Archiver ses codes-sources et logiciels : SoftWare Heritage

Julien Caugant

Archiver ses codes-sources et logiciels : SoftWare Heritage

13/12/2023

Julien Caugant
SCD Aix Marseille Université

CELLULE SCIENCE OUVERTE
SCD AMU

Présentation de la Cellule Science Ouverte (1)

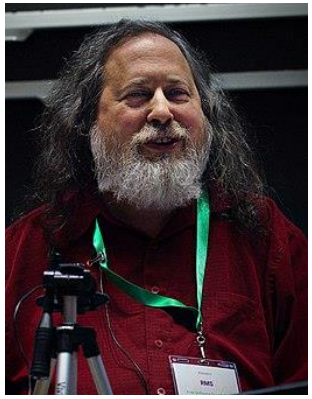
Sommaire

- Contexte
- Pourquoi archiver ses codes et logiciels?
- Comment archiver ses codes et logiciels?
- Points de vigilance

Contexte des codes et logiciels en Science Ouverte

Bref historique sur le logiciel libre

- Octobre 1976 : adoption de la loi du Copyright Act pour les logiciels
- 1984 : création du projet GNU par Richard Stallman, naissance du mouvement du logiciel libre et de la notion de copyleft en opposition au copyright



- 1985 : naissance des licences libres GNU
- 1998 : naissance du mouvement de l’open-source

Déclarations et appels en faveur de la Science Ouverte

Les déclarations et prises de position Initiative de Budapest pour l'Accès Ouvert [14-02-2002]

Déclaration de Berlin sur le Libre Accès à la Connaissance [22-10-2003]

Appel d'Amsterdam : plan d'action pour l'Open Science

Appel de Jussieu

Contexte européen Horizon Europe

European Open Science Cloud (EOSC)

Open Research Europe (ORE)

Coalition S et Plan S

Science Ouverte, données et codes

- 2011 : mission EtaLab, pour l'amélioration de l'action publique grâce aux données et à leur ouverture

« Le code source d'un logiciel détenu par une administration est un document administratif communicable. »

- 2018 : premier PNSO (Plan National pour la Science Ouverte), dont un des axes est de « Structurer et ouvrir les données de la recherche **»**
- 2021 : second PNSO dont le 3ème axe est d'ouvrir et promouvoir les codes sources produits par la recherche

PNSO 2 : axe sur les codes et logiciels

TROISIÈME AXE : OUVRIR ET PROMOUVOIR LES CODES SOURCES PRODUITS PAR LA RECHERCHE

7. **Valoriser et soutenir la diffusion sous licence libre des codes sources** issus de recherches financées sur fonds publics
8. **Mettre en valeur la production des codes sources** de l'enseignement supérieur, de la recherche et de l'innovation
9. Définir et promouvoir une **politique en matière de logiciels libres**

Pourquoi archiver ses codes et logiciels?

Le code : outil ou objet de recherche?

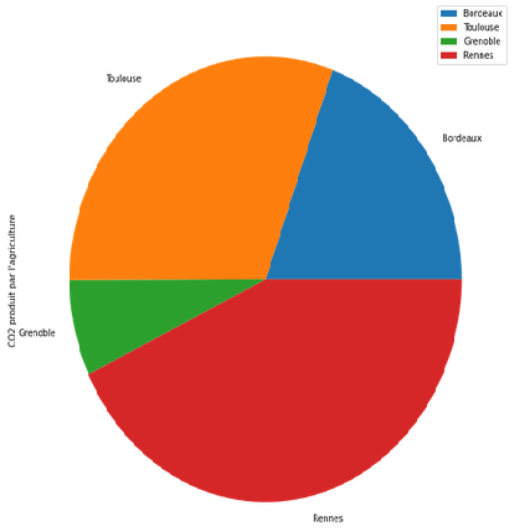
Entrée [12]:

```
# A partir du dataset initial, récupération des valeurs pour 4 villes
data_bordeaux = df.loc[(df["Commune"]=="BORDEAUX")].values[0, 2:12]
data_toulouse = df.loc[(df["Commune"]=="TOULOUSE")].values[0, 2:12]
data_grenoble = df.loc[(df["Commune"]=="GRENOBLE")].values[0, 2:12]
data_rennes = df.loc[(df["Commune"]=="RENNES")].values[0, 2:12]
```

