An ontology-based knowledge management framework for a distributed water information system

Qing Liu, Quan Bai, Corne Kloppers, Peter Fitch, Qifeng Bai, Kerry Taylor, Peter Fox, Stephan Zednik, Li Ding, Andrew Terhorst and Deborah McGuinness

ABSTRACT

With the increasing complexity of hydrologic problems, data collection and data analysis are often carried out in distributed heterogeneous systems. Therefore it is critical for users to determine the origin of data and its trustworthiness. Provenance describes the information life cycle of data products. It has been recognised as one of the most promising methods to improve data transparency. However, due to the complexity of the information life cycle involved, it is a challenge to query the provenance information which may be generated by distributed systems, with different vocabularies and conventions, and may involve knowledge of multiple domains. In this paper, we present a semantic knowledge management framework that tracks and integrates provenance information across distributed heterogeneous systems. It is underpinned by the Integrated Knowledge model that describes the domain knowledge and the provenance information involved in the information life cycle of a particular data product. We evaluate the proposed framework in the context of two real-world water information systems.

Key words | knowledge management, provenance

Qing Liu (corresponding author)
Corne Kloppers
Kerry Taylor
Andrew Terhorst
CSIRO ICT Centre, Tasmania 7000,
Australia
E-mail: *Q.Liu@csiro.au*

Quan Bai
School of Computing and Mathematical Sciences,
Auckland University of Technologies,
Auckland 1142,
New Zealand

Peter Fitch
Qifeng Bai
CSIRO Land and Water Black Mountain, ACT 2601,
Australia

Peter Fox
Stephan Zednik
Deborah McGuinness
Department of Computer Science,
Rensselaer Polytechnic Institute,
Troy, NY 12180,
USA

Li Ding
Qualcomm Inc., San Diego, CA 92121,
USA

INTRODUCTION

Hydrology is the science of water and is concerned with its states, storages and fluxes in location, time and phase. Given the increasing complexity of scientific problems, it has become difficult to allocate the necessary resources to solve problems within one organisation or at one site. Data collection and data analysis are often carried out in distributed heterogeneous systems. Compared with other data-oriented science communities, one of the distinctive aspects of the hydrologic science community (Tarboton *et al.* 2010) is that there is great emphasis on 'third party' data, i.e. data collected by other agencies. A significant challenge for domain users is to identify the right data for their purposes and to decide how and when to use that data. Furthermore, the community has placed too much attention on the networking of distributed sensing and too little on tools to manage, analyse and understand the data (Balazinska *et al.* 2007).

To appropriately interpret a data product generated by a 'third party', the users need to have a good understanding of the origin of data. Provenance refers to the sources of information, such as entities and processes, involved in producing or delivering an artefact (W3C Provenance Incubator Group 2010). For example, the provenance of a hydrological flow forecast may include: what sensors were used, what observations were processed for calibration, and which type and version of the hydrological model was

doi: 10.2166/hydro.2012.152

applied to produce the prediction. The provenance forms the knowledge base underpinning a data product.

Since provenance links artefacts and processes together, it allows the information flow that generates a particular data product to be scrutinised. This provides sufficient evidence to allow assumptions of the underlying decision-making to be understood and evaluated by politicians, senior advisors, domain experts and the public. Provenance can also be used to estimate data quality and reliability based on source data and data transformations. This is extremely important when using large numbers of datasets and analysis methods provided by multiple sources such as various interstate water agencies. Users can assess the truthfulness and robustness of data collection and analysis provided by third parties.

Even though provenance has been identified as one of the most promising methods to improve the transparency of data, there are challenges that must be addressed to make provenance information usable (W3C Provenance Incubator Group 2010). Most provenance systems developed by the computer science community are embedded in workflow systems (e.g. Zhao *et al*. 2003; Altintas *et al*. 2004; Barga & Digiampietri 2006; Scheidegger *et al*. 2008; Simmhan *et al*. 2008). During workflow execution, provenance is collected and stored for subsequent querying. Such provenance systems are tightly integrated with workflow systems, but many large scale applications in hydrology involve disparate sources and sub-systems. This makes it difficult to generate accessible and usable provenance from existing systems. In particular:

- Integrating provenance to represent the information life cycle of a data product creates an interoperability problem since different terms for artefacts and processes may be used by different sources. Provenance arising from disparate sources must be semantically linked.
- To describe the information life cycle of a data product, concepts that reflect the domain knowledge as well as the lineage relationships among artefacts and processes should be captured. Domain semantics are integral to making life cycle information understandable to diverse users and are not normally captured by workflow systems. Furthermore, the domain knowledge should be machine-readable and well defined for acceptance by the communities.
- Most of the existing provenance capture techniques assume a single system, either a workflow system, a data warehouse or an operating system. For large scale applications, provenance needs to be captured across distributed heterogeneous systems, then integratable and queryable.

In this paper, we address the interoperability problem for distributed water information systems by leveraging semantic technologies. The paper presents how to leverage an ontology-based approach to developing the knowledge model and the knowledge management framework for two real-world water information systems. The key contributions of the paper are as follows:

- The design of an Integrated Knowledge model (IKnow) for the water domain by aligning ontologies which represent the provenance information as well as the domain concepts presented in the information life cycle of a data product.
- A knowledge management framework in which knowledge generated by distributed heterogeneous systems could be harvested and integrated, grounded on the IKnow model.
- An evaluation of the knowledge management framework for two real-world water information systems, through the development of provenance queries concerned with the information life cycle of data products.

