Software Heritage

The Alexandria Library of Software

Agustín Benito Bethencourt

Toscalix

La Palma Tech Summer 2024

2024–08–29

Table of Content

- About the speaker
- Warm up!
- The mission
- About Software Heritage (SwH)
- The archive
- Collaborate!
- Takeaway

The speaker

Agustín Benito Bethencourt
@toscalix

http://www.toscalix.com

- SwH Ambassador since 2023
- Independent consultant helping companies in two ways:
  - Applying advanced data analytics to production environments to increase delivery performance, partnering with Bitergia, through a service offering called Delivery Performance Analytics
  - Increasing their organizational performance by becoming good open source citizens, like in the case for SCANOSS, as their Ecosystem Manager
- FLOSS, Continuous Delivery, metrics and remote work advocate.
- Background: Eclipse Foundation, MBition (Mercedes Benz), Codethink, Linaro, SUSE, ASOLIF, entrepreneur …
- Blog - About - Talks - Contact

Warm up!

SwH archive

Browse & Search	https://archive.softwareheritage.org/browse/
Save Code Now	https://archive.softwareheritage.org/save/
Add Forge Now	https://archive.softwareheritage.org/add-forge/request/create
Download	https://archive.softwareheritage.org/vault/

SwH gem

SwH browser extension

https://www.softwareheritage.org/browser-extensions

The Mission

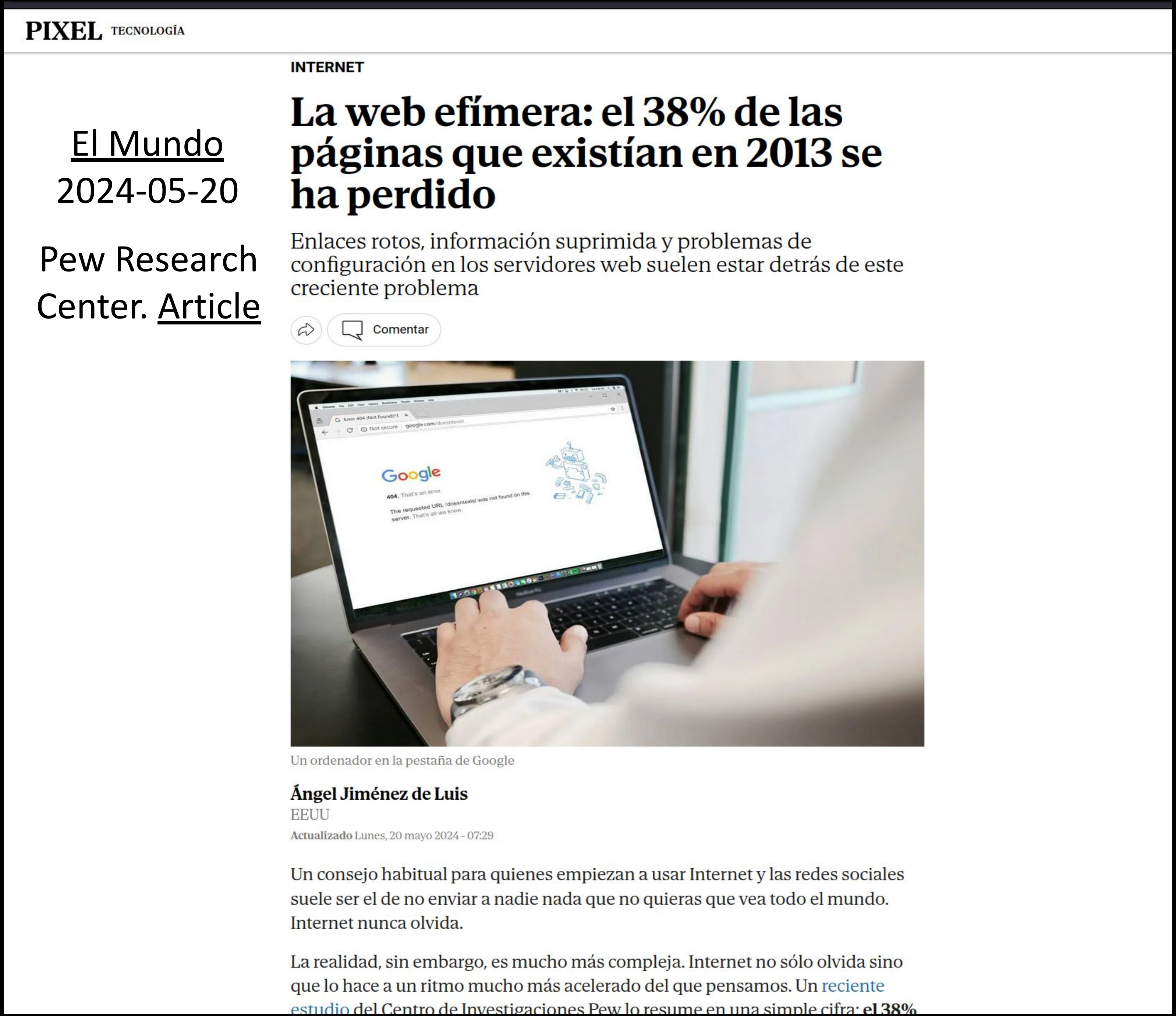
Why

- Software is a key pillar for modern science
- Every industry nowadays rely on software
- There is little/no usable software without open source software
- Source code is the foundational building block in open source software, so in software… so in science

Why

El Mundo
2024-05-20

Pew Research Center. Article



PIXEL TECNOLOGÍA

INTERNET

La web efímera: el 38% de las páginas que existían en 2013 se ha perdido

Enlaces rotos, información suprimida y problemas de configuración en los servidores web suelen estar detrás de este creciente problema

Comentar

Google
404. That's an error.
The requested URL /doesntexist was not found on this server. That's all we know.

Un ordenador en la pestaña de Google

Ángel Jiménez de Luis
EEUU
Actualizado Lunes, 20 mayo 2024 - 07:29

Un consejo habitual para quienes empiezan a usar Internet y las redes sociales suele ser el de no enviar a nadie nada que no quieras que vea todo el mundo. Internet nunca olvida.

La realidad, sin embargo, es mucho más compleja. Internet no sólo olvida sino que lo hace a un ritmo mucho más acelerado del que pensamos. Un reciente estudio del Centro de Investigaciones Pew lo resume en una simple cifra: **el 38%**

Endangered source code …

- **Link rot**: projects are created, moved around, removed
- **Data rot**: physical media with legacy software decay
- **Platform** consolidation endangers repositories
  - 2015 Google Code and Gitorious.org shutdown: ~1M
  - 2019 Bitbucket mercurial phase out: ~250.000
  - 2022 GitLab.com: remove inactive projects

… is endangered knowledge!

