International Environmental Modelling and Software Society (iEMSs)
2010 International Congress on Environmental Modelling and Software
Modelling for Environment's Sake, Fifth Biennial Meeting, Ottawa, Canada
David A. Swayne, Wanhong Yang, A. A. Voinov, A. Rizzoli, T. Filatova (Eds.)
http://www.iemss.org/iemss2010/index.php?n=Main.Proceedings

Hydrologists workbench: A governance model for scientific workflow environments

Box, Paul
CSIRO Land and Water, Sydney, Australia
paul.j.box@csiro.au

Abstract: Scientific workflows (SWF) are an emerging approach that enables scientists to compose and execute complex, distributed scientific processes. The approach is premised on the ability to compose, publish, share and reuse workflows across distributed communities of collaborating scientists. Scientific workflow software (SWFS) provides a technical framework to compose, publish, and reuse SWFs together with data, functionality and computational and other resources upon which they rely. Tools and components are built using service oriented architecture approaches with data and functionality exposed through services. Together, interoperable components from different initiatives form information infrastructure (which as socio-technical endeavours, present specific governance challenges. As workflows, and the resources upon which they rely are distributed and under the ownership of different people and organisations, an enabling governance framework is required. Governance comprises authority structures, roles, policies, processes, and mechanism that enable collective decision-making, and collaborative action to achieve common goals. This paper presents key challenges related to the governance of socio-technical aspects of scientific workflows and workbenches. These include the 'human activity systems' that enable the design, creation and sharing of workflows and the technical governance of workflows, and underlying resources throughout their lifecycles. The paper also presents a conceptual model for scientific workflow governance and discusses its application in the Hydrologists' Workbench (HWB) project.

Keywords: *scientific workflow software; governance; information infrastructure*

1. INTRODUCTION

Scientific workflows (SWF) are an emerging approach that enables scientists to set up and run *in silico* experiments through the composition and execution of a chain of scientific processes without the need for programming. Scientific workflow software (SWFS) provide a technical framework to create, publish, reuse and manage workflows, together with data, functionality, models, and computational and other resources upon which they rely.

Although the promise of scientific workflows is clearly recognised, significant challenges associated with the creation, sharing and execution of scientific workflows remain (Gil et al. [2007]). Deelman and Chervenak [2008] describe challenges associated with the workflow cycle and Goderis el al. [2005] identify a number of bottlenecks to the reuse of workflows, including the discovery and reuse of workflow fragments.
Given the collaborative, distributed, service-oriented nature of scientific workflow environments and the fact that workflows and the resources upon which they rely are often under custodianship and operation of different people and organisations, governance of scientific workflows is a critical issue. Shon et al. [2008], identify a number of dimensions of scientific workflows that require governance include reusability, workflow reproducibility and platform extensibility.

This paper describes key governance challenges associated with the application of SWFS in the hydrology domain, together with a conceptual framework used to inform governance solutions. The paper uses as a case study, the Hydrologists Workbench project, a scientific

workflow environment, for hydrological analysis and reporting. The paper describes and characterises some key SWF governance challenges associated with workflow design, and composition and the extension of the SWFS encountered though the HWB project. Finally, the paper presents a conceptual model to address governance in collaborative environments and briefly illustrates its application.

2. THE HYDROLOGISTS WORKBENCH

The Hydrologists Workbench (HWB) is a collaborative project undertaken by CSIRO and the Australian Bureau of Meteorology (the Bureau) (Cuddy and Fitch [2010]). The project aims to develop an integrated hydrological modelling desktop application built around scientific workflows. The HWB is intended to support scientific and reporting activities to meet the Bureau's expanded role under the Water Act 2007 encompassing the analysis, management and reporting of Australia's water resources information (DEWHA [2007]).

2.1. Trident

Following a review of candidate technology platforms, reported by Perraud et al. [2009] Trident was selected as SWFS platform to be used for the HWB. Trident is a suite of applications developed for the composition, management and execution of scientific workflows (Microsoft Corporation [2009]). Trident provides much of the core functionality for the HWB. However, HWB has a much broader scope than Trident and can be conceived as an integrated environment for hydrological scientific workflow design, composition, execution, publication and management. Thus the HWB also comprises the human activity system and associated mechanisms and tools that enable the collaboration necessary to create, share, execute and manage workflows, their component parts and the resources upon which they depend.

Trident is based on Windows Workflow Foundation, and is part of the .NET framework. Trident includes two applications: Composer for composition of workflows, and Management Studio for the management of workflows. Composer provides a visual design tool for composition of workflows by dragging configuring and connecting 'activities', the atomic building blocks for workflow composition. Both applications leverage a registry which handles the registration, management, discovery and access to workflows, activities, data products and other artefacts that comprise or support workflows.

