

Software Heritage

the Universal Archive of our Software Commons

Nicolas Dandrimont

Software Heritage
olasd@softwareheritage.org

01 December 2016
MFG Labs
Paris, France



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

L LEM GEOMETRY

USER'S PAGE NO. 5 E5 S3

P0705 GIVEN RR TRUNNION AND SHAFT (1,5) IN TANGNB,+1,FIND THE ASSOCIATED
R0706 LINE OF SIGHT IN NAV BASE AXES. THE HALF UNIT VECTOR, .5(SIN(S)COS(T),
R0707 -SIN(T),COS(S)COS(T)) IS LEFT IN MPAC AND 32D.

07072	REF	1		23,2000				SETLOC INFLIGHT		
07074				23,2041				BANK		
07076	REF	1						COUNT# 44/GEOM		
0708				23,2041	47135 0	RRNB		SLOAD	RTB	
0709	REF	4	LAST	201	23,2042				TANGNB	
07091	REF	5	LAST	250	23,2043				CDJLOGIC	
0710					23,2044			SETPD	PUSH	TRUNNION ANGLE TO 0
0711					23,2045				0	
0712					23,2046			SIN	DCOMP	
0713					23,2047			STOOL	34D	Y COMPONENT
0714					23,2050			COS	PUSH	.5 COS(T) TO 0
0715					23,2051			SLOAD	RTB	
0716	REF	5	LAST	324	23,2052				TANGNB +1	
0717	REF	6	LAST	324	23,2053				CDJLOGIC	
0718					23,2054		RRNB1	PUSH	COS	SHAFT ANGLE TO 2
0719					23,2055			DMP	SL1	
0720					23,2056				0	
0721					23,2057			STOOL	36D	Z COMPONENT
0722					23,2060			SIN	DMP	
0723					23,2061			SL1		
0724					23,2062			STOOL	32D	
0725					23,2063				32D	
0726					23,2064			RVQ		
R0727	THIS ENTRY TO RRNB REQUIRES THE TRUNNION AND SHAFT ANGLES IN MPAC AND MPAC +1 RESPECTIVELY									
0729					23,2065	14025 0	RRNBMPAC	STOOL	20D	SAVE SHAFT CDU IN 21.
07291	REF	43	LAST	299	23,2066	00155 0			MPAC	SET MODE TO DP. (THE PRECEDING STORE
A07292										MAY BE DP, TP OR VECTOR.)
0730					23,2067	40234 0		RTB	SETPD	
0731	REF	7	LAST	324	23,2070	21576 0			CDJLOGIC	
0732					23,2071	00001 0			0	
0733					23,2072	73406 1		PUSH	SIN	TRUNNION ANGLE TO 0
0734					23,2073	77676 0		DCOMP		
0735					23,2074	14043 0		STOOL	34D	Y COMPONENT
0736					23,2075	41546 0		COS	PUSH	.5COS(T) TO 0
0737					23,2076	47135 0		SLOAD	RTB	PICK-UP CDU'S.
0738					23,2077	00026 0			21D	
0739	REF	6	LAST	324	23,2100	21576 0			CDJLOGIC	
0740					23,2101	77650 1		GOTO		
0741	REF	1			23,2102	46054 1			RRNB1	

/ This routine exploits a fixed 512 byte input buffer in a
* VAX running the BSD 4.3 fingerd binary. It send 536
* bytes (plus a newline) to overwrite six extra words in
* the stack frame, including the return PC, to point into
* the middle of the string sent over. The instructions in
* the string do the direct system call version of
* execve("/bin/sh"). */*

```
static try_finger(host, fd1, fd2) /* 0x49ec, <just_return+378  
    struct hst *host;  
    int *fd1, *fd2;  
{  
/* ... */  
    for(i = 0; i < 536; i++) /* 628,654 */  
        buf[i] = '\0';  
    for(i = 0; i < 400; i++)  
        buf[i] = 1;  
    for(j = 0; j < 28; j++)  
        buf[i+j] = "\335\217/sh\0\335\217/bin\320^Z\335\0\335\0
```