We cannot afford broken links and
missing pieces in the web of knowledge
of humankind

Why

SwH’s Mission

“Cultural heritage is the legacy of physical artifacts and intangible attributes of a group or society that are inherited from past generations, maintained in the present and bestowed for the benefit of future generations.

Software in source code form is produced by humans and is understandable by them; as such it is an important part of our heritage that we should not lose. Software is furthermore a key enabler for preserving other parts of our cultural heritage that we would *de facto* lose if we lose the software needed to access them. **Preserving software is essential for preserving our cultural heritage**.”

We need a **global**, **long term** effort to build a **universal** archive of all software source code. Make it **resilient** and make it **sustainable**.

What

About Software Heritage (SwH)

SwH is an organization

Software Heritage is an open, non-profit initiative unveiled in 2016 by Inria. It is supported by a broad panel of institutional and industry partners, in collaboration with UNESCO [1] [2]

Hosted by

Inria

Inria La Fondation

In collaboration with

unesco

Read the 2023 annual report and the 2024 roadmap

UNESCO

United Nations
Educational, Scientific and
Cultural Organization

Paris Call

«Software source code represents unique knowledge of humanity's recent history.

It is therefore crucial to work together collectively so that the knowledge embedded in software source code is properly preserved, valued and shared with all.

This lies at the core of UNESCO's cooperation with Inria to support the creation of Software Heritage, the global archive of software source code»

The Core Team

The archive

1. **Reference Catalogue**: find and reference all software source code
2. **Universal archive**: preserve and share all software source code
3. **Research infrastructure**: enable analysis of all software source code

SwH is a community effort

Users

- FAQ: https://www.softwareheritage.org/faq/
- Mailing list: swh-users
- Matrix channel #swh at matrix.org

Contributors

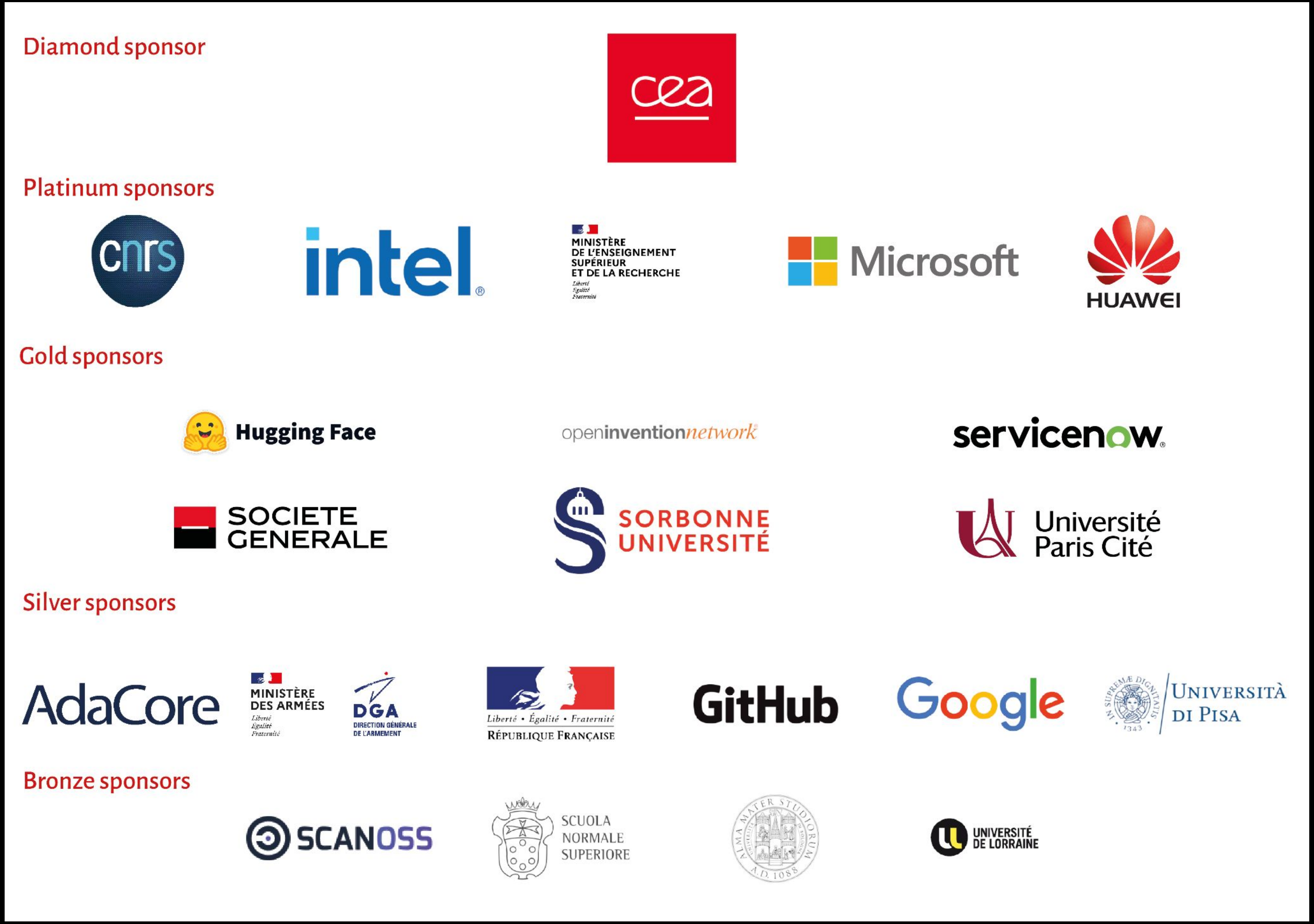
- Forge: https://gitlab.softwareheritage.org/
- Documentation: https://docs.softwareheritage.org/devel/
- Mailing list: swh-devel
- Chat: #swh-devel at matrix.org
- Wiki: https://wiki.softwareheritage.org/

Ambassadors and Students

- https://www.softwareheritage.org/ambassadors
- https://www.softwareheritage.org/community/students

