming large amounts of https://en.wikipedia.org/?curid=NNNNNN

# and ignoring 429 ratelimit responses, claims to respect robots:

# http://mj12bot.com/