TavernaProposal

Authors: Stian Soiland-Reyes, Shoaib Sufi (University of Manchester)

Contributor: Andy Seaborne (Apache foundation)

Submitted: 2014-09-23 (Originally published in Taverna wiki)

Abstract

Taverna is an open source and domain-independent suite of tools used to design and execute data-driven workflows.

Proposal

The Taverna suite includes:

- **Taverna Workbench**, a desktop application written in Java for graphically composing, editing and executing workflows composed of distributed Web services and local tools
- **Taverna Command Line Tool**, which allows execution of workflows from a command line
- **Taverna Server**, which provides a REST and SOAP API for executing workflows
- **Taverna Player**, a Web interface the Taverna Server written in Ruby towards, providing a high-level view of workflow executions and their results and allowing further integrations with other Ruby on Rails applications

Taverna allows browsing through and combining different service types in workflows, allowing them to integrate steps of arbitrary REST and SOAP Web services with command line tools (local and via SSH), scripts (Beanshell, R, Jython), and finally to visualize the results.

The goal of the Taverna suite is to help researchers to access distributed datasets and processing capabilities by the construction of (data) pipelines, and also to simplify the execution of these pipelines in various environments.

The Taverna suite of products is already successful and in wide use across different domains. The software is currently licensed as LGPL 2.1, with copyright owned by the University of Manchester. External contributors have all signed Apache-like CLAs.

Background

Taverna *workflows* coordinate inputs and outputs between computational processes and Web services. The workflow is designed in a graphical interface which shows the workflow as a series of boxes connected with arrows representing processes (i.e. executable services) and their data connections. Different *processes* in a workflow can be command line tools, REST and WSDL *Web services*; which are used for combining steps such as data acquisition, filtering, cleaning, integrating, analysis and visualization. Taverna calls these processes *"services*", as they generally are provided by remote (third-party) servers.These kind of computational workflows, also known as *pipelines* or *dataflows,* focus on the movement of data rather than the execution order of the underlying processes. Features such as *implicit iterations* (where an input list of values causes multiple process executions) and *par allel invocations* (independent processes are executed as soon as their data is available) are intrinsic to a dataflow system, not requiring any particular constructs by the workflow designer.As a *visual programming environment*, workflows aids collaboration and reuse of workflows. At the highest level, a workflow represents the conceptual level of an analysis, allowing understanding, discussion and communication of the overall analysis protocol. More detail can be revealed and modified for individual steps. At the individual process level, the workflow defines execution specifics such as operations, parameters and command line tools.Sharing of the workflow definitions allows re-use and re-purposing of the computational analysis. During workflow execution, *provenance* can be collected from every step, allowing deep inspection of intermediate values for the purpose of debugging and validation.

Rationale

There is a strong need to lower the barrier of entry to datasets and computational resources widely available on the Internet, to increase their use by researchers who understand the computational steps needed to produce their results, but who are not necessarily expert programmers. Taverna has already shown its success and popularity in a wide range of scientific disciplines.

Initial Goals

- Transition mailing lists to Apache (keep existing subscribers, but invite more)
- Taverna developer workshop (2014-10-30)
- Fully investigate/resolve incompatibly licensed dependencies
- Stage git repositories for move at https://github.com/taverna-incubator :
    - Update headers/metadata to indicate Apache License 2.0
    - Restructure git repositories (to ~ 10 repos?)
    - Rename Maven groupIds to org.apache.taverna.*
    - Rename packages to org.apache.taverna.*
- Move staged Github repositories to Apache git
- Automated builds in Apache's Jenkins
- Update to latest releases of Apache dependencies
- Propose updated release and testing procedure under Apache
- Moved Website and documentation

We intend to only release the current development version Taverna 3 under the Apache umbrella. 3.0 is not yet officially released - however the Taverna 3.0 Command Line can be released almost "as-is" after migration. The Taverna 3.0 Server is at beta quality, while the Taverna 3.0 Workbench is at alpha stage and would need to be stabilized to an initial beta release.

- Before first release: Maven Central releases of Taverna support libraries (e.g. taverna-scufl2 and taverna-databundle)
- First release: Apache Taverna Command Line 3.0 (OSGi-based)
- Release: Apache Taverna Server 3.0
- Release: Apache Taverna Workbench 3.0 beta
- Provenance exchange with relevant Apache products (e.g. Apache CXF->Taverna->CouchDB)
- Release: Apache Taverna Workbench 3.0

It is not yet decided if the current Workbench Editions will be carried over to Taverna 3, or if this can be solved by having a "Install extra plugin" step on first start-up of Apache Taverna. In any case, we imagine that some of these specializing editions will be maintained outside (but in collaboration with) the Apache project. This is particularly the case for the Astronomy edition as it depends on several LGPL/GPL libraries and is maintained by the AstroTaverna team.