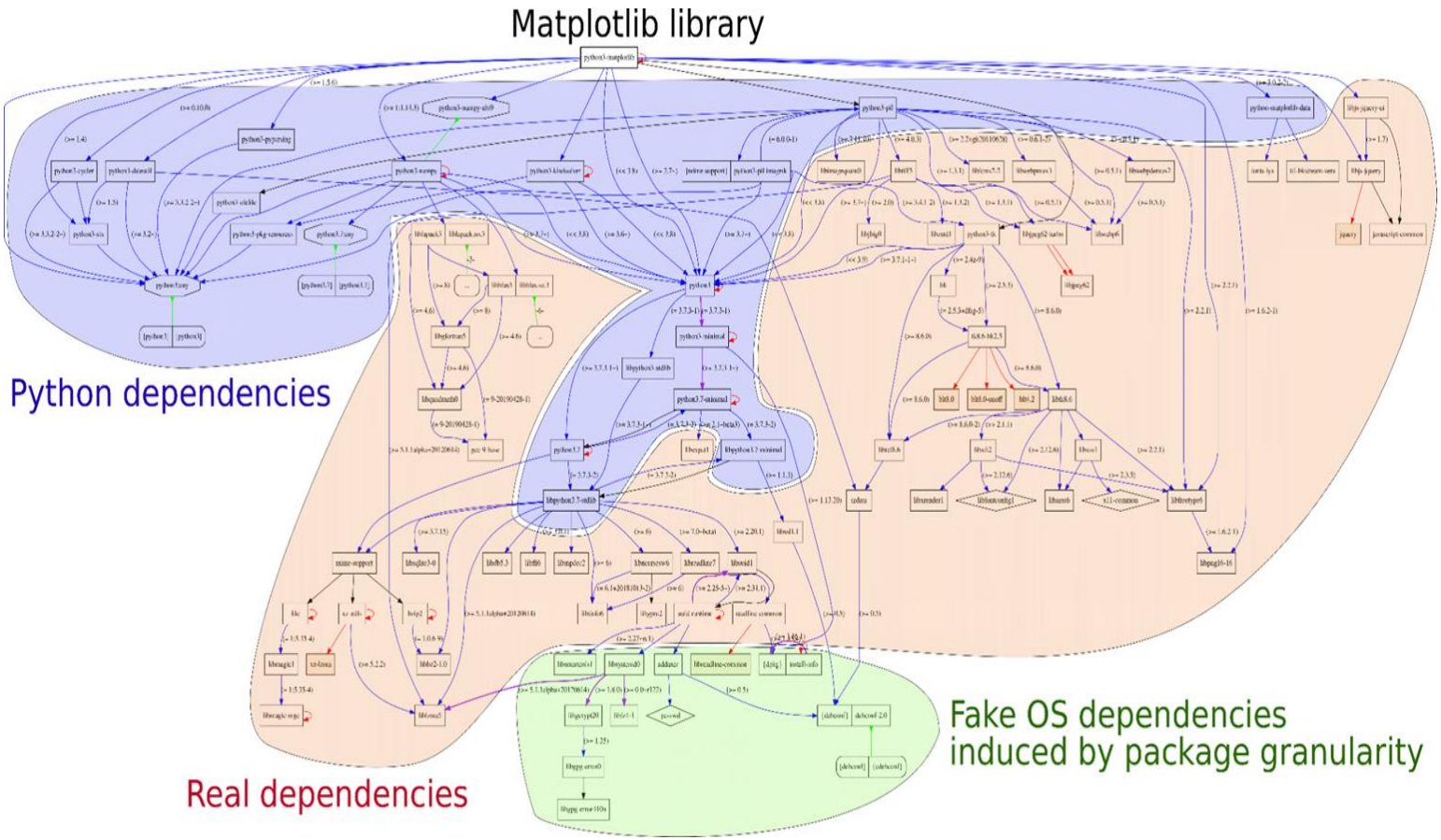
Entrée [13]:

