

Software Source Code Interest Group

Introduction

Roberto Di Cosmo (SWH and Inria)

`roberto@dicosmo.org`

April 2nd, 2019

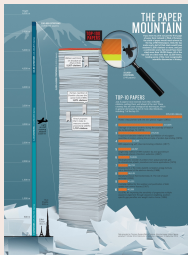


Software Heritage

THE GREAT LIBRARY OF SOURCE CODE

Why we are here

Software is *an essential component* of modern scientific research



Top 100 papers (Nature, October 2014)

[...] the vast majority describe experimental methods or software that have become essential in their fields.

<http://www.nature.com/news/the-top-100-papers-1.16224>

The *source code* is essential

- it contains the *real knowledge*,
- it is currently poorly accounted for

Reminder: the *source code* of a software artefact



“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

“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.”

Source code is not ... just data

executable and human readable knowledge (an all time new)

- written *by humans for humans*
- formats not really an issue: *text files are forever*

the development history is key to its understanding

- version history
- literate programming

complexity:

- large *web of dependencies*
- millions of SLOCs

Bottomline: software source code *is not just another* sequence of bits

Source code is *endangered*

Loosing precious legacy

foreclosures Google Code, Gitorious, now Codeplex

archives *off the record anecdotes*

you can use, and support, Software Heritage

Eu Copyright reform

- huge risk to software development and reuse
- more on this later

Bottomline

real need to raise awareness

Past and present activities

RDA 10

Montreal 9/2017

- motivations
- survey of ontologies
- metadata use cases

RDA 11

Berlin 3/2018

- identification of gaps in metadata

RDA 12

no meeting

RDA 13

- **updates** on ongoing activities
- **FAIR** for Software Source Code

Agenda

- ➊ Introduction (5m, done)
- ➋ Updates
 - Force 11 Software Citation IG (10m)
 - Software Source Code Identification WG (5m)
 - Software Heritage for Open Science: archive now open (10m)
 - Paris Call on Software Source Code (5m)
 - call for action EU Reform (5m)
- ➌ Group activity: what is FAIR for Software? (40m)
- ➍ Summary of results and wrap up (5m)

Group notes

<http://bit.ly/rda13scig>