



<b>Course Code &amp; Title</b>	LISS2114 Introduction to Web Scraping		
<b>Convenor(s)</b>	Caitlin Hogan		
<b>Institution</b>	QMUL	<b>Department</b>	Linguistics
<b>Academic Year</b>	2025-2026	<b>Term</b>	Summer
<b>Number of Sessions</b>	3	<b>Length of Session(s)</b>	2 hours each
<b>Day, Date</b>		<b>Start and End time</b>	<b>Room Location</b>
13/5/2026 20/5/2026 27/5/2026		10:00 - 12:00	Graduate Centre Room 201, Queen Mary, University of London <a href="#">Mile End Campus</a>
<b>Enrolment Link:</b>	Available to book on SkillsForge from <b>Monday 2 March 2026</b> . Click the <a href="#">course link</a> to log in and register. Questions? Visit our Training FAQ here: <a href="#">Frequently Asked Questions - LISS DTP (liss-dtp.ac.uk)</a>		

## Course Description

This course explores how to obtain information from the internet via web scraping, extracting data from online sources. We will explore the theoretical advantages of using web scraped data, the ethical aspects of web scraping, and how to scrape different kinds of data from the web. We will focus on simple, easy-to-implement strategies that do not require previous coding experience. These include the SimpleScraper and the Web Scraper, and we will also explore BootCaT tool to create a corpus via automated web scraping.

## Meet your course convenor(s)

Caitlin Hogan is a 4<sup>th</sup> year PhD candidate at Queen Mary University of London, whose thesis focuses on the language use and community norms of Korean pop music fans on X/Twitter. Her work combines corpus linguistics with discourse analysis and ethnography in order to understand dynamics of power, identity and policing amongst this community. She has published articles employing corpus-assisted discourse analysis using scraped data from Twitter in journals like Language@Internet and Social Media + Society.

## Learning Outcomes

- To be able to obtain data from a website of your choice using web scraping
- To be able to understand the ethical issues with web scraping, and how to present web scraped data in an appropriate way
- To know how to employ web scraped data in your own research



## Course Outline

### Session 1

Session 1 will review the theoretical background to web scraping, including examples of where data scraped from the internet has been used in social science research.

It will cover some easy-to-use web scraping tools and browser extensions, and feature tutorials on how to scrape data with these tools.

### Session 2

Session 2 is a practical session in which participants will be able to practice scraping sample data from an online source together with the rest of the class, with the convenor and other students there to help with troubleshooting. This allows students to practice scraping in a supportive environment to build their confidence.

### Session 3

Session 3 is an interactive, bring-your-own-data workshop session where students can bring data they hope to scrape, or error messages they get using BootCaT or one of the online scrapers. The convenor will assist with any problems students have.

## Reading list

No required readings, but optional readings are as follows

- Baroni, M., & Bernardini, S. (2004, May). BootCaT: Bootstrapping Corpora and Terms from the Web. In *LREC* (pp. 1313-1316).
- Daniel Glez-Peña, Anália Lourenço, Hugo López-Fernández, Miguel Reboiro-Jato, Florentino Fdez-Riverola, Web scraping technologies in an API world, *Briefings in Bioinformatics*, Volume 15, Issue 5, September 2014, Pages 788–797, <https://doi.org/10.1093/bib/bbt026>
- Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems*, 47, pp-pp. <https://doi.org/10.17705/1CAIS.04724>

## Pre course preparation

None required

## Eligibility

Students do not need any coding or programming experience to participate in this module.

## Number of students

Min: 2

Max: 30