# DESCRIPTIF DU COURS/COURSE DESCRIPTION

Over the last decade, new forms of data have become available through the spread of websites, digitization of public records, and the proliferation of mobile technologies. This has given rise to new techniques to collect and analyze data. This course is designed as a hands-on introduction into such techniques with ample exercises. By the end of it, you will have acquired the concrete skills to harness such data for your own work. Drawing on the programming language *R*, it will introduce you to several techniques to gather data from the web, analyze text, and related machine learning techniques. In doing so, we will discuss the ethical and methodological considerations that come with using modern, digital forms of data in the social sciences.

**Day 1: Introduction to R**

While this course is not a complete introduction into *R*, you are not required to know R beforehand. During day 1, I will teach you the programming skills needed for the rest of the course. This includes the fundamentals of *R* object classes/data types, data management, base *R* skills and related solutions from the tidyverse. At the end of this day, you will be capable of basic *R* programming, handling/managing of data frames, and to follow and apply the techniques we will introduce throughout the rest of this course. Of course, more training is required to master the language but you will get the intuitions and basics to speed up your learning as we go along.

**Day 2: Webscraping**

On day 2, I will introduce you to the structure of HTML and XML documents. Following this, I will teach you how to scrape static and dynamic websites, process the resulting data, including how to think of this process as a data management pipe. We will discuss issues that come with “big data” as well as ethical considerations in getting and using data from the web.

**Day 3: APIs**

Application programming interfaces (APIs) are a crucial cornerstone of accessing modern forms of data. You will be introduced to the concept of APIs as well as several techniques to successfully use APIs to gather and enrich

data using *R*. In particular, we will focus on RESTful APIs and processing JSON data format which is the standard data format for APIs. We will focus on how to stream and process data through APIs and discuss ethically correct usage of APIs.

**Day 4: Text analysis — Basic techniques and concepts**

One of the most promising new forms of data are digitalised texts. The rest of this course will therefore focus on introducing you to various techniques of using text in the social sciences. On day 4, we will lay the foundation for this, starting with basic and advanced techniques of string manipulations (e.g. cleaning and transforming strings, pattern matching using regular expressions, etc.). We will finish this day with introducing you to the logics of handling large-scale text data using what is called a corpus.

**Day 5: Text analysis — Advanced techniques**

On day 5, I will introduce you to more advanced techniques of handling text corpora. For this purpose, we will focus on the *R* package quanteda. You will learn how to prepare text corpora from files or other data sources and how to properly utilize methods to preprocess, clean, and manipulate text corpora. I will then introduce you to several different techniques to analyse such corpora, ranging from dictionary and sentiment analysis as well as classic and structural topic models.

## BIOGRAPHIE ENSEIGNANT/TEACHER BIOGRAPHY

Achim Edelmann is an Assistant Professor in Computational Social Science at Sciences Po. He holds a PhD in sociology from the University of Cambridge. Before joining the médialab, he habilitated in sociology at the University of Bern, was a visiting scholar at the University of California at Berkeley and a Postdoctoral Trainee at the Duke Network Analysis Center. Achim specializes in the sociology of culture, social networks, and computational methods. In his work, he increasingly collects and analyzes new forms of large-scale data