Upon installation, Trident has a limited set of pre-programmed activities that provide data access, transformation and flow control such as the 'if else' activity. Communities extend the basic functionality by developing custom activities to perform processing required for scientific workflows in specific domains. Much of the initial effort of the HWB project has focused on developing core functionality commonly required in the hydrology domain.

2.2. Integration with Geo-processing Workbenches

A key objective of the HWB is to interoperate with other workbenches and environments in use by target communities, to enable users to use the most appropriate tools for a specific task and orchestrate their execution through the HWB. As hydrological modelling, analysis and reporting has an intrinsic spatial dimension, the integration of HWB with geo-processing frameworks was a key research priority. The initial geo-processing framework targeted for integration was the ESRI ArcGIS desktop environment, a widely used geo-processing framework within the Bureau and CSIRO.

Integration with ArcGIS is based on Trident interacting with the geo-processing tools exposed through the ArcGIS geo-processing framework. To extend the default ArcGIS tool set and to create geo-processing workflow fragments custom Python scripts are written. These are exposed as tools in the ArcGIS geo-processing framework. A dedicated .Net activity is created to launch each custom (user defined) and default ArcGIS tool provided as part of the geo-processing framework. Default tool activities are generated by parsing xml

files that describe each tool, supplied as part of ArcGIS product. Activities for custom tools are developed using a hand crafted xml file that describes the tool and its parameters. Using this approach, the ArcGIS geo-processing framework provides registration management and access to the geo-processing tools.

2.3 Water Reporting Workflow Design and Composition

An initial focus area for HWB was the development of workflows to generate information products for monthly water situation reports which are under development by the Bureau. Although the workflows developed can be characterised as production rather than scientific workflows, the approaches to the design of workflows and required functionality are considered to be broadly applicable to the scientific workflow context.

The approach used for the design of workflows was loosely based on the work of Gil [2007] who identifies four stages of workflow design with attendant levels of workflow abstraction:

- Workflow sketches - used for initial workflow requirements specification
- Workflow templates - execution-independent specification of the processing steps, components to be used and the data flow between them
- Workflow instances - execution-independent with specification of input data
- Executable workflows – workflows instances assigned to resources for execution

This distinction provides a useful conceptual framework for approaching design and addressing reuse of workflows. In the context of HWB four levels of abstraction were used as the basis for design process. Firstly, Workflow sketches were compiled as UML activity diagrams. These were used to document existing functional components, identify functionality to the developed and for the refactoring of components to enable reuse across multiple workflows. At the next level of abstraction, workflow templates were created in Trident as generic reusable workflows that were intended for reuse. Workflow instances based on templates were created with parameter values including data sources. Finally, rather than executable workflows the project adopted the term workflow instance runs i.e. an execution of the workflow instance that produces concrete data outputs.

There were typically a number of iterations through each step of this design process to refactor components and workflows based on an improved understanding of required granularity and reusability. Perraud et al. [2010] provide an analysis of the appropriate level of granularity as part of evolving workflow design. At each step of the process a number of artefacts are created that express or are implementations of agreements about how a component should behave. These included, workflow sketches as UML activity diagrams, specification of .NET Trident activities and ArcGIS geoprocessing tools to be developed, development, and production versions of activities, workflow fragments and complete workflows and other artefacts such as configuration files, scripts, local data used as input to workflow activities and sample outputs which comprised part of the specification set.

Working in a collaborative environment across several agencies to extend the core functionality of Trident and to develop custom geo-processing tools in ArcGIS, required an overarching governance framework to ensure that agreements and their implementations were properly managed.

3. GOVERNANCE

To collaborate across organisational boundaries and build effective communities that are able to share and re-use information, processes, and knowledge, socio-technical information infrastructures are required. Aanestad et al. [2007] note that these infrastructures are intrinsically socio-technical endeavours and as such, require governance. Governance provides an overarching and enabling decision-making and accountability framework comprising authority structures, roles, policies, processes, and mechanisms that enable collective decision-making, and collaborative action to achieve common goals (Box and

Rajabifard [2009]). Governance provides oversight and an enabling framework for management activities and can be conceived as three interacting dimensions:

- the what – the scope of governance defined by the aspects of a communities' endeavour that are under governance
- the who – the key roles and relationships between stakeholders and the collective organisational structures through which governance is exercised
- the how – the mechanisms and processes of the human activity system through which governance operates and technical tools that support governance

3.1. Service Oriented Architecture (SOA) Governance

SWFS, is built around (local and web) services and thus many of the governance challenges and potential solutions associated with service oriented architecture (SOA) are relevant to SWF environment governance. The SOA approach is premised on the development, maintenance, discovery and use of interoperable services. These self-contained functional elements are designed to meet specific purposes and are able to interoperate. The publish, find, bind pattern, shown in Figure 1, provides the mechanism for the publication, discovery and use of services in a SOA.

