UimaProposal

Hello everyone -

We are submitting this proposal to the community for a new project in the incubator, and look forward to starting to work with this community.

This is a slightly modified and extended version of the proposal that has already been posted to general@incubator.apache.org. The whole mail thread can be found here.

If you don't feel like reading the whole thread, the main question that came up was:
this is all very well, but what does it really **do**? Attempts to answer that question where made here and here. We have since worked some of these into the proposal itself.

Proposal for Incubation Project: Unstructured Information Management Architecture - UIMA

Abstract

UIMA is a component framework for the analysis of unstructured content such as text, audio and video. It comprises an SDK and tooling for composing and running analytic components written in Java and C++.

Proposal: Unstructured Information Management Architecture framework

Unstructured Information Management applications are software systems that analyze large volumes of unstructured information in order to discover knowledge that is relevant to an end user. We propose UIMA, a framework and SDK for developing such applications. An example UIM application might ingest plain text and identify entities, such as persons, places, organizations; or relations, such as works-for or located-at. UIMA enables such an application to be decomposed into components, for example *"language identification"* -> *"language specific segmentation"* -> *"sentence boundary detection"* -> *"entity detection (person/place names etc.)"*. Each component must implement interfaces defined by the framework and must provide self-describing metadata via XML descriptor files. The framework manages these components and the data flow between them. Components are written in Java or C++; the data that flows between components is designed for efficient mapping between these languages. UIMA additionally provides capabilities to wrap components as network services, and can scale to very large volumes by replicating processing pipelines over a cluster of networked nodes.

This framework has already attracted a following among government, commercial, and academic institutions who previously developed analysis algorithms, but were unable to easily build on each other's works, and who want to be able to evolve their applications by independently upgrading parts, as better technology becomes available. Applications built with this framework are being used with plain text, audio streams, and image/video streams, identifying entities and relations, converting speech to text, translating into different languages, and determining properties of images.

The UIMA framework runs components in a flow, passing a common data object containing unstructured information (free text, audio, video, etc.) through the components. Each component examines the unstructured information and data added by other components, and adds data of its own. The framework mandates a standardized form of the data being passed, and a standardized form of the interfaces to the components.

We propose a project to develop, implement, support and enhance this framework (and, over time, other implementations) that comply with the UIMA standard (which has been submitted for standardization work within OASIS. Members of this community are encouraged to participate in that effort, as well; OASIS has an open approach to granting Technical Committee voting rights to members of OASIS, described here: http://www.oasis-open.org/committees/process.php#2.4.

The proposal includes both the framework, as well as tools to develop, describe, compose and deploy UIMA-based components and applications. The initial work will be based on the UIMA Version 2 framework code developed by IBM; snapshots of each release of this code are currently made available on SourceForge. The SourceForge versions would be stabilized in maintenance mode, if we are successful in moving to Apache. The framework is not specific to any IDE or platform, and does not depend on other middleware. Background:

Databases are core components of nearly all applications; they store information in structured tables. But more and more of the available digital data is unstructured (e.g. email, web documents, images, audio clips, video streams) with little information (metadata) attached to explain its content or context. Although many applications have been built to process unstructured data, they have either managed it as a BLOB or they have developed isolated applications for analyzing the content. In the absence of a standardized means for analytical applications to share insights extracted from the content, analytical applications cannot build upon one another. As a result, the industry has barely begun to tap the value locked in unstructured information.

Standardization is key to achieving component interoperability, with capabilities to mix components developed in different places and in Java, C++ and other languages. The Unstructured Information Management Architecture defines standards for component interoperability and application composition that will provide this needed unifying standard, and allow a variety of framework implementations to exist, while preserving the goal of unstructured information analytic component reuse.

UIMA was built to help developers create solutions that get more value from unstructured information more quickly and at lower cost by making it easy to reuse and combine analytic modules from different sources into new analytic applications. The architecture and the framework have been validated through work with USA's DARPA which is using it as a standard for key projects with several universities involved in advanced linguistics analysis, such as Carnegie Mellon, Columbia, Stanford and University of Massachusetts. Other companies, such as the Mayo Clinic and Sloan Kettering, are also building efforts around UIMA. In addition, over 15 software vendors, including companies such as Inxight, Attensity, ClearForest, Temis, SPSS, SAS, Cognos, Endeca, Factiva and others, announced plans to support UIMA.

The UIMA framework (binary and/or source code) has been downloaded over 8000 times from IBM alphaWorks (http://www.alphaworks.ibm.com/tech/uima) or SourceForge (http://uima-framework.sourceforge.net).