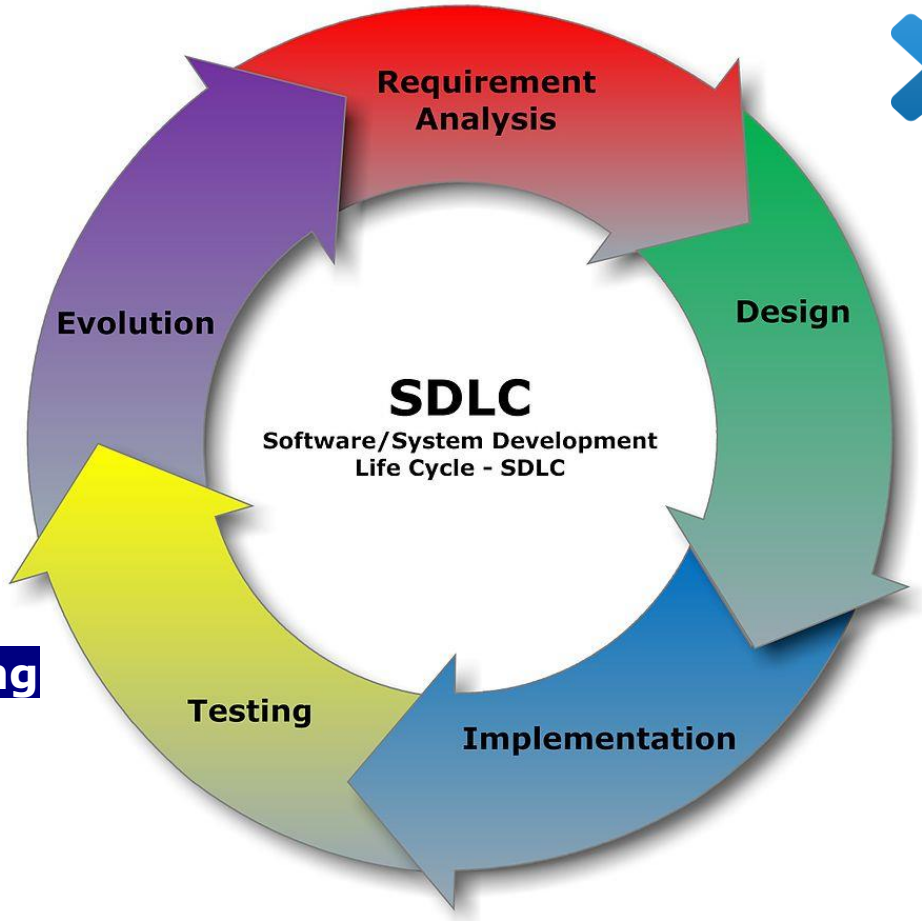
```
df_bis = pd.DataFrame({'CO2 produit par l\'agriculture': [data_borceaux[0], data_toulouse[0] , data_grenoble[0],data_re

                   index=['Bordeaux', 'Toulouse', 'Grenoble', 'Rennes'])
plot = df_bis.plot.pie(y='CO2 produit par l\'agriculture', figsize=(11, 11))
```



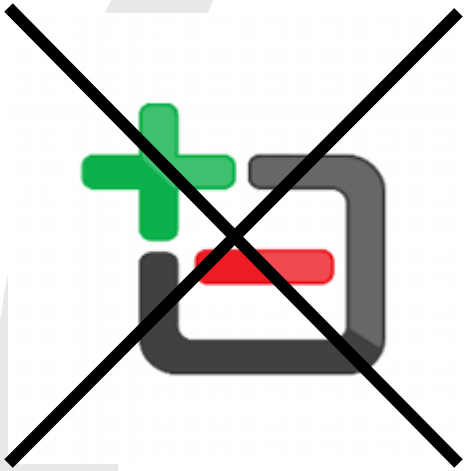
Un code pour les gouverner tous?



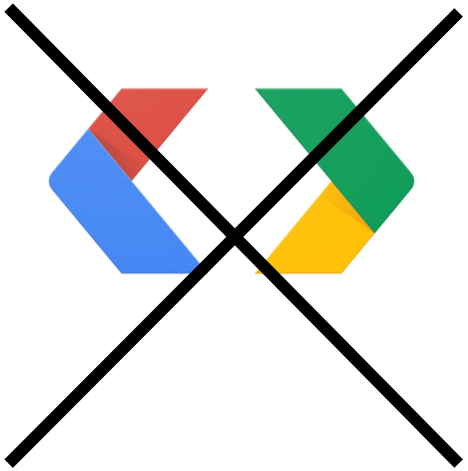


Stockage / Versioning

Le code, une donnée menacée?



2015
Gitorious



25 Janvier 2016
Google Code

2020
Mercurial

Posted: Thursday, March 12, 2015

When we started the Google Code project hosting service in 2006, the world of project hosting was limited. We were worried about reliability and stagnation, so we took action by giving the open source community another option to choose from. Since then, we've seen a wide variety of better project hosting services such as GitHub and Bitbucket bloom. Many projects moved away from Google Code to those other systems. To meet developers where they are, we ourselves migrated nearly a thousand of our own open source projects from Google Code to GitHub.

As developers migrated away from Google Code, a growing share of the remaining projects were spam or abuse. Lately, the administrative load has consisted almost exclusively of abuse management. After profiling non-abusive activity on Google Code, it has become clear to us that the service simply isn't needed anymore.

Beginning today, we have disabled new project creation on Google Code. We will be shutting down the service about 10 months from now on January 25th, 2016. Below, we provide links to migration tools designed to help you move your projects off of Google Code. We will also make ourselves available over the next three months to those projects that need help migrating from Google Code to other hosts.

- March 12, 2015 - New project creation disabled.
- August 24, 2015 - The site goes read-only. You can still checkout/view project source, issues, and wikis.
- January 25, 2016 - The project hosting service is closed. You will be able to download a tarball of project source, issues, and wikis. These tarballs will be available throughout the rest of 2016.