Current Status

Meritocracy

Taverna was initially created by the myGrid consortium in 2003. Since 2006, the majority of contributions to Taverna's core code-base, its architecture and direction have been led by staff at Tthe University of Manchester and the European Bioinformatics Institute (EMBL-EBI).

The project have benefited of a high-degree of extensions and integrations by other developers - but mainly in the form of plugins and integrations, including Taverna Online.

Taverna's developer community have unfortunately not had a culture of submitting patches that would warrant later commit access - perhaps due to its background in the science community. However contributors have been added as committers when the plugin becomes a part of the core distribution (e.g. External Tool plugin by Möller and Krabbenhöft and AstroTaverna by Garrido), or when their development has required patches to the existing code base.

Community

Taverna has an active community of plug-in developers and users. The developer mailing list (taverna-hackers@lists.sourceforge.net) has 248 members, the user mailing list (taverna-users@lists.sourceforge.net) has 370 members.

1500 users have registered as of 19 August 2014. Total downloads of all products since version 2.1 (released December 2009) is 35000.

Them Taverna Developers Workshop is being arranged for 30 October 2014 to bring together developers and integrators of Taverna. We want to encourage plug-in developers to participate further in the core development of Taverna as well, by introducing them to the code base and how to contribute.

Active steps to grow the communities of users and developers by targeting specific research domains such as the work by Kevin Benson on Taverna's use in the Heliophysics and Astrophysics community. Susheel Varma is helping increase the usage of Taverna within the Biomedical domain. Julián Garrido and his work on *AstroTaverna* is promoting Taverna within the IVOA Virtual Astronomy community. Sonja Holl and Björn Hagemeier's are targeting high performance computing.

Core Developers

What we currently consider to be the core Taverna Team is (in alphabetical order):

- Christian Brenninkmeijer (University of Manchester)
- Donal Fellows (University of Manchester)
- Robert Haines (University of Manchester)
- Aleksandra Nenadic (University of Manchester)
- Dmitry Repchevsky (Barcelona Supercomputing Center)
- Stian Soiland-Reyes (University of Manchester)
- Shoaib Sufi (University of Manchester)
- Vadim Surpin (Institute for Information Transmission Problems in Moscow)
- Alan Williams (University of Manchester)

The team consists of experienced developers who have worked on a multitude projects, particular within writing software for supporting scientists. The committers list (see below) includes additionally plugin developers whose contributions have become part of Taverna. Part of our desire to join the Apache Foundation is to recognise their effort and promote them into also being "core developers".

Alignment

Taverna dependencies include Apache Commons, Axis, Abdera, Batik, CXF, Derby, Felix, HttpComponents, Jena, log4j, Maven, POI, Velocity, Xerces, XMLBeans, Xalan, We use Tomcat for testing and deployment of the Taverna Server.As part of moving to Apache-compatible dependencies, Taverna will probably adopt OpenJPA to replace (LGPL) Hibernate.

Known Risks

Orphaned products

Most of the core developers are from the myGrid team at the University of Manchester, but are funded through a series of projects. Many of these projects incorporate Taverna, so the effort from Manchester is partially based on direct project requirements, but also partially on a volunteer effort for project maintenance and general development. The myGrid team has guaranteed funding until 2017.

The developers that are outside Manchester are generally funded for other activities, and so their effort to Taverna is to a greater extent a volunteer effort - although again project-specific requirements steer their effort (e.g. for a new Taverna plugin).

One of the reasons for our desire to move to the Apache Foundation is to formalise this volunteering/contribution effort so that it becomes obvious that it is not just the University of Manchester that is contributing to the core code base - and therefore reducing the impression that Taverna is vulnerable to Manchester's future funding and projects.

Inexperience with Open Source

Taverna has been an open-source project since its first release in 2003. Most of the contributors also have experience with working with and contributing to other open source projects (e.g. TCL, CXF, Jena), particularly as Taverna strongly relies on other open source tools. Most of the research projects which the myGrid members have participated in produces open-source software.

Homogeneous Developers

The committers' list includes many people from the myGrid group from the University of Manchester in United Kingdom - but these developers have been working on a range of distributed and European projects in the field of scientific software.

The other developers on the committers' list come from many different projects and institutions across the world, e.g. Russia, Canada, Germany and Spain.