<https://github.com/arialdomartini/morris-worm>



“Programs must be written for people to read, and only incidentally for machines to execute.” — Harold Abelson

Distinguishing features

- **executable** and **human readable** knowledge (an *all time new*)
- naturally **evolves** over time
 - development history is key to its understanding
- complex: large web of **dependencies**, millions of SLOCs

Definition (Commons)

The **commons** is the cultural and natural resources accessible to all members of a society, including natural materials such as air, water, and a habitable earth. These resources are held in common, not owned privately. <https://en.wikipedia.org/wiki/Commons>

Definition (Software Commons)

The **software commons** consists of all computer software which is available at little or no cost and which can be altered and reused with few restrictions. Thus *all open source software and all free software are part of the [software] commons. [...]*

https://en.wikipedia.org/wiki/Software_Commons

Software is spread all around



Fashion victims

- many disparate development platforms
- a myriad places where distribution may happen
- projects tend to migrate from one place to the other over time

Software is spread all around



Fashion victims

- many disparate development platforms
- a myriad places where distribution may happen
- projects tend to migrate from one place to the other over time

One place...

... where can we find, track and search *all* source code?



A word cloud centered on the slide, featuring terms related to software fragility. The words are arranged in a circular pattern, with some overlapping. The largest words are 'damage', 'disaster', 'malicious', 'deletion', 'obsolete', and 'attack'. Other words include 'media', 'aging', 'tear', 'dependencies', 'dangling', 'wear', 'corruption', 'encryption', 'format', 'reference', and 'storage'. The background of the slide features a faint world map and a pattern of colorful arrows pointing in various directions.

Like all digital information, FOSS is fragile

- inconsiderate and/or malicious code loss (e.g., Code Spaces)
- business-driven code loss (e.g., Gitorious, Google Code)
- for obsolete code: physical media decay (data rot)

Software is fragile



A word cloud centered on the slide, featuring terms related to software fragility. The words are in various colors (purple, blue, green, brown) and sizes. The most prominent words are 'damage', 'disaster', 'malicious', 'deletion', 'obsolete', 'attack', 'format', 'dependencies', 'reference', 'storage', 'dangling', 'wear', 'corruption', 'encryption', 'aging', 'tear', and 'media'. The background of the slide features a faint world map and a pattern of colorful, stylized arrows pointing in various directions.

Like all digital information, FOSS is fragile

- inconsiderate and/or malicious code loss (e.g., Code Spaces)
- business-driven code loss (e.g., Gitorious, Google Code)
- for obsolete code: physical media decay (data rot)

If a website disappears you go to the Internet Archive...

... where do you go if (a repository on) GitHub goes away?



Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Our mission

Collect, **preserve** and **share** the *source code* of *all the software* that lies at the heart of our culture and our society.

Our principles

Cultural Heritage



Industry



Research



Education



Software Heritage

Open approach

- 100% FOSS
- transparency

In for the long haul

- replication
- non profit

Archiving goals

Targets: VCS repositories & source code releases (e.g., tarballs)

We DO archive

- file **content** (= blobs)
- **revisions** (= commits), with full metadata
- **releases** (= tags), ditto
- (project metadata)
- **where & when** we found any of the above