Rationale

We believe that moving the UIMA framework development to the Apache development community will lead to faster innovation, better integration with other open source software, and broader adoption of UIMA, accelerating the industry's ability to get the most value from text, audio, and video content. The UIMA framework is becoming attractive to developers who want to build components; we believe that having UIMA on Apache will encourage the development of a basic set of open source components that will jumpstart these developers' efforts. One of the first components we see possible synergy with is a search component based on Apache Lucene that would enable semantic search. We like the concept of the Lucene Sandbox as a way to encourage innovation around UIMA, and would envision something similar for this project.

Initial Goals

Some initial work we see in the incubator includes the following:

- redoing the parts of the tooling that were done as derivative works of Eclipse source code, to enable everything to be licensable under the Apache license
- extending the framework to better support "scale-out"
- extending the framework to better align with the emerging UIMA Standards work
- extending the framework to support XMI-based SOAP and/or other service interfaces
- extending the framework to support OSGi-based approaches to componentization and packaging
- exploring embeddings of the framework within other interested Apache projects, including synergies with Lucene
- providing aids to the community to migrate from previous versions of the framework to the Apache version
- setting up community support: hosting a facility similar to the Lucene Sandbox to encourage innovation and experimentation; establishing a wiki and some process to allow better documentation to be developed by the community, and linking our existing XHTML documentation via an XSL transform to Apache FOP

Current Status

Meritocracy

Meritocracy seems to us an ideal way to grow the community of developers around UIMA, it being a controlled, rational way to give those who positively contribute, more ability to directly contribute. This approach also gives contributors one of the best reasons to join the community of volunteers - to be recognized for the merit of their contributions.

Community

Currently, the UIMA Framework development is being done by IBM, with input from a group of early adopters in industry and government. Going forward, we see IBM continuing to support several committers working on it. We have already begun talking with other people outside of IBM that have expressed interest in contributing towards the development. This includes members of academic institutions, people working for some of the software vendors that have announced plans to support UIMA, and others from companies that have expressed interest since initial announcements about our open source plans. Multiple non-IBM people have already expressed desires to become committers.

Core Developers

The previous core developers of UIMA are Adam Lally, Thilo Goetz, Marshall Schor, Edward Epstein, Jaroslaw Cwiklik and Thomas Hampp. Many others have also contributed. The developers come from both the Research and Development parts of IBM.

Alignment

UIMA has significant synergy with search applications, and we expect to see integration with Lucene in the future. UIMA makes use of the Apache Portable Runtime (APR) for C++ support. It is designed to be embeddable into other frameworks, such as web application servers. Part of UIMA is Eclipse-based tooling. We use ANT for build scripting. UIMA has support for various language bindings including C++ and Java; we also have more limited bindings for Perl, Python, and TCL. UIMA uses Web Services as part of its approach to wiring up components in its domain. It makes use of XML services such as Xerces and Xalan.

The development of UIMA has been based on merit with open discussion among a distributed team of developers, from both Research and Development organizations.

License

The current license for the source code is CPL, with a small number of files licensed under the EPL (Eclipse Public License), because these were created as "derivative works" of existing Eclipse open source code. When the code base is moved to Apache, it will be relicensed under the Apache license, except for the small number of files licensed under the EPL as derivative works of Eclipse source files. We plan to work in the incubator to redo these parts, so the entire offering can be licensed under the Apache license.

The distribution for the C++ enablement layer includes open source components ICU (a Unicode package) which has its own license. We plan to work with community to properly make use of this non-Apache licensed component. Our current vision for the future of UIMA has it aligning with and incorporating other standards-based open source components/protocols, some of which may have licensing other than the Apache license (for example, the Xml Metadata Interchange (XMI), and the EMF ECore Model from Eclipse); we will work with the community in figuring out how to move forward on this.

Other IP

When we requested OASIS to set up a Technical Committee chartered to develop a platform-independent specification for text and multi-modal analysis, we specified that it be set up under the "RF on Limited Terms" mode of the OASIS IP Policy. "RF" means Royalty Free, and the Limited Terms means companies that are working with us on the Technical Committee are restricted in adding additional terms.

These are the most liberal terms and make any Essential Claims available to ALL and ROYALTY FREE. For the details please refer to:

- http://www.oasis-open.org/who/ipr/ipr_faq.php
- http://www.oasis-open.org/who/intellectualproperty.php

Ultimately of course, there is always a risk that someone in the world holds a patent that can be claimed as Essential. The most any standards organization can do is govern the behavior of those who participate in its work and publicly document the licensing commitment of all participants.

Known Risks

Orphaned Software

UIMA has been in active development for 5 years. The community of users has steadily grown, and there are now significant commercial and research organizations actively using it. UIMA is embedded in IBM software products and is delivered through IBM services engagements. IBM has developers assigned to it, and is continuing to support its development. In addition, several people outside of IBM have already expressed interest in working on UIMA, and have been providing IBM with initial feedback. One of the objectives of starting this Apache project is to provide a meritocratic structure for those people to begin more actively contributing to UIMA.