Reliance on Salaried Developers

Development of Taverna is mainly performed as part of the developers' salaried work, but funded through many different projects at several institutions (see above). These projects do not generally have "contribute to Taverna" as their main goals - so therefore in many ways the effort is still volunteer-based - contributing to Taverna as a way to support one's own work.

From our experience of running Taverna over the last 10 years, new contributors will continue to join as Taverna becomes an ingredient in new projects, while existing contributors more slowly fade out of their involvement. Often existing contributors and users gives the personal link to the new contributors.

Relationships with Other Apache Products

Apache already contains projects that seem relevant to Taverna.

Apache Airavata http://airavata.apache.org/ is a software framework for executing and managing computational jobs and workflows on distributed computing resources. Taverna's concern is not as much job coordination, but more of a data flow between services. Airavata's XBaya Workflow Suite can export workflows in Taverna 1 format SCUFL, but could be updated to work with Taverna 3's SCUFL2 format.

Apache ODE https://ode.apache.org/ is a WS-BPEL workflow engine. BPEL as a workflow language is quite verbose compared to dataflow languages like Taverna, and is additionally bound to a particular protocol (SOAP). Nevertheless, a sub-section of Taverna workflows could in theory run on the Apache ODE engine - and the Taverna 3 Platform API has facilities for plugging in alternative workflow engines. We have previously considered Apache Hadoop as one such alternate engine for executing a different subset of workflows with local command line tools.

Apache OODT http://oodt.apache.org is a scientific data processing and data management system. OODT has a workflow manager, a file manager, and a resource manager, along with client-side frameworks including an automatic remote file acquisition system; automated crawler, and science algorithm wrapping facility. OODT and Taverna could benefit from cross pollination.

Apache Pig https://pig.apache.org/ is a high-level language for creating Map-Reduce programs for Apache Hadoop. There already exists third-party efforts to convert Taverna Workflows to Hadoop and Pig - https://github.com/umaqsud/taverna-to-pig https://github.com/schenck/taverna-to-hadoop (thus making a graphical interface for building Apache Pig workflows) - and part of the Apache Taverna effort would be to invite these to join the project.

Apache Storm http://storm.incubator.apache.org/ is a distributed real-time computation framework. Experiments are under development to use Taverna as a front-end for creating Apache Storm workflows - http://markmail.org/message/zg5ylo2aucpwfc5j

Apache has several popular frameworks for building REST/SOAP web services (Apache CXF, Apache Clerezza), data services (Apache Jena, Apache Hive, Apache CouchDB) and specific workflow engines (Apache Oozie for Hadoop, Apache ODE for WS-BPEL). Taverna as a general REST and SOAP service client can be used for combining, testing and demonstrating such services.

An Excessive Fascination with the Apache Brand

Taverna is a long-running project (since 2003) with an existing user- and developer base across the academic world. Our main motivation for moving to Apache is to further encourage an open development process and engage existing and new developers to contribute to the core code base. We also want to ensure long-term continuity of the Taverna products, and for its future directions to be decided by the whole Taverna community rather than one of the parties involved.

Documentation

Taverna's documentation is available from http://www.taverna.org.uk/documentation/taverna-2-x/, including the extensive user manual, tutorials and videos.

The developer documentation includes developer tutorials for working with Taverna's source code and creating plugins.

Initial Source

Taverna's source code is available from the 'taverna' github team account: https://github.com/taverna/. These 85 git repositories reflect the current modules of Taverna's plugin system after recently transitioning from Google Code SVN at http://taverna.googlecode.com/svn/taverna/. The history of Taverna's code base goes back to being hosted in CVS at SourceForge http://taverna.cvs.sourceforge.net/, transitioned as of http://taverna.googlecode.com/svn/archived/cvs2svn-2008-09-25/. Note that reasonable steps have been made to preserve commit history when moving between version control system, this has not always been achieved when moving between modules and refactoring larger Java packages. Some source files might therefore in git have initial commits like "Moved from /taverna/utils/trunk" referring to SVN paths.

One of the reason for many repositories is that we rely on Apache Maven and a plugin system (since Taverna 3 OSGi-based) where different modules have different version numbers and release cycles (e.g. tags/branches). This is essential for the plug-in support of Taverna as the plug-ins depend on the semantic versioning of the APIs and required implementations.

It is however in our current plans to merge repositories that have similar release cycles and greatly reduce the number of repositories, to about 10 repositories that would be imported to Apache's Git server.