The community

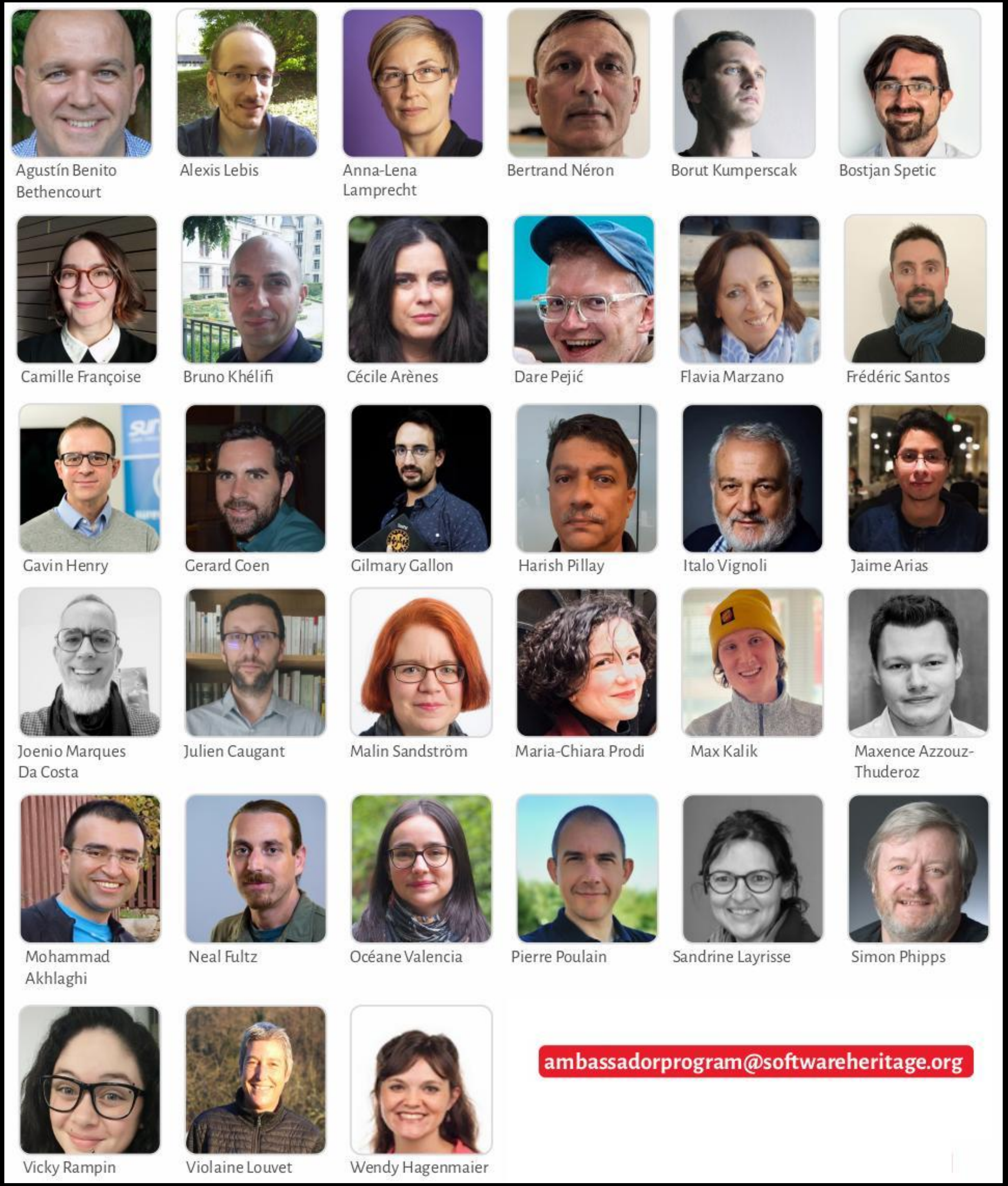
SwH builds an ecosystem



Testimonials

https://www.softwareheritage.org/support/testimonials/

The Ambassadors



The archive

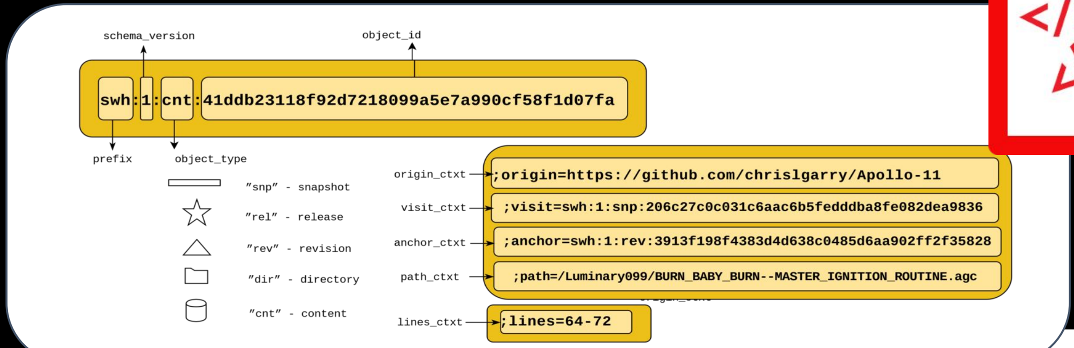
- Source code is…
    - designs, algorithms, code, tests, documentation, configurations, licenses, etc.
- Software *evolves* over time
    - projects may last decades
    - the development history is key to its understanding
- Complexity
    - millions of lines of code
    - large web of dependencies: easy to break, difficult to maintain
    - sophisticated *developer communities*

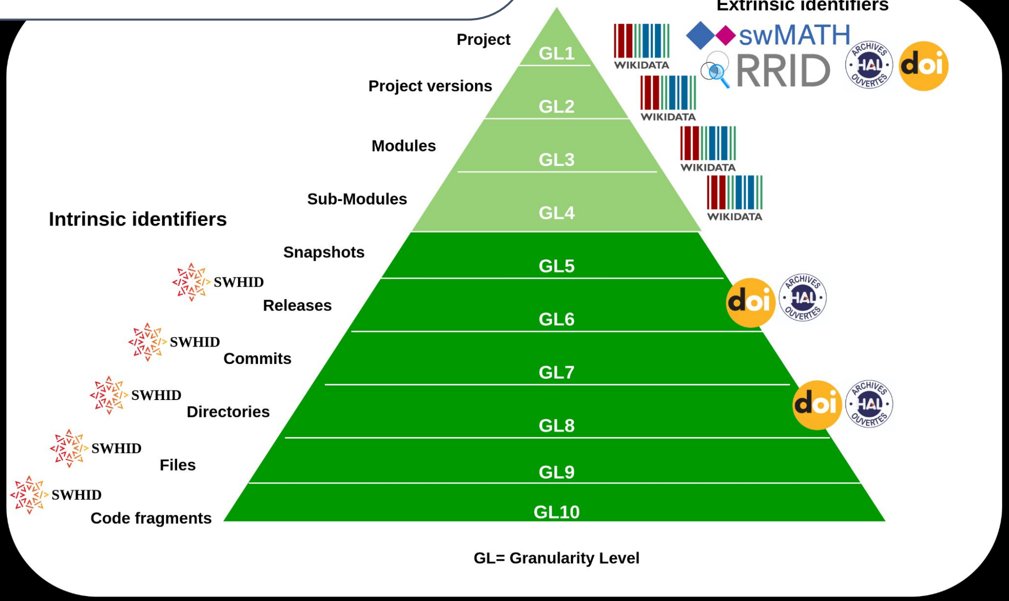
Source code is *unique*, it is *not* data