Google will continue to provide Git and Gerrit hosting for certain projects like Android and Chrome. We will also continue maintaining our mirrors of projects like Eclipse, kernel.org and others.

Sunsetting Mercurial support in Bitbucket

April 21, 2020 | 3 min read

Denise Chan

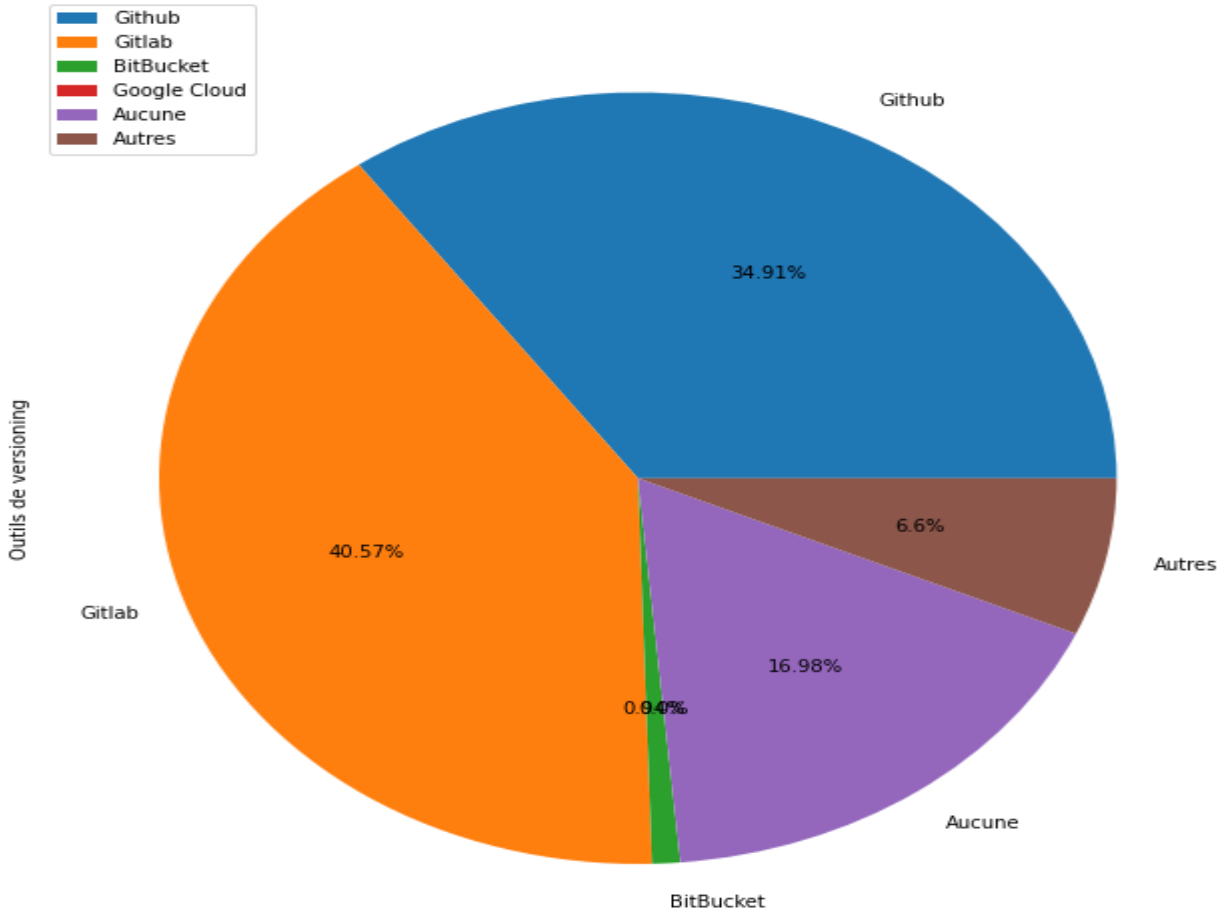
[Update Aug 26, 2020] All hg repos have now been disabled and cannot be accessed.

[Update July 1, 2020] Today, mercurial repositories, snippets, and wikis will turn to read-only mode. After July 8th, 2020 they will no longer be accessible.

Quels intérêts pour le chercheur ou le développeur?

