

Sciences humaines, sciences sociales à l'ère numérique

L'analyse des données et l'urbanisation d'outils

Stéphane Pouyllau, ingénieur de recherche CNRS - Professeur Attaché Univ. Evry - Paris Saclay

Méthodes, techniques, outils...

Notions abordées dans les cours à venir...

La mise en données (méthodes de modélisation des connaissances)

L'analyse des données et l'urbanisation d'outils

La pérennité de l'information

La publication Web à l'heure de l'IA générative et à la nécessité du *Low-Tech*

...

L'analyse des données et l'urbanisation d'outils

Les points essentiels

- Mise en données et documentation d'informations scientifiques
- Définition des traitements et explicitation (documentation) des algorithmes à utiliser
- Mise en œuvre technique et scientifique (le choix des outils)
- Publication des résultats, des données et des codes



Série "Severance" S1 E2, l'équipe de Mark S. dans la salle de raffinage des données de Lumon.

openrefine.org

OpenRefine is a powerful free, open source tool for working with messy data: cleaning it from one format into another; and extending it with web services and data.

Main features

- Faceting
- Clustering
- Reconciliation
- Infinite undo/redo
- Privacy
- Wikibase

python.org

The Python™ website features a search bar and navigation links for Python, PSF, Docs, PyPI, Jobs, and Community.

Functions Defined

```
# Python is Filenumber series up to n
def fib(n):
    """ Print Fibonacci series up to n """
    a, b = 0, 1
    while a < n:
        print(a, end=' ')
        a, b = b, a+b
    print()
    fib(1000)
```

Python is a programming language that lets you work quickly and integrate systems more effectively. [Learn More](#)

Get Started, **Download**, **Docs**, **Jobs**

cran.rstudio.com

The Comprehensive R Archive Network (CRAN) provides precompiled binary distributions of the base system and contributed packages. It includes sections for Download and Install R, Source Code for all Platforms, and Questions About R.

Download and Install R

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, so you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. This should be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2024-10-31, File of Leaves) [R-4.4.2.tar.gz](#), read [what's new](#) in the latest version.
- The CRAN directory [srcbase-prelease](#) contains R alpha, beta, and rc releases as daily snapshots in time periods before a planned release.
- Between releases, the same directory [srcbase-prelease](#) contains snapshots of current patched and development versions. Please read about [new features](#) and [bug fixes](#) before filing corresponding feature requests or bug reports.
- Alternatively, daily snapshots are [available here](#).
- Source code of older versions of R is [available here](#).
- Contributed extension packages.

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our [an frequently asked questions](#) before you send us an email.

Supporting CRAN

- CRAN operations, most importantly hosting, checking, distributing, and archiving of R add-on packages for various platforms

LCHS Concept Label sample csv

14550 rows

ConceptLabel	example
Fictional character	3328 Convention of Fictional character
English	1373 English (language)
Vietnam	1298 Vietnamese (language)
Craft	1223 Craft
Italy	1151 Franciacorta (Italy)
New York, N.Y.	1018 Daily News Building (New York, N.Y.)
African people	976 Kenyan (African people)
Iran	1487 Cultural Revolution (Iran)
Alewife	723 Andover Islands Reserve (Alewife)
Extinct city	719 Sumurun (Extinct city)
France	709 Fortuna (France)
Aric	691 Food Hotel Aric (Arist)
India	684 Chittagong (India)
Germany	652 Einfachheit (Germany)
Russia	649 Zelenovskiy Forest (Russia)
Japan	649 Sakai Region (Japan)
Qing	645 Kunshan (Qing)
China	638 Kunshan (People's Republic of China)
English	626 Children's stories, British (English)
Game	621 Trivii (Game)
Mexico	571 Reserva Ecológica Cuyabá (Mexico)
Scotland	570 Harris (Scotland)
Abbas	552 Firdausi (Abbas)
Ishio	507 Orchard Training Area (Ishio)
Islamic law	498 Game laws (Islamic law)
Tzotzil	492 Camp Walker (Tzotzil)
Computer program language	466 Computer program language
Iran	445 Rock-Roll Music (Iran)
Psi	439 Mitbrook Marsh (Psi)
B.C.	429 Featherstone (B.C.)
N.Z.	429 Puketi Rhododendron Trust Garden (N.Z.)

localhost

RStudio Launcher

Users/stephanepeouillyau/Documents/cours/Evry/2024

- Notebook
- Console
- Python 3 (ipython)
- Terminal
- Text File
- Markdown File
- Python File
- Show Contextual Help

RStudio

R version 4.4.2 (2024-10-31) "Pile of Leaves"

R is Free software and comes with ABSOLUTELY NO WARRANTY.

