StanbolProposal

Apache Stanbol incubation proposal

Status

Proposal has been accepted: http://s.apache.org/stanbol.vote - see you soon at http://incubator.apache.org/stanbol

Abstract

Apache Stanbol is a modular software stack and reusable set of components for semantic content management.

Proposal

Stanbol components are meant to be accessed over RESTful interfaces to provide semantic services for content management. The current code is written in Java and based on the OSGi modularization framework, but other server-side languages might be used as well.

Applications include extending existing content management systems with (internal or external) semantic services, and creating new types of content management systems with semantics at their core.

The architecture of the current (alpha-level) code consists of four layers:

- Persistence: services that store (or cache) semantic information and make it searchable
- Lifting: services that add semantic information to “non-semantic” pieces of content
- Knowledge models and reasoning: services that enhance the semantic information
- Interaction: intelligent user interface management and generation

Background

Stanbol comes out of the IKS project (Interactive Knowledge Stack, http://iks-project.eu/), a research project funded by the European Community (EC) which aims to create a semantic content management software stack.

One of the goals of IKS is for its software to survive the 4-year funding period of the EC, which ends in 2012.

Developing its code in the open at the Apache Software Foundation, and growing a community before IKS funding runs out, is the best way to ensure the sustainability of the Stanbol software.

For more background information, some articles and tutorials on FISE, which was the first usable IKS module, can be found in the “FISE links” section of http://wiki.iks-project.eu/index.php/FISE

Rationale

Content Management Systems (CMS) can benefit from semantic add-ons in a number of ways, including more intelligent linking, automatic or semi-automatic tagging of content, enhanced user interactions based on intelligent and dynamically adaptable user scenario modeling, etc.

However, many CMS vendors and developers are not aware of or skilled enough in semantic technologies to make effective use of them. Research in semantic technologies often happens in academic circles which might not make their findings available in a way that’s easily consumable by today’s CMS vendors and developers.

Some big companies are using semantic technologies behind the scenes to provide powerful services, but that technology is usually not accessible to smaller vendors.

Stanbol aims to bridge these gaps by providing CMS vendors and developers with easy to integrate semantic components that add value to their offerings.

At the same time, more experimental advanced semantic applications will be built on the Stanbol stack, with the medium-term goal of enabling pure semantic-based content management and other applications.

Initial Goals

- Import the existing IKS code.
- Clean up as needed to take advantage of Apache infrastructure (Hudson continuous builds, etc.)
- Replace up any dependencies that do not fulfill Apache licensing criteria.
- Create the Stanbol website and migrate IKS website/wiki information to it as needed.
- Make a first release and publicize it to start growing the community.

Current Status

Meritocracy

As IKS is an EC research project with funding, it does not formally operate as a meritocracy.

However, due to the open source way of working adopted by the consortium, an informal meritocracy has emerged within IKS.

We estimate that adapting to the ASF's meritocratic way of working will be easy for the initial set of Stanbol committers, as the differences to the current way of working are not dramatic.

Community

The IKS project plan includes an important effort to build a community around the software that it produces. Several community workshops have already taken place, attended by more than 40 European CMS developers and vendors.

See http://wiki.iks-project.eu/index.php/Workshops for more info.

A community is emerging around IKS, and moving to the Apache project governance model should help grow it - also by reassuring community members that the software will continue to be available and maintainable once the IKS EC funding runs out.

Core Developers

The IKS consortium consists of seven academic research groups and six "industrial partners", companies active in the CMS space.

See http://iks-project.eu/team/team for the list.

The current IKS software has been written by a group of about a dozen developers from this consortium, with few external contributions until now. Members of the Clerezza community have contributed some key pieces, and ties between both communities are strong.

Alignment

As many Apache projects have something to do with content management, obvious synergies exist, which should allow us to grow the community from inside the ASF as well as from the outside.

Known Risks

Orphaned Products

The IKS code as it stands now might be orphaned when the EC funding of IKS runs out at the end of 2012.

That's why we want to move to Apache now, to have a bit more than two years to make Stanbol independent of its EC funding.

Inexperience with Open Source

The IKS team includes a number of very experienced Open Source developers, along with people doing their first open source contributions.

Since the IKS consortium started writing code early this year, we have had ample opportunity to bring everybody up to speed as to how open source works, and we're confident that the initial committers will quickly adapt to the ASF's way of working.

Homogeneous Developers

The current developers are spread amongst the IKS consortium partners, with no dominant company or organization.

Reliance on Salaried Developers

Until the end of 2012, the work of IKS consortium members is funded by the consortium, so there is a "common boss" problem, and we can assume that most or all of that work is salaried.

Moving software development to the ASF, and especially growing a community to include committers from outside the IKS consortium, should help reduce or eliminate this risk. Even IKS partners using the software in their products will help reduce the "common boss" problem, as both the IKS and the partner company will have a need for Stanbol software.

Relationships with Other Apache Products

The IKS software is written as a set of OSGi components and runs on Apache Felix, using the launcher from Apache Sling.

It also uses several key components from the Apache Clerezza incubating project, along with a number of other Apache libraries. Several Clerezza committers have been contributing in IKS workshops, without being part of the IKS consortium.

Clerezza in turn uses Jena, which is also joining the Apache Incubator.

Lucene/Solr will be used for indexing and search.

We also expect to use software from or collaborate with Mahout, Tikka, Jackrabbit, UIMA and Chemistry.

An Excessive Fascination with the Apache Brand

The brand is not what makes the difference for the IKS team, the motivation is the opportunity to build and grow a community.

