Software Source Code Focus Group

Introduction

Roberto Di Cosmo (INRIA), Neil Chue Hong (SSI)

roberto@dicosmo.org

April 7, 2017
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

Question 1: where is the real knowledge?
"The source code for a work means the preferred form of the work for making modifications to it."

The source code is essential!

```
Hello World

Program (source code)
/* Hello World program */
#include<stdio.h>
void main()
{
    printf("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
```
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)

```c
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threethirds = 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 * ( threethirds - ( x2 * y * y ) ); // 1st iteration
    // y = y * ( threethirds - ( x2 * y * y ) ); // 2nd iteration, this can be removed
    return y;
}
```

Net. queue in Linux (excerpt)

```c
/*
 * SFB uses two B[][] : 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 32-bit 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 "blist" */
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.”
Question 2: is source code ... 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 (demo?)
- literate programming

**complexity:**
- large *web of dependencies*
- millions of SLOCs

**Bottomline:** software source code *is not just another* sequence of bits
Question 3: are we taking care of it?

No universal catalog

The Knowledge Conservancy Magic Triangle

- Articles: HAL, ArXiv, 100s of inst. repositories
- Data: Zenodo, Figshare, 100s of various repositories
- Software:
  - Question 4: does GitHub fit the bill?
  - Question 5: can we avoid useless duplication of efforts?

No universal archive
Question 6: what are key properties for Software Source code archives?

### Availability
- *all the history of all the software*
- no restrictions (technical, legal, … ) on *content* or *metadata*

### Traceability
- *persistent* and *intrinsic* identifiers: no DOI, no URL, no middle man, no dangling pointers!
- full *provenance* information

### Uniformity
- one *standard* metadata structure, *irrespective of the origins*
- *uniform* naming *schema*
Question 7: metadata
- what kind of ontology exist for software?
- what would be appropriate for Source Code?

Question 8: relation to professional software development
- is scientific software different from, say, usual open source software?
- can we learn from the open source developers?

Question 10: taking action
- is RDA a good place to take specific action concerning software and its source code?
- if yes, is there ongoing work already, and how should we proceed?
Reminder

**RDA BoF page**

https://www.rd-alliance.org/software-source-code-focus-group-rda-9th-plenary-bof-meeting

**Working document used during the session**