R is a collaborative project with many contributors; Type "contributors()" for more information and "citation()" for how to cite R or R packages in publications.

Type "help()"/"?some.function", "help('help')", or "help('quit')", for on-line help.

Type "q()" to quit R.

[Workspace loaded from ~/RData]

Open Source Components

Copy Version Ok

43 B

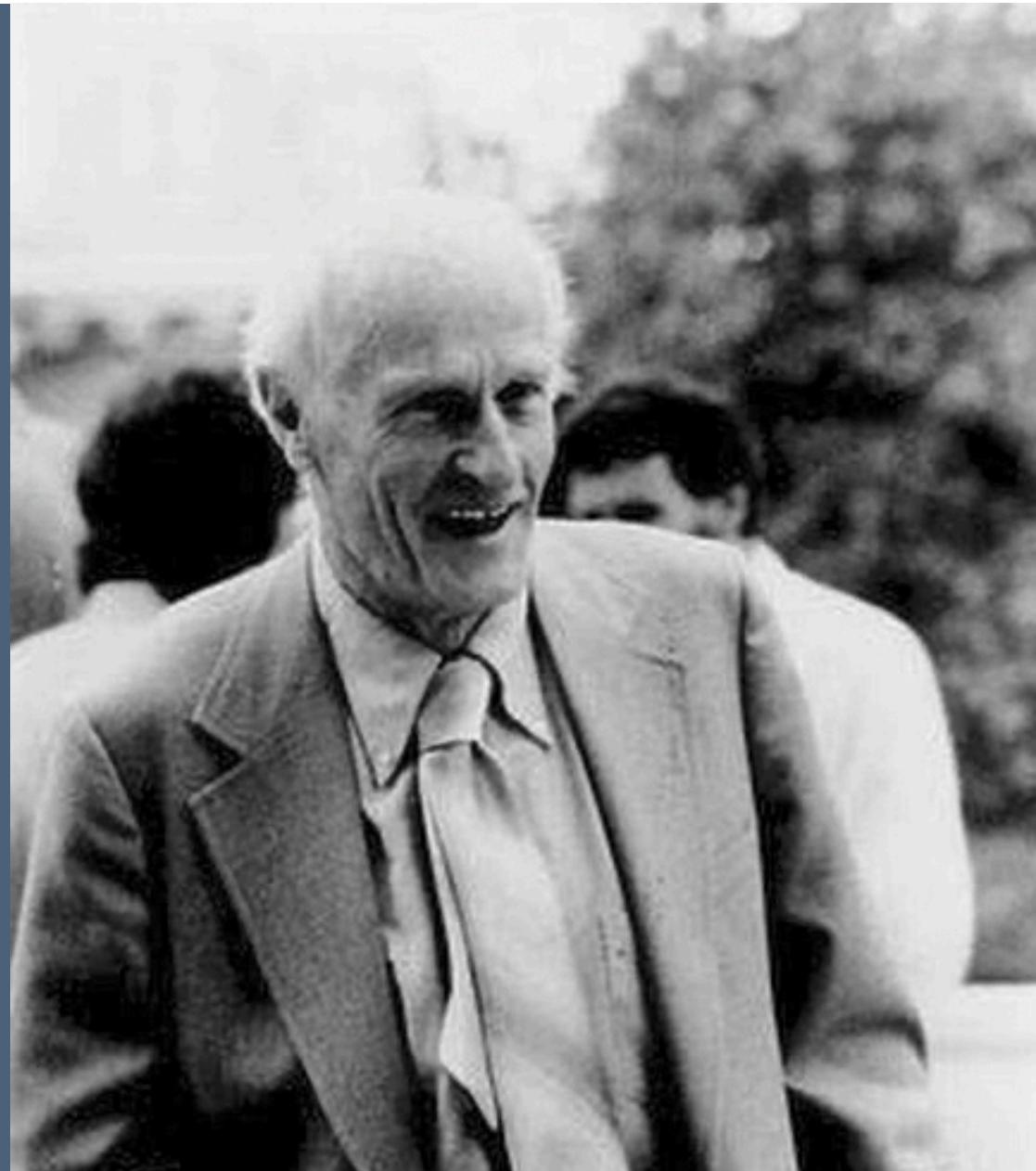
Langages et IDE

Langages	<i>Regular expression (Regex)</i>	Python	R
IDE	Open Refine	JupyterLab	R Studio

