From Open Source to Software Heritage

Building collaboration infrastructures

Roberto Di Cosmo

roberto@dicosmo.org

November 13th, 2018
Enterprise Architecture Days
Société Générale

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

Outline

1. Introductions
2. Free and Open Source Software
3. Understanding the tidal waves
4. The Software Heritage initiative
5. Building for the long term
6. Conclusion

Short Bio: Roberto Di Cosmo

Computer Science professor in Paris, now working at INRIA

- *30 years* of research (Theor. CS, Programming, Software Engineering, Erdos #: 3)
- *20 years* of Free and Open Source Software
- *10 years* building and directing structures for the common good

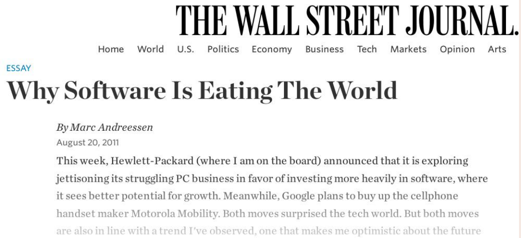
1999	*DemoLinux* – first live GNU/Linux distro
2007	*Free Software Thematic Group*
	150 members 40 projects 200Me
2015	*Software Heritage* at INRIA
2018	*National Committee for Open Science*, France

Outline

1. Introductions
2. Free and Open Source Software
3. Understanding the tidal waves
4. The Software Heritage initiative
5. Building for the long term
6. Conclusion

Software is eating the world…

Business



THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Arts

ESSAY

Why Software Is Eating The World

By Marc Andreessen

August 20, 2011

This week, Hewlett-Packard (where I am on the board) announced that it is exploring jettisoning its struggling PC business in favor of investing more heavily in software, where it sees better potential for growth. Meanwhile, Google plans to buy up the cellphone handset maker Motorola Mobility. Both moves surprised the tech world. But both moves are also in line with a trend I've observed, one that makes me optimistic about the future

Software companies

outperform or buy out

hardware companies

Marc Andreesen, 2011

Technology

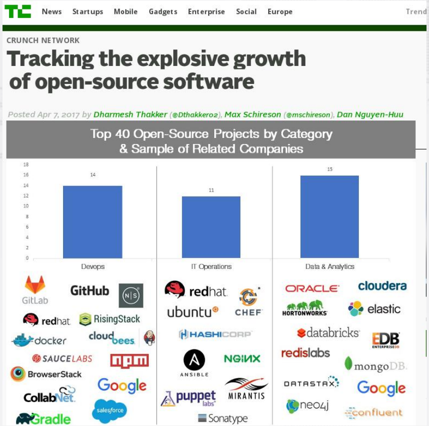
Software Defined Everything

Hardware gets commoditised

Software becomes the new value!

Worldwide Software Defined Everything (SDE) Market to grow at a CAGR of 27.9% over the period 2016–2022 to aggregate $143.35 billion by 2022

… Open Source is eating the Software World



CRUNCH NETWORK

Tracking the explosive growth of open-source software

Posted Apr 7, 2017 by Dharmesh Thakker (@Dthakkero2), Max Schireson (@mschireson), Dan Nguyen-Huu

Top 40 Open-Source Projects by Category & Sample of Related Companies

Open Source Software

can be openly (re)used, modified, (re)distributed, *with full access to its source code!*

Now it is worth billions…

Microsoft acquires GitHub

7Bn$

IBM acquires RedHat

34Bn$

Outline

1. Introductions
2. Free and Open Source Software
3. Understanding the tidal waves
4. The Software Heritage initiative
5. Building for the long term
6. Conclusion

Free Software: 30 years in a nutshell

Three Main Waves (and layers)

First 15 years: 1984-1998 — The early revolution

focus *freedom* for users and (especially) developers

keyword free software

Second 15 years: 1999-2014 — Progressive industry adoption

focus software quality and reduced cost

keyword open source (20th anniversary!)

The third wave: 2015-… — Ecosystems, strategic alignment

focus community and organisation

keyword governance

Key resources : competency, and adoption

We really are in a knowledge economy!

- competencies
- talent
- network
- adoption
- mindshare

Bottomline

The infrastructure for (open) collaboration is the new competitive advantage!

Outline

1. Introductions
2. Free and Open Source Software
3. Understanding the tidal waves
4. The Software Heritage initiative
5. Building for the long term
6. Conclusion

Source code matters!

> "The source code for a work means the preferred form of the work for making modifications to it."
>
> — GPL Licence

Hello World

Program (excerpt of binary)

```
4004e6:  55
4004e7:  48 89 e5
4004ea:  bf 84 05 40 00
4004ef:  b8 00 00 00 00
4004f4:  e8 c7 fe ff ff
4004f9:  90
4004fa:  5d
4004fb:  c3
```

Program (source code)

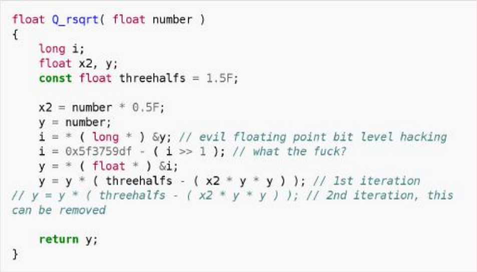
```
/* Hello World program */

#include<stdio.h>

void main()
{
    printf("Hello World");
}
```

Software Source Code is *special*

Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.) 1985

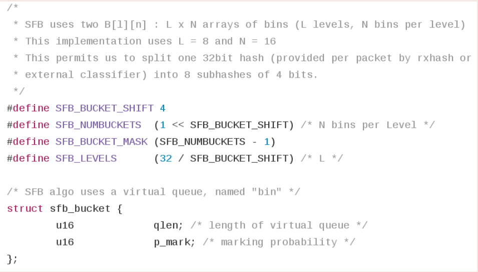
"Programs must be written for people to read, and only incidentally for machines to execute."