Recently, the W3C Provenance Working Group is working on defining a language, Provenance Data Model (PROV), for exchanging provenance information among applications. The model is domain-agnostic but is equipped with extensibility points allowing further domain-specific and application-specific extensions to be defined (W3C Provenance Working Group 2012). As it matures, we will investigate how to adopt the Provenance Data Model to develop the Integrated Knowledge model.

The rest of the paper is organised as follows. In the section on 'Related work', we introduce the existing provenance models and systems. Provenance requirements from the hydrology domain are analysed in 'Requirements analysis'. In 'The integrated knowledge model', we introduce the ontology-based knowledge information model that

captures not only the lineage relationships between artefacts and processes but also the domain knowledge generated in the information life cycle of flow forecasts. 'The knowledge management framework' describes the framework to harvest and integrate the knowledge generated by heterogeneous systems. In 'Model and system evaluation', the framework is implemented and evaluated over the two real-world water information systems. We conclude the paper in the final section by summarising the proposed approach and pointing towards future work.

RELATED WORK

In this section, we first review the existing provenance information models developed by the computer science community, then general provenance systems. Finally we review application-specific systems.

Provenance models

The semantic web offers a technology stack for information retrieval and reasoning. It is well recognised as an effective infrastructure to enable information sharing across applications and the web. A provenance model is a representation of the artefacts, processes and their relationships involved in the information life cycle of data. The semantics of provenance is encoded using Resource Description Framework Schema (RDFS) (W3C 2004) or Web Ontology Language (OWL) (W3C 2004). Users can inspect the ontologies to understand the provenance semantics and domain-independent reasoning tools can be employed. There are a number of efforts (McGuinness *et al*. 2007; Sahoo *et al*. 2008; Hartig & Zhao 2010; Moreau *et al*. 2011) which use semantic web-based approaches to model provenance. An extensive review (W3C Provenance Incubator Group 2010) has been conducted by the W3C Provenance Incubator Group.

General provenance systems

Provenance systems are typically designed to capture and manage provenance within the scope of a given computing platform. Most of the existing provenance systems are domain-independent but platform-dependent: (a) database systems (Cui *et al*. 2000; Bhagwat *et al*. 2004); (b) operating systems (Vahdat & Anderson 1998; Muniswamy-Reddy *et al*. 2006); and (c) workflow systems (Zhao *et al*. 2003; Altintas *et al*. 2004; Barga & Digiampietri 2006; Kim *et al*. 2008; Scheidegger *et al*. 2008). For data products generated by heterogeneous systems, a platform-dependent approach is not able to provide a complete information life cycle description of provenance.

The Karma Provenance Framework (Simmhan *et al*. 2008) and the Provenance Aware Service Oriented Architecture (PASOA) (Groth *et al*. 2005) both support provenance capture across different systems.

Several comprehensive surveys (Bose & Frew 2005; Simmhan *et al*. 2005a, b; Buneman & Tan 2007; Tan 2007; Freire *et al*. 2008) have been conducted on data provenance and workflow provenance from different perspectives. In Moreau (2010), 453 provenance papers are reviewed and benefits identified in eScience, curated databases and semantic web.

Domain-specific systems

Some provenance-aware systems have been built for specific domains, such as bioinformatics, healthcare and geoscience.

For the sensor domain, the Earth System Science Server (ES3) (Dozier & Frew 2009) is a software environment for satellite image processing, with operating-system-based provenance management capability. Liu *et al*. (2010b) propose a Provenance-Aware Virtual Sensor System in which the aggregated data from remote sensors is stored using a local repository. Then a set of virtual sensor transformation workflows are executed and the provenance is recorded using the Open Provenance Model (OPM). The computations are organised in a centralised environment. In de Lange (2010), a provenance-aware sensor network for real-time data analysis is developed in which a custom query language is used to ease query specification. The work focuses on the query framework. A provenance-based indexing method to make sensor data searchable was developed in Ledlie *et al*. (2005). Park & Heidemann (2008) explored a process of transforming online sensor data and sharing the filtered, aggregated or improved data with others in Sensornet Republishing. However, most of the systems are either

platform-specific and cannot work in distributed heterogeneous environments, or only focus on the part of the information life cycle of data products.

In the geosciences domain, the use of the components for metadata generation and propagation to augment geospatial data provenance have been explored (Yue *et al.* 2010). Patni *et al.* (2010) developed a provenance ontology to link sensors with observed phenomena. It extends the Provenir upper level ontology (Sahoo & Sheth 2009) to model domain-specific provenance. The Semantic Provenance Capture in Data Ingest Systems (SPCDIS) (Zednik *et al.* 2009) uses the Proof Markup Language (PML) (McGuinness *et al.* 2007) to model provenance of image data processing.

In this work, we present a generic knowledge model and knowledge management framework for the water domain. It captures the complete information life cycle of data products in a distributed heterogeneous environment. In Liu *et al.* (2010a), we first presented the idea for building a provenance-aware Hydrologic Sensor Web. This work develops the idea through to a concrete approach.

