NutchProposal

Proposal for new project Nutch

Doug Cutting – cutting at apache dot org

---

(0) rationale

Nutch is web search software. It builds on the Apache Lucene search library, adding a crawler, web database (including full link graph), plugins for various document formats, user interface, etc. It is currently used by sites such as http://search.creativecommons.org/, http://library.cornell.edu/, and the Internet Archive.

Nutch is a two-year-old open source project, currently hosted at Sourceforge and backed by its own non-profit organization. The non-profit was founded in order to assign copyright, so that we could retain the right to change the license. We have now determined that the Apache license is the appropriate license for Nutch and no longer require the overhead of an independent non-profit organization. Nutch's board of directors and its developers have both been polled and support a move to the Apache foundation.

We anticipate that Nutch will join the recently proposed search.apache.org top-level project, with Lucene and its various ports.

(0.1) criteria

Meritocracy:

Nutch's developers are already comfortable operating as a meritocracy. Nutch's current developer policies are a bit more informal than that of Apache, but, then, there have never been any notable conflicts to resolve.

Community:

Nutch has an established and active developer community.

Core Developers:

Nutch has four active committers who are experienced open source developers.

Alignment:

Nutch currently users the following Apache projects: Ant, Lucene, Xerces, POI, commons.

(0.2) warning signs

Orphaned products:

Nutch is not an orphan. It has the same corporate sponsors that it has always had.

Inexperience with open source:

Nutch's committers are experienced with open source.

Homogenous developers:

Nutch's committers do not all share an employer or nation. All decisions are made openly on public mailing lists.

Reliance on salaried developers:

Nutch has no salaried developers.

No ties to other Apache products:

Nutch has strong ties to Lucene.

A fascination with the Apache brand:

Nutch has a strong brand already. While the Apache brand will enhance that, that is not a primary motivation for Nutch to join Apache.

(1) scope of the subprojects

All code is currently licensed under a variant of the Apache License 1.0. The developers have approved a move to the Apache 2.0 license and a re-assignment of copyright to the Apache Foundation. We have signed Contributor License Agreements on file for all developers.

(3) identify the ASF resources to be created

(3.1) mailing list(s)

- nutch-dev
- nutch-commits
- nutch-user
- nutch-agent

(3.2) Subversion or CVS repositories

- https://svn.apache.org/repos/asf/incubator/nutch

(3.3) Jira

- Nutch (NUTCH)

(4) identify the initial set of committers

- Doug Cutting (Lucene committer)
- Michael Cafarella (current Nutch committer at Sourceforge)
- Andrzej Bialecki (current Nutch committer at Sourceforge)
- John Xing (current Nutch committer at Sourceforge)
- Sami Siren (current Nutch committer at Sourceforge)

(5) identify apache sponsoring individual

- Erik Hatcher, Champion and Mentor
- Doug Cutting, Mentor
  (as defined in http://incubator.apache.org/incubation/Roles_and_Responsibilities.html)