Talkoot: Drupal Extensions to Create Online Collaborative Portals for Earth Science

Rahul Ramachandran1, Sunil Movva1, Manil Maskey2, Ajinkya Kulkarni2, Helen Conover2
University of Alabama in Huntsville
rramachandran@itsc.uah.edu

Christopher Lynnes
NASA/GSFC

Brian Wilson
NASA/JPL

Abstract

On the emerging "Social Web," millions of people offer their knowledge online in a collective knowledge system comprising an active community of motivated members posting problems and solutions in blogs, forums, mailing lists, collaborative portals and other Web 2.0 technologies. These technologies complement formal means of sharing knowledge via conferences and published papers, where it is impossible to share all the research details, and where negative results are rarely included. A small but growing number of scientists and researchers are beginning to harness these Web 2.0 technologies as a transformative way of doing science. With the advent of Service Oriented Architectures, the model of chaining services to create analysis workflows provides the research community unprecedented opportunity to collaborate, sharing their workflows with one another, reproducing and analyzing research results, and leveraging colleagues' expertise to expedite the process of scientific knowledge discovery. A crucial component needed for this unprecedented level of cooperation within the research community is a reusable, extensible and customizable environment for building collaborative "open science" portals for managing these shared analysis workflows. This paper describes the design and the development of Talkoot, a customizable "software appliance" to build collaborative portals for Earth Science services and analysis workflows.

1. Introduction

On the emerging "Social Web," millions of people offer their knowledge online in a collective knowledge system comprising an active community of motivated members posting problems and solutions in blogs, forums, mailing lists, collaborative portals and other Web 2.0 technologies [ref]. A small but growing number of scientists and researchers are beginning to harness these Web 2.0 technologies as a transformative way of doing science. Since communication is at the heart of science, these technologies provide researchers easy mechanisms to critique, suggest, and share ideas, data and algorithms. These technologies complement formal means of sharing knowledge via

conferences and published papers, where it is impossible to share all the research details, and where negative results are rarely included [ref].

At the same time, science software developers have embraced Service Oriented Architectures (SOA) and Systems of Systems such as GEOSS [Ref]. Data processing, analysis, mining and visualization algorithms are being converted into publicly available web services, allowing researchers access to large suites of algorithms for data processing and science analysis. This model of chaining services to create analysis workflows provides the research community unprecedented opportunity to collaborate, sharing their workflows with one another, reproducing and analyzing research results, and leveraging colleagues' expertise to expedite the process of scientific knowledge discovery. In many cases, the output of one workflow can be an input to others, leading to chained workflows with components shared by two or more researchers. A crucial component needed for this unprecedented level of cooperation within the research community is a reusable, extensible and customizable environment for building collaborative "open science" portals for managing these shared analysis workflows. Current collaborative portals (e.g., MyExperiment.org) have been one-time development efforts for specific science domains that cannot be easily extended beyond their initial features or reused by other science domains.

This paper describes Talkoot (finish for barn raising), a software toolkit to build collaborative portals for Earth Science services and analysis workflows. Talkoot provides a solution to easily build customizable collaborative portals that will greatly facilitate community collaboration in science analysis. The key feature of Talkoot is that researchers (not just information technologists) will be able to build collaborative sites around service workflows within a few hours. We envision online communities coming together, much like Finnish "talkoot," to build a shared collaborative research space. Talkoot software appliance consists of a set of modules that can be mixed and matched based on the end user requirement. In addition, Talkoot is applicable to many different science domains, mission teams, research projects and organizations.

2. Related Work

1.1 *HubZero*

HubZero [Ref] is a collaboration platform for scientific collaboration developed at Purdue University. It was designed to support to create an online community for the Network for Computation Nanotechnology. The objective of the platform is to connect theorists who develop simulation tools with the users of the tool. HubZero business model has evolved and is now provided as a "Platform as a Service" with growing number of Hubs supporting different domain. The simulation tools are exposed to the end user via the browser within a container using a virtual network computing client. This approach controls both access and quota limitation on the backend computational resources. Users are allowed to upload their own code but the code under goes a whetting process that involves testing before it is made available on the hub. The collaboration elements within HubZero include the ability for multiple users to view the same session and a community forum to share and discuss ideas.