In hydrology, Shu *et al.* (2012) present some principles for provenance representation which we also follow here, and develop a multi-level OWL ontology for streamflow forecasting use case. That paper has chosen to extend a different generic provenance ontology: OPM in that paper and PML in our case, demonstrating some significant consequential differences. In both papers, the use of a generic provenance ontology together with multiple ontologies for domain modelling demonstrates extensibility and reusability. The work of Shu *et al.* (2012) shares one of the two use cases we present here, but the domain modelling there is considerably more extensive and detailed, providing a case study for ontology design that leverages reasoning services for provenance at the general, domain and use-case levels. In this paper we have focused on identifying and aligning well-known partial domain ontologies, which should enhance future interoperability and also usability for people and tools familiar with the terms. In this paper, we go beyond the modelling and querying capability of Shu *et al.* to embed it within two different distributed workflow architectures, incorporating provenance harvesting, a provenance ontology and visualisation applied to two working systems.

PROVENANCE REQUIREMENTS ANALYSIS IN THE WATER DOMAIN

In this section, we analyse the general requirements for representing the life cycle of data products relevant to the water domain. Based on this analysis, we present general provenance requirements.

User requirements analysis

From discussions with various domain experts, requirement analyses were performed from three different perspectives: role, knowledge acquired and system involved. In the following, we expand these perspectives.

Role classification

To understand the user groups, the roles they play in providing and consuming provenance and the provenance questions to be asked around the data provided, we classify the water domain users into four groups. The intention is not to list all the provenance questions for each group but to present a representative selection.

Hydrometrist. A hydrometrist is responsible for measuring the hydrological cycle including rainfall, groundwater characteristics, water quality and surface water flow characteristics using gauging stations and instruments. There is a wide range of factors that can influence the quality of the base data managed by hydrometrists. Given a measurement, the exemplar questions to be asked include: (a) what type of gauge gave the reading and its accuracy and frequency; (b) who deployed the gauge; and (c) which agency does the gauge belong to?

Data analyst. A data analyst in this context is responsible for preparing the data for hydrological modelling. Understanding the preprocessing can help a hydrologic modeller to decide if the processed data are appropriate to be used for the model. Questions include: (a) what quality checking method was used for validation; (b) what gridding algorithm and what version were applied; and (c) what observations were used for the gridding algorithm?

Hydrologic modeller. A hydrologic modeller studies the behaviour of hydrologic systems to understand hydrological processes. The knowledge used by hydrologic modellers is important for water managers to make appropriate

decisions. Questions to be asked include: (a) given a prediction, what hydrologic simulation model and what version were applied; (b) when and how was the model calibrated; and (c) if the flow forecast did not reflect the actual reading, how was the flow forecast produced?

Water manager. Since water security is a major challenge for society, the main activities as a water manager are to plan, distribute and manage the use of water resources based on the currency of water information and prediction results. Questions to be asked of water managers include: (a) is the decision appropriate for a particular circumstance and (b) how reliable is the information being used to make the decision?

Knowledge classification

As we can see from the role analysis, the provenance of a data product in the water domain may involve multiple user groups, heterogeneous datasets, tools and/or systems. In order to answer the provenance questions discussed previously, a knowledge model is required to describe the provenance involved. By examining the above provenance questions required, we classify the knowledge into three categories as follows.

Domain concepts. Since provenance represents the history of a piece of information, we argue that the domain concepts as an important part of the history should be captured in the knowledge model. In the water domain, the exemplar representative concepts are gauge, observation and flow forecast. Each concept has its own domain-specific meaning and represents rich domain knowledge. For example, a gauge is a sensing device with properties such as location, accuracy and frequency. Such knowledge may help data analysts to make good decisions about the quality of collected data. On the other hand, an observation is an act associated with a discrete time instant or period through which a number, term or other symbol is assigned to a phenomenon (Fowler 1998). It involves application of a specified procedure, such as a sensor, instrument, algorithm or process chain (Cox 2010). By modelling and propagating domain concepts, the knowledge management framework will enable users to discover and re-use the domain knowledge.

Computational methods. The data are accessed and analysed through various computational methods. These methods should be described appropriately to enable users to understand the analysis done to the related data, for example, what parameters were used when the hydrologic model was calibrated.

Lineage relationship. In many cases, a data product is generated by a chain of methods. The lineage relationship among the methods should be modelled to describe the data product's causality graph which captures the dependences between the computational methods. As part of data quality information, ISO 19115–2 (Cox 2009) defines a lineage metadata tag to provide information about the processes involved to produce the data in the dataset. However, the description uses free text and does not readily support the automatic processing of provenance information (Yue *et al*. 2010).

System classification

Provenance can be generated by heterogeneous systems. We classify them into the following four groups which represent most of the existing hydrological working environments.

Database. The observation data recorded by gauges are mainly stored in relational databases which are accessible via web services or specific applications through query interfaces.

Workflow system. A workflow management system provides a visual environment to design, execute and re-use scientific workflows. In our context, data analysts use workflow systems to process observations retrieved from the databases.

Standalone application. There are some applications and tools that are designed to solve hydrological domain problems. For example, simulation models use gauge observations to generate various predictions.