Langages

Regex

- Une expression régulière ou expression rationnelle ou expression normal est une chaîne de caractères qui décrit, selon une syntaxe précise, un ensemble de chaînes de caractères possibles
- Stephen Cole Kleene (1909-1994), mathématicien et logicien américain



Langages

Regex

Exemples d'expressions régulières

Expression régulière	Mots décrits	Mots non décrits
détecté	« détecté »	« détect », « détecta », « détectés », « »
ex (a?elælélé)quo	« ex équo », « ex equo », « ex aequo » et « ex æquo »	« ex quo », « ex aiquo », « ex aeko », « ex æéquo »
^Section .+	« Section 1 », « Section 22 », « Section A », ...	« voir Section 1 », « Sectionner »
6,66*\$	« 6,6 », « 6,666 », « 6,6666 », ...	« 6,66667 »,
[1234567890]+([1234567890]+)?	« 2 », « 42 », « 0,618 », « 49,3 », ...	« 3, », « ,75 » , « »

Languages

Regex

<https://regex-generator.olafneumann.org>

The screenshot shows a web browser window for the Regex Generator at <https://regex-generator.olafneumann.org>. The interface is divided into two main sections:

- Step 1:** A yellow header bar with the title "Regex Generator" and the subtitle "Creating regular expressions is easy again!". Below this is a text input field labeled "Paste a sample text." containing the log entry: "2020-03-12T13:34:56.123Z INFO [org.example.Class]: This is a #simple #logline containing a 'value'.". A small explanatory note below the input says: "Give us an example of the text you want to match using your regex. We will provide you with some ideas how to build a regular expression." The URL in the address bar is partially visible as "https://regex-generator.olafneumann.org/".
- Step 2:** A black header bar with the number "2" and the question "Which parts of the text are interesting for you?". Below this is another text input field with the same log entry. To the right of the log entry is a visual representation of the text where each word or significant character is highlighted with a different color (e.g., red, blue, green, yellow) in a pixelated grid pattern.



Langages

Regex

fr.wikipedia.org

Expression régulière

Pour les articles homonymes, voir [régulier](#) et [rationnel](#).

En [informatique](#), une **expression régulière** ou **expression rationnelle**¹ ou **expression normale**^{note 1} ou **motif** est une **chaîne de caractères** qui décrit, selon une syntaxe précise, un **ensemble** de chaînes de caractères possibles. Les expressions régulières sont également appelées **regex** (un **mot-valise** formé depuis l'anglais *regular expression*). Les expressions rationnelles sont issues des théories mathématiques des **langages formels** des années 1940. Leur capacité à décrire avec concision des **ensembles réguliers** explique qu'elles se retrouvent dans plusieurs domaines scientifiques dans les années d'[après-guerre](#) et justifie leur adoption en [informatique](#). Les expressions régulières sont aujourd'hui utilisées pour programmer des logiciels avec des fonctionnalités de lecture, de contrôle, de modification, et d'analyse de textes ainsi que dans la manipulation des langues formelles que sont les [langages informatiques](#).

Les expressions **régulières** ont la qualité de pouvoir être décrites par des formules ou motifs (en anglais *patterns*) bien plus simples que les autres moyens.²

Histoire

Dans les années 1940, [Warren McCulloch](#) et [Walter Pitts](#) ont décrit le système nerveux en modélisant les neurones par des **automates** simples. En 1956, le logicien [Stephen Cole Kleene](#)^{3,note 2} a ensuite décrit ces modèles en termes d'**ensembles réguliers** et d'**automates**. Il est considéré comme l'inventeur des expressions régulières⁴. En 1959, [Michael Rabin](#) et

58 langues

Sommaire masquer

Début

Histoire

Utilisation

> Principes

Standards

> Classe de caractères

Fonctions avancées

> Notations : implémentations et standardisation

Expressions régulières et Unicode

Implémentations et complexité algorithmique

> Notes et références

> Voir aussi

Rechercher sur Wikipédia

Rechercher

Faire un don Créer un compte Se connecter ...

