



## Programme des séances 2018-2019

### Séance 1 : **6 décembre 2018** – Ronan Ysebaert (RIATE) et Hugues Pécout (CIST) " *Introduction au logiciel R et aux chaînes de traitements reproductibles* "

Cette séance s'adresse aux personnes souhaitant s'initier au logiciel-langage R et aux démarches d'analyse reproductibles. Elle s'articulera tout d'abord autour d'une présentation de R, de son environnement Rstudio et des atouts de cet outil (polyvalence, extensibilité, reproductibilité...).

Dans un second temps seront présentées plusieurs bibliothèques utiles pour couvrir une chaîne de traitement, depuis l'import et la mise en forme des données, en passant par leur analyse et leur représentation (analyse statistique, spatiale, textuelle, cartographie, etc.) jusqu'à leur export et leur documentation. Un exemple simple, concret et pratique clôturera la séance.

### Séance 2 : **7 février 2019** – Elodie Baril (Ined) et Arnaud Bringé (Ined))" *Organiser et traiter plus facilement ses données avec R : utiliser le package dplyr* "

La gestion de données sous R apparaît souvent compliquée et peut constituer un frein pour des utilisateurs habitués au management de données sous d'autres logiciels statistiques ou des débutants en R. Cette séance s'adresse aux personnes souhaitant s'initier à cette syntaxe claire et cohérente, sous formes de verbes permettant de traiter et manipuler facilement les données contenues dans une ou plusieurs tables R.

Ce package, proposé en 2016, fait partie du tidyverse (<https://www.tidyverse.org/>) et a l'avantage de présenter une grammaire unifiée. Il semble aujourd'hui incontournable pour qui souhaite traiter ses données avec R en toute autonomie.

Grâce à sa syntaxe plus « naturelle » dplyr permet de sélectionner des lignes/colonnes, créer des variables, regrouper de données et enfin calculer des statistiques descriptives. Des fonctions servent également à renommer les variables, ordonner les données, fusionner des tableaux très simplement. Elles sont en général plus rapides que leur équivalent sous R de base et sont adaptées au traitement de données de grande dimension.

Nous verrons à l'aide d'un exemple concret comment préparer ses données, expliciter les éléments de la grammaire de dplyr, enchaîner les opérations élémentaires pour constituer un fichier d'analyse et enfin calculer facilement des statistiques agrégées.

### Séance 3 : 4 avril 2019 – " Introduction à Python pour les utilisateurs de R "

**Matthieu Viry** (LIG - Univ. Grenoble Alpes / CNRS / Grenoble INP)

La présentation dressera un aperçu de l'écosystème destiné au calcul scientifique, à la statistique et à la manipulation de données géo spatiales, notamment en montrant la correspondance entre bibliothèques incontournables et façon de faire appropriées dans les deux langages.

Enfin l'accent sera mis sur la grande polyvalence du langage Python, notamment lorsqu'il s'agit d'effectuer des opérations transversales à l'analyse de données (création d'applications Web, opérations relatives au système d'exploitation, etc.).

### Séance 4 : 6 juin 2019 – "L'analyse textuelle en R avec R.temis"

**Milan Bouchet-Valat** (Ined), **Antoine Chollet** (Ensaï) et **Bénédicte Garnier** (Ined)

**R.temis** est un nouveau paquet R dédié à l'analyse textuelle, qui prend la suite de l'interface graphique RcmdrPlugin.temis (R.TeMiS). Contrairement à son prédécesseur, R.temis fournit des fonctions à utiliser dans des scripts (par exemple dans RStudio). Il s'adresse donc aux utilisateurs déjà familiers avec R, tout en se voulant accessible. Le principe du paquet est de faciliter les étapes essentielles de l'analyse textuelle (importation, traitement et analyses) en s'appuyant au maximum sur les paquets existants (tm, FactoMineR, explor, igraph...) pour permettre aux utilisateurs de réaliser les analyses dont ils ont besoin sans les enfermer dans une méthode particulière.

**R.temis** prend en charge les méthodes suivantes :

- importation de corpus au format .csv, .txt, Alceste, Factiva, Europresse et LexisNexis
- suppression des mots vides, lemmatisation automatique modifiable manuellement
- découpage des textes en paragraphes
- construction de sous-corpus à partir de termes
- bilan lexical, spécificités, co-occurrences, concordances
- nuage de mots
- analyse des correspondances sur tableau lexical entier ou agrégé
- classification
- graphes de mots

Après une présentation des principes du logiciel et des principales méthodes, nous proposerons une illustration pratique à partir d'un extrait du corpus EuroBroadMap, constitué de réponses d'étudiants à la question « Quels sont les mots que vous associez le plus à l'« Europe » ? Choisissez 5 mots au maximum ».