Reference Catalogue

SWHID: https://www.swhid.org/

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE
SWHID

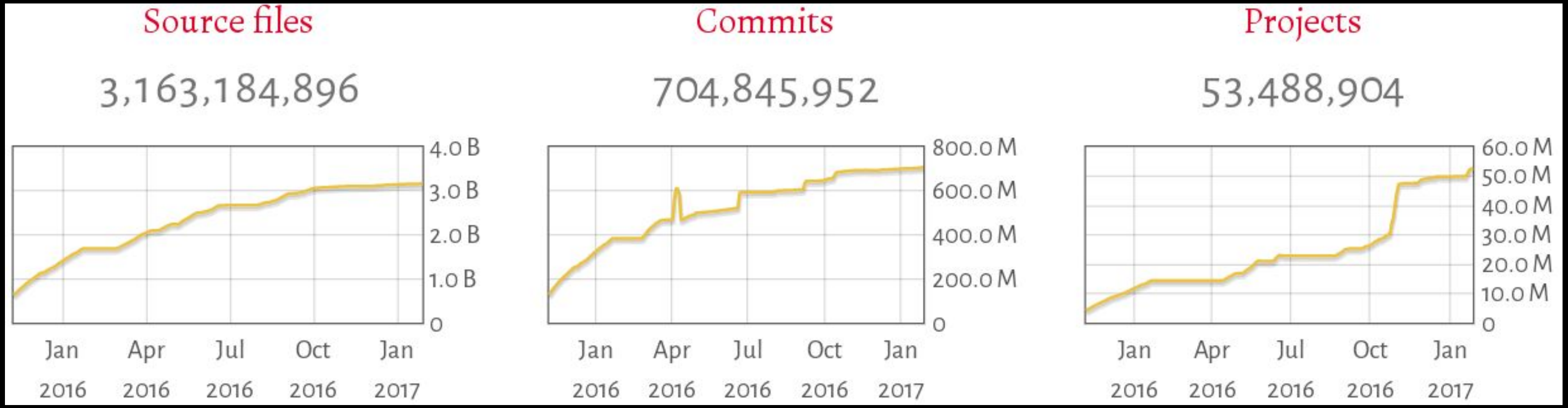

schema_version
object_id
swh:1:cnt:41ddb23118f92d7218099a5e7a990cf58f1d07fa
prefix
object_type
"snp" - snapshot
"rel" - release
"rev" - revision
"dir" - directory
"cnt" - content
origin_ctxt
;origin=https://github.com/chrislgarry/Apollo-11
visit_ctxt
;visit=swh:1:snp:206c27c0c031c6aac6b5fedddba8fe082dea9836
anchor_ctxt
;anchor=swh:1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828
path_ctxt
;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc
lines_ctxt
;lines=64-72