Apparence masquer

Taille du texte

Petite

Standard

Grande

Largeur

Standard

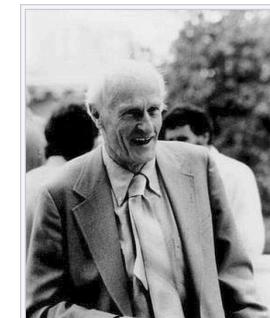
Large

Couleur (bêta)

Automatique

Clair

Sombre



Langages

Python

- Guido van Rossum, créateur de Python, en 1989-1991, Pays-Bas puis USA
- Python est un langage de programmation interprété
- Il est multiplateformes
- Il favorise la programmation impérative structurée, fonctionnelle et orientée objet



Languages

Python

<https://lectures.scientific-python.org>

The screenshot shows a web browser window displaying a scientific Python tutorial page. The URL in the address bar is <https://lectures.scientific-python.org>. The page title is "1.2. The Python language". The page content includes a sidebar with links to "Previous topic", "Next topic", "This Page", and "Quick search". A main text block starts with "We introduce here the Python language. Only the bare minimum necessary for getting started with NumPy and SciPy is addressed here. To learn more about the language, consider going through the excellent tutorial <https://docs.python.org/3/tutorial>. Dedicated books are also available, such as [Dive into Python 3](#)". Below this, a green box contains the text "Python is a programming language, as are C, F..." followed by the Python logo. The page is part of the "1. Getting started with Python for science" section, which itself is under the "Scientific Python Lectures" main menu.

Scientific Python Lectures » 1. Getting started with Python for science » 1.2. The Python language

previous | next

Previous topic
Next topic
This Page
Quick search

Go **thon for scientific computing**

We introduce here the Python language. Only the bare minimum necessary for getting started with NumPy and SciPy is addressed here. To learn more about the language, consider going through the excellent tutorial <https://docs.python.org/3/tutorial>. Dedicated books are also available, such as [Dive into Python 3](#).

Python is a programming language, as are C, F...

1.2.1. First steps

1.2.2. Basic types

- [1.2.2.1. Numerical types](#)
- [1.2.2.2. Containers](#)
- [1.2.2.3. Assignment operator](#)

Langages

R

- *Ross Ihaka* et Robert Gentleman en 1993, université d'Auckland
- R est un langage de programmation et un logiciel libre
- Adapté aux sciences statistiques et au traitement de données de type tableau et séries Il favorise la programmation impérative structurée, fonctionnelle et orientée objet



Langages

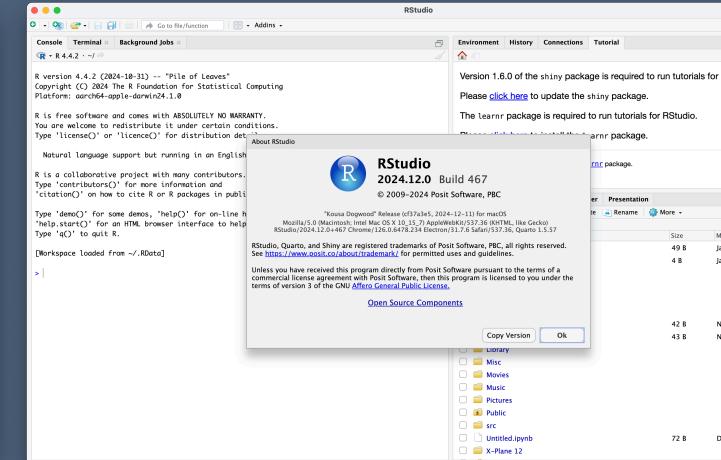
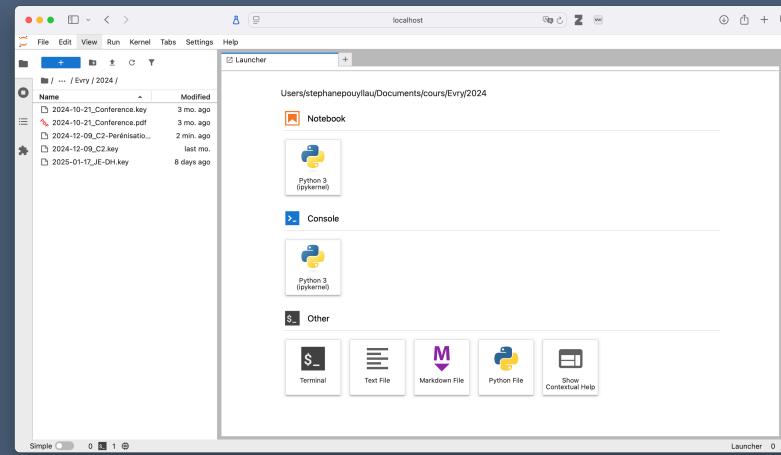
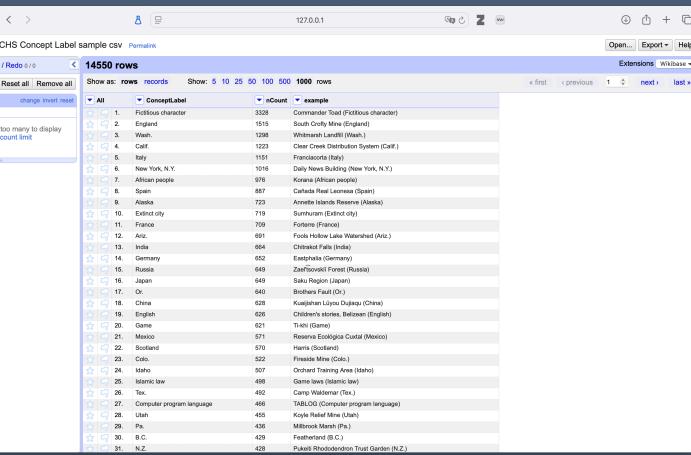
R