1.2 *MyExperiment.org*

myExperiment is a social website for discovering, sharing and curating Scientific Workflows with the intent to improve scholarly knowledge cycle and to reduce time to discovery [DeRoure]. The goal of myExperiment.org is allow researchers to share not just the digital materials but also the methods (workflows) associated with the research. It serves as a public repository for collecting scientific workflows from different workflow systems. It build the conceptual notion of an e-Laboratory consisting of digital research objects.

2 Open Source Framework

Content Management Systems (CMS) are currently being used to effectively manage complex publications. CMSs already offer a wide variety of basic features that are required by any collaboration site, including management of content and the content creation process, management of users and their roles, ability to import and export content, content syndication, personalized content based on user profiles, extensibility, version management, archiving and publishing capabilities. There are many open source CMS available [3] with Drupal being one of the top three. Drupal already accounts for more than 293,000 installations on the web [http://drupal.org/project/usage/drupal]. Drupal requires a minimal software stack consisting of PHP, Apache Server and MySQL database. Most web hosting services provide this stack as a minimum, some even provide Drupal scripts that allows automated installations. Drupal is a robust and can be customized and extended with well documented API's. Like most well conceived Web content management systems, Drupal separates content from presentation through the templates (PHP, xHTML and CSS).

Drupal allows the site administrator to setup a site by installing the core Drupal application and then customize the site by adding functionality via third party modules. Once a Drupal site is set up, non-technical users can easily add content and maintain these sites. The basic data model holding content in Drupal is called a node. These corresponds to information presented on a page and can be created, edited and deleted by authors with the right permission. Modules such as Content Creation Kit (CCK) can be used to extend the Node data model. Using CCK, site administrators can add additional fields, file uploads, references to other nodes or pages etc.

Drupal-based solution offers many key advantages. Content authoring can be done with flexible access controls, a useful feature to have for a collaborative site. It provides all the Web 2.0 features such as forums, blogs, newsletters, wikis, quizzes, polls, sweepstakes, and other social networking capabilities. It provides built in caching for improved performance. It provides a role based permission system allowing site administrators to create many different roles that can be assigned to different users. Drupal's built in user registration and profile capability along with browser based administration make it easy to use and manage. Finally, it has over 1,500 modules that can be used to customize a portal without having to write a single line of code.

3 Talkoot Extensions

Well designed CMFs such as Drupal are shipped with basic core functionality, and this functionality can be extended without impacting management of the core code base by the installation of additional modules [8]. The core typically includes a library of common functions and modules that provide basic functionality like user management, access control, taxonomy, templating, session management etc. An example module stack is presented in the Fig 1. Talkoot extends Drupal's functionality with a series of modules specific to Earth Science for registering, creating, managing, discovering, tagging and sharing Earth Science web services and workflows for science data processing, analysis and visualization. Some of the key Talkoot module are described below:

3.1 *Research Notebook*

Talkoot defines a special content type to support sharable "science notebooks. Any knowledge creation activity by the scientist is automatically captured and stored in an online Research Notebook. Text, images, links, embedded workflows, data used, chat logs, event information, and other user content can be stored within a research notebook. Any content within the research notebook can be searched based on time, type and tag assigned during creation.

<< Sunil – what about the fields? What information does each field hold? Can it be extended by CCK? More details need>>

3.2 *Workflow Suite*

The workflow suite consists of a set of module, with each module providing specific functionality. These module are

3.2.1 Workflow

Workflow is core module within the Workflow suite. It defines a specialized content type and allows users to create, edit and share Science workflows as Drupal nodes. Workflow module provides a both a teaser view for any science workflow node and web form interface to execute a workflow. The teaser view can be used in conjunction with the View's module to create succinct summary displays for all the workflows cataloged at a site. The module also executes the workflow on a defined BPEL engine by invoking the SOAP WS interface.

<< Sunil – what about the fields? What information does each field hold? Can it be extended by CCK? More details need>>

3.2.2 Workflow Queue

The Workflow Queue module implements a Job queue to track all the executing workflows. In addition, it provides an API to other modules to add workflow executions

to the queue. This module is responsible for polling the BPEL engine at regular intervals to check the status of the executing jobs and updating the status within the database.

