ODFToolkitProposal

The ODF Toolkit

Abstract

The ODF Toolkit is a set of Java modules that allow programmatic creation, scanning and manipulation of OpenDocument Format (ISO/IEC 26300 == ODF) documents. Unlike other approaches which rely on runtime manipulation of heavy-weight editors via an automation interface, the ODF Toolkit is lightweight and ideal for server use.

The ODF Toolkit is currently hosted by the ODF Toolkit Union and is licensed under the Apache 2.0 license.

Proposal

To move the following components from the ODF Toolkit Union to a single "ODF Toolkit" project at Apache:

- Simple Java API for ODF: http://simple.odftoolkit.org/
- ODFDOM: http://odftoolkit.org/projects/odfdom/pages/Home
- ODF Conformance Tools: http://odftoolkit.org/projects/conformancetools/pages/Home

(We'd be open as well to a catchier name. We've been calling it "The ODF Toolkit", prefaced always with "The". Or individually by component name. But "The Apache ODF Toolkit" or "Apache ODF Toolkit" are ponderous.)

In addition to migrating the code, we would migrate the website, tutorials, samples, Bugzilla data, and (if feasible) the mailing list archives. We would also seek to transfer the odftoolkit.org domain name to Apache.

While under incubation we will merge these projects into a single SDK with three layers:

- Package layer, representing the ZIP + Manifest container file of an ODF document. This structure is shared by other document formats, such as EPUB
- DOM Layer, a schema-generated layer that maps 1:1 with the ODF schema. This uses Apache Velocity as the templating engine.
- Convenience layer: an intuitive, high level API for use by app developers who are not familiar with ODF XML, but who have basic knowledge at the level of a word processor user.

Background

The ODF Toolkit Union was jointly announced by Sun and IBM at the OpenOffice.org Conference in Beijing, November 2008. The idea was to create a portfolio of tools aimed at accelerating the growth of document-centric solutions. The Open Document Format specification is large and complex. Most developers simply do not have the time and energy to master the 1,000-page specification By providing programming libraries, with high level APIs, the ODF Toolkit offers an means to reduce the difficulty level, and encourage development of innovative document solutions.

Rationale

During the recent OpenOffice incubation proposal discussions, the mention of possible moving the ODF Toolkit to Apache was met with enthusiasm.

Apache is emerging as the leading open source community for document related projects. The ODF Toolkit would have a good deal of synergy with other Apache projects, including the ODF Toolkit's dependency on Apache XML tools like Xerces, to possible multi-format applications with POI libraries to pipelining ODF with SVG and PDF rendering with Batik, FOP or PDFBox. Getting these various document processing libraries in one place, under a compatible permissive license would be of great value and service to users-developers interested in combining these tools for their specific project requirements.

Last, but not least, there is obvious synergy with Apache OpenOffice, as a prominent office suite supporting the ODF format.

The ODF Toolkit is already licensed under Apache License, Version 2.0, enabling a smooth transition.

Current Status

Meritocracy

We understand the intention and value of meritocracy at Apache. The initial committers are familiar with open source development. A diverse developer community is regarded as necessary for a healthy, stable, long term ODF Toolkit project.

Community

The ODF Toolkit is developed by a small set of core developers, though the community extends to include a broad set of application developers who use the code and contribute bug reports, patches and feature requests.

Although there are some open source projects that use these components directly, such Apache Directory Studio and GNU Octave, to support ODF import /export, it is more typical for these kinds of libraries to be used by application developers in small, ad-hoc document automation and data wrangling applications.

Core Developers

The coders on the existing ODF Toolkit will comprise the initial committers on the Apache project. These committers have varying degrees of experience with Apache-style open source development, ranging from none to being committers on other Apache projects..

Alignment

Along with the technical synergies described earlier, Apache aligns well due to its license and emphasis on meritocracy.

Known Risks

Orphaned products

The risk, as in most projects, is to grow the project and maintain diversity. This is a priority that is keenly desired by the community.

Inexperience with Open Source

The initial developers include experienced open source developers, including committers from other Apache projects. Although the majority of proposed committers do not have Apache experience, they do have open source experience.

Homogeneous Developers

The ODF Toolkit Union was created by IBM and Sun (later Oracle) who provided the majority of its engineering resources as well as its direction. Moving this project to Apache enables a new start. We intend to engage in strong recruitment efforts in order to further strengthen and diversify the community.

Reliance on Salaried Developers

When we look at sponsored developers, with the ability to work on this project full time, IBM currently has more committers. We believe that this situation will change, as the project grows in incubation.

Relationships with Other Apache Products

Several potential areas for collaboration with other Apache projects have been suggested, including:

Apache POI which is similar library, focused on Microsoft Office format documents

Apache Tika is a generic toolkit for extracting text and metadata from various file formats.

Apache PDFBox is a Java library for working with PDF documents. If not direct code sharing over the Java / C++ divide, then at least sharing of PDF know-how and perhaps things like test cases between these projects would be great.

We are interested in further exploring these options.

A Excessive Fascination with the Apache Brand

Our primary interest is in the processes, systems, and framework Apache has put in place around open source software development more than any fascination with the brand.

Documentation

There is documentation for the Simple Java API for ODF project, including a Cookbook, and JavaDoc:

http://simple.odftoolkit.org/cookbook/

http://simple.odftoolkit.org/javadoc/index.html

For the ODFDOM, there is a good overview documenting the project here: http://odftoolkit.org/projects/odfdom/pages/ProjectOverview

A 3rd party introductory tutorial here: http://www.langintro.com/odfdom_tutorials/

Initial Source

Will come from the ODF Toolkit Union, the latest stable source, plus any work in-progress

External Dependencies

We do not believe that we have any external dependencies other than Apache Xerces, Xalan, Velocity (a build-time dependency), Java 6 and the ODF schemas (also a build-time dependency)

Cryptography

We are currently working on adding support for digital signatures and encryption of documents. The project will complete any needed export control paperwork related to these features.

Required Resources

The following mailing lists:

- `odf-dev@incubator.apache.org` - for developer discussions
- `odf-users@incubator.apache.org` - for users discussions
- `odf-commits@incubator.apache.org` - for Subversion commit messages
- `odf-private@incubator.apache.org` - for PPMC issues, but only where privacy is required

Other resources

A source code repository, preferable git

An issue tracker

A wiki

A website

Initial Committers

Name	Email	Affiliation	iCLA
Rob Weir	robweir at apache dot org	IBM	yes
Biao Han (Devin)	hanbiao at cn dot ibm dot com	IBM	yes
Svante Schubert	svante dot schubert at gmail dot com	Individual	
Ying Chun Guo (Daisy)	guoyingc at cn dot ibm dot com	IBM	yes
Don Harbison	dpharbison at apache dot org	IBM	yes
Andy Brown	andy at the-martin-byrd.net	Individual	yes
Dave Fisher	wave at apache dot org	Individual	yes
Juergen Schmidt	jsc at apache dot org	Individual	yes

Sponsors

Champion

Sam Ruby

Nominated Mentors

- Nick Burch
- Yegor Kozlov

Sponsoring Entity

The Apache Incubator