<https://rzine.fr>

The screenshot shows a web browser window displaying the Rzine website. The URL 'rzine.fr' is visible in the address bar. The page features a header with the Rzine logo (a blue cube with a white 'R'), navigation links for 'Publications', 'Documentation', 'À propos', and social media icons. Below the header, there's a large image of the Rzine logo and the text: 'Rzine', 'Articles de méthodes pour les Sciences Humaines et Sociales', and 'ISSN 2743-8791 - Revue de documents computationnels'. A descriptive paragraph explains the journal's focus on methods of analysis in SHS using reproducible and didactic software. Another paragraph highlights the use of notebooks for executable articles. A third paragraph mentions open access, peer-reviewed evaluation, and software-based editorial processes. A blue button labeled 'CONSULTER LES ARTICLES PUBLIÉS' is present. The footer contains copyright information: '© Rzine | ISSN 2743-8791 | Mis à jour le 22-01-2025 | Crédits'.

IDE

Integrated development environment



IDE

Integrated development environment

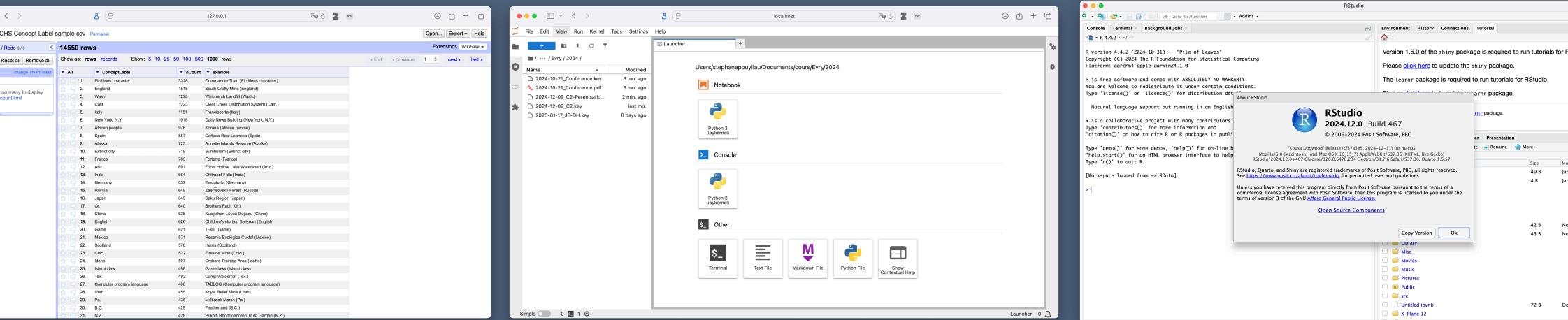
Open Refine : <https://openrefine.org/>

JupyterLab : <https://jupyter.org/>

R Studio : <https://posit.co/>

IDE

Integrated development environment



Démo...

<https://github.com/spouyllau/Cours-Master-Histoire-Evry-2024---C05>

Urbanisation d'outils

Résultante de l'analyse des données

- Définition : Choisir et définir, pour un processus d'analyse et de traitement de données, une chaîne d'outils, dialoguant automatiquement ou pas, afin de procéder à la conception d'un système d'information.