GL= Granularity Level

Universal archive

Nov’16



Source files

3,163,184,896

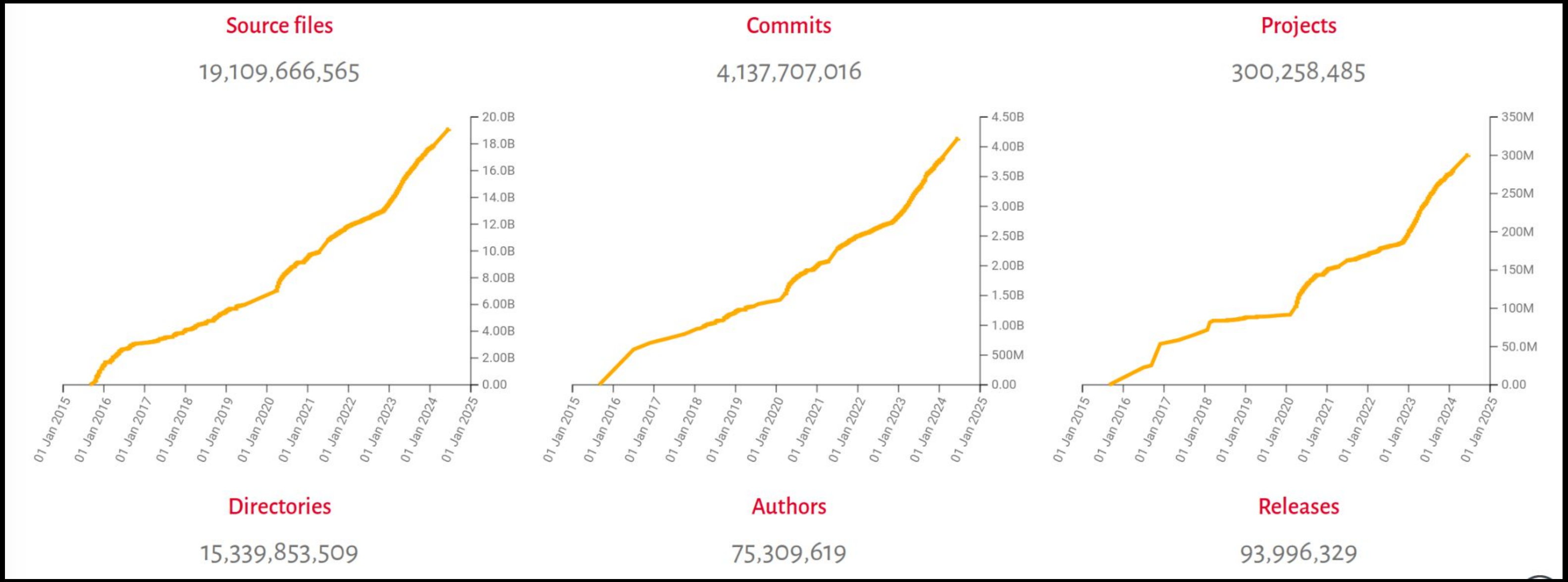
Commits

704,845,952

Projects

53,488,904

Jun’24



Source files

19,109,666,565

Commits

4,137,707,016

Projects

300,258,485

Directories

15,339,853,509

Authors

75,309,619

Releases

93,996,329

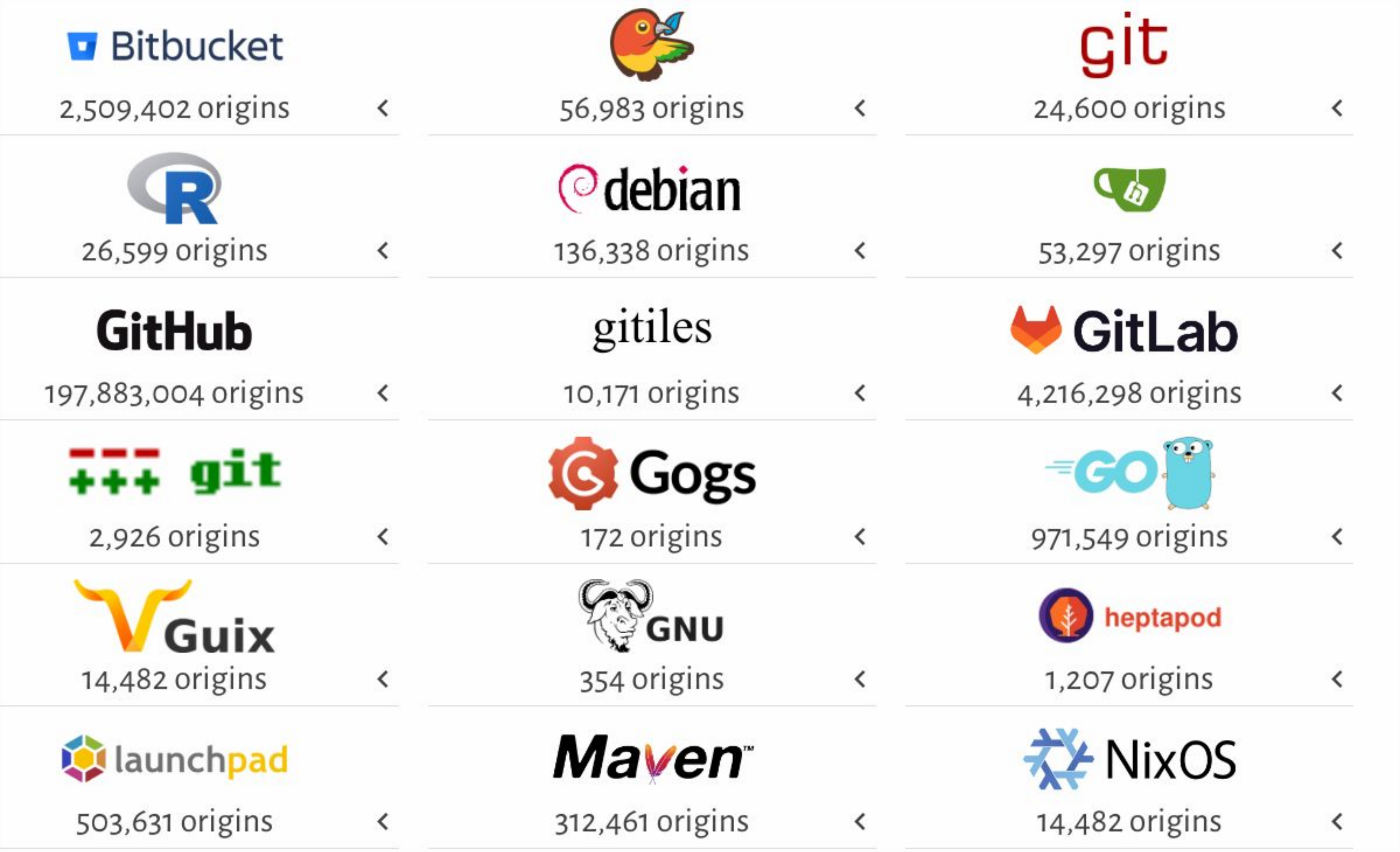
Harvest and archive:

- docs.softwareheritage.org/#landing-preserve
- save.softwareheritage.org
- deposit.softwareheritage.org
- softwareheritage.org/swhap/

Howto:

- HOWTO archive and reference your code
- Unlock the Power of Software Heritage Archive