Web service. A web service is a software system which provides an API for managing and/or retrieving information. For example, the Sensor Observation Service (SOS) is one of the web services developed by the Open Geospatial Consortium (OGC) (http://www.opengeospatial.org/standards/sos) for publishing and retrieving observation data.

Operating system. Processes are executed using command line interpreters to invoke a sequence of system functions. Commonly, the commands are recorded as scripts in files to complete some repetitive tasks.

Provenance challenges and requirements

In our experience with various hydrological working systems, it is impossible to answer most of the provenance questions since the knowledge of information life cycle of data is not modelled and propagated. To answer the provenance questions discussed above, a systematic approach needs to be developed to: (a) harvest the knowledge generated by different user groups; (b) make the provenance understandable by both human and machine; (c) link provenance generated by heterogeneous systems together to form the information life cycle of a data product; and (d) make the information life cycle accessible and queryable to the user groups. Therefore, a knowledge model is required and should satisfy the following key requirements:

- The model should be able to describe the information life cycle of a data product in a precise way that includes the domain concepts, the computational methods involved and the lineage relationship. This is essential to improve data transparency and enable interoperability for discovering and re-using provenance knowledge.
- Since data analysis could be conducted by different communities, which span multiple disciplines, the knowledge covered by the information life cycle of a data product could be large. To make the knowledge manageable, the knowledge model should be designed in a modular way. This will have the benefits of efficient query, easy maintenance, good understandability and re-usability.
- At the system level, the knowledge management framework should minimise the impact on the existing data management systems and should be able to harvest and integrate the provenance in a distributed environment.

In the next two sections we will present the knowledge model and the knowledge management framework that satisfies the above requirements.

THE INTEGRATED KNOWLEDGE MODEL

In this section, we introduce the Integrated Knowledge (IKnow) model.

IKnow model overview

While many general provenance models have been developed with distinctive features, a provenance model must be suited to answer the provenance questions generated by specific domain users. Without modelling the domain knowledge, the data could not be interpreted and understood properly.

We develop a new knowledge model to describe the knowledge involved in the information life cycle of data products. The ontology-based approach is to serve as the key enabling technology. An ontology is a formal, explicit specification of a shared conceptualisation (Gruber 1993). It is the representation of knowledge of a domain, where a set of objects (classes) and their relationships (properties) are described by vocabularies with constraints. By applying an ontology-based approach, the knowledge is expressive and computer-interpretable which are essential for distributed heterogeneous systems for knowledge sharing and management.

It is important that the knowledge model is interoperable with the existing domain ontologies, which capture the knowledge generated by the communities. Furthermore, we need to answer provenance questions regarding the knowledge generated from various domains. This requires that the provenance knowledge can be described as if it resides in a unified source. In other words, all the selected ontologies need to be integrated seamlessly. Therefore, the IKnow model is developed to link to the existing ontologies by providing clear alignments to the concepts involved.

By examining the knowledge classification based on the requirements analysis, the IKnow model includes three types of ontologies: (a) the domain concept ontologies; (b) a computational method ontology; and (c) a general provenance ontology to capture the lineage relationships of the information life cycle.

Therefore, the IKnow model captures not only the lineage of water information products but also domain concepts. It is designed using a modular approach to effectively represent the knowledge shared by various communities and improve the interoperability that supports meaningful knowledge exchange among different sources.

Next we discuss our approach to develop the ontology-based Integrated Knowledge model for the water domain. We use a concept map diagram (Novak & Canas 2008) to

show the main concepts captured by the source ontologies and the relationships between concepts.

Ontology selection

We evaluate the existing water domain knowledge representations based on availability and capability. Four ontologies are selected as the source ontologies to develop the IKnow model. In this subsection, we describe the main capabilities of the four ontologies. Some of the selected key concepts and their relationships are highlighted to demonstrate the idea. The complete ontology is larger than that presented in the figure. The linkage among the source ontologies will be discussed in the next subsection.

WaterML (WML)

In the water domain, WaterML (Taylor *et al.* 2010) is an emerging standard which describes an information model and format for the publication of water observations in XML. It makes use of the Observations and Measurements (O&M) standard (Cox 2010) and specialises it with harmonised definitions of water observation concepts to describe the relevant aspects of hydrological observations. We approximate the WaterML 2.0 as an RDFS model, which we use as the main method to describe observational data.

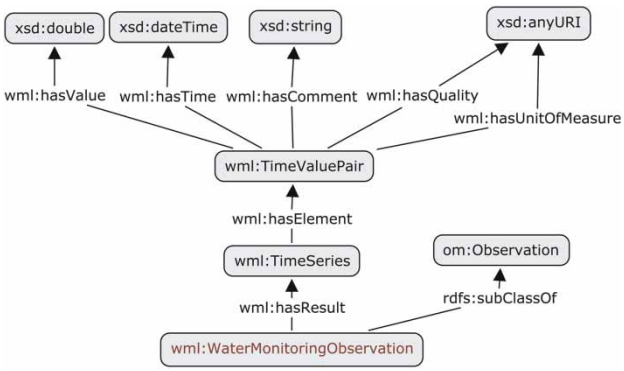

Figure 1 | A fragment of the WaterML RDFS.

In Figure 1, *wml:WaterMonitoringObservation* is a subclass of *om:Observation* that is developed in the O&M standard. It is a *Time Series* which can be described by a *Time Value Pair*. The *wml:hasResult* is modelled as a constraint on the use of *om:result*. WaterML also provides the capability to describe the *Quality* and *Unit of Measure* of observations through linking to other ontologies.

