A Prototype Provenance Management System for a Continuous Flow Forecasting System

Corné Kloppers, Chris Peters, Qing Liu, Quan Bai, Peter Taylor, Christian Malewski, Heiko Mueller, Andrew Terhorst, Brad Lee, Stephan Zednik, Patrick West, and Peter Fox

1 Tasmanian ICT Center, Australian Commonwealth Scientific and Research Organization (CSIRO), Hobart, Australia
2 Rensselaer Polytechnic Institute (RPI), Troy, NY USA

1 Introduction

Limited freshwater resources in many parts of Australia have led to a highly regulated system of water allocation. Poor situation awareness can result in over-extraction of water from river systems, compromising river ecosystems. To increase situation awareness, the Tasmanian ICT Centre is developing a continuous flow forecasting system based on the Open Geospatial Consortium (OGC) Sensor Web Enablement (SWE) standards. SWE standards provide an interoperability layer over existing observation systems, numerical models and processing systems. A prototype Hydrological Sensor Web has been established in the South Esk river catchment in northeastern Tasmania. The Sensor Web aggregates sensor assets owned and operated by multiple agencies. Observations from these assets drive a rainfall-runoff model that predicts river flow at key monitoring points in the catchment. The intention is to use this information to manage water restrictions on a more proactive basis.

The generation of hydrological information such as predicted river flows, involves complex interactions between instruments, simulation models, com-
putational facilities and data providers. Correct interpretation of information produced at various stages of the information life-cycle requires detailed knowledge of data creation and transformation processes. Such provenance information allows hydrologists and decision-makers to make sound judgments about the trustworthiness of hydrological information. Managing provenance information is an inherent part of the prototype South Esk Hydrological Sensor Web. Two major challenges had to be overcome to build the provenance management system: modelling provenance and tracking provenance information.

2 Provenance Information Model

Provenance information must be expressed using terminology familiar to water resource managers and hydrologists. The provenance model for a continuous flow forecast system must cover: (1) knowledge of the water domain, (2) information processing and (3) data lineage. Our provenance information model integrates four interlinguas to satisfy these requirements: A hydrological profile of the ISO 19156 encoding standard for observations and measurements known as WaterML 2.0 [1], the W3C Semantic Sensor Network Sensor Ontology [2], the Proof Markup Language (PML) [3], and the Process Ontology [4]. By integrating domain knowledge with provenance and process information, the resulting provenance information model enables water domain researchers and water resource managers to analyse and understand how observations and derived data products were generated.

3 Provenance Tracking

Tracking provenance is a challenge because individual steps in the continuous flow forecast system do not handle provenance information explicitly. Individual rainfall observations from different sensor sites are retrieved using the OGC Sensor Observation Service (SOS). These observations feed into a Kepler scientific workflow that generates a rainfall surface consumed by the rainfall-runoff model. The model outputs are published via another SOS.

To track provenance information in such a distributed heterogeneous system of software components, we built a Provenance Management System (PMS) for warehousing provenance information. The architecture of the
Figure 1: Overview of the continuous flow forecast workflow and the provenance management system
PMS is divided into four major components: Harvesting, Capturing, Storing and Querying. The overall system architecture is shown in Figure 1.

**Harvesting.** The PMS relies on two sources for provenance information: (1) workflow execution log files and (2) sensor information exposed via a SOS interface. Two types of provenance harvesters were implemented: a log file harvester and a SOS harvester. The harvesters parse the data from the respective sources and extract intrinsic provenance artefacts. Each harvester generates data exchange documents in JSON (JavaScript Object Notation) format. The data exchange documents are then posted to the RESTful web service.

The benefits of the chosen harvesting approach are threefold: the existing flow forecast workflow did not have to be modified, the harvesting components have zero runtime impact on the actual workflow execution and the harvesting process can be executed independently.

**Capturing.** Provenance information retrieved by the harvesters is encoded as JSON documents and pushed to a RESTful web service, where the provenance information is translated into PML (RDF graph) and saved to a persistent database. A harvester specific adaptor transforms the JSON document to PML (RDF triples) based on the Provenance Information Model. The resulting RDF graph is called a provenance trace. Each provenance trace is stored in a RDF database via a Jena interface as a separate named graph. Named RDF graphs allow grouping of related RDF triples and are therefore a very convenient separation for each provenance trace.

**Storage.** Provenance information is stored in AllegroGraph, a triplestore database, in the form of an RDF graph. A triplestore is a purpose-built database for the storage and retrieval of RDF data. AllegroGraph includes a rich set of useful enterprise features, e.g., backup and point-in-time recovery, scales to billions of triples while maintaining superior performance and it supports SPARQL queries and RDFS++ reasoning.

**Querying.** Querying provenance information stored in an RDF database is done by developing predefined Simple Protocol and RDF Query Language (SPARQL) queries. It provides a standard query language and data access protocol for use with provenance information encoded in the Resource Description Framework (RDF) format.
4 Conclusion

In this paper, we investigate how to model and track provenance information for a continuous flow forecasting system. The provenance model and its tracking mechanisms are presented. Harvesting, Capturing, Storage and Querying can be seen as distributed and loosely-coupled components. Our architecture provides a scalable, adaptable, and domain agnostic approach for provenance enablement. Whilst our description of PMS is tailored towards a specific flow forecast system most of the approaches are generic enough to be adopted and applied to other workflows or domains.

References