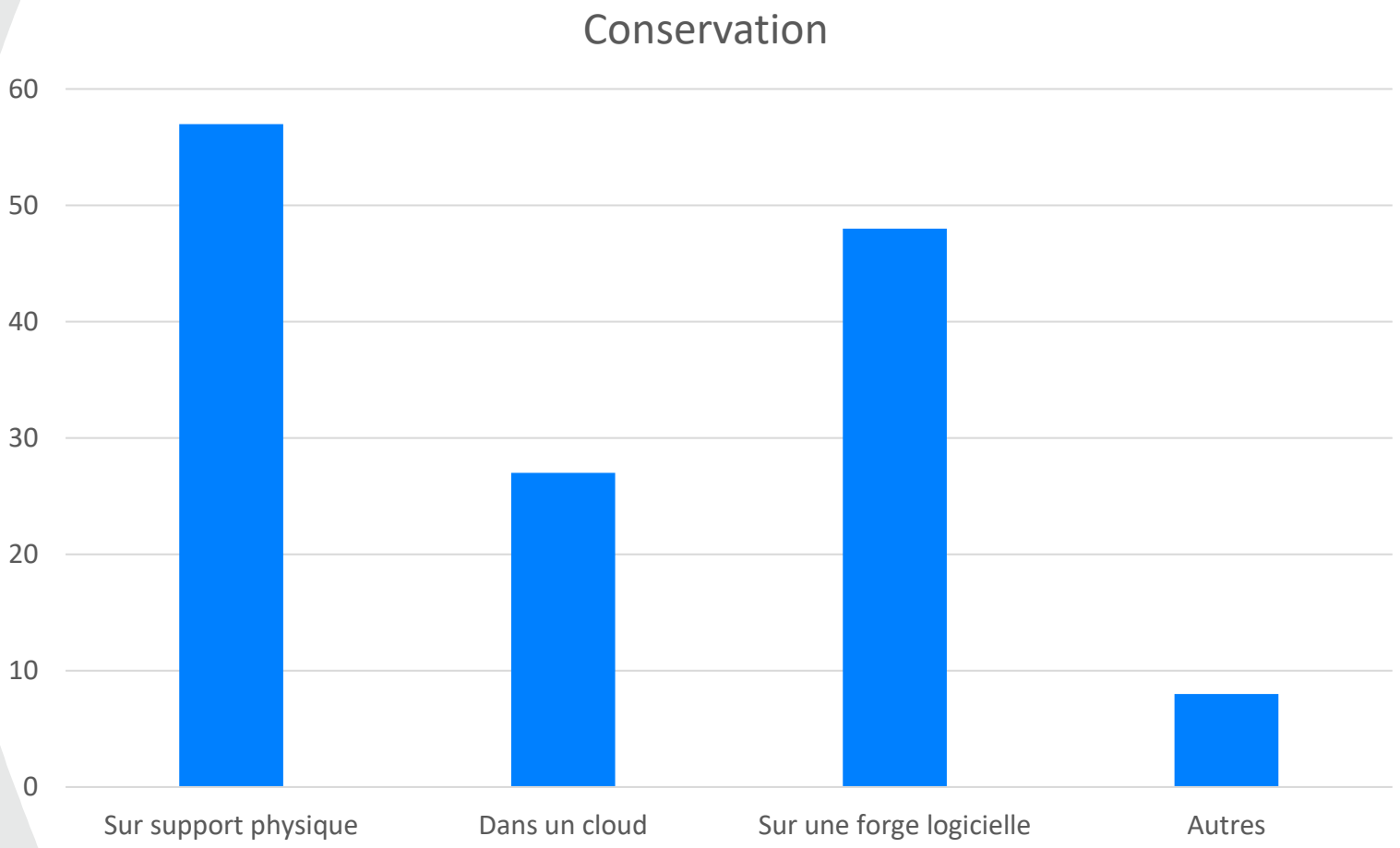
- Conservation pérenne
- Partage du code
- Reproductibilité et transparence
- Citation

Enquête AMU (2023) : Outils de versioning



- **Autres :** SourceSup de RENATER
- SourceForge.net
- Dataverse

Enquête AMU : Conservation du code à long terme



*** Question à choix multiples, plusieurs réponses possibles**

Comment archiver ses codes et logiciels?

SoftWare Heritage

- Initiative à but non lucratif lancée en 2016, avec le soutien de l'INRIA, par Roberto Di Cosmo et Stefano Zacchiroli, en partenariat avec l'UNESCO ainsi qu'avec des acteurs du monde scientifique et de l'industrie (CNRS, Intel, Microsoft…)
- Cette archive contient aujourd'hui plus de 17 milliards de fichiers sources, et plus de 269 millions de projets logiciels

