CorinthiaProposal

Corinthia Podling Proposal

Abstract

Corinthia is a toolkit for converting between and editing common office file formats, with an initial focus on word processing. It is designed to cater for multiple classes of platforms - desktop, web, and mobile - and relies heavily on web technologies such as HTML, CSS, and JavaScript for representing and manipulating documents. The toolkit is small, portable, and flexible, with minimal dependencies. The target audience is developers wishing to include office viewing, conversion, and editing functionality into their applications.

The file format conversion library is implemented in highly-portable C, and can be easily embedded in native applications, with bindings for other programming languages planned. The library allows two-way conversion between different formats, and avoids irreversible loss of content or formatting unsupported in a target format by updating the source format in a way that makes only the minimal changes necessary.

The editing library is implemented in JavaScript, and runs in a browser runtime - either an actual web browser, or a web view embedded in a native app. It follows the philosophy of responsive design, popular on the web, where layout of a document is automatically adapted to suit the screen size and orientation, enabling the same content to be viewed on mobile phones, tablets, and desktop systems. All layout is handled by the browser's own engine; the editor works solely with the document's HTML structure and CSS styles. Currently the editor only operates in an embedded web view, but we plan to have it run in all major web browsers, and provide a clean API for easy integration into various native apps.

Importantly, Corinthia document viewing and editing is on the intermediate form (HTML & CSS), limited to common, widely-supported features. Corinthia is not a comprehensive substitute for format-specific authoring, editing, and final-form printing/production software. It is intended to complement, not compete with, major office suites.

Identification and confirmation of inter-convertible features of different formats for dependable import and export involves development of extensive test documents in the different formats. There is profiling of the extent to which standardized formats are supported in practice, with identification of deviations and implementation-dependent choices that impact convertibility.

Proposal

The goal of Corinthia is to provide a responsive design editor as well as a toolkit that enacts a defined conversion between different office document formats. Responsive design fits the layout as needed, tablet or desktop. The editor is a lightweight editor - an extension and not a replacement for the desktop editor.

Many office document programs claim to read/write to the ISO open standards for office documents, OpenDocument Format (ODF) and Office Open XML (OOXML), but do not document which parts are left unimplemented. Furthermore, the standards have a large number of "implementation defined" parts, making real-world congruence chancy. The Corinthia toolkit wants to put this unacknowledged aspect into the open and provide "compliance sheets" for document formats, as known from industry computer protocols.

Corinthia aims at generating a large set of test documents, which can be used to verify the "compliance sheets". The code can work as test case for other applications (or entities tendering for OOXML/ODF based systems) as well.

The base of Corinthia and its toolkit is the library DocFormats, which converts between different office document file formats. Currently it supports .docx (part of the OOXML specification), HTML, and LaTeX (export-only). In addition to this is an editing library, which allows manipulation of the HTML files in a web browser or embedded web view, and can be used in conjunction with DocFormats to edit documents in all supported formats.

The design of DocFormats is based on on the idea of bidirectional transformation (BDT), in which a specific document (the original file in its source format) is converted into an abstract document (in the destination format). A modified version of the abstract document can then be used to update the specific document in a non-destructive manner, keeping intact all parts of the file which are not supported in the abstract format by modifying the original file rather than replacing it.

Descriptions of BDT can be found in:

Aaron Bohannon, J. Nathan Foster, Benjamin C. Pierce et. al. Boomerang: Resourceful Lenses for String Data. Technical Report MS-CIS-07-15 Department of Computer and Information Science University of Pennsylvania. November 2007. (http://www.cis.upenn.edu/~bcpierce/papers/boomerang.pdf

Benjamin Pierce. Foundations for Bidirectional Programming. ICMT2009 - International Conference on Model Transformation. June 2009. (http://www.cis.upenn.edu/~bcpierce/papers/icmt-2009-slides.pdf)

The short term goal of the project is to have an easy-to-integrate library that any application can use to embed support for a range of different file formats, and use the parsing, serialisation, and conversion facilities for various purposes. These include editors, batch conversion tools, web publishing systems, document analysis tools, and content management systems. By abstracting over different file formats and using HTML as a common intermediate format, one can just code an application to that end, and let DocFormats take care of conversion to other formats.