Semantic sensor network ontology (SSN)

SSN was developed recently by the W3C Semantic Sensor Network Incubator Group (Lefort *et al.* 2011). It is a general model to encode the capabilities and operations of sensor assets. It provides a framework for describing sensors. This makes it as a good candidate to describe stream gauges.

In Figure 2, *Sensor* produces *ObservationValue* and its capabilities such as *Accuracy*, *Frequency*, *Response Time*, etc., can be described by the corresponding class.

Process ontology (PO)

PO is part of the OWL for Services (OWL-S). OWL-S is an ontology of services developed by the OWL Services Coalition (Martin *et al.* 2007). It aims to enable web services

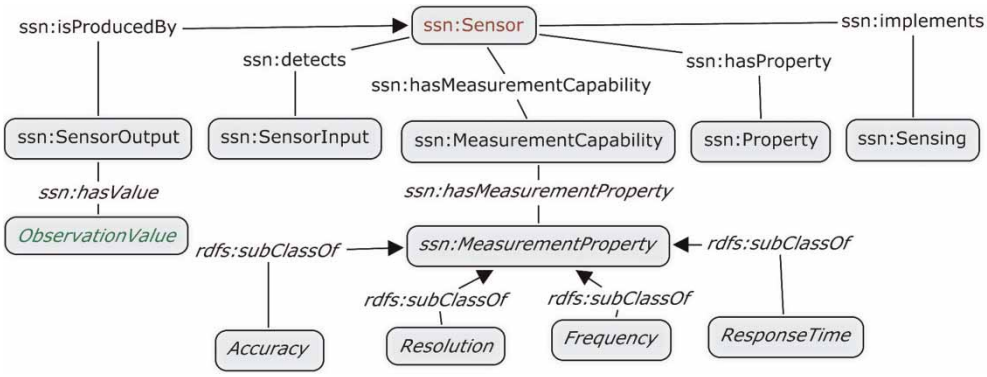

Figure 2 | A fragment of the semantic sensor network ontology.

with particular properties to be discovered, invoked, composed and monitored. OWL-S includes an upper ontology and three sub-ontologies: (a) the service *profile* for advertising and discovering services; (b) the *process model*, which gives a detailed description of a service's operation; and (c) the *grounding*, which provides details on how to interoperate with a service via messages. As part of our IKnow model, we use the process model for the representation of the computational methods.

Figure 3 shows a fragment of the concepts to describe a *Process*: *Input*, *Output*, *Parameter, Result*, etc. These concepts are further modelled in the corresponding class.

Proof markup language (PML)

PML is a semantic-web-based provenance representation (McGuinness *et al.* 2007). It is defined through three core OWL ontology modules: (a) a Provenance module (pmlp) to support provenance-related entity annotation; (b) a Justification module (pmlj) to support lineage relation annotation; and (c) a Trust module to support trust annotation.

Figure 4 shows some of the main concepts described in pmlp and pmlj. *pmlj:NodeSet* can be regarded as a virtual container that contains a conclusion *pmlp:Information* and the derivation *pmlp:InferenceStep* of the conclusion. *pmlp:InferenceStep* uses *pmlp:InferenceRule* to describe what method is applied and uses *pmlp:hasAntecedentList* to capture the derivation relations (dotted line). *pmlp:Source* captures the data acquisition from the data source. The *pmlp:Query* is a representation of provenance questions.

Figure 5 shows the main idea of how PML represents an information flow. Each rectangle represents a *pmlp:NodeSet* by which the information transformation and its output are described. The information transformation can be a computational method (e.g. Process 1 in Figure 5) or a direct assertion (e.g. Observation 1). The rounded rectangles describe where the information comes from, such as an organisation or an instrument as in our example. The lineage relationships among information transformations are modelled in a 'backtrack' fashion through *pmlj:hasAntecedentList*.

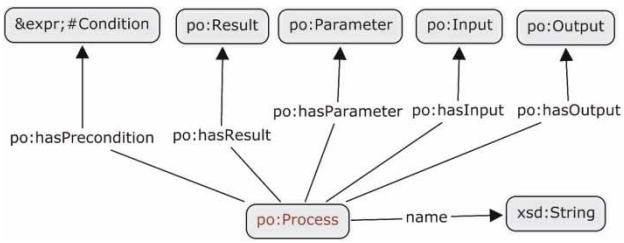

Figure 3 | A fragment of the process ontology.

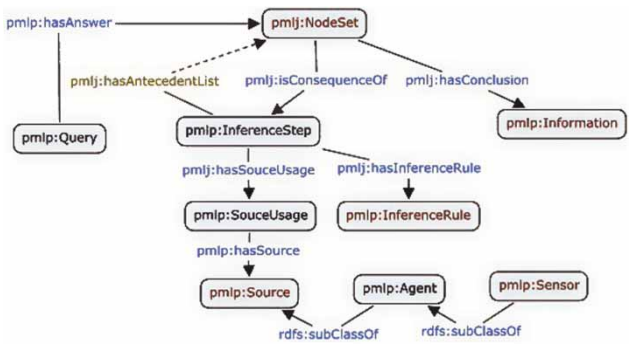

Figure 4 | A fragment of the proof markup language.