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION

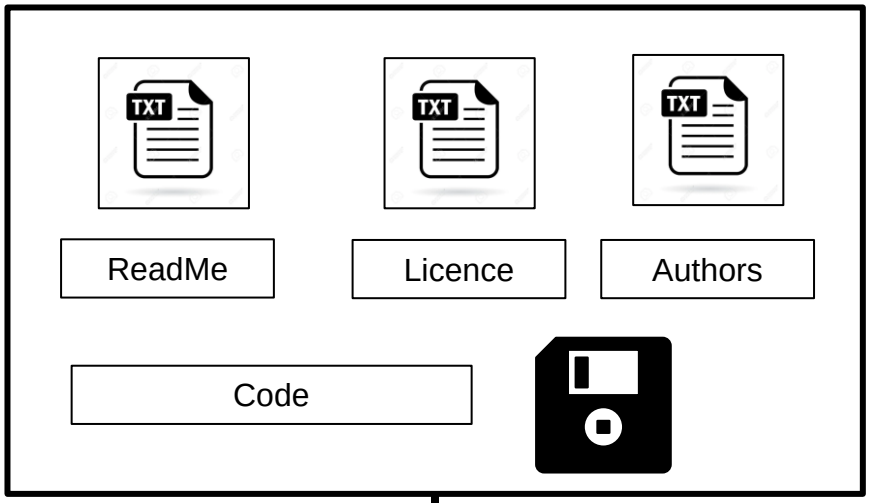
Les missions de SWH

SoftWare Heritage (SWH) a vocation à répondre à 4 défis majeurs : *récolter, organiser, préserver et rendre accessible :*

1. *Récolter : SWH va moissonner les forges logicielles pour collecter et conserver les codes-sources. Il est aussi possible de déposer son code directement dans l'archive SWH ou bien dans HAL*
2. *Organiser : l'archive SWH va conserver le code de manière pérenne et permettre de signaler ce dernier sur HAL. Inversement, on peut déposer son code et le décrire sur HAL, et celui-ci sera transfert vers SWH. Pour permettre ce dépôt, certains fichiers sont nécessaires en supplément de ceux du code*
3. *Préserver : pour garantir un archivage pérenne, SWH repose sur une infrastructure dite distribuée et multipartite*
4. *Rendre accessible : les codes déposés (le code-source et non le code compilé) sur l'archive sont librement consultables et téléchargeables, s'inscrivant pleinement dans le cadre de la politique nationale de Science Ouverte. Une fois déposés, un SWHID leur est attribué afin qu'ils puissent être clairement identifiés et facilement retrouvables. Le dépôt dans HAL permet également de faciliter la recherche et la citation. On peut même citer une partie de code uniquement!*

Bitbucket 2,444,377 origins	56,982 origins	git 23,400 origins
26,217 origins	debian 135,119 origins	38,408 origins
GitHub 187,808,011 origins	gitiles 9,988 origins	GitLab 4,129,099 origins
git 1,707 origins	Gogs 122 origins	GO 874,791 origins
GNU 354 origins	heptapod 1,180 origins	launchpad 497,363 origins
Maven 312,483 origins	NixOS 13,731 origins	npm 3,469,890 origins
4,959 origins	Packagist 297,475 origins	fedora PAGURE 67,588 origins
Phabricator 201 origins	pub.dev 46,611 origins	python 489,946 origins
SOURCEFORGE 381,154 origins	stagit 208 origins	

* Si dépôt via SWH, une fois l'archivage effectué, il faut signaler le dépôt via son SWHID sur HAL



Mon code est-il dans une forge ou un site de gestionnaire de versions de logiciel?

Oui

Je crée mon fichier codemeta.json, et je le dépose sur ma forge avec mon code et les autres fichiers