The medium term goal of the project is to have a series of end-user applications (separate from the library itself), including an editor and file conversion tool. These will serve as examples of how the libraries can be used.

And ultimately to have a touch based UI for office documents.

It is also a goal to cooperate with other open source projects, in terms of getting input from them as well as providing APIs for their use. Corinthia is meant to be easy to understand and work with, making it more approachable for a range of projects.

Background

The document conversion library and the editing library have been shipping as components of UX Write on the iOS app store since February 2013. Both components have undergone continued development since that time. As far as UX Write is concerned, they provide a stable and reliable codebase.

As an open source project, Corinthia is completely new, in the sense that it is now moving from a single-developer commercial project, to an open, community-based project. We believe that this is the most beneficial path forward for the technology, to enable it to be developed to its full potential, and made available to anyone who needs to deal with multiple document formats or provide editing functionality on web, desktop, or mobile platforms.

Rationale

Corinthia's approach to providing an editor and thoroughly-documented conversion of office documents is perfectly aligned with Apache's mission to produce software for the public good. There is further benefit in documented tests demonstrating where implementations of standard formats deviate for any reason, identifying where interoperability and inter-conversion is improvable.

The project has potential to grow by collaboration with other projects, tapping mutual interests and identifying cases for improved interoperability, providing new architectures and design philosophies available to supplement existing implementations.

Introducing Corinthia in the Apache family of projects provides ready availability and participation with the diverse community of experienced Apache contributors under convenient familiar conditions.

Initial Goals

The initial and most important goal is to enlarge the community consisting of developers, testers, and people who know the standards in depth.

There are four technical goals:

- Cleanup core, to make it easy to add filters (format converters)
- Complete the ODF filter
- Produce an editor based on JavaScript & HTML which can be embedded in mobile apps or used in a Web browser
- Develop additional tests and compliance sheets for supported file formats

Our initial goals might not be big visions, but we prefer something reachable, and then make bigger goals as we grow.

Current Status

Meritocracy

Some of the initial committers are already part of Apache, and those who are not are becoming familiar with working in "the Apache way".

Community

Our community could be larger, and committers from AOO and others have shown interest in the project, but we have preferred to stay a stable, but very active group until we are part of Incubator.

Apache/Incubator provides a lot of tools (e.g., mailing lists) and community practices ~~The Apache Way~~ that enable community engagement and growth.

Core Developers

Peter Kelly,

Jan Iversen,

Svante Schubert,

Dennis E. Hamilton,

Alignment

Corinthia has commonalities with Apache OpenOffice (AOO), but is not competing. AOO is a desktop product and integrated suite and Corinthia is a lightweight editor and a developer product (library).

Corinthia has a document API as do Apache POI and the incubating Apache ODF Toolkit, but the focus is different. Corinthia targets a conversion library and an editor. POI and ODF Toolkit provide APIs for processing documents and are both Java-based.

Sharing test documents in standard document-file formats with projects like AOO and POI is a valuable opportunity.

Known Risks

The biggest risk Corinthia faces is failing to attract a larger community (not only developers but also testers and documenters). Actions have been taken and will continue to minimize the risk:

- Contact to student projects (in particular Capstone)

- Talks at ApacheCons

The project uses existing technologies, so there are no real technological risks.

There is of course a risk that nobody wants to use the project, but the fun building the community and project make this risk bearable.

Orphaned Products

None

Inexperience with Open Source

All initial committers have worked several years with open source.

Homogenous Developers

The initial committers are geographically distributed across the world. Half of the initial developers are experienced Apache committers and all have experience in working in distributed development communities.

The original source has already been partly refactored by other developers to make sure knowledge is spread among multiple people.

Reliance on Salaried Developers

No committers are being paid to participate.

Peter Kelly and Louis Suarez-Potts have a company that has added a commercial editor for iPad & iPhone on top of the library.

