Introduction à l'extraction de données sur internet avec R

Thomas Delclite

Statbel (office national belge de statistique)
thomas.delclite@economie.fgov.be

L'usage de plus en plus massif d'internet incite les chercheurs et statisticiens à produire des bases de données issues d'internet : listing de prix, analyse de réseaux sociaux, suivi de réaction à des sujets d'actualité, etc. Pour ces sujets, l'information est partiellement disponible car si l'usager d'un site peut « voir » les informations, les sites web ne permettent souvent pas d'extraire directement l'ensemble des données. Théoriquement, il serait possible de naviguer sur le(s) site(s) web et d'extraire manuellement les informations, mais la limitation tient au fait que ce travail est long et sujet à erreur de report.

L'objectif de l'atelier est, en trois heures, de former à l'extraction de données à l'aide du logiciel R. Nous verrons ensemble comment, à l'aide de programmes simples, localiser, extraire et structurer des données issues de sites internet simples.

À la fin de l'atelier, les participant.e.s seront capables d'identifier les données disponibles sur un site internet, de programmer une routine capable d'extraire automatiquement les données et de les structurer sous un format propice à l'analyse. L'analyse des données n'est pas abordée dans cet atelier.

Il est nécessaire de posséder un ordinateur et d'installer R et Rstudio (disponibles gratuitement) avant l'atelier. Une connaissance de R est utile mais pas indispensable.

L'atelier sera divisé en 4 parties :

1- Analyse du code source d'une page web et langage Xpath
2- Package XML de R
3- Extraction d'une page et structuration
4- Automatisation du programme

Note bibliographique

Thomas Delclite travaille à Statbel, l'office national belge de statistique, depuis 2015. Au sein du service de méthodologie, il s'occupe du tirage de différents échantillons pour des enquêtes sociales, effectue le calage des résultats de ces enquêtes et estime la variance de leurs indicateurs-clefs. Par ailleurs, Thomas Delclite enseigne à l'université de Lille l'extraction de données sur Internet à raison d'un séminaire de 18 heures par an.