Introduction to Software Heritage (SwH)

The Alexandria library for code

Agustín Benito Bethencourt

Toscalix

16, April, 2024 FOSSNorth 2024

Table of Content

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

The speaker

Agustín Benito Bethencourt
@toscalix

http://www.toscalix.com

- SwH Ambassador since 2023
- Current activity: independent consultant
  - Delivery Performance Analytics helping organizations. in:
    - Establishing a BI strategy applied to the production of software defined products.
    - Optimising SDLC processes by applying state of the art principles and practices.
  - Ecosystem manager at SCANOSS
- FLOSS, agility, Continuous Delivery, metrics and remote work advocate.
- Background: Eclipse Foundation, MBition (Mercedes Benz), Codethink, Linaro, SUSE, ASOLIF, entrepreneur …
- Blog - About - Talks - Contact

The Mission

Why

- Software is a key pillar for science
- There is no open science without OSS.
- This is also true for most industries nowadays
- Source code is the foundational building block in OSS.
- Proprietary code beyond EoL is worth preserving (previously made open source or public domain).

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

The 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

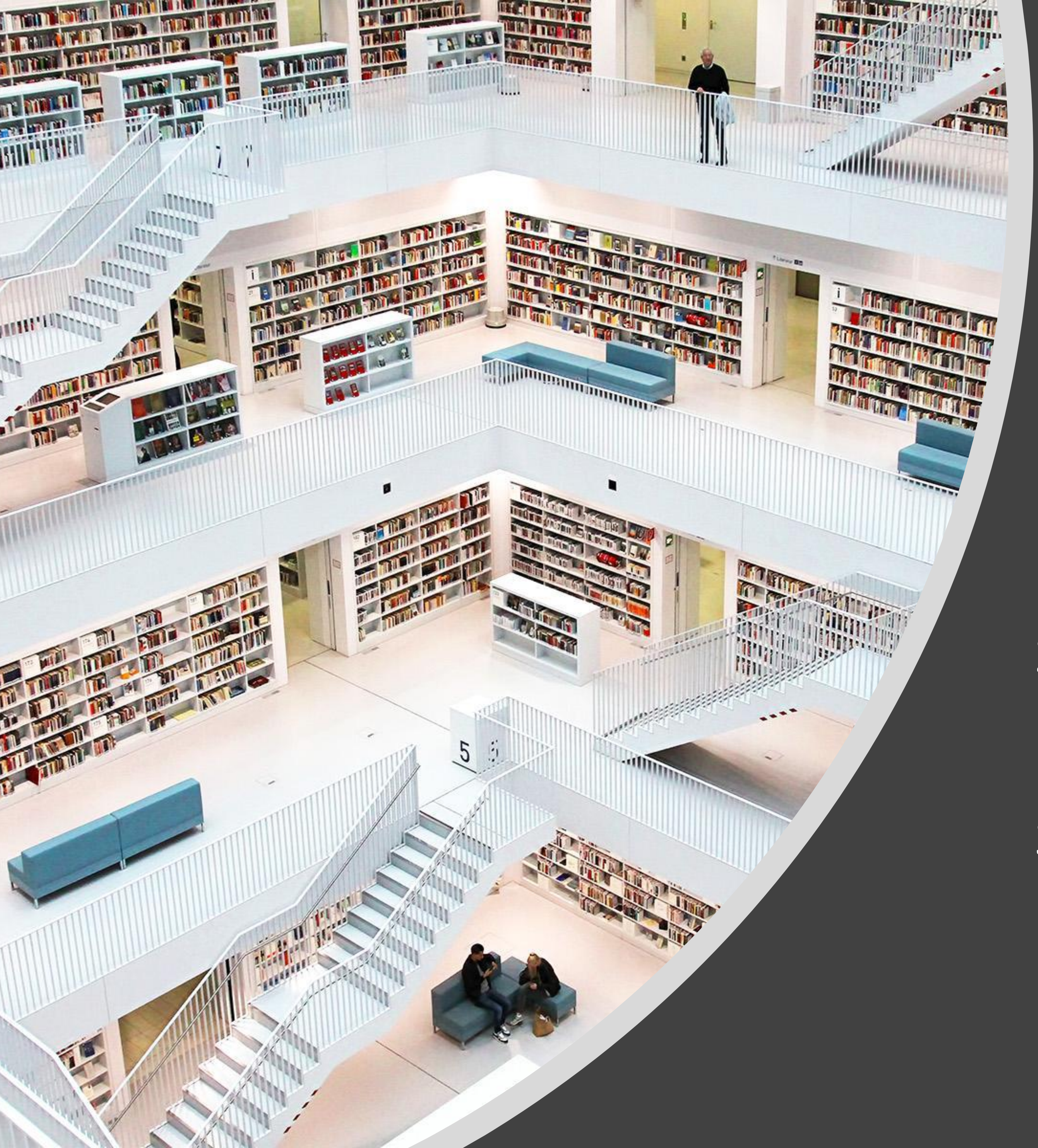
Paris Call

«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

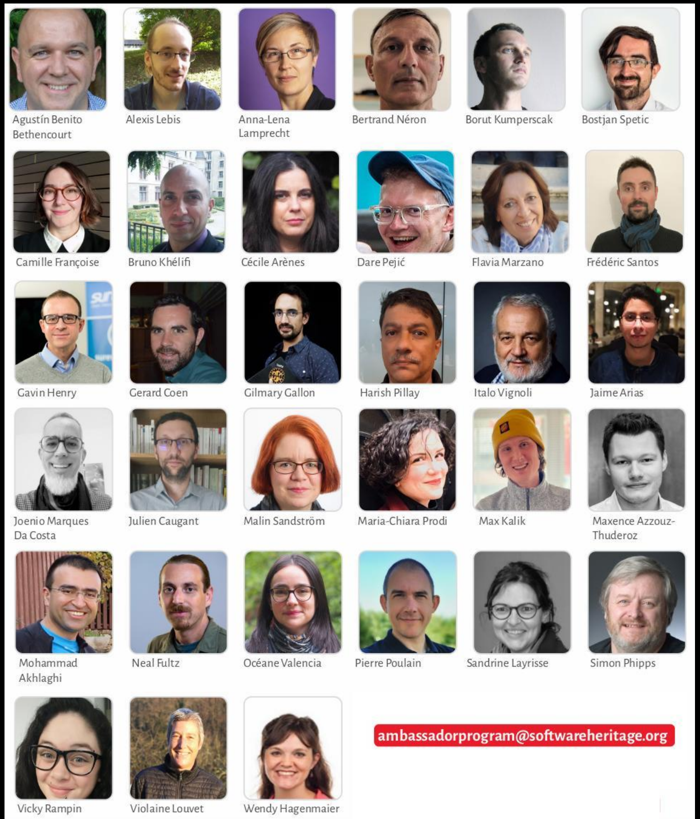
