Open-Source Software in the Scientific World

Case study: Triana Software

Adina Riposan\textsuperscript{1}, Ian Taylor\textsuperscript{2}

\textsuperscript{1}Military Technical Academy, and Contact Net Ltd, Bucharest, Romania
\texttt{adina.riposan@contactnet.ro}

\textsuperscript{2}School of Computer Science, Cardiff University, UK, and Center for Computation and Technology, LSU, USA, \texttt{ian.j.taylor@cs.cardiff.ac.uk}

Resume

\textit{Triana Workflow Environment} (\url{http://www.trianacode.org/}) was created as an Open-Source scientific environment, and more than 500 applications have been further developed based on Triana in support to scientific groups around the world, for example in: \textit{radio astronomy}, astrophysical simulations, galaxy visualization, gravitational wave analysis, data mining, biodiversity problems, grid-enabled medical simulations, environmental science. The Triana environment is used for problem solving and orchestrating flows of operations/services, both in fine-grained dataflow applications and as a course-grained distributed workflow system.

Workflow is implicit in scientific algorithms that specify a series of inter-dependent operations to be executed, to connecting such algorithms in a series of derivations, which when aggregated perform some higher-level task. Based on Triana, the \textit{Alchemist Infrastructure} was further designed, to provide for a new paradigm in search and discovery of distributed resources, based on multimodal workflows and the coupling of metadata fusion and social tagging with the more traditional index-based search techniques.

More recently, we have approached the application of Triana and the Alchemist in the field of \textit{Biomedical research}, in support for biomedical image and spectral data discovery, where metadata generated from user and automatic annotations is fused and \textit{semantics} can be included as part of the query. The initial use-case is the utilization of the Triana/Alchemist infrastructure in \textit{Diabetic Retinopathy} research and clinical trials, for the early detection and prevention of retinal disease and investigational drug discovery.

1. Introduction

Triana Workflow Environment (\url{http://www.trianacode.org/}) has been created as an Open-Source environment for scientific communities, and many other software components have been developed around Triana and its capabilities during several stages of evolution. These open-source systems and the underlying technologies have provided substantial support to the scientific groups in their work in various fields such as \textit{astrophysics} and \textit{gravitational waves research}, \textit{audio research}, and more recently \textit{biomedical research} – from the initial scientific workflows building for composing domain- and application-specific algorithms, to the deployment of advanced mechanisms for search and discovery of distributed resources, based on multimodal workflows and the coupling of metadata fusion and social tagging with the more traditional index-based search techniques.
Furthermore, these open-source software components provide for the development of commercially valuable Web Services-based workflow decision support systems and search mechanisms – as customised commercial implementations – to exploit the extensively used service-oriented architecture (SOA) infrastructure undertaken within the Triana Workflow Environment.

2. Triana Software for Scientific Workflows

Triana Workflow Environment, initially funded through the GEO 600 project (http://www.geo600.uni-hannover.de/), was created as an intuitive tool for enabling quick-look data analysis of gravitational wave data, an area where it is still used. Under the GridOneD project (www.gridoned.org), Triana was extended into the distributed computing domain by creating graphically intuitive interfaces and corresponding mechanisms for distributing its components across peer-to-peer (P2P) and Grid computing environments, and it is now considered one of the pre-eminent workflow tools in scientific distributed computing worldwide [1], [2].

The existing Triana Web Services framework has been used in a number of other international research projects such as GridLab (http://www.gridlab.org/), DataMiningGrid (http://www.datamininggrid.org/), BiodiversityWorld (http://www.bdworld.org/), GEMSS (http://www.ccrl-nece.de/gemss/index.html), ReSC (http://www.resc.rdg.ac.uk/projects.php), for Web Services composition, and can switch between Grid and P2P environments. WSPeer [3] has been Triana's Web Services toolkit for the past three years and many projects have used this combination to specify their distributed course-grained service workflows. More than 500 components have been developed based on Triana workflows, for example, Triana has been in this capacity for: radio astronomy, astrophysical simulations, galaxy visualization, data mining, biodiversity problems, grid-enabled medical simulations, environmental science. P2PS [4] has been used as the underlying P2P environment during this time and it has been successfully used in many domains, such as gravitational wave analysis, audio processing and distributed music information retrieval (MIR), distributed P2P simulations and e-health.