Inexperience with Open Source

The individuals working on this software have background as IBM software developers. While many of them have experience working with open source software, none of them has had extensive experience contributing to other open source software. However, IBM as an organization has extensive experience contributing to open source projects and will make available resources to provide guidance to the developers working on this project.

Homogenous Developers (work for same company?)

Currently all the developers work for IBM, although they come from different geographically dispersed organizations within IBM. We will reach out during the incubation time to get others to contribute; we have already received interest from several parties.

Reliance on salaried developers

Currently the developers are paid employees of IBM.

Relationships with Other Apache Products

We make use of several Apache components (SOAP / Web Services, XML (Xerces, Xalan), languages (Perl), scripting languages (ANT), Apache Portable Runtime. In addition, UIMA has been embedded in other frameworks, such as web application servers, and integrated with search engines. We are exploring Lucene extensions that could take advantage of UIMA processed data. We are currently investigating and prototyping some software packaging concepts based on OSGi; the Apache Incubator project Felix may have relevance as we go forward. The documentation is being moved to XHTML and plans to use Apache FOP for producing PDF reference materials.

An Excessive Fascination with the Apache Brand

UIMA is already being adopted by a wide cross section of users, both commercial and academic, world-wide. Our experience shows that analytic modules can be reused and combined through UIMA making it easier and faster for developers to build new analytic applications for specific industries or domains. Given the diversity of content and analytics that will be required to address the multitude of opportunities - from military intelligence to quality assurance to contact center analytics – growing this infrastructure so that it better aligns with other major Open Source communities should help accelerate industry's ability to get value from content assets.

We believe that the Apache community of developers has the experience, background, visibility, and synergistic resources to encourage and foster a vibrant developer community around this project.

Documentation

There is a combination Introduction, Conceptual Overview, Tutorial, Tools and Framework User's Guides and References, downloadable from http://dl.alphaworks.ibm.com/technologies/uima/UIMA_SDK_Users_Guide_Reference_2.0.pdf

Scope of the project

The project will develop implementations of the UIMA architecture (which is concurrently being submitted to the OASIS standards process), supporting the breadth of platforms that developers working in this field are using, including Java, C++, Perl, Python and TCL; and utilities and tooling to support component and application developers and assemblers / packagers. It will initially include the Java UIMA framework for UIMA Version 2 (you can see a snap shot of the Version 2 release Source{{`Forge; the delivered code would this code base plus normal incremental bug fixes and improvements), plus additional components (mainly documentation and test cases, which are not currently on Source}}`Forge). Over time, the project is expected grow to include supporting various embeddings and integrations with other Apache components such as search engines and web application frameworks.

Over time, we envision the project becoming an umbrella for related open-source around UIMA, including things like open-source pre-annotated corpora, and hosting a facility similar to the Lucene Sandbox to encourage innovation and experimentation.

The UIMA framework is primarily a set of libraries (in Java, C++, Perl, etc.), test cases, and UIMA utilities and tools (scripts, plugins, executables, etc.) used to build, test and debug UIMA analytic components. The tooling includes several Eclipse platform plugins.

Initial source

The source currently is maintained in IBM internal software control systems, with a copy of each release placed on SourceForge. At the time of launch, we plan to contribute the latest version of the code base (with some renaming of package prefixes to reflect apache.org), test cases, build files, and documentation, under the terms specified in the ASF Corporate Contributor License. We plan to donate the existing C++ enablement layer and the support for Perl, Python, and TCL a few months later than the initial donation; this delay is to give us time to finish preparing that code base for Open Source.

ASF resources to be created

Mailing lists:

- uima-dev
- uima-commits
- uima-user (we already have a substantial user community and expect them to turn up at Apache soon after we've hopefully been accepted into the incubator)

For other resources such as Subversion repository, JIRA etc. we hope for guidance from our mentors.

Initial Set of Committers

- Michael Baessler (mba@michael-baessler.de)
- Edward Epstein (eddie_epstein@aewatercolors.com)
- Thilo Goetz (twgoetz@gmx.de)
- Adam Lally (alally@alum.rpi.edu)
- Marshall Schor (msa@schor.com)

Sponsor

We are requesting the Incubator to sponsor this. Our current vision is that it will become a top level project (other projects that develop UIMA components could become subprojects, for instance).

Mentors

- Sam Ruby (ruby@apache.org)
- Ken Coar (coar@apache.org)
- Ian Holsman (lists@holsman.net)

Section 6: Open Issues for Discussion