Quake 2 source code (excerpt)



```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    x2 = number * 0.5F;
    y = number;
    i = * ( long * ) &y; // evil floating point bit level hacking
    i = 0x5f3759df - ( i >> 1 ); // what the fuck?
    y = * ( float * ) &i;
    y = y * ( threehalfs - ( x2 * y * y ) ); // 1st iteration
// y = y * ( threehalfs - ( x2 * y * y ) ); // 2nd iteration, this
can be removed

    return y;
}
```

Net. queue in Linux (excerpt)



```
/*
 * SFB uses two B[l][n] : L x N arrays of bins (L levels, N bins per level)
 * This implementation uses L = 8 and N = 16
 * This permits us to split one 32bit hash (provided per packet by rxhash or
 * external classifier) into 8 subhashes of 4 bits.
 */
#define SFB_BUCKET_SHIFT 4
#define SFB_NUMBUCKETS  (1 << SFB_BUCKET_SHIFT) /* N bins per Level */
#define SFB_BUCKET_MASK (SFB_NUMBUCKETS - 1)
#define SFB_LEVELS      (32 / SFB_BUCKET_SHIFT) /* L */

/* SFB algo uses a virtual queue, named "bin" */
struct sfb_bucket {
        u16             qlen; /* length of virtual queue */
        u16             p_mark; /* marking probability */
};
```

Len Shustek, Computer History Museum

"Source code provides a view into the mind of the designer."

Software Heritage, in a nutshell

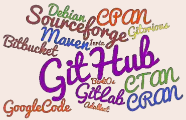
Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Collect, preserve and share the *source code* of *all the software*

Preserving our heritage, enabling better software and better science for all

Reference catalog



find and reference all the source code

Universal archive

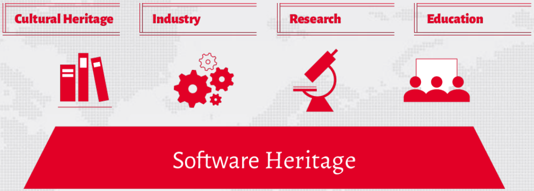
damage
storage
disaster
reference
deletion
media
malicious
obsolete
attack
aging
dependencies
tear
dangling
wear
corruption
encryption
format

preserve all the source code

Research infrastructure

enable analysis of all the source code

A principled infrastructure



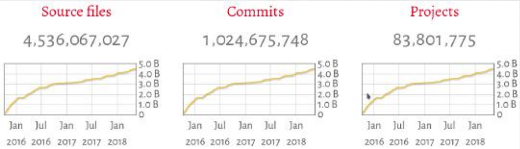
Cultural Heritage

Industry

Research

Education

Software Heritage



Source files

4,536,067,027

Commits

1,024,675,748

Projects

83,801,775

Technology

- transparency and FOSS
- replicas all the way down

Content

- intrinsic identifiers
- facts and provenance

Organization

- non-profit
- multi-stakeholder

The *graph* of Software Development

Access to all the history of software development!

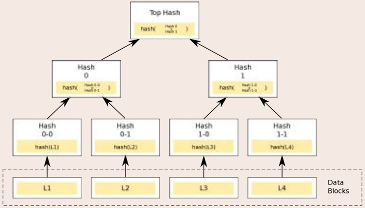
- lookup by content hash
- browsing: "wayback machine" for archived code
  - http://archive.softwareheritage.org/api
  - http://archive.softwareheritage.org/browse/search
- download: `wget` / `git clone` from the archive
- deposit of source code bundles directly to the archive

… and much more …

the world's software development *graph* is here!

The *blockchain* of Software Development

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



Combination of

- tree
- hash function

Classical cryptographic construction

- widely used (e.g., Git, blockchains, IPFS, . . .)
- built-in deduplication
- provides intrinsic, unforgeable identifiers at all levels

Software Heritage is a blockchain for source code!

Outline

1. Introductions
2. Free and Open Source Software
3. Understanding the tidal waves
4. The Software Heritage initiative
5. **Building for the long term**
6. Conclusion

Growing Support

Landmark Inria Unesco agreement, April 3rd, 2017

Inria
INVENTEURS DU MONDE NUMÉRIQUE

UNESCO

Sharing the vision

Software
GitLab
ENGINEERING
eclipse
OW2
fsfe
nlnet
SIGSOFT
Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
INAPI
THE LINUX FOUNDATION
ADULLACT
INFORMATICS EUROPE
Computer History Museum
software freedom conservancy
Software Freedom Law Center
AdaCore
gandi.net
TDF
FREE SOFTWARE FOUNDATION
creative commons
SIF
open source initiative
openinventionnetwork

Contributing to the mission

Inria
INVENTEURS DU MONDE NUMÉRIQUE

>= 100Ke/year

Microsoft
intel
SOCIETE GENERALE

>= 50Ke/year

>= 25Ke/year

HUAWEI
Google

>= 10Ke/year

DANS
NOKIA Bell Labs
ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA
GitHub
UQÀM
FOSSID

Outline

1. Introductions
2. Free and Open Source Software
3. Understanding the tidal waves
4. The Software Heritage initiative
5. Building for the long term
6. Conclusion

Come in, we're open!

Software Heritage

www.softwareheritage.org @swheritage

Library of Alexandria of code

- recover the past
- structure the future

A CERN for Software

- build better software
  - for industry
  - for society as a whole

Becoming a sponsor

https://sponsorship.softwarheritage.org