utilitR: une documentation utile pour R

Lino Galiana* Olivier Meslin†

Résumé

La documentation `utilitR` est un projet *open source*, collaboratif, destiné à tout utilisateur de `R` dans le cadre d'un usage courant d'analyse de données. Le projet `utilitR` se distingue d'une documentation classique en proposant une approche par tâche (importer, manipuler des données, réaliser des graphiques, cartes, etc.), sur données réelles, avec des recommandations précises sur les packages à privilégier et des conseils sur les bonnes pratiques ou la démarche à adopter face à un problème.

Les fonctionnalités les plus avancées de l'écosystème `R Markdown` (principalement `bookdown`, `blogdown` et `pagedown`) sont mobilisées afin de produire, avec les mêmes fichiers sources, un site *web* actualisé en continu (https://www.utilitr.org), une version paginée de chaque chapitre pour une impression simplifiée et une version complète de la documentation dans un PDF de plus de 350 pages.

`utilitR` sert également de laboratoire aux méthodes d'intégration et de déploiements continus à l'Insee. En acculturant une vingtaine de contributeurs à mener un projet open source exigeant, `utilitR` préfigure les futures méthodes de diffusion des travaux statistiques.

Mots-clefs : Documentation – Open Source – R Markdown

Développement

Pour accompagner la transition vers des langages *open source* du Système Statistique Public, le projet `utilitR` a émergé de manière spontanée afin de proposer une documentation à partir de cas d'usages. L'objectif de la documentation est de proposer un ensemble de recommandations et de conseils pour l'usage de `R` sur données réelles dans un cadre de travail collectif. Pour cela, un package facilitant l'accès aux données du site `insee.fr`, à savoir `doremifasol`, a vu le jour. Ce dernier permet que l'ensemble des des tâches évoquées (importer des données, manipuler celles-ci avec le `tidyverse` ou `data.table`, communiquer avec `ggplot` ou `R Markdown`...) soient illustrées avec des données communément utilisées par les statisticiens ou *data scientists*.

La documentation s'organise autour de quatre axes principaux :

- Mener un projet statistique avec `R` ;
- Importer des données avec `R` ;
- Manipuler des données avec `R` ;
- Produire des sorties avec `R`.

La première partie présente l'écosystème `R` et `RStudio` favorisant la conduite d'un projet d'analyse avec `R` dans un cadre collaboratif. En particulier, elle comporte une présentation extensive de `Git`. La deuxième partie se focalise sur l'import de données à partir de divers formats (CSV, SAS, Excel, bases de données...). La troisième partie présente les principales approches pour manipuler des données (`data.table`, `tidyverse`, données spatiales) et se focalise ensuite sur des tâches courantes (nettoyage de données textuelles, jointures, statistiques sur données d'enquêtes...). La dernière partie se concentre sur la valorisation du travail de données: visualisation, production de sorties `R Markdown`...

Dans le champ de la statistique publique, la démarche `utilitR` est originale. Il s'agit d'une démarche spontanée, collaborative et ouverte. Outre la mise à disposition des fichiers sources, `Github` sert à héberger l'ensemble des débats entre contributeurs. Les recommandations et orientations stratégiques sont discutées dans le dépôt et ouvertes à des contributeurs au-delà du système statistique public. Au sein de l'Insee,

*Insee, lino.galiana@insee.fr
†Insee, olivier.meslin@insee.fr

l'approche est originale car elle ne repose pas sur une organisation hiérarchique. En interne, des chefs peuvent soutenir l'initiative, ses valeurs (exposées dans un manifeste public) et l'aider à disposer de moyens ou d'un appui interne. Ces sponsors ne contrôlent cependant pas le contenu de la documentation et ses orientations stratégiques qui sont du ressort du groupe des contributeurs. Cette organisation ouverte et collaborative sert de laboratoire aux futurs projets de l'institut adoptant une démarche *open-source*. En acculturant une vingtaine de contributeurs à mener un projet *open source*, `utilitR` préfigure les futures méthodes de diffusion des travaux statistiques.

Le processus de publication du contenu est exigeant et repose sur les méthodes de type CI/CD. Afin de tester la reproductibilité des exemples disponibles dans la documentation dans de multiples contextes, une image `Docker` est construite avec toutes les dépendances nécessaires. Les mêmes fichiers sources servent à construire le site web (à partir de `bookdown`) et une version PDF (à partir de `pagedown`). Un package de *templates*, disponible sur `Github`, sert à contrôler le comportement de `R Markdown`. L'originalité de cette approche est qu'elle s'appuie, y compris pour la version PDF, exclusivement sur du HTML+CSS et non sur `LateX`. Le site web est déployé en continu avec `Netlify`. Ce processus d'automatisation a permis quelques innovations dans l'éco-système `R Markdown`, notamment une évolution du fonctionnement de `pagedown` permettant de ne plus être limité par la taille du document et ainsi envisager la généralisation de livres ne reposant plus sur `LaTeX`.