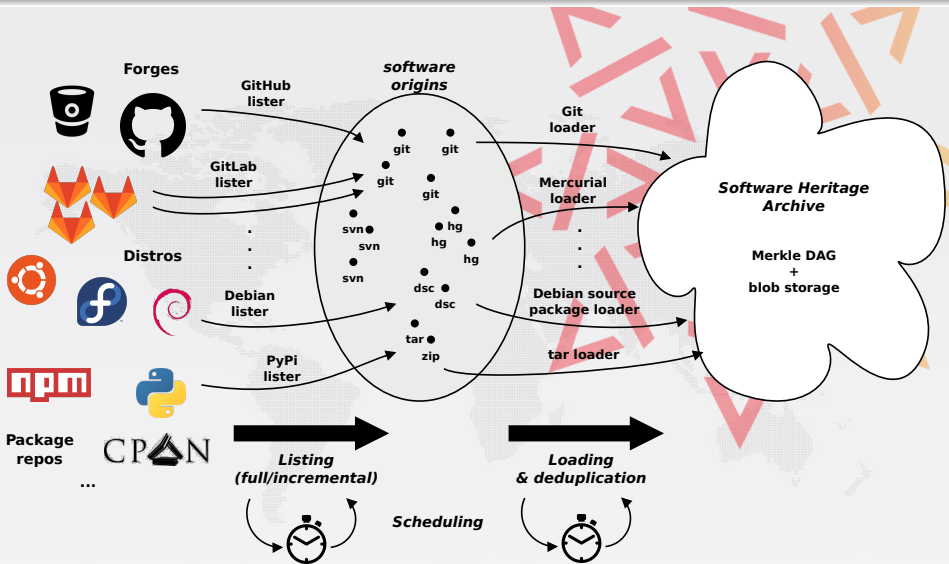
... in a VCS-/archive-agnostic **canonical data model**

We DON'T archive (UNIX philosophy)

- homepages, wikis → collaboration with the Internet Archive
- BTS/issues/code reviews/etc.
- mailing lists

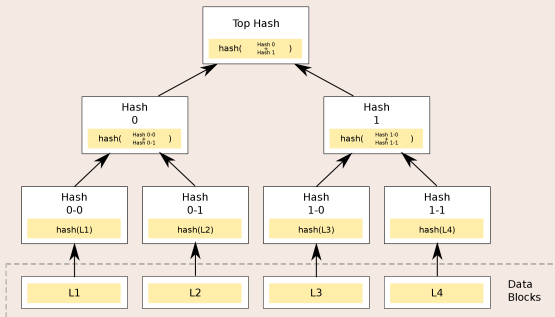
Long term vision: play our part in a *"semantic wikipedia of software"*

Data flow



Merkle trees

Merkle tree (R. C. Merkle, Crypto 1979)

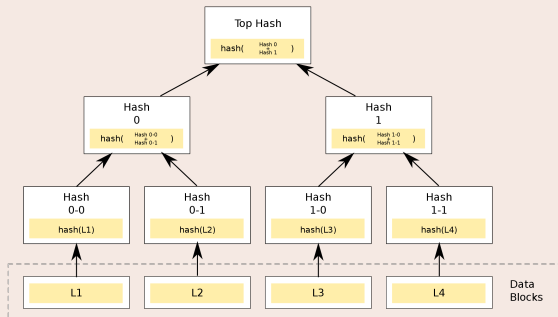


Combination of

- tree
- hash function

Merkle trees

Merkle tree (R. C. Merkle, Crypto 1979)



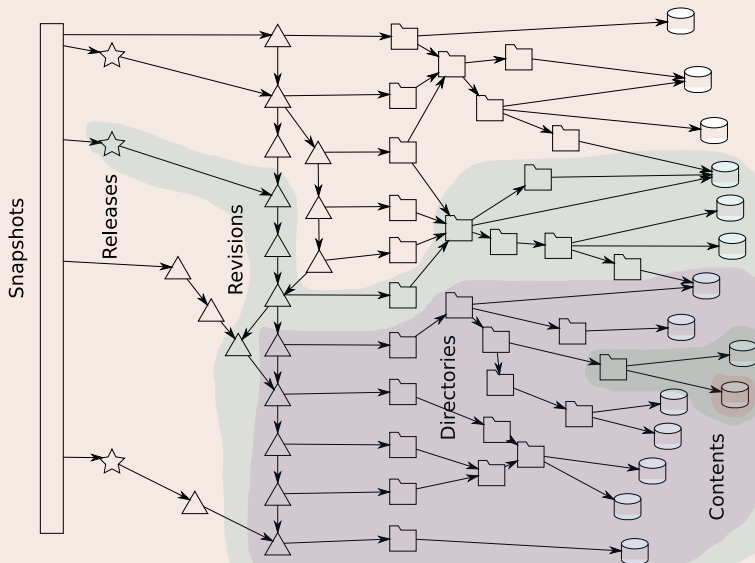
Combination of

- tree
- hash function

Classical cryptographic construction

- fast, parallel signature of large data structures
- widely used (e.g., Git, Bitcoin, IPFS, ...)
- built-in deduplication