« Save Code Now » sur SWH et génération du SWHID

Le code est archivé dans SoftWare Heritage*

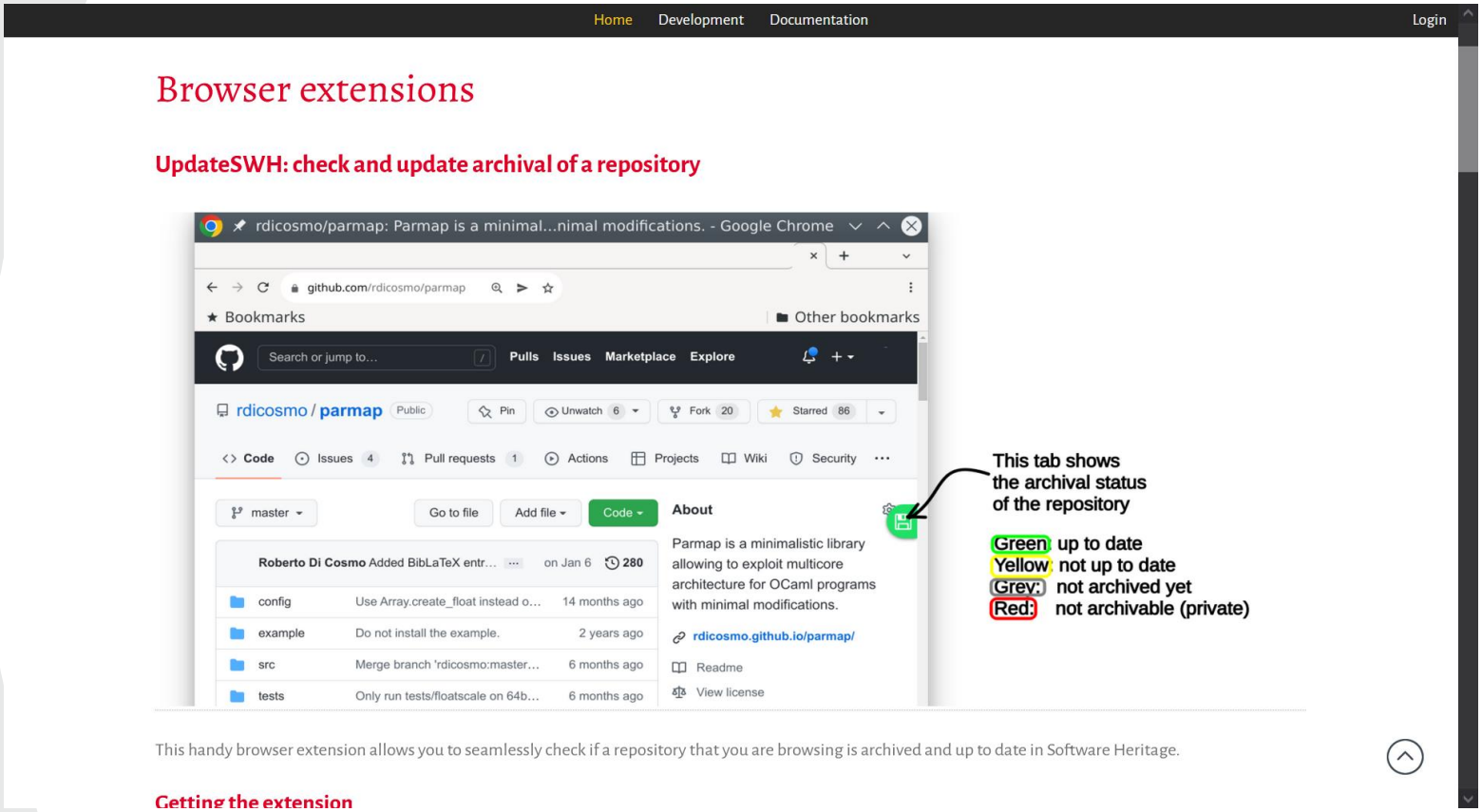
Non

Je place mes fichiers dans une archive (zip, tar…)

