Any23Proposal

Any23

Abstract

The following proposal is about *Anything To Triples* (shortly Any23) defined as a Java library, a Web service and a set of command line tools to extract and validate structured data in RDF format from a variety of Web documents and markup formats. Any23 is what it is informally named an *RDF Distiller*.

Proposal

Any23 "Anything to Triples" is a library written in Java 6 and released under the Apache 2.0 License. It provides a set of extractors for scraping semantic markup (such as Microformats, RDFa and Microdata) from several sources (HTML4, XHTML5, CSV), a set of data validations, a set of parsers and writers to handle the main RDF transport formats (RDFXML, Ntriples, NQuads, Turtle). The library provides a command line tool for dealing with data extraction, conversion and validation, and a REST service implementation. The library is plugin based, allowing the hot loading of new extractors and validators. Any23 enables third-parties developers to access structured data from Web pages without the need of implementing ad-hoc scraping techniques. In this sense, Any23 will relieve developers from build complex solutions when developing data acquisition pipelines and processes targeted to semantically marked-up Web data.

Background

Any23 has been initially developed at DERI (Digital Enterprise Research Institute), as main component of the RDF extraction pipeline used in Sindice (the Semantic Web Index), now is evolved in joint effort with FBK (Fondazione Bruno Kessler). At present time the Any23 official developers page contains all the documentation, while the code is maintained on Google Code. An official up-to-date showcase demo is also available.

Rationale

Provide and maintain a robust, standard and updated library for extracting and validating semantic markup from heterogeneous sources would provide large benefits to the entire Open Source Community. Researchers and academic projects are adopting RDF related technologies from years while the industry is actually moving toward Semantic Web technologies with more concreteness. Several industry initiatives related to the Web of Data are taking place in the these months. Schema.org, for example, is an initiative sponsored by Google Inc, Yahoo Inc and Microsoft Corporation to structure the data in a harmonized way on HTML5 pages. Schema.org leverages on the HTML5 Microdata native specification. OpenGraphProtocol is the open standard sponsored by Facebook Inc to include metadata in HTML page headers. OpenGraphProtocol, initially based on RDFa, allows to describe the content of a Web page and its underlying vocabulary could be directly represented using RDF.

Current Status

Meritocracy

The historical Any23 team believes in meritocracy and always acted as a community. Mailing list, open issue tracker and other communication channels have always been adopted since its first release. The adoption in a larger community, such as Apache, is the natural evolution for Any23. Moreover, the Apache standards will enforce the existing Any23 community practices and will be a foundation for future committers involvement.

Core Developers

In alphabetical order:

- Davide Palmisano <dpalmisano at gmail dot com>
- Giovanni Tummarello <giovanni dot tummarello at deri dot org>
- Michele Mostarda <michele dot mostarda at gmail dot com>
- Richard Cyganiak <richard at cyganiak dot de>
- Reto Bachmann-Gmuer <reto at apache dot org>
- Simone Tripodi <simonetripodi at apache dot org>
- Szymon Danielczyk <danielczyk.szymon at gmail dot com>
- Tommaso Teofili <tommaso at apache dot org>

Alignment

Main aim of the project is to develop and maintain a fully flavored semantic markup distiller that can be used by other Apache projects that need an RDF extraction tool. The Any23 library core is written using the following Apache libraries.

- Apache Commons Lang
- Apache Commons HTTP Client
- Apache Commons Codec
- Apache Tika
- Apache Commons CLI

- Apache POI

The Any23 service is targeted to run within any compliant Servlet container like Tomcat.

Known Risks

Orphaned Products

The increasing number of Any23 adopters and the raising interest for Semantic Web related technologies let us believe that there is a minimal risk for this work to being abandoned from the community. Moreover Any23 has already been used in production by Sindice.com and other DERI projects for years.

Inexperience with Open Source

All of the committers have experience working in one or more open source projects inside and outside ASF.

Homogeneous Developers

The list of initial committers are geographically distributed across Europe with no one company being associated with a majority of the developers. Many of these initial developers are experienced Apache committers already and all are experienced with working in distributed development communities.