Ontology alignment

The knowledge generated in the information life cycle should be queryable by users. This requires that the provenance knowledge resides as if it is organised in an unified ontology.

The selected source ontologies are complementary in some concepts, and overlapping in others. To bring together the ontologies seamlessly, an ontological alignment is needed in order to provide interoperability with the systems

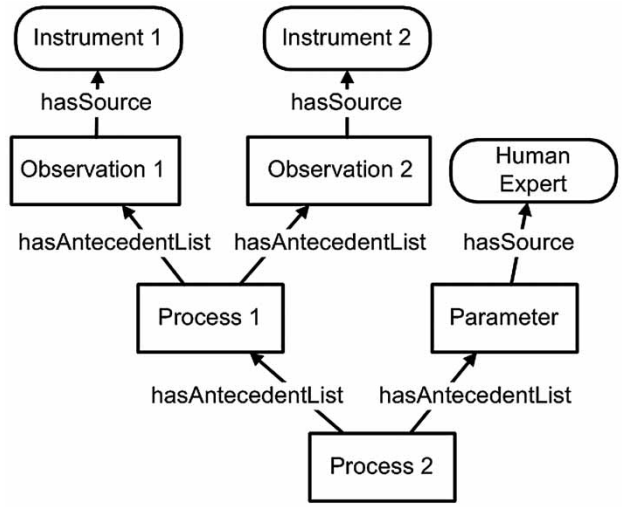

Figure 5 | An example of PML trace.

using them. In other words, the ontologies are aligned with each other for the purpose of exchanging information instead of developing a 'complete' new ontology. We take two steps during the ontology alignment: (a) identify the correspondences among the source ontologies and (b) build mechanisms to represent these correspondences.

Identify the correspondence

We find that that some concepts presented in the provenance knowledge can be represented by several source ontologies. Therefore we need to select appropriate representations from the source ontologies which can provide richer semantic capabilities in our context.

Lineage relationship. We select the PML as the central ontology to link the other source ontologies together. In particular, the PML justification module is selected to describe the lineage relationships among the computational methods and the PML provenance module is used as a bridge to link to the domain concepts and the computational methods.

Domain concepts. As discussed, the main domain concepts to be captured include the knowledge of gauge instruments and the observations generated. Given the information richness captured by Semantic Sensor Network Ontology, *ssn:Sensor* is selected to describe the capability of gauges that generate observations. The corresponding concept captured in PML is *pmlp:Sensor*.

We use *wml:WaterMonitoringObservation* to describe time series generated by some computational methods or sensors. The corresponding concept captured by the PML is *pmlp:Information* which provides a general space to let users specify the information semantics.

Note that some concepts are captured by both domain ontologies. For example, the Semantic Sensor Network Ontology supports the observation concept (*ssn:Observation*) which could be another option to describe time series. We choose to use the *wml:WaterMonitoringObservation*. This is because it specialises O&M with harmonised definitions of water observation concepts as mentioned before. O&M is defined as a standard model for the exchange of observation acts and results. However, the relationship between a sensor and its observations could be retrieved through the PML justification module which will be illustrated later.

Computational methods. The class *pmlp:InferenceRule* describes the execution methods that generate the domain knowledge. However, its describing capability is very limited. *po:Process* provides a very rich description of process. Therefore, a correspondence between *pmlp:InferenceRule* and *po:Process* is developed.

Represent the correspondence

The corresponding concepts identified among the source ontologies need to be aligned with each other to provide global provenance terminology. In general, there are several approaches to represent the correspondences. One simple way is to use *owl:equivalentClass* to imply that the corresponding concepts from two different ontologies have the same meaning. In OWL, this implies that every individual of one class is also an individual of the other class. We believe this is too strong a statement and it may be not true for many cases. For example there is a class *ssn:ObservationValue* in SSN to describe the value of the result of an observation. In WaterML the value of an observation is described by the class *wml:TimeValuePair*; if we define these classes as equivalent, we imply that all *ssn:ObservationValue* individuals are also *wml:TimeValuePairs* (not true) and that all WML observations are generated from sensors (also not true).

Since the the concept defined in PML is very generic, we define a mapping whereby domain classes are subclasses of the more generic PML classes. Specifically, the *wml:WaterMonitoringObservation*, *ssn:Sensor* and *po:Process* are defined as subclasses of *pmlp:Information*, *pmlp:Sensor* and *pmlp:Method Rule*, respectively (see Figure 6).

Although PML is defined as the central ontology to align all the other three source ontologies, there are also some correspondences among the other source ontologies. Figure 6 shows the conceptual alignment among the four source ontologies based on the correspondence identified.