Contributors

- 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 Ambassadors



"Pursuing our roadmap for the archive requires significant resources. We welcome companies, institutions, and individuals who would like to join our sponsorship program and sustain the Software Heritage."

The community

SwH builds an ecosystem

Diamond sponsor

cea

Platinum sponsors

cnrs

intel

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

Microsoft

HUAWEI

Gold sponsors

Hugging Face

openinventionnetwork

servicenow

SOCIETE GENERALE

SORBONNE UNIVERSITÉ

Université Paris Cité

Silver sponsors

AdaCore

MINISTÈRE DES ARMÉES

DGA DIRECTION GÉNÉRALE DE L'ARMEMENT

Liberté • Égalité • Fraternité RÉPUBLIQUE FRANÇAISE

GitHub

Google

UNIVERSITÀ DI PISA

Bronze sponsors

SCANOSS

SCUOLA NORMALE SUPERIORE

ALMA MATER STUDIORUM A.D. 1088

UNIVERSITÉ DE LORRAINE

Testimonials

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

The archive

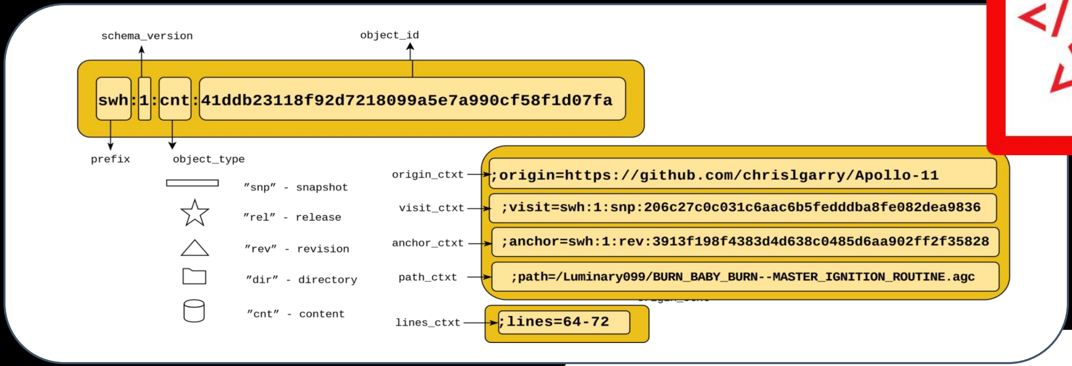
- Software *evolves* over time
  - projects may last decades
  - the development history is key to its understanding
- Complexity
  - millions of lines of code
  - large web of dependencies
    - easy to break, difficult to maintain
    - *research software* a thin top layer
  - sophisticated *developer communities*
- The human side
  - design, algorithm, code, test, documentation, community, funding and so many more facets …