Reliance on Salaried Developers

To the best of our knowledge, the biggest part of the initial committers is being paid to develop code for this project due to the adoption of Any23 in their organizations infrastructures. In any case, some of the core historical developers (some of them no longer getting paid from the original companies behind Any23) are still committing even if Any23 is not employed in their actual organizations. Any23 has already proven its capability to attract external developers.

Relationships with Other Apache Products

In the last years, other projects have been under ASF incubation process relying on the Semantic Web technology stack, such as Apache Clerezza, Stanbol and Jena. This could be seen as a proof of the consolidation and the adoption growing tendency of such technologies. Apart the specificity of those projects, sharing the same underlying stack, Any23 could be employed in every projects needing a reliable framework to access structured semantic markup. Any23 core could be easily released also as a Apache Nutch Plugin and then, used to handy fill SAIL-compliant triple stores.

An Excessive Fascination with the Apache Brand

Even if the Any23 community recognizes the power and the attractiveness of the ASF brand, we are absolutely aware of our already established role in the wider Semantic Web developers community. Any23 already proved its reliability in closely support all the new specifications coming from the Microformats communities, our major contributors in term of opened issues about new feature requests. Furthermore, we are convinced that we can enthusiastically bring inside the ASF new and fresh energies in order to improve our visions, insights and knowledge about the other projects and, most important, to have the possibility of enlarge our small community with talented and passionate developers.

Documentation

Any23 Documentation

1. Any23 Project Homepage
2. Any23 Developer Homepage
3. Any23 Live Demo

Any23 Related Specifications

1. RDF
2. HTML5
3. RDFa
4. Microdata
5. Microformats
6. RDF/XML
7. Turtle
8. N-Triples
9. N-Quads

Any23 Other documentation

1. Any23 presentation on Slideshare

Initial Source

The intial source comprises code developed on GoogleCode licensed under the Apache License 2.0 (to be contributed under Grant from Giovanni Tummarello for Any23).

Source and Intellectual Property Submission Plan

Source code will be moved from GoogleCode space inside the SVN space of the podling.

External Dependencies

All the external dependencies (and their licenses) used by Any23 follows:

- Nekohtml (Apache 2.0)
- OpenRDF Sesame (BSD-style license)
- Jetty (Apache License 2.0 and Eclipse Public License 1.0)
- Java Simple Plugin Framework (new BSD License)
- Boilerpipe (Apache License 2.0)
- slf4j (MIT License)
- junit (Common Public License - v 1.0)
- Mockito (MIT License)

Cryptography

The project does not handle cryptography in any way.

Required Resources

- Mailing lists
  - any23-private (with moderated subscriptions)
  - any23-dev
  - any23-user
  - any23-commits
- Subversion directory
  - https://svn.apache.org/repos/asf/incubator/any23
- Website
  - Confluence (ANY23)
- Issue Tracking
  - JIRA (ANY23)

Initial Committers

Names of initial committers - in alphabetical order - with current ASF status:

- Chris Mattmann <mattmann at apache dot org> (Member)
- Davide Palmisano <dpalmisano at gmail dot com> (ICLA signed)
- Giovanni Tumarello <giovanni dot tummarello at deri dot org> (ICLA signed)
- Lewis John McGibbney <lewismc at apache dot org> (PMC Member)
- Michele Mostarda <michele dot mostarda at gmail dot com> (ICLA signed)
- Paul Ramirez <pramirez at apache dot org> (Member)
- Reto Bachmann-Gmuer <reto at apache dot org> (Committer)
- Szymon Danielczyk <danielczyk.szymon at gmail dot com> (ICLA signed)

Sponsors

Champion

- Chris Mattmann <mattmann at apache dot org> (Member)

Nominated Mentors

- Chris Mattmann <mattmann at apache dot org>
- Nick Kew<niq at apache dot org>
- Paul Ramirez <pramirez at apache dot org>

- Simone Tripodi <simonetripodi at apache dot org>
- Tommaso Teofili <tommaso at apache dot org>

Sponsoring Entity

- Tika PMC

Other interested people (in alphabetical order)