Transfert SWH

Dépôt sur HAL de l'archive, et renseignement des métadonnées

Plugin SWH pour faciliter le dépôt



+ Identifiant pérenne pour le code
+ Possibilité de signaler un morceau de code

Points de vigilance

1. Veiller à la clarté et à la précision de son code

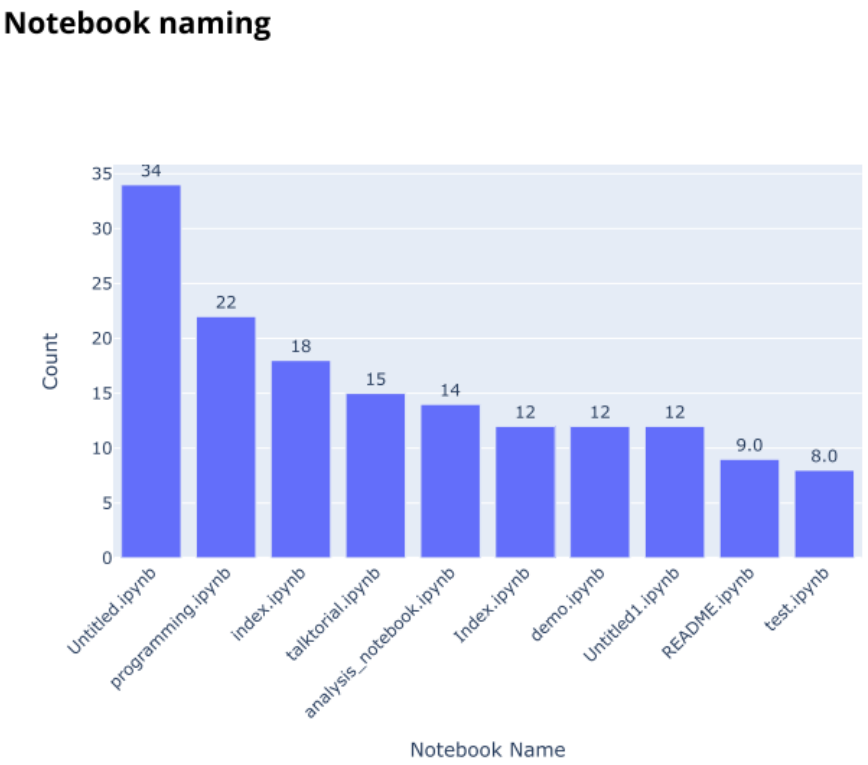
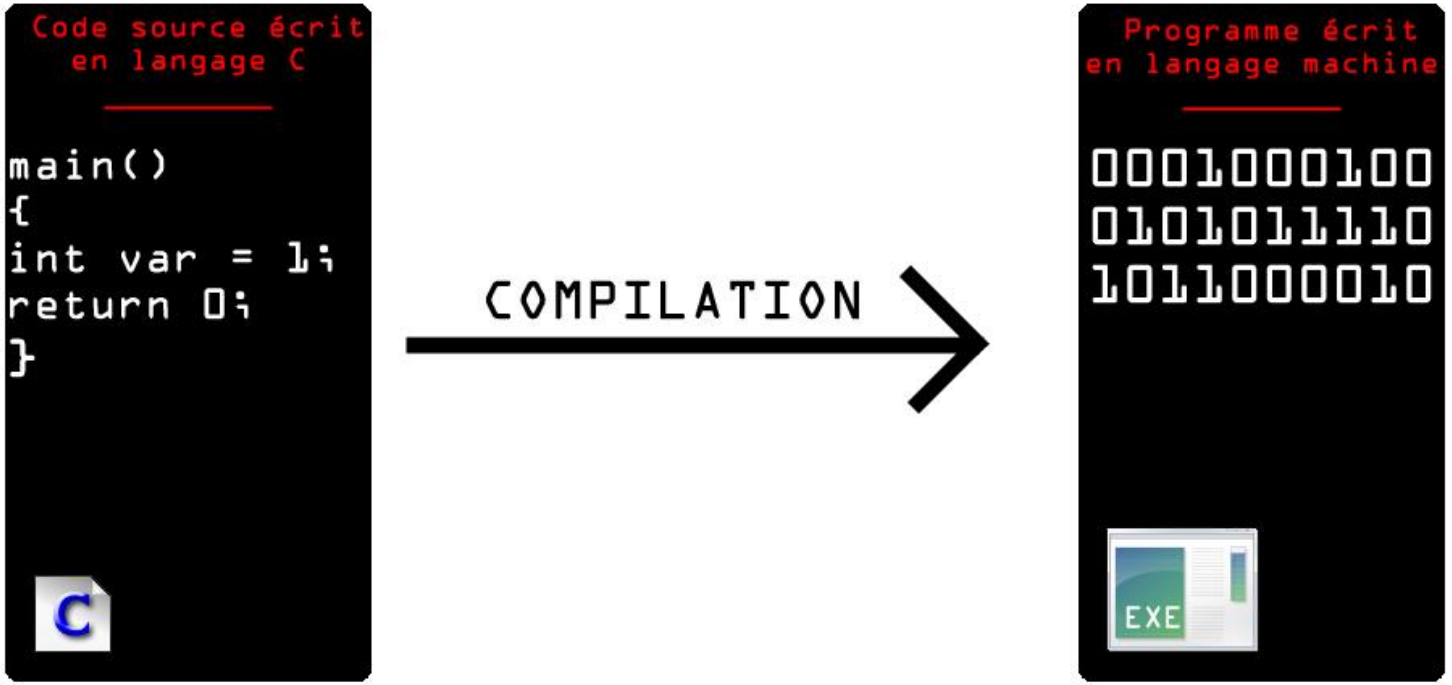


Figure 12. Most frequent notebook titles.

2. Déposer son code-source et non le code compilé



3. Une forge logicielle n’est pas une archive

Hébergement ≠ archivage

Et après?

- Reconnaissance du développement de codes comme activité de recherche (avec dépôt dans une archive)
- Accompagnement des chercheurs lors des procédures de dépôts, et modération locale
- Mise en place d'un réseau de référents?
- Organisation de journées thématiques? D'ateliers virtuels?
- Autres besoins (forge logicielle AMU?)?

Quelques ressources

- Guide de dépôts et bonnes pratiques sur HAL : https://hal.inria.fr/hal-01872189v2

- Obtenir des informations sur les licences logicielles : https://spdx.org/licenses/

- Générer un fichier metadata.json : https://codemeta.github.io/codemeta-generator/

- Bilan de l'enquête sur les pratiques de développement à AMU : https://oaamu.hypotheses.org/4854

Nous contacter : scd-science-ouverte@univ-amu.fr