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

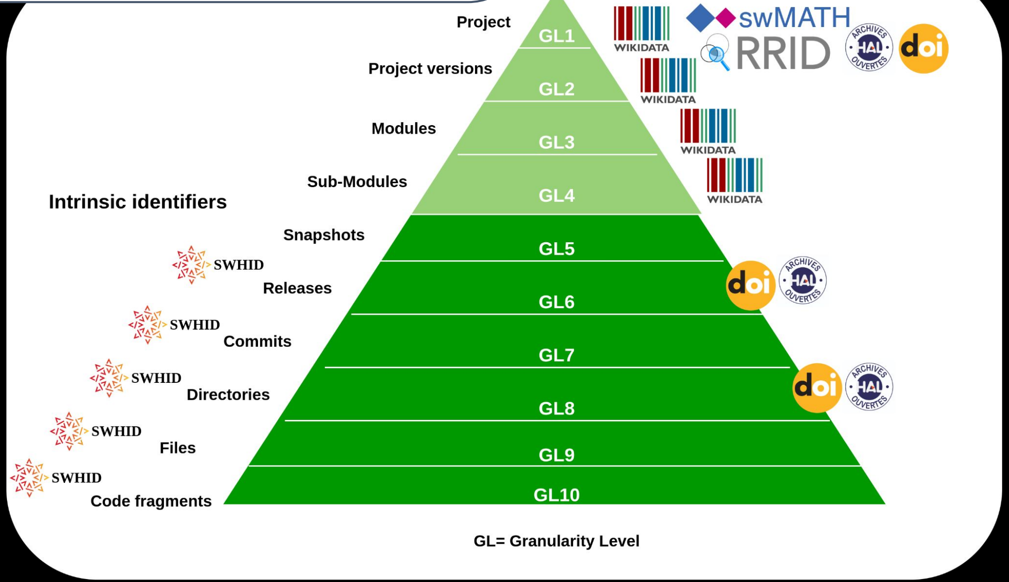
Reference Catalogue

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
visit_ctxt
anchor_ctxt
path_ctxt
lines_ctxt
;origin=https://github.com/chrislgarry/Apollo-11
;visit=swh:1:snp:206c27c0c031c6aac6b5fedddba8fe082dea9836
;anchor=swh:1:rev:3913f198f4383d4d638c0485d6aa902ff2f35828
;path=/Luminary099/BURN_BABY_BURN--MASTER_IGNITION_ROUTINE.agc
;lines=64-72

