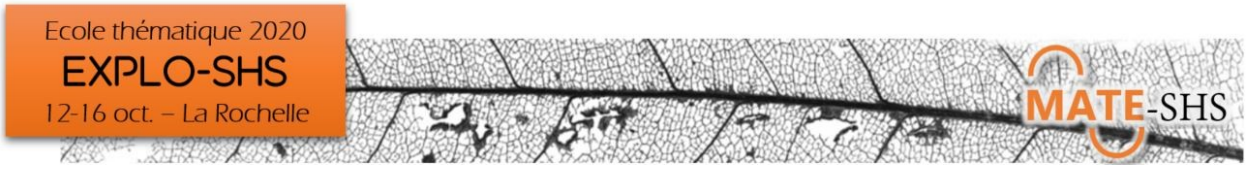

Massification des données, structures et langages du web : ce qu'il faut savoir avant de se lancer

Frédéric Vergnaud, Antoine Mazières (en visio) et Benjamin Ooghe Tabanou

Focus, jeudi 15 octobre 8h30-9h15

Présentation générale

Afin de donner aux stagiaires quelques repères fondamentaux sur le lexique et les technologies qui seront utilisés au cours des trois ateliers du jeudi matin, ce focus se propose dans un premier temps de jeter les bases du fonctionnement d'Internet, du web et de la structuration (HTML) et forme (CSS) d'une page web.

Dans un second temps, nous examinerons la manière de cibler et d'extraire des éléments de ces pages (scraping) grâce au modèle de page DOM et au langage de requête XPATH qui permet de s'y déplacer.

Enfin, dans une troisième partie, nous verrons en quoi le crawling se distingue du scraping et comment la structure hypertextuelle du web reposant sur les liens peut permettre de nouvelles formes d'analyse.