The Triana environment is used for problem solving and orchestrating flows of operations or services, both in fine-grained dataflow applications and as a course-grained distributed workflow system [1]. Workflow is implicit in scientific algorithms that specify a series of operations to be executed, to connecting such algorithms in a series of derivations, which when aggregated perform some higher-level task. Workflows can be simple and contain a few components, or complex, where logic-based support can be integrated to make intelligent decisions about the dynamic evolution of the particular workflow. Workflows involve tasks with inter-dependencies expressed and handled by a computation flow, the chain of elementary tasks being not necessarily linear and requesting a graph of interconnected tasks. The workflow management is data driven, as the scientific experiments need to process large datasets and the scheduler responsible for distributing the computational load should take into account the input dataset as well as the workflow graph topology.

Triana allows to spontaneously create and run data analysis algorithms on the data at its source. Such a component-based object-oriented approach allows scientists to easily create new algorithms that conform to an agreed and defined set of data types and can adapt to different internal parameters, and makes it easy to create individual user interfaces for each component to allow the modification of its internal parameters [1]. Such features coupled with the graphical approach encourages reusability of existing units and allows simple type-safe orchestration of data analysis pipelines on-the-fly without the need for code-level reconfiguration. Triana is much more than a workflow or dataflow system, it enables project scientists to design and create systems than connect together a number of software components, whether they are local,
legacy applications, distributed P2P, Web or WS-RF services, or whether they wish to integrate their project with Grid computing capabilities, such as GRAM, GridFTP, etc.

3. The Alchemist

The Alchemist [5] is a generic peer-to-peer architecture that provides a peer-to-peer (P2P) overlay coupled with a data fusion workflow environment (Triana [1]) for the searching and discovery of distributed resources. The Alchemist infrastructure considers an alternative approach to the search engines widely used today, the new paradigm allowing users to proactively push information into the “search database”, respectively onto a set of dynamically selected super-peers that collectively form the distributed search database and use idle cycles on participants’ computers to process information and perform search queries.

The Alchemist is designed to be generic and can host a number of different peer-to-peer applications or other frameworks. The Alchemist builds a search- and index-specific infrastructure that can be used, based upon application-specific algorithms, to catalogue information in a distributed network. The algorithms are composed as workflows, allowing developers to extend current search schemes to create finer-grained matching, filtering of results, and weigh parameter values. The Alchemist is initially applied in audio/video data search by using the underlying toolkit to provide multimodal workflows for the manipulation of different kinds of audio and video metadata; the data-fusion capability allows high-level search mechanisms to be built on top of the framework, by coupling index-based searches with social tagging techniques and metadata fusion, thus no longer confining to the lowest common denominator of current ranking and indexing systems.

The Alchemist - as a generic infrastructure - is a project at Cardiff University that is creating a framework to provide a P2P layer for supporting pluggable network discovery and caching overlays coupled with the ability to execute distributed workflows [5]. The framework is being built on an existing middleware system, called WSPeer [3], which provides a SOAP messaging layer (using Web or WS-RF Services [6]) within a P2P network that supports a super-peer topology of rendezvous or advert caching peers to support the scalability of the discovery and access to information on the network as a whole (it scales with growing demand and has no central point of failure or control). This system already interfaces with existing Grid middleware (e.g. Globus) through Web Services interfaces but is also able to provide access to such capabilities within a decentralised environment. For example, rather than relying on centralised discovery mechanisms e.g. UDDI or similar, we can search super peers for WSDL files to provide access to the distributed services. In this way, the network can cope with far more transient participants, it can support exchanging roles and it is capable of scaling proportionately with the number of peers.