Universal archive

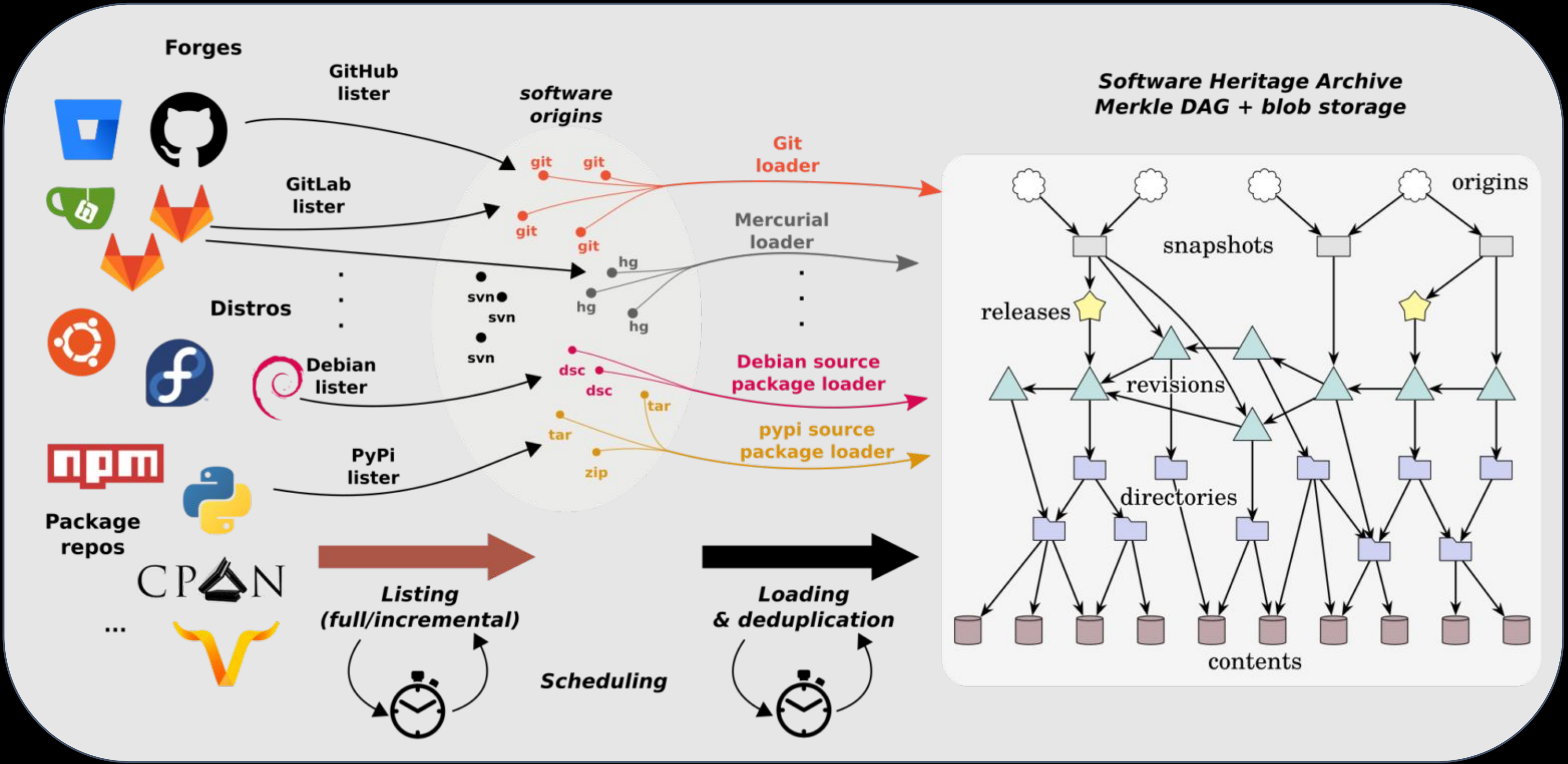


Bitbucket 2,509,402 origins	56,983 origins	git 24,600 origins
R 26,599 origins	debian 136,338 origins	53,297 origins
GitHub 197,883,004 origins	gitiles 10,171 origins	GitLab 4,216,298 origins
git 2,926 origins	Gogs 172 origins	GO 971,549 origins
Guix 14,482 origins	GNU 354 origins	heptapod 1,207 origins
launchpad 503,631 origins	Maven 312,461 origins	NixOS 14,482 origins