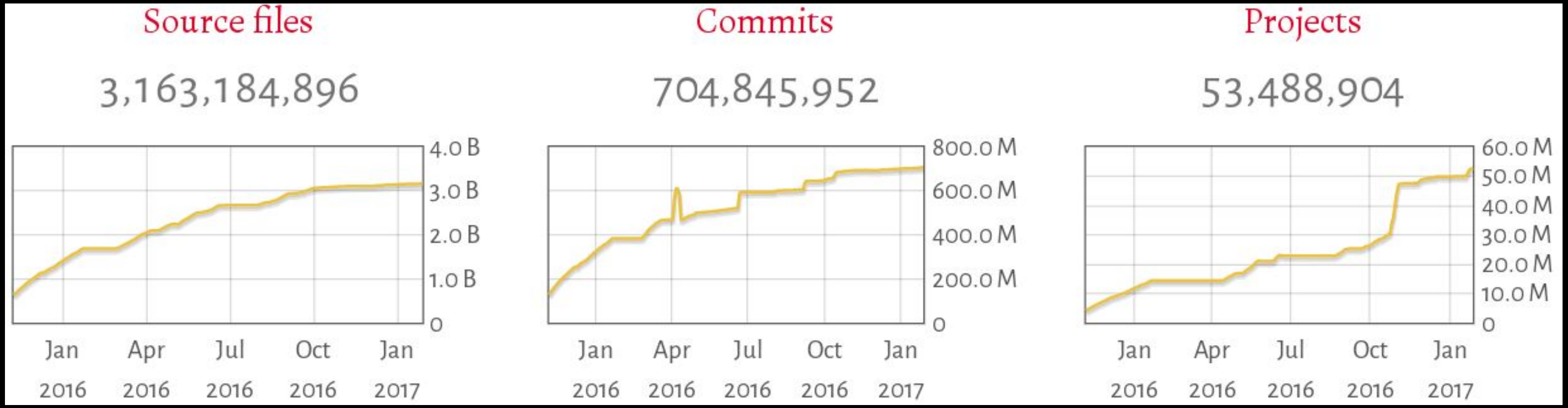
Extrinsic identifiers



GL= Granularity Level

Universal archive

Nov’16



Source files

3,163,184,896

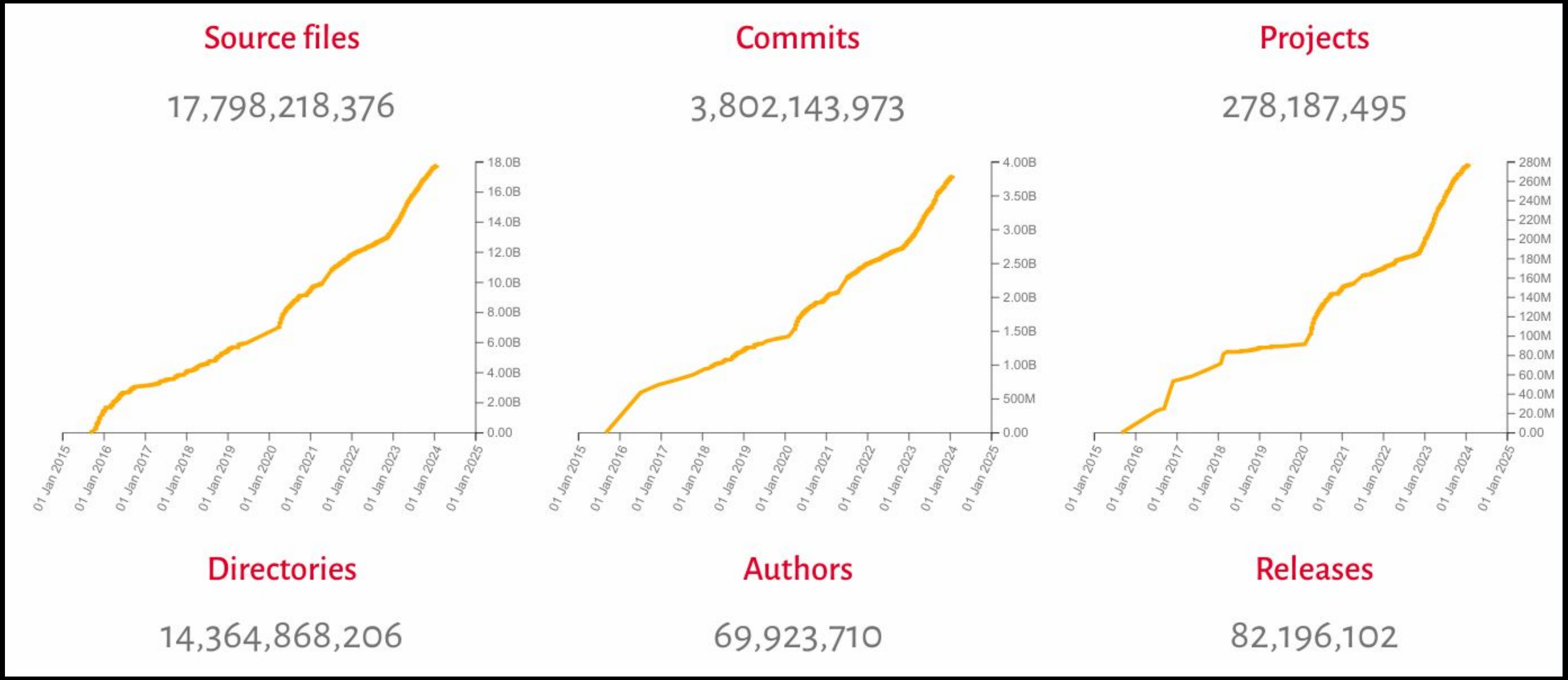
Commits

704,845,952

Projects