Documentation

Existing components are documented at http://wiki.iks-project.eu/ and http://code.google.com/p/iks-project/w/list but that information is still incomplete due to the alpha status of most of that software.

Initial Source

http://code.google.com/p/iks-project/

External Dependencies

Appendix A contains the list of Maven groupIds of dependencies of the various Stanbol modules.

Most of those are compatible with ASF requirements (http://apache.org/legal/resolved.html) but an extensive check is needed, to remove/change any non-compatible ones.

Required Resources

Mailing Lists

- stanbol-private (moderated subscriptions)
- stanbol-dev
- stanbol-commits

Subversion Directory

- http://svn.apache.org/repos/asf/incubator/stanbol

Issue Tracking

- JIRA (STANBOL)

Other Resources

We will probably request a wiki once the podling is setup, and access to a Hudson continuous build server.

Initial Committers and affiliations

The following people are members of the IKS consortium, see http://iks-project.eu/team/team for a description of their organizations:

- Aldo Gangemi (aldo DOT gangemi AT cnr DOT it) - CNR
- Andreas Filler (andreas DOT filler AT hs-furtwangen DOT de) - Hochschule Furtwangen
- Andreas Gruber (andreas DOT gruber AT salzburgresearch DOT at) - Salzburg Research
- Benjamin Nagel (bnagel AT hotmail DOT de) - University of Paderborn
- Bertrand Delacretaz (bdelacretaz AT apache DOT org) - Adobe
- Enrico Daga (enrico DOT daga AT cnr DOT it) - CNR
- Fabian Christ (christ DOT fabian AT googlemail DOT com) - University of Paderborn (s-lab)
- Henri Bergius (henri DOT bergius AT iki DOT fi) - Nemein - original co-author of Midgard
- Jean-Michel Pittet (jmp AT adobe DOT com) - Adobe
- Joerg Steffen (steffen AT dfki DOT de) - DFKI
- Massimo Romanelli (romanell AT dfki DOT de) - DFKI - W3C MMI Working Group, USDL Incubator Group
- Olivier Grisel (ogrisel AT nuxeo DOT com) - Nuxeo - committer on Nuxeo, contributor to Mahout, http://github.com/ogrisel and http://bitbucket.org/ogrisel
- Ozgur Kilic (ozgur AT srdc DOT com DOT tr) - SRDC
- Rupert Westenthaler (rupert DOT westenthaler AT gmail DOT com) - Salzburg Research
- Sebastian Germesin (sebastian DOT germesin AT dfki DOT de) - DFKI
- Stefane Fermigier (sf AT fermigier DOT com) - Nuxeo SA
- Szabolcs Grunwald (szaby DOT gruenwald AT salzburgresearch DOT at) - Salzburg Research
- Tilman Becker (becker AT dfki DOT de) - DFKI
- Valentina Presutti (valentina DOT presutti AT cnr DOT it) - CNR
- Walter Kasper (kasper AT dfki DOT de) - DFKI - OpenLogos project lead, http://openlogos-mt.sourceforge.net/
- Wernher Behrendt (wernher DOT behrendt AT gmail DOT com) - Salzburg Research
- Wolfgang Maass (wolfgang DOT maass AT unisg DOT ch) - University of St. Gallen

The following initial committers are not members of the IKS consortium:

- Andreas Kuckartz (a DOT kuckartz AT ping DOT de)
- Ard Schrijvers (ard AT apache DOT org)
- Tommaso Teofili (tommaso AT apache DOT org)

Sponsors

Champion

- Bertrand Delacretaz (bdelacretaz AT apache DOT org)

Nominated Mentors

- Ted Dunning (tdunning AT apache DOT org)
- Ross Gardler (rgardler AT apache DOT org)
- Upayavira (upayavira AT apache DOT org)
- Isabel Drost (isabel AT apache DOT org)

Sponsoring Entity

Apache Incubator.

Appendix A: list of dependencies

Here's the list of Maven groupIds of the current Stanbol dependencies, omitting org.apache.* and commons-* groupIds but including transitive dependencies.

```
asm
com.aetrion.flickr
com.beetstra.jutf7
com.drewnoakes
com.googlecode.json-simple
com.hp.hpl.jena
com.ibm.icu
com.sun.jersey
com.sun.xml.bind
dom4j
edu.smu.tspell
eu.iksproject
hermit
info.aduna.commons
it.unimi.dsi.fastutil.chars
javax.activation
javax.mail
javax.servlet
javax.ws.rs
javax.xml
javax.xml.bind
javax.xml.stream
jetty
jtidy
junit
local.jrdf
log4j
mysql
net.fortuna.ical4j
net.fortuna.mstor
net.sf.jacob-project
net.sf.kxml
net.sourceforge
net.sourceforge.juniversalchardet
org.antlr
org.bibsonomy
org.bouncycastle
org.clojars.thnetos
org.codehaus.castor
org.codehaus.jackson
```

```
org.codehaus.jettison
org.codehaus.woodstox
org.freemarker
org.hsqldb
org.htmlparser
org.jaudiotagger
org.jdom
org.json
org.mockito
org.mortbay.jetty
org.nsdl.mptstore
org.openrdf.sesame
org.ops4j.base
org.ops4j.pax.exam
org.ops4j.pax.runner
org.osgi
org.samba.jcifs
org.scala-lang
org.semanticdesktop.aperture
org.semanticdesktop.nepomuk
org.semanticweb.owlapi
org.semweb4j
org.slf4j
org.textmining
org.wymiwyg
owl-link
owlapi
ronaldhttpclient
stax
trove
xerces
xmlpull
```