Google Summer of Code 2014
Student application
Tor Project: Search Engine for Hidden Services

Juha Nurmi
juha.nurmi@ahmia.fi

I would like to develop ahmia.fi - search engine for hidden services. It needs a lot love and care. I have founded, developed and maintained ahmia.fi and would like to continue doing so. I published the source code of ahmia.fi.

1. What project would you like to work on? Use our ideas lists as a starting point or make up your own idea. Your proposal should include high-level descriptions of what you're going to do, with more details about the parts you expect to be tricky. Your proposal should also try to break down the project into tasks of a fairly fine granularity, and convince us you have a plan for finishing it. A timeline for what you will be doing throughout the summer is highly recommended.

I would like to work on the project Search Engine for Hidden Services.

I would like to develop ahmia.fi as a free software (see a short presentation about ahmia.fi https://ahmia.fi/static/presentation/#(1)). I have been developing and maintaining ahmia.fi search engine. It needs a lot love and care.

Ahmia.fi is making the Tor network accessible in many different ways: listing hidden services, gathering their descriptions and providing a full text search.

During my GSoC I have planned to the implement various new key features to ahmia.fi.

Search development

Full text search development

- Popularity tracking (catch users clicks and tell YaCy the popular pages): development of a popularity tracking feature for ahmia.fi and Integration of that feature with YaCy API (providing stats for popular pages and suggestions for relevant results)
    - Checking out the backlinks too: can we get useful popularity data from the backlinks inside the .onion domains or do we have to check the backlinks from the public WWW
    - 2 workweek
- Use an another crawler to search .onion pages from the public Internet
    - Search new .onion domains from different online sources
    - This is an excellent case to test open source crawlers like Heritrix and Apache

Nutch
    - 2 workweeks
- Public open YaCy back-end for everyone
    - Let's make our YaCy network open so anyone can join their YaCy nodes
    - This way we could get real P2P decentralization
    - ahmia.fi is a free software and the back-end YaCy network should be free to everyone; also, we will get voluntary YaCy nodes this way
    - Share installation configuration package that joins a YaCy node to ahmia.fi's nodes
    - 1 workweek

Better edited HS descriptions

- Design and development of a more useful and complete UI including more complete and exhaustive descriptions and details (e.g., show the whole history of descriptions and let the users edit it better)
    - Requires security conscious design
    - Expose some of popularity/backlinks information to users, in case that lets them pick results more safely
    - 1 workweek
- ~~Comment and vote about the content (safe/unsafe)~~
    - ~~Ahmia.fi needs a commenting and rating systems for hidden services~~
    - ~~It is useful to gather a user's knowledge about the sites~~
    - ~~1 workweek~~

Tor browser friendly version of the ahmia.fi

- ~~Development of a JavaScript free version of ahmia.fi~~
    - ~~1 workweek~~
- ~~Search API~~
    - ~~1 workweek~~
- Hidden service mirror for ahmia.fi
    - Shared SQL database and YaCy back-end
    - 1 workweek

Information about hidden services and their content: Automated statistics and visualizations

- Development of an Analytics feature
    - As the result of the indexing Tor network's content ahmia.fi can produce an authoritative and exact quantitative research data about what is published through the Tor network
    - 1 workweeks
- Automated visualizations
    - it is very practical to visualize the data
    - what these hidden services are? number of web server, IRC servers, BitTorrent

trackers…
    - word clouds: we can even cluster which hidden services are close to each other and show some connections
    - I generated some SVG pictures of the backlinking between .onion sites
        - ZOOM out to see these huge pictures:
        - https://ahmia.fi/static/visuals/gephi.svg
        - https://ahmia.fi/static/visuals/visualRDF.svg
    - 1 workweeks
- Show cached text versions of the pages
    - there has been cached text versions of the pages but I had to remove them
    - the problem is non-trivial: there are a lot of ways to inject pictures and harmful JavaScript to the text cache
    - when I found that someone even injected images using only URL schema I had to take down the text cache (data:[<MIME-type>][;charset=<encoding>][;base64],<data>)
    - 1 workweek

API development

In addition, ahmia.fi provides RESTful API to integrate other services to use hidden service description information (see https://ahmia.fi/documentation/). Hidden services can integrate their descriptions directly to the hidden service list (see https://ahmia.fi/documentation/descriptionProposal/). Ahmia.fi knows which hidden services are online and you can use the API to check hidden service's online status. This API should be maintained to keep it general and simple. Furthermore, ahmia.fi uses this API internally.