53,488,904

Jan’24



Source files

17,798,218,376

Commits

3,802,143,973

Projects

278,187,495

Directories

14,364,868,206

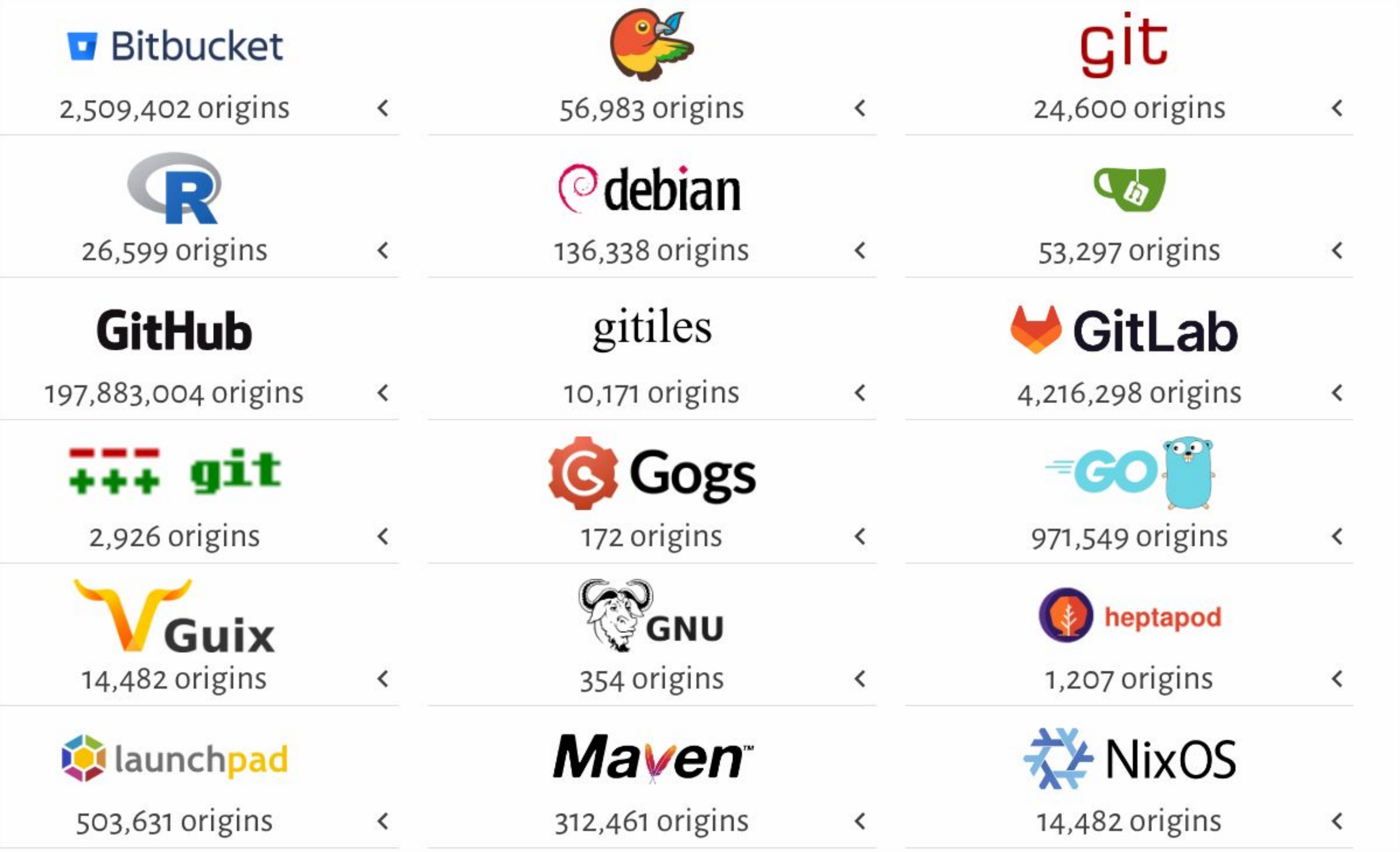
Authors

69,923,710

Releases

82,196,102