<< SUNIL - how is the job exectuted?? FIFO? What happen the job fails to execute? How is the user notified when the job finishes?? Is this dependent on ODE or is this module generic enough??>>

3.2.3 Workflow Clone Module:

The Clone module in Drupal is used to allow users make copies of existing nodes and also be able to edit the copy. In addition, the ownership is set to the copier and both the menu and URL aliases reset. The Clone module inserts "Clone of" into the title to remind the user that they are not editing the original content rather the copied version. Workflow clone module implements the Clone module hooks. It allows users to import any published and shared workflow into their research notebook. Once the workflow is within their notebook, the users have full privileges to edit input parameters for the workflow, modify the steps within the workflow and execute the workflow.

3.2.4 Workflow Composer

There are multiple technologies for describing workflows (BPEL, SensorML, SciFlo), tools for composing workflows (Mining Workflow Composer, SciFlo's VizFlow, Flex based clients) and different engines (SciFlo engine, BPELPower, Active BPEL, Apache ODE) to orchestrate the invocation of services within a workflow.

<<SUNIL – need details here. Such as implemented in FLEX (why), supports Workflows in BPEL? ODE??. Design based on XBaya with modifications?? Key features – drag and drop, preconfigured to existing WSDLs and BPEL engine endpoint?? Auto completion? How to do store parameters? What gets saved in xwf? What gets saved in the database?? >>

3.3 *Experiment*

Experiment module is a node type module that facilitates the creation of experiments within a Research Notebook. Hence, a user can create many experiments within their research notebook and each experiment storing all relevant information for that specific experiment such as data used, workflows, logs etc. Workflows and different runs of the workflows can be stored within an experiment along with all the service parameters, input data files and the output generated by the workflows. Experiments can be shared with other users to foster collaboration. This allows complete knowledge sharing rather than just bits and pieces.

3.4 *Data Content*

This module keeps track of all the files in the users account on the computation cluster. It captures all the information about these files as Drupal nodes. Hence, the Data Content module effectively stores all the metadata associated with these files with the node. The users can provide additional tags to describe these files. This stored information allows users to easily search and find pointers to the actual files on the computational resource.

<< SUNIL – again details? Required fields?? What else?? How is it populated?? Does the user have to create content? Or does is it created from the workflow details??>>

3.5 *Data Search*

This Talkoot module provides users data discovery and access within the portal. The module can query different data catalogs to find online files of interest. The listing of these files and their online locations is returned to the user. It allows user to import discovered data into their research notebook and also simultaneously moves these online data files to the computational cluster utilizing the sFTP module. The module has hooks that can be implemented by sub-modules to provide plug and play search extensions for different catalogs. The current extensions have been implemented for Mirador Search Service, Global Hydrology Resource Center (GHRC) data catalog and the MODAPS. Mirador is a Earth Science data search tool developed by the GES DISC providing search interface access to all the data holding within the GSFC DAAC. GHRC in another NASA DAAC holding data from different microwave instruments such as SSM/I and AMSU. Data search extension to search MODIS dataset is done via MODAPS service API and allows users to easily search for online MODIS data from any Talkoot instance. A data search query formulated via the Talkoot interface simultaneously searches these three archives and the results are then aggregated before being presented to the user.

3.6 *Drupal sFTP*

The sFTP module is used to import data from different distributed locations to the back computational resource interfacing with the Talkoot instance. It works in conjuciton with the Data content module and automatically tracks all the files being imported. It allows data import via HTTP, FTP or from a local machine onto the computational resource utilizing the sFTP protocol. Since the typical Earth science data transfers are large, the module implements a cron hook to provide asynchronous transfer capability. Thereby, the users don't have to wait on their web page until the transfer has been completed. It also prevents the large transfers from timing out.

3.7 *Service Registry*

This module will allow end users to register new SOAP or REST-based web services into their collaboration portal, parsing information from the individual WSDL service descriptions during the registry process. Well-known service protocols (OGC WMS, WFS and WCS, OpenDAP, opensearch) will be labeled to make them machine-callable by workflows. The module will allow users to add other metadata to these services

including semantic tags for efficient discovery and use of these services by others. The services registry will be open search-able by keywords and semantic facets.