Both Semantic Sensor Network Ontology and WaterML provide the semantics of observation concept. However, observations could be generated either by gauges or by computational methods such as gridding algorithms. Here we intend to distinguish the sources of observations. The observation generated by gauges are captured by *iknow:*

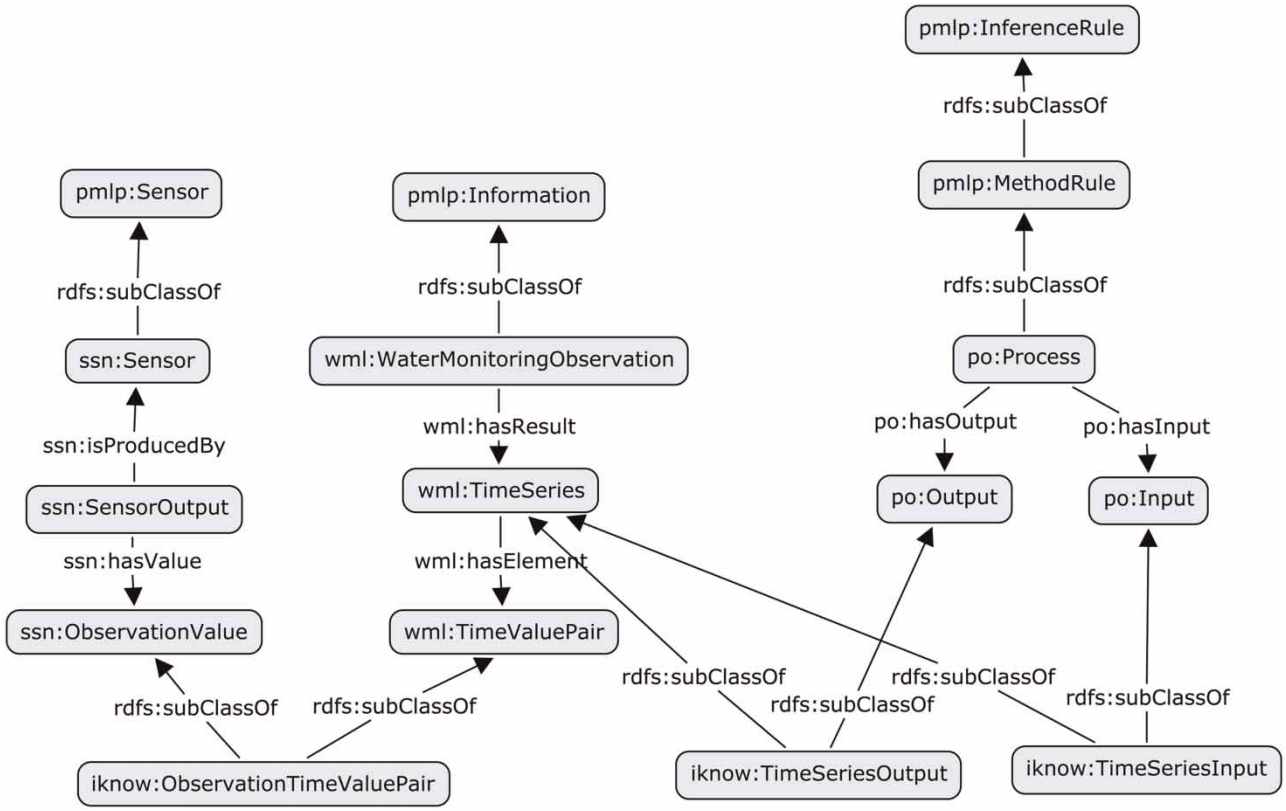

Figure 6 | A fragment of the IKnow model.

ObservationTimeValuePair which is a sub-class of *ssn:ObservationValue* and *wml:TimeValuePair*. The observations generated by computational methods are described using *wml:TimeValuePair*.

The correspondences between WaterML and the Process Ontology show that some time series act as inputs or outputs of computational methods. The *iknow:TimeSeriesInput* and *iknow:TimeSeriesOutput* are introduced to describe the above concepts. They are defined as sub-classes of *wml:TimeSeries* and *po:Input*/*po:Output*, respectively.

Through building the correspondences among the four ontologies, navigation paths are created from PML to WaterML, SSN and PO in the IKnow model. It captures the knowledge presented in the information life cycle of data products. This will enable retrieval of provenance across systems and allow queries to be answered across domains. As long as the concepts involved in the information life cycle of data products are covered by the IKnow model, users are able to generate different provenance graphs from the model.

Based on the above analysis, we can see the IKnow model is a generic model that can be easily extended to link to other domain ontologies through linking the PML-based concepts, such as *pmlp:Information*, *pmlp:Source* and *pmlj:MethodRule*, to domain concepts. Since data analysis can be conducted by different communities which span multiple disciplines, the knowledge covered by the information life cycle of a data product should be diverse and domain-rich. The IKnow model is designed using a modular approach to effectively manage the knowledge involved. This brings benefits of not only efficient querying, easy maintenance, good understandability and re-usability, but also improved interoperability that supports meaningful knowledge exchange among different sources.

THE KNOWLEDGE MANAGEMENT FRAMEWORK

In this section, we introduce the knowledge management framework. It provides knowledge harvesting across heterogeneous systems, knowledge aggregation, storage based on the IKnow model integration and querying capability.

While there are some *ad hoc* solutions for provenance-aware applications, we believe that a systematic approach in dealing with provenance generated by heterogeneous systems is important for the water domain. This will have a benefit for not only the application knowledge management, but also the system knowledge management.

Figure 7 illustrates the architecture of our knowledge management framework. At the bottom of the figure, we show a generic information flow generated by heterogeneous systems. To enable knowledge produced by the various systems to be collected and used, the knowledge management framework includes the following main components.