The Alchemist infrastructure has been initially used in the existing DART project (Digital Audio Retrieval using Triana), in which Triana is the workflow manager that a scientist can use to analyse audio (MP3 files). Once the analysis has been designed, the workflow (an XML file and associated Java code) can be published to the network and uploaded to the package repositories as illustrated. Each participant in the network then periodically queries the network in order to see if new workflows are available. When discovered, users download the workflows and Jar files from the distributed overlay of package repositories and begin the new analysis. In DART, there can be potentially thousands of MP3 files on each user's machine, which are analysed by using collaborative filtering and audio content analysis workflows. Once analysed, metadata is returned and stored on a layer of super peers. DART users can then query the super peers for recommendations – note therefore that the super peers here are more sophisticated than normal super peers as they not only cache information but then are capable of comparing it too, and the Alchemist framework makes it easy for these types of roles to be defined.
4. The Path Towards Biomedical Research

In the Biomedical research field, the Alchemist infrastructure provides an innovative mechanism for biomedical image data search and spectral data discovery, where metadata generated from user and automatic annotations is fused and semantics can be included as part of the query [7]. Biomedical audio-visual content and associated metadata can be discovered and retrieved, allowing also the visualization of particular regions of interest within the images, or anomalies in the patterns of spectral data. While content-based retrieval is an active field of research, multiple modality search (using data from multiple sources and media, and data from different levels of biological organisation) can give deeper insights into the nature of biological entities and the processes they are involved in. Triana - as a scientific workflow environment - has been used in Bioinformatics projects as a Grid-based problem-solving environment for collaborative exploration and analysis of global biodiversity patterns.

Within the context of Biomedical sciences, the Alchemist has a significant potential to support (i) distributed biomedical communities focused on a specific disease process (ii) disease-oriented collaborative studies which share large datasets (iii) integrative biology projects that need to analyse inter-related information at different levels (e.g. clinical, cellular, molecular and genomic) (iv) population-based studies (e.g. clinical trials in diabetes or cancer). For biomedical images, current practice usually involves searching databases containing (i) patient data repositories (ii) case-oriented reference atlases (i.e. dynamic information across spatial and temporal scales of abstraction) or (iii) training collections (documented biomedical images, either anonymised individual or averaged data and training datasets). By using the Alchemist for disease-oriented studies, P2P caching could support the search, selection and aggregation of similar biomedical and spectral datasets. Caching, using rendezvous peers (as in WSPeer [2]), can be adapted to cache similar requests for resources and thus store hits on nodes which are within closer proximity to the particular group of researchers interested in the file.

The initial use-case in the field of biomedical research is the utilization of the Alchemist infrastructure in Diabetic Retinopathy research and clinical trials, for the early detection and prevention of retinal disease and investigational drug discovery [7]. We use the Alchemist to create multimodal diabetic retinopathy workflows, in order to combine data from several distributed resources (horizontal integration) and different biomedical levels (vertical integration) for advanced retinopathy detection and prevention.

Mobile support is also provided for the Alchemist infrastructure. Clinicians can benefit greatly from the possibility of using a mobile device to initiate search operations, retrieve medical audio/visual data and associated records, or operate data transfer between different static repositories, in a controlled and secure way. We also address the possibility of using mobile devices for workflow management by the remote control of the enactment engine from a mobile user interface, for the data fusion of different kinds of data and metadata and the integration of algorithms designed under a common problem-solving environment [8].

5. Acknowledgements

We would like to thank our sponsors, PPARC, EU and EPSRC for supporting these R&D efforts through the GridOneD, GridLab and EP/C006291/1 projects, respectively. We would also like to thank the researchers and programmers (Eddie Al-Shakarchi, Andrew Harrison, Ian Kelley, Matthew Shields and Ian Wang) at Cardiff University for their contribution in realising these software components.
References