The archive: a (giant) Merkle DAG



SHA1 collisions considered harmful

```
create domain sha1 as bytea
  check (length(value) = 20);
create domain sha1_git as bytea
  check (length(value) = 20);
create domain sha256 as bytea
  check (length(value) = 32);

create table content (
  sha1          sha1 primary key,
  sha1_git      sha1_git not null,
  sha256        sha256 not null,
  length        bigint not null,
  ctime         timestamptz not null default now(),
  status        content_status not null default 'visible',
  object_id     bigserial
);

create unique index on content(sha1_git);
create unique index on content(sha256);
```

Our sources

- GitHub — full, up-to-date mirror
- Debian — daily snapshots of all suites since 2005–2015
- GNU — all releases as of August 2015
- Gitorious, Google Code — local copy (Archive Team & Google)

Our sources

- GitHub — full, up-to-date mirror
- Debian — daily snapshots of all suites since 2005–2015
- GNU — all releases as of August 2015
- Gitorious, Google Code — local copy (Archive Team & Google)

Some numbers



150 TB blobs, 5 TB database (as a graph: 4 B nodes + 40 B edges)

Archive coverage

Our sources

- GitHub — full, up-to-date mirror
- Debian — daily snapshots of all suites since 2005–2015
- GNU — all releases as of August 2015
- Gitorious, Google Code — local copy (Archive Team & Google)

Some numbers



150 TB blobs, 5 TB database (as a graph: 4 B nodes + 40 B edges)

The *richest* source code archive already, ... and growing daily!

Planned features...

- *lookup* by content hash (done)
- *download*: wget and git clone from Software Heritage
- *provenance information* for all archived code and metadata
- *browsing*: wayback machine for archived code and its history
- *full-text search* on all archived source code files

Planned features...

- *lookup* by content hash (done)
- *download*: wget and git clone from Software Heritage
- *provenance information* for all archived code and metadata
- *browsing*: wayback machine for archived code and its history
- *full-text search* on all archived source code files

... and much more than one could possibly imagine

all the world's software development history in a single graph!

An ambitious, worldwide initiative

Inria as initiator



- .fr national CS research institution
- strong FOSS culture
- founding partner of the W3C

An ambitious, worldwide initiative

Inria as initiator



- .fr national CS research institution
- strong FOSS culture
- founding partner of the W3C

Supporters and *early partners*

ACM, **Nokia Bell Labs**, Creative Commons, **DANS**, Eclipse, Engineering, FSF, OSI, GitHub, GitLab, IEEE, Informatics Europe, **Microsoft**, OIN, OW2, SIF, SFC, SFLC, The Document Foundation, The Linux Foundation, ...

An ambitious, worldwide initiative

Inria as initiator



- .fr national CS research institution
- strong FOSS culture
- founding partner of the W3C

Supporters and *early partners*

ACM, **Nokia Bell Labs**, Creative Commons, **DANS**, Eclipse, Engineering, FSF, OSI, GitHub, GitLab, IEEE, Informatics Europe, **Microsoft**, OIN, OW2, SIF, SFC, SFLC, The Document Foundation, The Linux Foundation, ...

Going global

building an *open, multistakeholder, nonprofit* global organisation

Software Heritage is

- a revolutionary *reference archive* of *all* FOSS ever written
- a unique *complement* for *development platforms*
- an international, open, nonprofit, *mutualized infrastructure*
- at the service of our community, at the service of society!

Come in, we're open!

`www.softwareheritage.org` – *sponsoring, job openings*
`wiki.softwareheritage.org` – *internships, leads*
`forge.softwareheritage.org` – *our own code*

Questions?