We suggest that this would be the first phase of the incubator project, to prepare and stage the merged repositories to https://github.com/taverna-incubator

Taverna source code uses the package names (and children packages):

- net.sf.taverna - since Taverna 2
- uk.org.taverna - new from Taverna 3
- org.taverna (sic) - Taverna Server

Some contributed code uses package names depending on their originating projects:

- org.purl.wf4ever.provtaverna
- org.biomart.martservice

We intend to release only the upcoming Taverna 3.0 version under the Apache umbrella (not 2.x) - therefore, according to semantic versioning rules http://semver.org/, the transition period of the Apache Incubator would be the best (and possibly only) chance to rename Java packages and Maven groupIDs to *org.apache.taverna.** Under OSGi the packaging and JAR goes hand-in-hand (several JARs don't normally provide the same package), and therefore any package rename would be done together with the repository restructuring.

Source and Intellectual Property Submission Plan

- Taverna source code from http://github.com/taverna/ (to be staged as a reduced list of repositories at https://github.com/taverna-incubator)
    - (c) University of Manchester.
    - Signed Apache-like CLAs for all external contributors.
    - Current license is LGPL 2.1 (and GPL3 for one domain-specific download), as sole copyright holder Manchester will change this to Apache License 2.0
    - Check-out-all-and-build meta project - https://github.com/taverna/taverna-build
- *taverna.org.uk* domain - registrant University of Manchester

- http://www.taverna.org.uk/ content (c) University of Manchester
- http://dev.mygrid.org.uk/wiki/display/tav250/ Confluence wiki content (c) University of Manchester
- http://dev.mygrid.org.uk/wiki/display/developer Confluence wiki content (c) University of Manchester

The details of intellectual property submission will be worked out together with myGrid project manager Shoaib Sufi and the University of Manchester's Contracts Office.

As University of Manchester is the copyright holder of all the Taverna Source code (either directly or through signed CLAs), we are able to change the license to Apache License 2.0 wholesale.

External Dependencies

Taverna, as an integrating workflow system, has a fairly large number of dependencies - the latest 2.5.0 Core Workbench distribution has 517 JARs (although many of those are duplicates in different versions)

We are intending for our first Apache-based release to be Taverna 3, which has already reduced this dependency list.

We have performed an analysis of Taverna 3 dependencies - this list should be complete for the dependencies (and their transitive dependencies) of Taverna Workbench.

The internal dependencies that are managed by Taverna/myGrid would need to be part of the transition to Apache so that their license can change from LGPL 2.1 to Apache License 2.0. As we will change groupId at the same time to *org.apache.taverna*, it should be fairly trivial to ensure that no JARs from the original Taverna repositories are included in the first Apache releases, as

1. They are only available from the Taverna Maven repository
2. Their groupId (*net.sf.taverna*/*uk.org.mygrid*/*uk.org.taverna*) would be easy to identify in the distribution folder.

We know that some of the external dependencies are licensed as LGPL, and for AstroTaverna, some dependencies are licensed as LGPL. As Apache License is incompatible with *GPL (but not vice versa), the general solution we suggest for this is to either:

- Try to use alternative non-LGPL dependencies, aka. Apache JPA instead of Hibernate
- Keep module that requires LGPL dependency as a separate Taverna plugin, maintained and published independently at Github (e.g. https://github.com/wf4ever/astrotaverna).

In our analysis of Taverna's third party licenses we have identified the incompatible GPL/LGPL dependencies and suggested a resolution that will be performed as part of incubation.

We found a list of dependencies with unknown licences (not declared through Maven). Part of incubation is to fully resolve this list as it could be hiding additional incompatible dependencies. (In many cases, simply using a newer version will include licensing information.)

Cryptography

Taverna uses these cryptography dependencies:

- BouncyCastle
- OpenJDK builds with the default JCE full/strong encryption policy (bundled in installer)

Taverna utilise these to form of an encrypted keystore (storing username/password and client certificates for third-party services accessed by the designed workflow) with corresponding user interface, and additionally binds to Java's SSL support to provide UI and command line options for security interactions, e.g. accepting new server certificates, or asking for username/passwords for HTTP Basic Authentication (which can then be stored in the keystore).

Required Resources

Taverna currently relies on a mixture of infrastructure hosted for free by third-parties (e.g. Github, SourceForge, GoogleCode, Launchpad, Bitbucket) and infrastructure hosted by the myGrid group at the University of Manchester (Jenkins, Jira, Confluence, Wordpress).

Mailing lists

Existing mailing lists for Taverna are hosted at Sourceforge with archives at markmail. See http://www.taverna.org.uk/about/

- commits@taverna.incubator.apache.org (replacing taverna-cvs@lists.sourceforge.net)
- private@taverna.incubator.apache.org (replacing support@mygrid.org.uk - to a lesser degree as we would want to encourage openness)
- dev@taverna.incubator.apache.org (replacing taverna-hackers@lists.sourceforge.net , 240 members)
- users@taverna.incubator.apache.org (replacing taverna-users@lists.sourceforge.net , 370 members)

Git repositories

The Taverna community would prefer to keep using git and Github, and we would request for experimental writable git repositories http://www.apache.org/dev/writable-git with mirroring to Github.