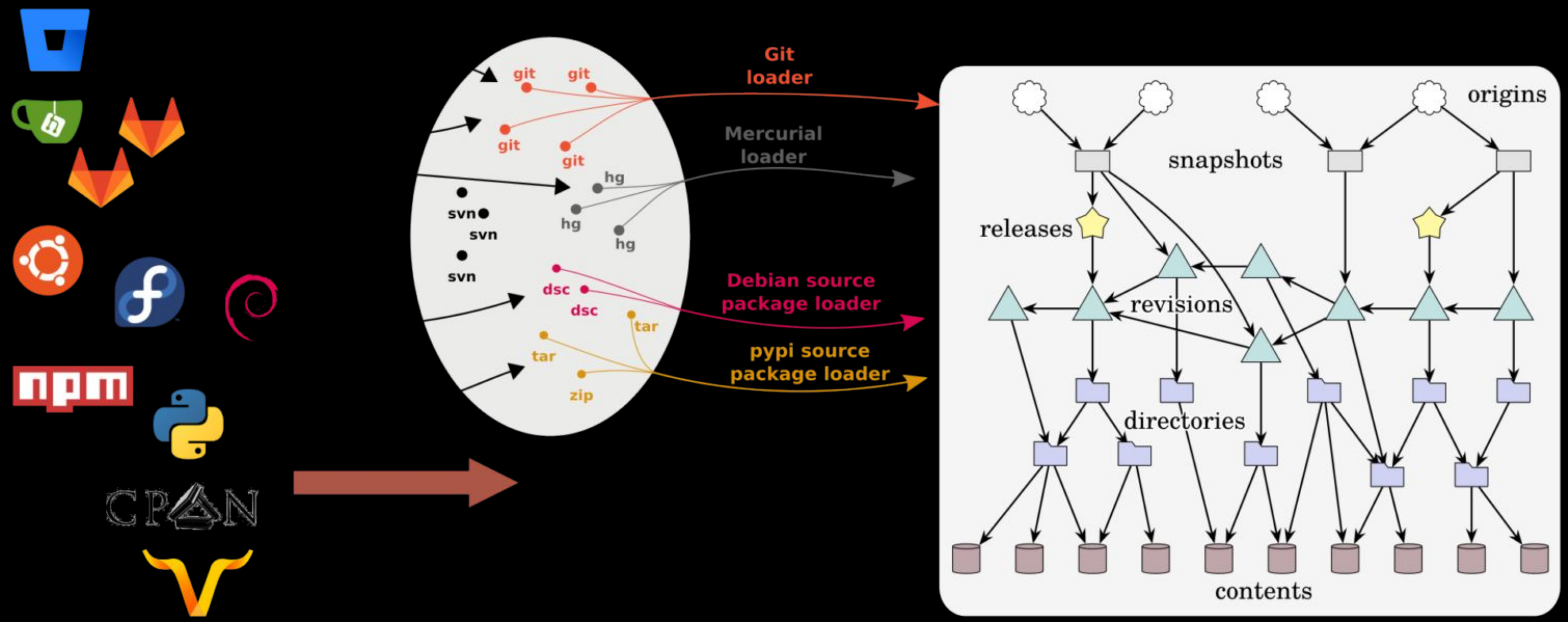
Truly universal



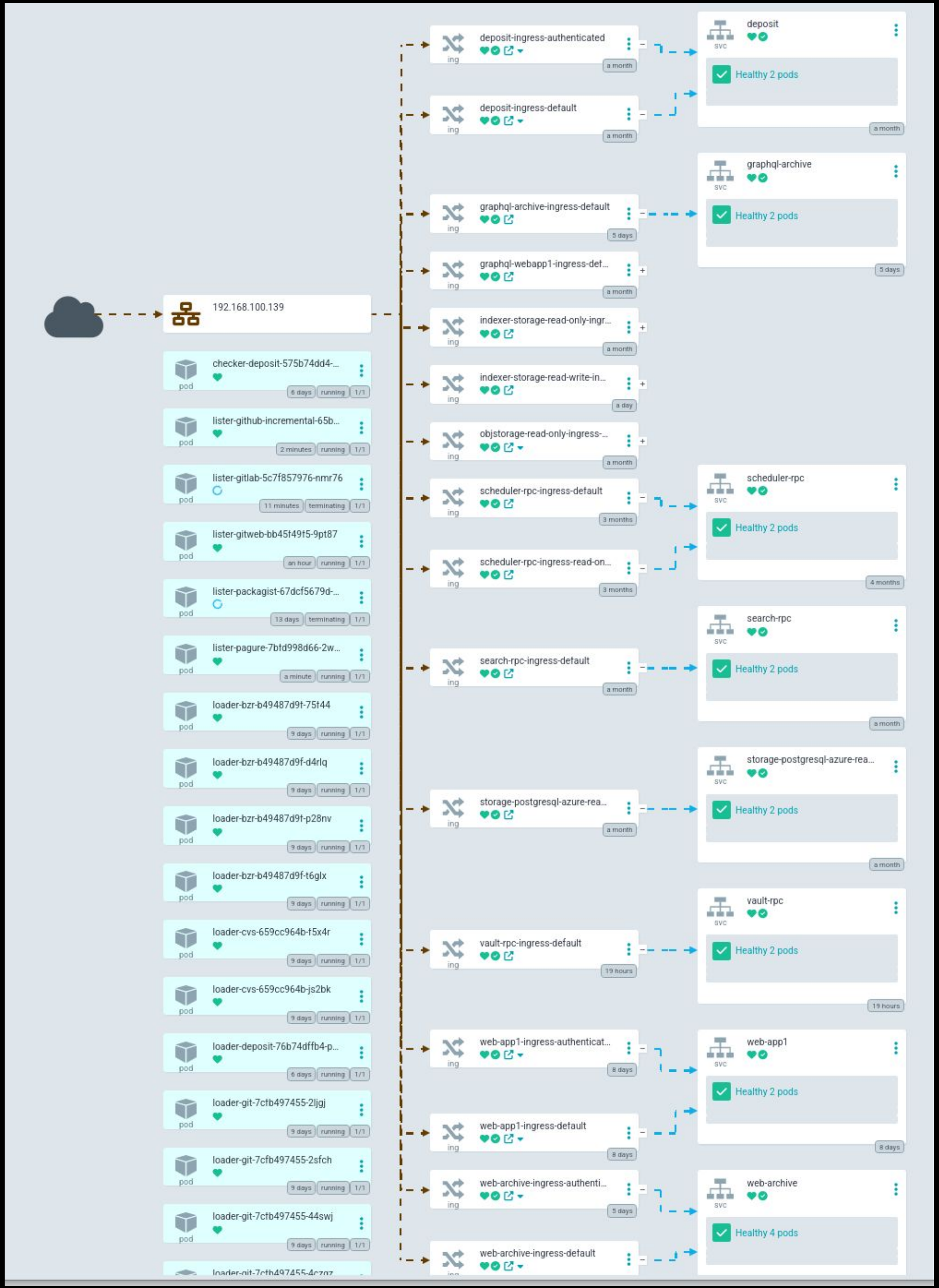
Bitbucket 2,509,402 origins	56,983 origins	git 24,600 origins
R 26,599 origins	debian 136,338 origins	53,297 origins
GitHub 197,883,004 origins	gitiles 10,171 origins	GitLab 4,216,298 origins
git 2,926 origins	Gogs 172 origins	GO 971,549 origins
Guix 14,482 origins	GNU 354 origins	heptapod 1,207 origins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