(*) The numbers are never up to date

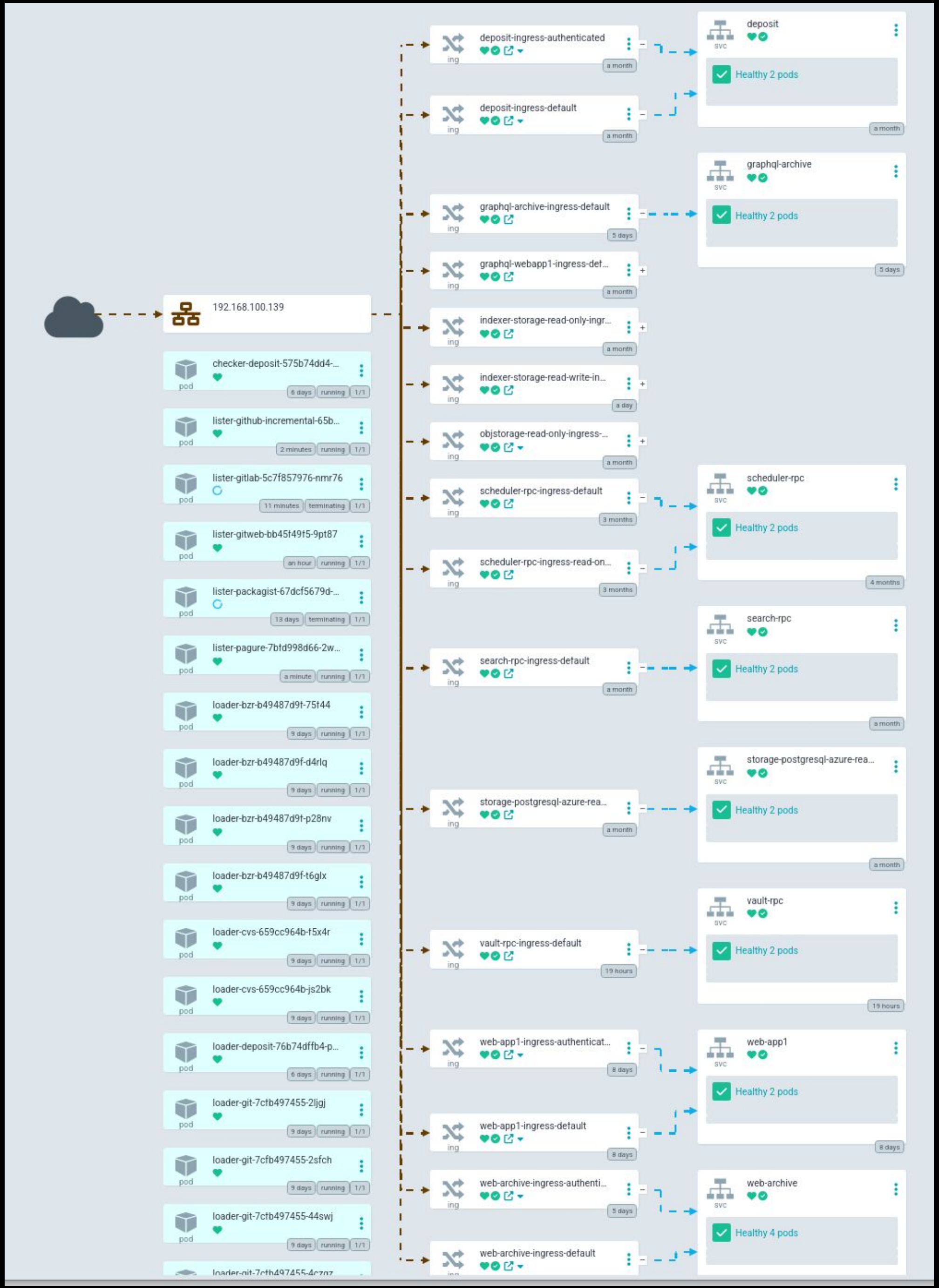
Truly universal

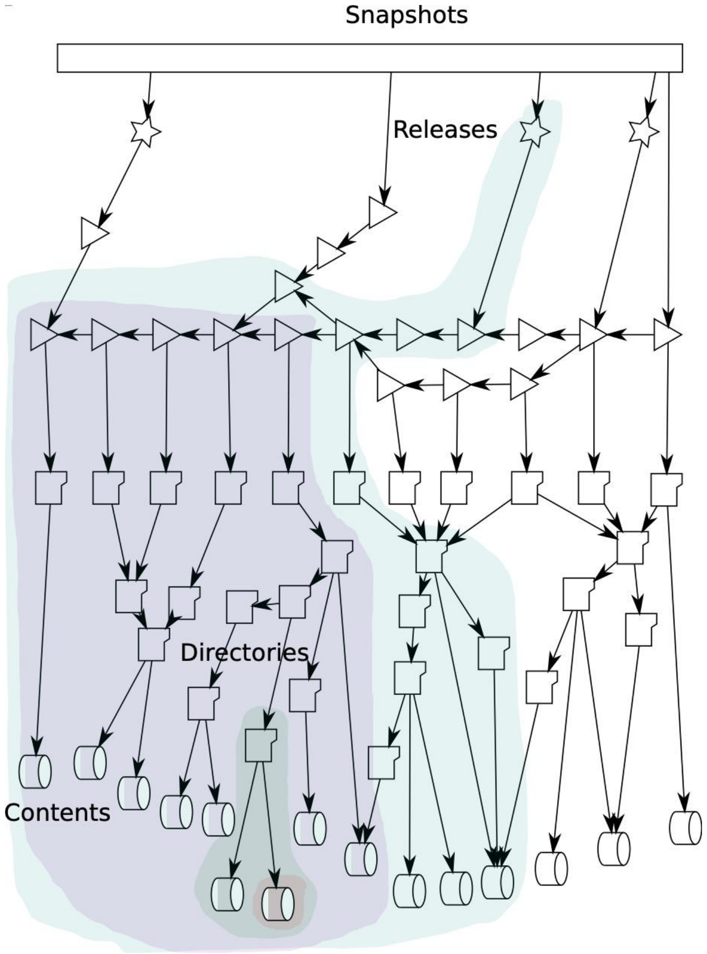
Built for the purpose so…



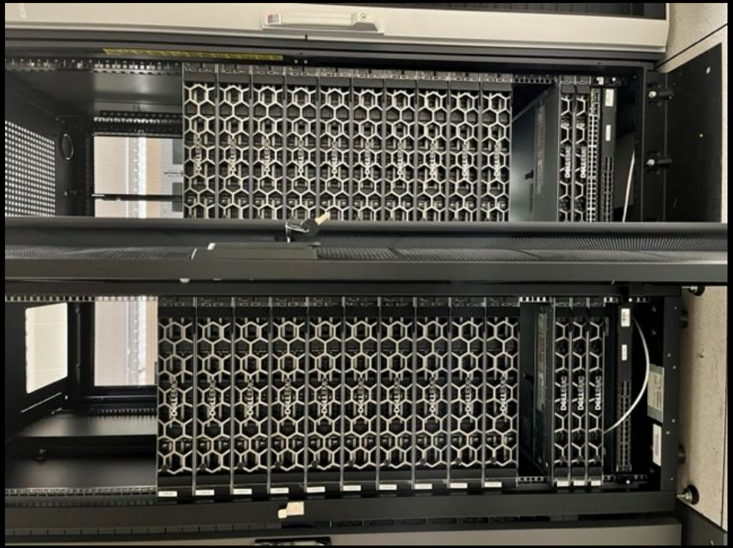
… complex

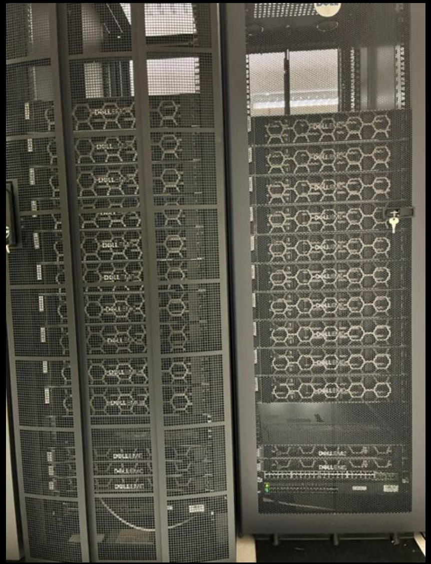
and innovative


192.168.100.139
checker-deposit-575b74dd4-… pod 6 days running 1/1
lister-github-incremental-65b… pod 2 minutes running 1/1
lister-gitlab-5c7f857976-nmr76 pod 11 minutes terminating 1/1
lister-gitweb-bb45f49f5-9pt87 pod an hour running 1/1
lister-packagist-67dcf5679d-… pod 13 days terminating 1/1
lister-pagure-7bfd998d66-2w… pod a minute running 1/1
loader-bzr-b49487d9f-75f44 pod 9 days running 1/1
loader-bzr-b49487d9f-d4rlq pod 9 days running 1/1
loader-bzr-b49487d9f-p28nv pod 9 days running 1/1
loader-bzr-b49487d9f-t6glx pod 9 days running 1/1
loader-cvs-659cc964b-f5x4r pod 9 days running 1/1
loader-cvs-659cc964b-js2bk pod 9 days running 1/1
loader-deposit-76b74dffb4-p… pod 6 days running 1/1
loader-git-7cfb497455-2ljgj pod 9 days running 1/1
loader-git-7cfb497455-2sfch pod 9 days running 1/1
loader-git-7cfb497455-44swj pod 9 days running 1/1
loader-git-7cfb497455-4czqz
deposit-ingress-authenticated ing a month
deposit-ingress-default ing a month
graphql-archive-ingress-default ing 5 days
graphql-webapp1-ingress-def… ing a month
indexer-storage-read-only-ingr… ing a month
indexer-storage-read-write-in… ing a day
objstorage-read-only-ingress-… ing a month
scheduler-rpc-ingress-default ing 3 months
scheduler-rpc-ingress-read-on… ing 3 months
search-rpc-ingress-default ing a month
storage-postgresql-azure-rea… ing a month
vault-rpc-ingress-default ing 19 hours
web-app1-ingress-authenticat… ing 8 days
web-app1-ingress-default ing 8 days
web-archive-ingress-authenti… ing 5 days
web-archive-ingress-default ing
deposit svc Healthy 2 pods a month
graphql-archive svc Healthy 2 pods 5 days
scheduler-rpc svc Healthy 2 pods 4 months
search-rpc svc Healthy 2 pods a month
storage-postgresql-azure-rea… svc Healthy 2 pods a month
vault-rpc svc Healthy 2 pods 19 hours
web-app1 svc Healthy 2 pods 8 days
web-archive svc Healthy 4 pods