The repositories would be named taverna-*, as the current repositories on the github team: https://github.com/taverna/. This repository organization is styled equivalent to the git repositories of cordova-* and couchdb-*.

Exactly how repositories are split/merged is open for discussion - it is part of our current plan to reduce the number of repositories by merging common modules with a similar release cycle - this could be done at an early phase of the incubation period.

Issue Tracking

JIRA Taverna (TAV)

Existing issues in Taverna 3's current JIRA - http://dev.mygrid.org.uk/issues/browse/T3 - should be imported - but its current list of Modules should be further agreed.

Other Resources

- Wiki spaces in Confluence https://cwiki.apache.org/confluence - importing the most recent Taverna-related spaces and documentation from http://dev.mygrid.org.uk/wiki/spacedirectory/view.action?startIndex=24
- Jenkins - replacing myGrid Jenkins at http://build.mygrid.org.uk/ci/
- Maven repository at https://repository.apache.org/ - replacing myGrid artifactory http://repository.mygrid.org.uk/
- File-based Web space for Plugin Update Site - replacing http://updates.taverna.org.uk/ and http://www.mygrid.org.uk/taverna/updates/
- Home pages - to be transitioned from from http://www.taverna.org.uk/ (Wordpress)
- Binary distribution download hosting, about ~8 GB pr release, replacing http://www.taverna.org.uk/download/ (currently downloads are hosted by http://launchpad.net/ and https://bitbucket.org/)

Initial Committers

The initial list of committers reflect the current list of active developers at the Github team: https://github.com/orgs/taverna/people (Note that not all of these have made their membership public on Github)

Name	E-mail	Apache CLA?	Apache account
Alan R Williams	alaninmcr googlemail.com	Confirmed	alaninmcr
Aleksandra Nenadic	a.nenadic manchester.ac.uk	Confirmed	anenadic
Christian Y. Brenninkmeijer	brenninc cs.man.ac.uk	Confirmed	brenninc
David Withers	david.withers gmail.com		withers ? (need CLA)
Dmitriy Repchevsky	dmitry.repchevski bsc.es	Confirmed	redmitry
Donal K. Fellows	donal.k.fellows manchester.ac.uk	Confirmed	dkf
Finn Bacall	finn.bacall manchester.ac.uk		fbacall ? (need CLA)
Hajo Nils Krabbenhöft	hajo krabbenhoeft.de		fxtentacle ? (need CLA)
Ian Dunlop	ian.dunlop manchester.ac.uk	Confirmed	ianwdunlop
-Ingo Wassink -	-ingo.wassink nedap.com -	*Opted out*	
Julián Garrido	tetrarquis gmail.com	Confirmed	jgarrido
Mark Wilkinson	markw illuminae.com		markwilkinson ? (need CLA)
Luke McCarthy	elmccarthy gmail.com		elmccarthy ? (need CLA)
Robert Haines	rhaines manchester.ac.uk	Confirmed	hainesr
Shoaib Sufi	shoaibsufi gmail.com	Confirmed	shoaibsufi
Steffen Möller	moeller inb.uni-luebeck.de		smoe ? (need CLA)
Stian Soiland-Reyes	stian soiland-reyes.com	Confirmed	stain
Stuart Owen	stuzart gmail.com	Confirmed	stuzart

In addition to the Core Team (mentioned earlier), this list also reflects Taverna's existing meritocracy as it includes plugin developers whose contributions have been merged into the main code base. We acknowledge that not all of these are likely to continue as "Core" developers, but would like to encourage that during the Incubating process.

Affiliations

The majority of the initial committers are employed by University of Manchester as part of the myGrid team, including responsibilities for contributing to and supporting Taverna. http://www.mygrid.org.uk/about-us/people/core-mygrid-team/.

Dmitriy Repchevsky is employed by the Barcelona Supercomputing Center, including responsibilities for contributing to Taverna. Steffen Möller is employed by University of Lübeck. Julián Garrido is employed by Instituto de Astrofísica de Andalucía.

Sponsor Champion

Andy Seaborne

Nominated Mentors

- Andy Seaborne
- Chris Mattmann
- Suresh Srinivas
- Suresh Marru

- Marlon Pierce

Offers of participation, not formally a mentor:

- Michael Joyce

Sponsoring Entity

The Incubator.