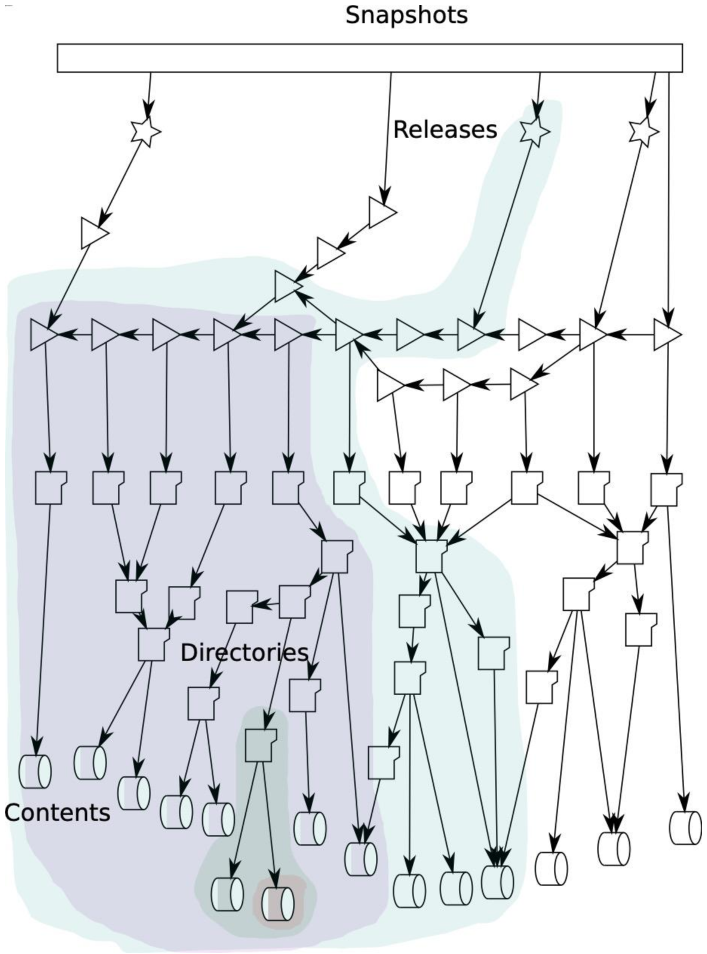
(*) The numbers are never up to date

Built for the purpose…



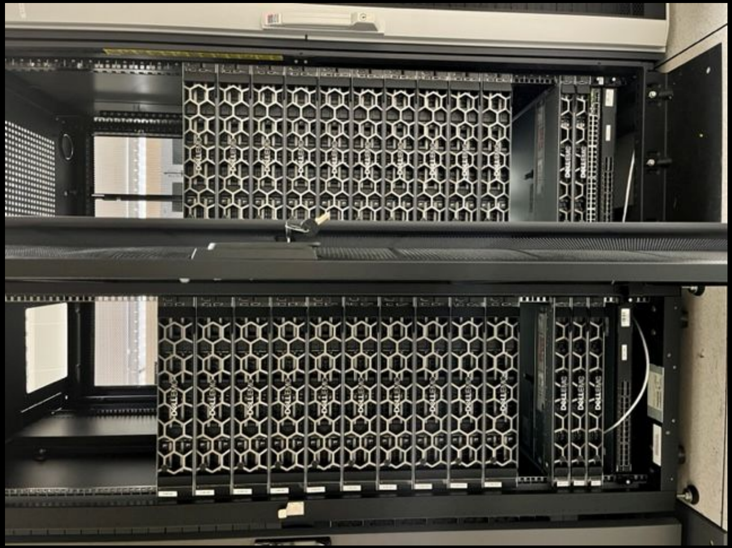
… so complex and innovative

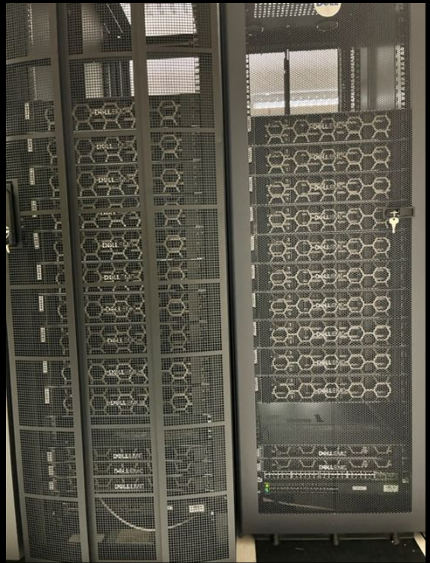

192.168.100.139
checker-deposit-575b74dd4-…
pod
6 days
running
1/1
lister-github-incremental-65b…
pod
2 minutes
running
1/1
lister-gitlab-5c7f857976-nmr76
pod
11 minutes
terminating
1/1
lister-gitweb-bb45f49f5-9pt87
pod
an hour
running
1/1
lister-packagist-67dcf5679d-…
pod
13 days
terminating
1/1
lister-pagure-7btd998d66-2w…
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





Harvest and archive:

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

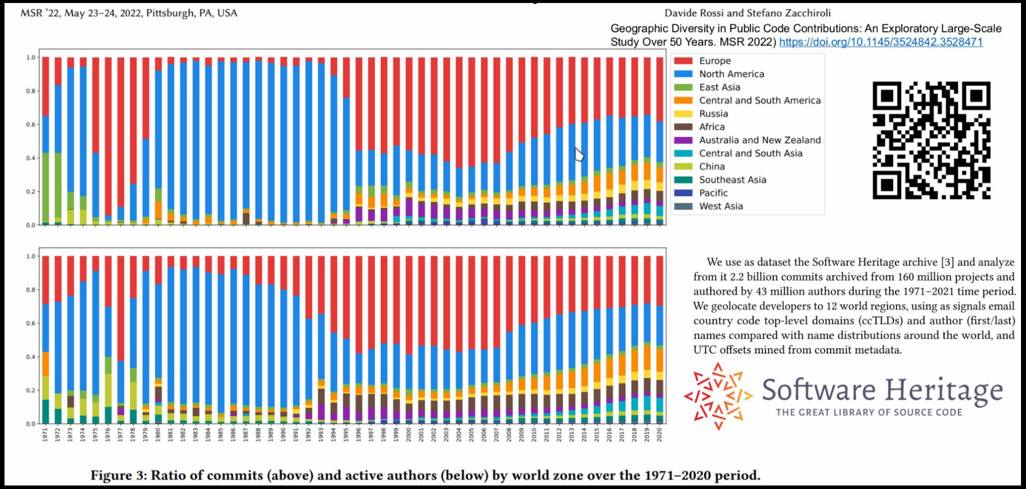
Howto:

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

Universal archive

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 Leandro von Werra, Machine Learning Engineer @ Hugging Face and Harm de Vries, Lead of the LLM lab @ ServiceNow



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.

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

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/

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/

Thank you

Introduction to Software Heritage (SwH)

The Alexandria library for code

Agustín Benito Bethencourt

Toscalix

16, April, 2024 FOSSNorth 2024