<<SUNIL – need words here!!>

4 Talkoot Dependencies and Issues

<< SUNIL – Apache ODE, Asynchronous issues?? Other points of failure???>

5 Talkoot Instance: Data Mining Solutions Center

5.1 *Objectives*

5.2 *Creation and Sharing of Workflows*

5.3 *Creation of Science Stories backed by Data and Workflows*

High level architecture of the demonstration portal built using Talkoot module is depicted in Figure 2. This test portal is using data mining and specialized image processing services deployed on a backend compute infrastructure of a seventy node cluster with an network attached storage. The Drupal core requires a minimal software stack that includes an Apache webserver, MySql database and PHP. This stack is available online as one click install in LAMP or WAMP bundles. Talkoot modules extend the Drupal by providing Earth science specific functionality and are installed by dropping them in the sites/modules folder. They can then be configured via the browser by the site adminstrator. User access to both put data and execute the workflows on the cluster is managed through Drupal.

The prototype portal uses the Research Notebook Module. This module collates all the work done by scientists. This includes not only their workflows but also the other posts made on the portal such as ones in forums, comments, chats, etc. A snapshot of a scientist's Research Notebook is shown in Figure 3.

The workflow module allows scientists to create science workflows on the portal. This module allows users to describe their workflow using simple forms such as title and short description. The users can tag their workflow using taxonomies or free tagging. An workflow submission form example is presented in Figure 4.

The workflow module also allows the scientists to launch a Flex based workflow composer using the Compose Workflow button (as seen in Fig. 4). The workflow composer is developed in Flex and works completely within the web browser (Fig 5). The workflow composer automatically loads services registries and to show the list of available services to the scientists. Scientists can compose a workflow by dragging individual service on to the workflow canvas and connecting multiple services together. Once the scientists are satisfied with the workflow, they have the option of either saving the workflow on their local machine or on the portal itself. Saving the workflow on the portal automatically deploys the workflow to the workflow engine. Scientists can specify the parameters needed for individual services within the workflow such as location of the

input data and then execute the workflow on the cluster. Completion notifications are sent automatically to the scientists when the workflows have finished execution. These notifications are recorded in their research notebook.

The key objective of Talkoot is to allow collaboration among scientists by providing the ability to share and reuse workflows. Therefore instead of building a new workflow from the beginning, a scientist can search shared published workflows on the portal based on tags, annotations and keywords. Once the scientist finds a suitable workflow, he or she can clone the workflow into their individual research notebook. The clone button (as shown in Figure 6) is displayed for all the workflows that have been published. Once the workflow has been cloned, the scientist can view this workflow within the composer, modify the workflow to suit his or her research requirements, save and publish the workflow.

The prototype portal also uses a Browse FTP module specially developed for Talkoot. This module allows scientists to create restricted directories on the cluster and upload data files that can be used by their workflows.

6 Lessons Learned

7 Summary

The Internet has enabled a new phase in collaboration among science researchers, such as the use of wikis for community research. However, wikis are limited to text, static images and hyperlinks, providing little support for collaborative data analysis. We are developing Talkoot as a novel open science environment that integrates many Web 2.0 technologies into a turnkey software appliance for constructing collaborative portals, representing a major improvement in science collaboration tools. More importantly, with the ability to share and execute analysis workflows, Talkoot portals can be used to do collaborative science in addition to communicate ideas and results, foster the development of collaborative online communities to manage science workflows, and create a growing world of directly reproducible science to support exploratory science, informal group review, and even formal peer review. This paper describe our vision for Talkoot and the status of current prototype. Details about the project including a movie demonstrating the current prototype can be found at [http://miningsolutions.itsc.uah.edu/mws/]. Talkoot prototype portal can be accessed here [http://miningsolutionsdev.itsc.uah.edu/talkoot/].

Talkoot supports the importance of social collaboration as stated by DeRoure.

1. By pooling and sharing workflows we can truly accelerate science and reduce time and other redundancy.
2. By combining data, workflows with the final results as show in the Science Story example, scientific discovery is accelerated by promoting transparency, reproducible research.

8 Acknowledgements

This work has been funded by NASA ACCESS Grant # NNX08AT90A. We would also like to acknowledge *Dr. Frank Lindsay, Steve Berrick and Martha Maiden* for supporting and guiding this effort.

9 References