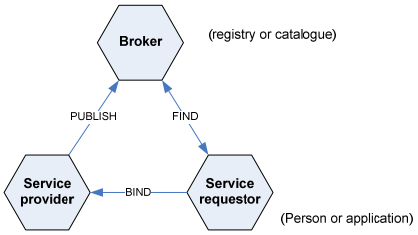

Figure 1. The publish, find, bind pattern

Services, although under the control of different owners, are interdependent, necessitating collaboration between owners, developers, operators, and users of the service across departmental and organisational boundaries (Josuttis [2007]). In addition to the technical tools for service management, a governance framework is also required to ensure consistency, and predictability of interdependent services (Stanek [2006]). SOA governance provides the business context for the design, development and operation of services and addresses related aspects of the service lifecycle; design-time and run-time. Design-time governance relates to the environment in which services and other components are designed, developed, tested and approved for publication. Run-time governance addresses the governance of operational aspects of SOA including service discovery, access monitoring, security and management.

In SOA, registers (or lists) of resources and registries (the systems used to manage them) play a vital role in publication, discovery and use of community resources. A range of registers of such things as code, users and permission, standards, and other resources that the community care about, are essential to the sustained operation and growth of an initiative. These artefacts document community agreements and enable the discovery and use of the resources necessary to develop, maintain, operate and grow the infrastructure. Thus registries and the registers they manage play a critical role in supporting governance.

3.2 HWB Governance Challenges

Efforts to extend the HWB through the development of activities in Trident and tools in ArcGIS and compose workflows to meet the specific requirements of the hydrology domain, have informed an understanding of key challenges that require resolution through governance. These challenges which relate to functionality, data, people and computation resources across both design-time and run-time domains are presented in Table 1. Run-time domain in this context includes the composition and execution of workflows in Trident and other workbenches.

Table 1. Governance design–time and run time issues

Domain	Design-time	Run-time
Functionality	- identification, documentation and prioritisation of workflows to be developed - identification and specification of functionality required for the hydrological domain - development lifecycle management for Trident activities and ArcGIS tools - specification and management of integration with other workbenches - development lifecycle management for other development environments	- discovery, interpretation and reuse of workflows, workflow fragments and activities - sharing and exchange of workflows and activities - discovery, interpretation and use of functionality in other workbenches (e.g. ArcGIS)
Data	- agreement on common data types and exchange formats - identifying provenance tracking and persistence requirements for data sources, intermediate data sets - management of final products - managing interactions with data providers and data sources under external governance arrangements	- Discovery, interpretation, and access to data sources

People	- assignment and managing of decision rights and roles in relation to:
	  - governance processes
	  - extending workbench functionality through the creation and registration of activities and tools
	  - design and composition of workflows
	  - managing interactions with external stakeholders and 'communities' to enable cross-community development and re-use of workflows and components
		- management of permissions related to workflow and component access composition, reuse and execution

Computational resources		- assignment of workflows to computational resources for execution

3.3. Run-time Governance Capabilities

Trident offers some capabilities that support governance. These capabilities are underpinned by a registry used to register, manage and access workflows, activities, data products and other artefacts that comprise or support workflows. The aspects of scientific workflows governance that are addressed through the Trident registry are:

- **functionality -** registration, discovery, management and use of activities that have been developed and the workflows into which they are composed
- **data -** registration of data sources and data provenance tracking
- **people** - management of user permissions to access registries, workflows, and activities
- **computation resources** - registration and management of computation resources for workflow execution

Likewise, ArcGIS through its geo-processing framework provides a number of governance capabilities that, as with Trident, address run-time governance needs. These tools support the registration, discovery management and use of tools. The ArcGIS ArcCatalog application also provides some tools that support data source registration and management.

Although Trident (and to a lesser extent, ArcGIS) support many required aspects of HWB run-time governance, significant aspects of a governance solution are missing. Firstly, the 'human activity system' comprising the overarching institutional and process framework that enable a community to work together is missing. This system sets standards and policies, creates processes, assigns roles and permissions and interacts with external and related governance mechanisms. Secondly, a range of mechanisms and tools required to support governance absent. These include:

- the design-time environment – the governance of the entire development lifecycle for functional components from requirements through publication to retirement
- run-time environment for HWB components not registered in Trident such as ArcGIS tools

Trident registry is only able to support governance of things that are registered in Trident. As many components (workflows, models and blocks on functionality) will be developed and stored outside of Trident there is a need to register and govern these artefacts.