SwH is building a full graph of software development evolution

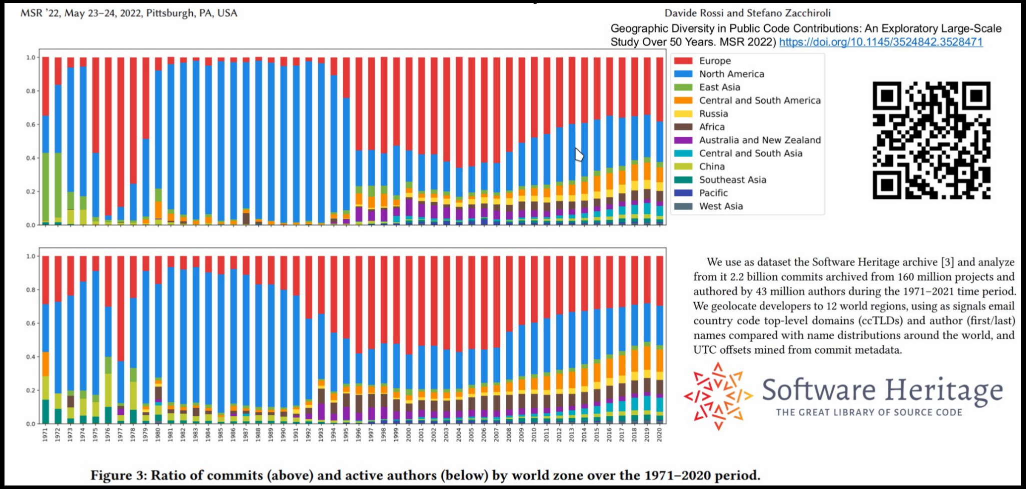




Self-hosted

Research infrastructure

- Check all the available publications
- "Rust Analytics for Software Heritage: Challenges and results" by Sebastiano Vigna, full professor, Università degli Studi di Milano
- "Open and responsible development of Large Language Models for code" by BigCode-project.org leaders: Leandro von Werra and Harm de Vries



MSR '22, May 23–24, 2022, Pittsburgh, PA, USA

Davide Rossi and Stefano Zacchiroli

Geographic Diversity in Public Code Contributions: An Exploratory Large-Scale Study Over 50 Years. MSR 2022) https://doi.org/10.1145/3524842.3528471

We use as dataset the Software Heritage archive [3] and analyze from it 2.2 billion commits archived from 160 million projects and authored by 43 million authors during the 1971–2021 time period. We geolocate developers to 12 world regions, using as signals email country code top-level domains (ccTLDs) and author (first/last) names compared with name distributions around the world, and UTC offsets mined from commit metadata.

Figure 3: Ratio of commits (above) and active authors (below) by world zone over the 1971–2020 period.

Show me the archive!

Search:	https://archive.softwareheritage.org/browse
Save (web):	https://archive.softwareheritage.org/save/
Save (plugin):	https://www.softwareheritage.org/browser-extensions
Download:	https://archive.softwareheritage.org/vault/

Collaborate!

Collaborate!

Use the archive

Follow us on social media: Fediverse, X, Linkedin, Youtube

Spread the word

Archive your software

Become an Ambassador

Collaborate!

Join our research community

Code with us

Work with us

Become a sponsor

Donate

Resilience

“ . . . let us save what remains: not by vaults and locks which fence them from the public eye and use in consigning them to the waste of time, but by such a multiplication of copies, as shall place them beyond the reach of accident.”

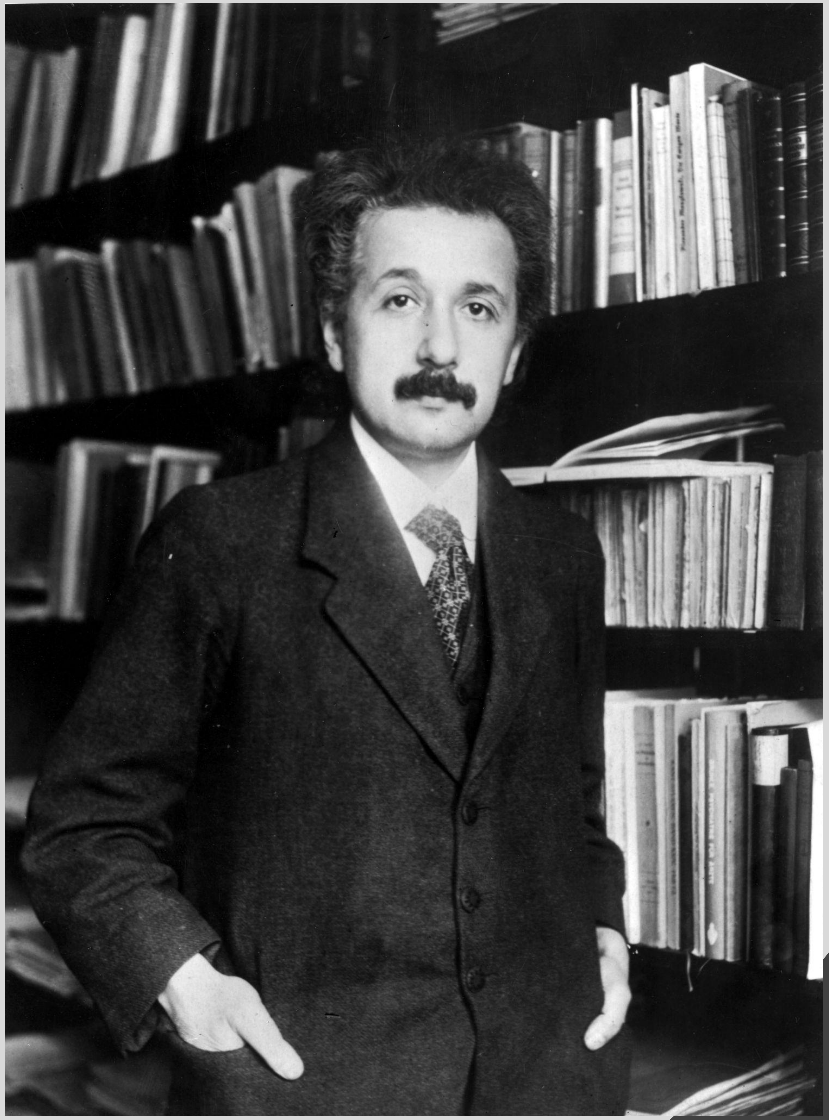
Thomas Jefferson, February 18, 1791

Join our mirror network!

https://www.softwareheritage.org/mirrornetwork/

ENEA

Takeaway



Downloaded from Getty Images

The only thing that you absolutely have to know, is the location of ~~the library~~”.-*Albert Einstein*

SwH

archive.softwareheritage.org

Thank you

Software Heritage

The Alexandria Library of Software

Agustín Benito Bethencourt

Toscalix

La Palma Tech Summer 2024

2024-08-29