Talbot
Pierre

Proposal : “Checks and hashes project” for the Boost C++ Library

Abstract

The collection of check and hash functions are a suite of template files whose aim to facilitate the validation of codes and numbers following the international standard.
The hash functions are an easy way to make a fingerprint of a stream of data.

Personal Details

Name: TALBOT Pierre
College/University: College of Robert Schuman, Libramont.
Course/Major: Computer Science.
Degree Program (B.Sc., M, Sc., PhD, etc.) : Bachelor degree
Email: pierre.talbot.6114@herslibramont.be
Homepage: -
Availability: I'm free during the week but have courses until the 11th May and then exams until the 10th June. Usually free until midnight.
How much time do you plan to spend on your GSoC? In the beginning and until my exams ends, I will work on it more or less 3 hours a day and 12 hours the week-end. After my exams ending the 12th June, I will work hard and will devote my summer to the GSoC.
What are your intended start and end dates? I'll really start around the 12th June and I want to provide a complete library, so I would like to spend much more time than the GSoC on it.
What other factors affect your availability (exams, courses, moving, work, etc.)? Furthermore, I will go on holiday 3 days during July so I'll not be available.

Background Information

Please summarize your educational background (degrees earned, courses taken, etc.).

1997-2003 : Primary school.
2003-2009 : Secondary school.
2009 – 2010 : College of computer science, first year.
2010 - now : College of computer science, second year.

Please summarize your programming background (OSS projects, internships, jobs, etc.).

I programmed some school projects (Naval Battle, Pong,...) and the last summer programmed a website. I haven't professional experience yet. I coded some programs for algorithmic contest.

Please tell us a little about your programming interests.

I really like everything about the algorithmic and the network, I like taking part in any computing event (Prologin, Informatics Olympiad, ...).

Please tell us why you are interested in contributing to the Boost C++ Libraries.

I'm very keen to make some complete and useful tools for programmer. I love the idea that the tools I coded will be used by thousands of other programmers. Overall, this organization is well known, reputed and used among a large panel of programmers, writing a library for the Boost C++ Libraries would be an honor.

What is your interest in the project you are proposing?

I'm interested in algorithmic and this project is linked to my interest center. I'm accustomed to testing code during computing event.

Have you done any previous work in this area before or on similar projects?

No, it would be a first experience.

What are your plans beyond this Summer of Code time frame for your proposed work?

This project has the capability of being expandable so I'll continue to work on it in my free time. If for whatever reason I couldn't carry on with the project I would find a substitute. If

we want that the project survive, we must work on it over the years. It will be a growing library.

Please rate, from 0 to 5 (0 being no experience, 5 being expert), your knowledge of the following languages, technologies, or tools:

- C++ : 3
- C++ Standard Library : 3
- Boost C++ Libraries : 3
- Subversion : 1
- Git : 3

What software development environments are you most familiar with (Visual Studio, Eclipse, KDevelop, etc.)?

Eclipse, but I often code with VIM and compile in console.

What software documentation tool are you most familiar with (Doxygen, !DocBook, Quickbook, etc.)?

I used Doxygen for personal projects.

Project Proposal

Project description

Since the international development, every product, bank account or even packet which travelled on the network need to be controlled. For instance, in an Ethernet network, every packet is controlled with a Control Redundancy Check (CRC) to avoid errors of transmission.

A lot of domains, if not all, use computing system, so they need to validate the input. In the marketing world, every consumer product is marked with a Global Trade International Number (GTIN). The GTIN can be the International Standard Book Number (ISBN) for books, the European Article Numbering (EAN) for daily products and so on.

In the bank and financial world, the bank account numbers always have an international identifier called International Bank Account Number (IBAN). The credit cards number have their own check algorithm (often the Luhn algorithm), but use different number formats. For instance, we should provide checks for the most known credit card such as Visa, MasterCard and American Express (AMEX).

The development of computer has led to large amount of data to be stored. For example, the employees records of a company or the students results of a school. Even the health and care services need to store the data of their patients. Those data should be organized to provide a fast access to a single record, and a simple way to do it are the hash tables. And the secret ingredient of the hash tables are the hash functions, a great idea would be to propose the most used.

Another application of the hash functions is the cryptographic sector, we remember that the database of the website Skyblog[i] was hacked one year ago. They stored the logins and passwords of the users in clear, so the hackers used those easily. This kind of drama situation can be avoided by storing encrypted passwords.

All these domains are more or less closely linked to informatics, the programmers will always have to check the input. They will never trust the validity of information. On the other hand, the hash functions find many more applications that the two examples here above. The programs using this two previous tools are numerous. It's for those reasons that we should provide a check and hash library to complete the Boost C++ Library.

Objectives

This library has different purposes :

- Provide a better control of input and a simple way to hash data with the most known hashes algorithms.

- Offer a flexibility with template function, so the data could come from files, users input, …
- Remove the boring work from the programmer who makes sure that the input data respects the official standards. Stop losing time reading specifications !
- Propose functionalities which transform old numbers to the newest version (for instance : ISBN 10 to ISBN 13).

Overall the ultimate goal is to construct a library fully documented and tested. During the summer, we'll create a fast useable library even if we don't have time to implement every function.

Resources

Even in other languages, a uniform check and hash library is rare, but the codes of these checks are spread out all over the Internet. Therefore, we can re-use the algorithms, and even if there are written in other language, we can use those to understand the concept.

There is also a large amount of applets or online checks coded in Javascript, but again it isn't useful for an integration into a program. We can also mention that the Boost library already has a CRC algorithm[ii] which can be a good point to start with if we want to modify it.

Outline

The priority is to code the most known and used checks and hashes, the following should be written, tested and documented before the end of the summer. I added suspension point because this project has the incredible advantage to be expandable. The most known checks and hashes will be coded first.

Checks :

- **International standard numbers and codes** : ISBN, ISSN, GTIN (EAN, UPC,…), IBAN,…
- **Credit card numbers** : Visa, MasterCard, AMEX, …
- **Check algorithms** : Luhn, Verhoeff,…
- **Control redundancy checks** : CRC16, CRC32, CRC for Ethernet, …
- **Checksums** : Parity check, modulo check, …

Hashes : Message Digest 5 (MD5), Secure Hash Algorithm 1 (SHA-1),…

-

Proposed Milestones and Schedule

Please provide estimated milestones and schedule for completing the proposed work.

27 April – 24 May

- Familiarize myself with the revision control system *Subversion*.
- Starting the project while studying of the Boost tools :
    - Boost.Build and jam file.
    - The documentation system Quickbook.
    - Boost.Test and more particularly the Unit Test Framework (UTF).
- This period will also be dedicated to speak with the Boost community and my mentor.

25 May – 15 July

Write, test and document a draft of the check library.

15 July : MID-TERM EVALUATION DEADLINE

16 July – 31 July

Write, test and document the hash functions MD5 and SHA-1 256 bits.

1 August – 12 August

Finalize the checks and hashes library

13 August – 20 August

Complete review of the project to correct the very last details.

[i] http://www.zataz.com/news/20256/skyrock--skyblog--piratage.html
[ii] http://www.boost.org/doc/libs/1_46_1/libs/crc/