4. A REGISTRY BASED FRAMEWORK FOR GOVERANCE

To address the key governance challenges and provide missing governance capabilities, a governance framework is required. Key requirements for the framework are that it be:

- based on existing information infrastructure governance approaches
- practical, lightweight and commensurate with the resources under governance
- scalable and evolvable
- focus on the governance requirements of technical components that need to be managed for discovery, reuse and sustained interoperability
- consistent with the inbuilt registration capabilities of Trident and ArcGIS geo-processing framework, supplemented with additional registries where necessary

It was determined that the conceptual model of governance encapsulated in the ISO 19135 standard Procedures for Registration of Geographic Items [2004] be used as the basis for HWB governance. The standard articulates the use of registers (lists) and registries (systems that manage lists) together with a defined roles and processes related to register creation and management. In this model, the registers define the scope of governance, the processes and roles describe how governance is exercised and by whom. The UML diagram shown in Figure 2, highlights the key roles and relationships related to register and registry ownership management and use.

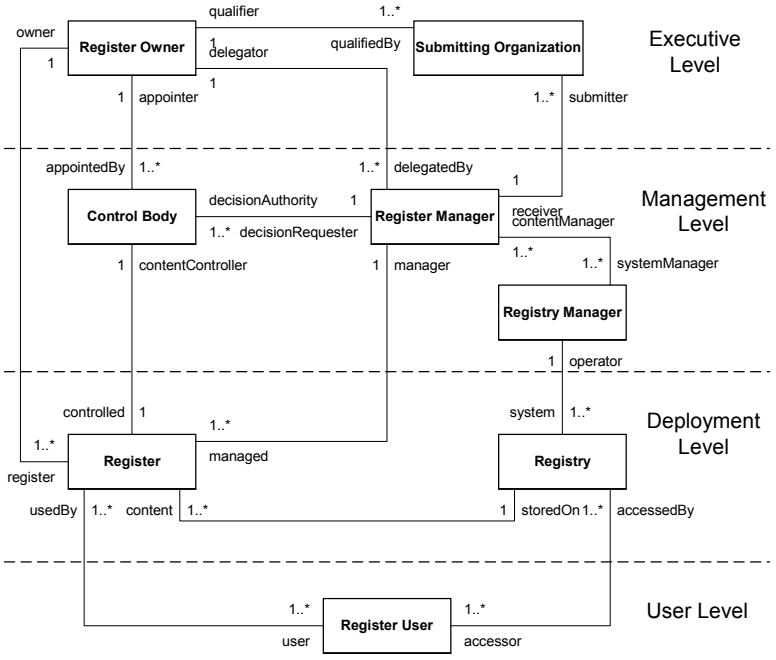

Figure 2. IS0 19135 - Registration Roles

Using this approach, the things that a community cares about and must manage to ensure the achievement of collective goals, can be conceptualised as a number of registers - lists of things such as agreements and resources (people, data, and technology). Users are able to access registers and find information that enables them to create, access or use common components that together constitute the collective system.

At the core of this approach are two processes; the creation and the assignment of roles related to register management and the registration process. The registration process involves a number of roles that together implement governance. A register owner determines who has authority to make submissions (submitting organisation) to the register, to adjudicate submission requests (control body) and to manage the registers (register manger) and the registry systems used to manage them (registry manager).

The governance model implicit in this standard can be applied to collective endeavours that need to register, manage, discover and reuse common information artefacts that are critical to the coherence of collective effort. Using this approach, a governance regime can be developed through the creation and management of registers and the assignment of roles related to their management and use.

4.1 Model Application

This conceptual model was used to develop a framework to address technical aspects of HWB governance. 'Technical governance' deals with 'technical' artefacts that the community cares about and which must be governed. These artefacts either specify an agreement about how some aspect of a component will behave or are a component that is based on or implements an agreement. For each set of such artefacts to be governed, a register is created and registration roles identified in ISO19135 assigned with respect to register management and the registration process.

A number of registers were identified as being part of the HWB governance framework including: a data register, to log data used for development and testing; a code register, to manage code (under version control); an ArcGIS tool register to manage geo-processing tools and Trident activity, workflow and users register. Registry capabilities of a number of tools were used to implement the registers. For example the Trident registry was used to register activities, workflows and will be used to register users and execution nodes. ArcGIS was used for registering geo-processing tools. In addition version control software (subversion) and underlying repositories were used to register and manage code used as the basis for Trident activities and ArcGIS and other tools. The activity register evolved from a an initial excel-based solution to a software development management tool (JIRA) which enabled detailed tracking of the entire development lifecycle of functionality.