Relationships with Other Apache Projects

Corinthia has/will have a relation to at least the following projects:

- **AOO**, core developers have said on dev@ that for targeting mobile platforms, a rewrite of AOO would be better than building on top of the existing sources. It is our hope to have long and beneficial interaction with AOO.
- **Httpd**, we would like to make a module that on the fly presents odf/ooxml documents as pure HTML.
- **POI**, Corinthia library is similar to POI, but simpler, more generic, and written in C instead of Java. We hope to be able to share know-how as well as test cases.

Corinthia is based on document standards which are used by numerous high-profile projects. We would like to cooperate with the projects to exchange knowledge.

Documentation

The current documentation can be found at https://github.com/uxproductivity/Corinthia/wiki

The project is aware that this is work in progress and there is special attention on this task.

Initial Source

The source originated as part of the UX Write product. The file format conversion library, DocFormats, and the JavaScript + HTML5 editor are now under Apache License version 2 (ALv2).

The iOS-specific code from UX Write is not part of the grant and remains closed source.

Source and Intellectual Property Submission Plan

Source code will be moved from the GitHub uxproductivity/Corinthia repository when the incubator repository is set up. The content of the repository will be included in a cCLA grant from Peter Kelly. All original contributors are aligned with movement to an Apache Podling.

External Dependencies

The current source includes two third-party libraries to which minor modifications have been made(https://github.com/uxproductivity/Corinthia/tree/stable/DocFormats/3rdparty/external).

- **minizip**, a layer on top of zlib, http://www.winimage.com/zLibDll/minizip.html
- **w3c-tidy-html5**, an HTML5 manipulation library, https://github.com/w3c/tidy-html5

The changes made to the original sources are undocumented. The plan is to have pristine sources with documented changes in some manner. The existing licenses are ALv2 compatible and are honored. Any changes in the adaptation for Corinthia that are meaningful patches for the original code will be contributed upstream.

Furthermore, Corinthia depends on

- libxml2, http://xmlsoft.org
- zlib, http://www.zlib.net
- Simple DirectMedia Layer (SDL), http://www.libsdl.org/
- Showdown, http://github.com/showdownjs/showdown

Cryptography

Corinthia does not use cryptography nor does it delivery any cryptographic functions.

Required Resources

Mailing Lists

- corinthia-dev@incubator.apache.org for general discussions
- corinthia-private@incubator.apache.org for private discussions

- corinthia-issues@incubator.apache.org for issue-tracker notifications

Subversion Directory

- https://svn.apache.org/repos/asf/incubator/corinthia
  for management of the incubator web site

Git Repository

Our current git repository is on GitHub (https://github.com/uxproductivity/Corinthia).

It will be migrated to

- ASF Git repository incubator-corinthia.git

And mirrored at github through the apache organization.

Issue Tracking

A JIRA issue tracker is requested.

Other Resources

- Wiki: We are currently using github wiki, we would like to move to an Apache supported wiki (preferable mediawiki), before our documentation becomes difficult to move.
- Buildbot: We would like to be able to build/test on OSX, Windows and Ubuntu.
- Code signing: It is desirable to provide OS-compliant digital signatures with project-created convenience binaries of compiled libraries, utilities, and client applications.
- Web: We would like, if possible, to have the home page corinthia.incubator.apache.org (just raw html please).
- Blog: We would like to have a blog, preferable wordpress.

Initial Committers

Dennis E Hamilton, orcmid@apache.org

Dorte Fjalland, dorte@casacondor.com (ICLA recorded)

Jan Iversen, jani@apache.org

Louis Suárez-Potts, louis@apache.org

Peter Kelly, peter@uxproductivity.com (ICLA recorded)

Svante Schubert, svanteschubert@apache.org

Affiliations

None

Sponsors

Champion

Jan Iversen

Nominated Mentors

Jan Iversen (IPMC)

Daniel Gruno (IPMC)

Sponsoring Entity

Incubator IPMC.

---- ⚠️ **FINAL** ⚠️
This proposal is now complete and has been submitted for a VOTE.