Integration with softwares that are using hidden services

- Integration with Tor2web
    - Thanks to our suggestion recently, Tor2web has implemented a feature that provides secure and anonymous statistics within a day. I want to implement an automatic fetch and handling of this data
    - Ahmia.fi should fetch these and add each new .onion page to its database
    - Child pornographic is a plague for the Tor network and a well designed and authoritative entity may be useful for provide some filtering lists. To this aim we are currently handling manually a filter list already integrated with Tor2web and in use on quite all the nodes of the Tor2web network (https://ahmia.fi/policy/, https://github.com/globaleaks/Tor2web-3.0/issues/25). In collaboration with Tor2web I want to develop an efficient and automated system to handle and share a filtering information in a secure manner
    - after Freedom hosting take down there have been only few Child porn sites
    - 1 workweek
- Development of a Content Abuse Signaling feature in order to allow fast handling of

abuse comments; I want to implement a Callback API in order to publish this data to Tor2web nodes in real-time
    - we would also like to get automated signal from the Tor2web nodes when they are banning some site so ahmia.fi can also ban that site if neccessery
    - we are only sharing the MD5Sum of the banned domain
    - 1 workweek

- Globaleaks integration
    - Currently, GlobaLeaks informs ahmia.fi to index new hidden services
    - Ahmia.fi could extend the visibility of Globaleaks on the search results
    - Together with GlobaLeaks: RESTful API according to Globaleaks' needs
    - 1 workweek

Total estimated amount of work is 13 weeks.

2. Point us to a code sample: something good and clean to demonstrate that you know what you're doing, ideally from an existing project.

Working search engine: https://ahmia.fi/search
The source code of the ahmia.fi: https://github.com/juhanurmi/ahmia

3. Why do you want to work with The Tor Project in particular?

I would love to support human rights. I believe that human rights are important because without them life would be controlled by somebody else and people could not make decisions themselves.

In practice, free software is one way to support human rights. In particular, Tor Project is providing this kind of free software I would love to support.

Anonymity is an important right in order to support freedom of speech and defend human rights. I have been actively contributing to the Tor Project since 2010 by implementing the first public search engine for hidden services, ahmia.fi, and by running a very fast exit relay and by maintaining filtering list and tor2web.fi. I have significant hands-on competence with Tor and search engines.

Moreover, I am planning to join to torservers.net and launch several fast exit relays in Finland.

4. Tell us about your experiences in free software development environments. We especially want to hear examples of how you have collaborated with others rather than just working on a project by yourself.

As a Linux user, I have been using and supporting free software over ten years.

I am a contributor to Callimachus open source project (a framework for data-driven applications based on Linked Data). Callimachus aims to make Semantic Web applications easier to create.

I am a Fellow member of Hermes Center for Transparency and Digital Human Rights; I have built a minimal integration API between my search engine and their software: GlobaLeaks (an open source project aimed at creating a worldwide, anonymous, censorship-resistant, distributed whistleblowing platform) and Tor2web (an open source project aiming to allow transparent Internet exposure of websites running on Tor Hidden Services).

I was a volunteer and a lecturer in Observe, Hack, Make 2013: A five day international hacker festival in the Netherlands. There I presented ahmia.fi project to the other hackers.

Also, I am a member of the OKF Finland Open Science Work Group (OKF). The OKF is a hub for community-driven activities around open science to advocate standards of openness in Finnish academia and facilitate transfer of knowledge between academic institutions and wider society. I am pushing researchers to publish their source codes with proper licensing.

5. Will you be working full-time on the project for the summer, or will you have other commitments too (a second job, classes, etc)? If you won't be available full-time, please explain, and list timing if you know them for other major deadlines (e.g. exams). Having other activities isn't a deal-breaker, but we don't want to be surprised.

Yes, full-time.

6. Will your project need more work and/or maintenance after the summer ends? What are the chances you will stick around and help out with that and other related projects?

I am already maintaining the ahmia.fi search engine and going to continue doing so.

7. What is your ideal approach to keeping everybody informed of your progress, problems, and questions over the course of the project? Said another way, how much of a "manager" will you need your mentor to be?

Using familiar messaging systems, such as Email, IRC and Jabber. I am going to publish weekly updates to the tor-dev mailing list. Weekly online meeting with the mentor is sufficient.

I can travel to Italy to meet Globaleaks and Tor2web developers if it is necessary and helps to develop the API.

8. What school are you attending? What year are you, and what's your major/degree/focus? If you're part of a research group, which one?

I am a Ph.D student at the Tampere University of Technology. My major is semantic computing. Since 07.2010 I have been working at the department of mathematics, Intelligent Information Systems Laboratory. First as a research assistant and then after master's degree (1.7.2013) I have been working as a project researcher and a lecturer.

9. How can we contact you to ask you further questions? Google doesn't share your contact details with us automatically, so you should include that in your application. In addition, what's your IRC nickname? Interacting with us on IRC will help us get to know you, and help you get to know our community.

E-mail: juha.nurmi@ahmia.fi
Jabber: elephant@jabber.fi
Twitter: @AhmiaNews
OTR Fingerprint:65FE90B9E3D7DCF29398516CC01DED21DD31256D

10. Are you applying to other projects for GSoC and, if so, what would be your preference if you're accepted to both? Having a stated preference helps with the deduplication process and will not impact if we accept your application or not.

This is the only project I am applying to.

11. Is there anything else that we should know that will make us like your project more?

This is what I would really like to do. I have spent a lot of time to help Tor. Building a search engine for the hidden services is relevant and useful for the whole community.