5. CONCLUSIONS

The governance model presented in this paper provides a conceptual framework for a governance solution. The way in which the model is applied will to a certain extent be determined by the implementation environment and organisational and IT governance regimes within which the SWFS is used. The governance model is intended to provide an over-arching framework for the human activity system within which the system operates. Wherever possible, the inherent capabilities of SWFS and other interacting workbenches be used to manage the functionality, data and users to create an integrated governance solution. Where necessary the governance framework can be implemented using registry capabilities provided by a variety of software.

Experience in implementing this governance model within a small collaborative team for the HWB project has shown that establishing and effectively operating governance, entails a significant overhead cost. This investment may in some cases be difficult to justify as the benefits of collaboration are not evident until such times as reuse begins in earnest. In order

to realise the longer-term promise of increased efficiencies based upon the creation, sharing and reuse of activities and workflow fragments leading to faster scientific discovery, change in working practices are also required. Collaborating team members who may be used to working alone or in silos to meet their own needs, must move to more collaborative models in which interoperable pieces of processing, functionality and the workflows into which they are composed, are designed, developed and maintained in a manner that enables discovery and reuse. The complex interwoven socio-technical nature of information infrastructures within which SWF are embedded, and the behavioural aspects of communities are critical aspects of scientific workflow environments. An improved understanding of these phenomena will inform approaches to creating and sustaining successful collaborative SWF environments and thus warrant further investigation.

ACKNOWEDGEMENTS

This work is part of Hydrologists' Workbench project in the Water Information Research and Development Alliance between CSIRO's Water for a Healthy Country Flagship and the Bureau of Meteorology.

REFERENCES

Aanestad, M., E. Monteiro, and P. Nielsen, Information infrastructures and public goods: Analytical and practical implications for SDI. *Information Technology for Development* 13, no. 1: 7-25, 2007.

Box, P., and A. Rajabifard, SDI governance: Bridging the gap between people and geospatial resources. In Proceedings of the 11th GSDI Conference, GSDI, Rotterdam, The Netherlands: GSDI, June 15-19 2009.

Cuddy, S., and P. Fitch, Hydrologists workbench: a hydrological domain workflow toolkit, [this conference], 2010.

Deelman, E., and A. Chervenak, Data management challenges of data-intensive scientific workflows. In *Proceedings of the IEEE International Symposium on Cluster Computing and the Grid (CCGRID'08)*, 687–692, 2008.

DEWHA (Department of Environment, Water, Heritage and the Arts), the Water Act 2007, 2007

Gil, Y., E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers, Examining the challenges of scientific workflows. *Computer* 4, no. 12: 24–32, 2007.

Gil, Y., Workflow Composition: Semantic Representations for Flexible Automation. In *Workflows for e-Science*, Springer, 244-257, London, 2007.

Goderis, A., U. Sattler, P. Lord, and C. Goble, Seven Bottlenecks to Workflow Reuse and Repurposing. In *The Semantic Web – ISWC 2005*, Lecture Notes in Computer Science, Springer Berlin, 323-337, Heidelberg, 2005.

ISO (International Organization for Standardization), ISO 19135:2005 Geographic information - Procedures for registration of items of geographic information, International Standards Organisation (ISO), 2004.

Josuttis, N., *SOA in Practice: The Art of Distributed System Design*. O'Reilly Media, 352pp., Sebastopol, California, 2007

Microsoft Corporation, Project Trident: An Introduction, Microsoft Project Trident: A Scientific Workflow Workbench Version 1.0a – https://connect.microsoft.com/trident (Last accessed 2010-03-03). July 9, 2009.

Perraud, J.-M., B. Qifeng, and P. Fitch, Briefing paper - Determination of technological platform for the Hydrologists' Workbench, CSIRO: Water for a Healthy Country National Research Flagship, 2009.

Perraud, J.-M., B. Qifeng, and D. Hehir, On the appropriate granularity of activities in a scientific workflow applied to an optimization problem, [this conference], 2010.

Shon, J., H. Ohkawa, and J. Hammer, Scientific workflows as productivity tools for drug discovery. *Current Opinion in Drug Discovery and Development* 11, no. 3: 381, 2008.

Stanek, R., Governance, the key to SOA success. Loosely coupled - http://www.looselycoupled.com/opinion/2006/stanek-gov0517.html, (Last accessed 2010-05-03) May 17, 2006.