Harvester service

Tracking provenance is challenging because heterogeneous systems do not support provenance explicitly. The provenance generated by various systems are hidden behind various sources. The two types of sources we encountered are: (a) executed log files generated by systems and (b) databases that exposed the provenance via Web Service interfaces. Customised log harvesters and service harvesters are developed respectively. They are executed to extract intrinsic provenance artefacts from the above two sources. Each harvester generates provenance documents in JSON (JavaScript Object Notation) format and passes the JSON documents to the Data Conversion Service.

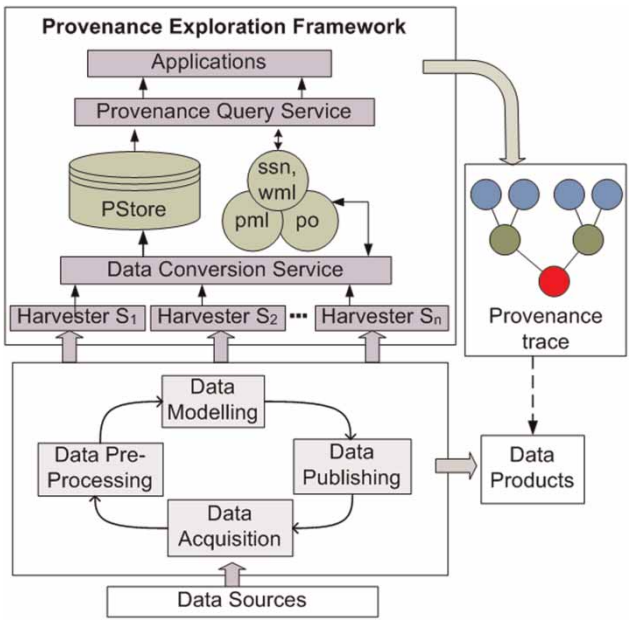

Figure 7 | The knowledge management framework.

A strategy is required to deal with missing provenance. For the two cases studied in this paper, data are generated using a fixed time interval. Therefore, provenance can be harvested based on that pre-defined time interval accordingly. The system is able to identify missing data because the provenance trace structure is known. In this case, the Data Conversion Service is not able to construct an appropriate trace in JSON format and a notification can be generated. Imperfection in provenance is one of the important research topics to be studied in future.

To reconstruct the information life cycle of a data product, we need to stitch together the provenance generated by independent distributed systems. The PASOA (Groth *et al*. 2005) provides the communication protocols to identify dependence among distributed resources. This will be in our future implementation plan. In our current work, the information life cycle of a data product can be reconstructed based on the pre-determined trace structure and the temporal order. We assume that the computational methods are all executed within the same time zone and that the timestamps are generated by synchronised clocks.

Data conversion service

The harvested provenance encoded in JSON documents are pushed into a RESTful web service. They are assembled into RDF instance data based on the designed Integrated Knowledge model. The resulting RDF graph is called a provenance trace. Each executed workflow produces one provenance trace.

Storage

The provenance trace is stored in the AllegroGraph RDF repository via a Jena interface as a named graph. Named

graphs allow grouping of related RDF triples and are therefore a very convenient separation for each provenance trace. AllegroGraph supports Simple Protocol and RDF Query Language (SPARQL), RDFS++ and Prolog reasoning from numerous client applications. AllegroGraph uses disk-based storage, enabling it to scale to billions of triples while maintaining superior performance (http://www.franz.com/agraph/allegrograph3.3//).

Query service

Knowledge, stored in an RDF database, is queried using a separate tool set via the SPARQL. SPARQL is a standard query language and data access protocol for retrieving data encoded in the RDF format. Queries can be phrased according to the terms and structure of the Integrated Knowledge model.

Application

Various applications and visualisation methods can be built on top of the Query Service. In Figure 7, a provenance trace is visualised at top right which shows the information life cycle to generate that particular data product.

The harvester services are designed to minimise the impact on the existing heterogeneous systems. They also minimise the runtime impact on the process execution since the harvesting processes can be executed independently. The knowledge generated by heterogenous systems is harvested and integrated, offering simple but powerful knowledge management functionality. This framework provides a scalable and adaptable approach for knowledge enablement.

MODEL AND SYSTEM EVALUATION

In this section, we evaluate the Integrated Knowledge model and the proposed framework. The model is evaluated from the perspective of how the information life cycle of a data product is represented in a precise way to enable querying the domain knowledge and lineage relationship. The key criteria for framework evaluation is that not only the distributed provenance information can be harvested and integrated but also the impacts on the existing systems are miminised.

A generic prototype of knowledge management framework for the water information system is developed for our two user groups to support knowledge capturing in distributed environments.

The two user groups we engaged presented different types of knowledge acquisition and, therefore, have different issues and challenges of knowledge management. We use the first user scenario to explain how cross-domain knowledge is linked through the IKnow model. For the second user scenario, we demonstrate some limitations of the model and present our approach. For each scenario, we first discuss the user groups and their working scenarios and then present the IKnow model and the systems.

User group 1: Department of Primary Industries, Parks, Water and Environment, Tasmania (http://www.dpiw.tas.gov.au)

The Department of Primary Industries, Parks, Water and Environment (DPIPWE) is responsible for the sustainable management and protection of Tasmania's natural and cultural assets for the benefit of Tasmanian communities and the economy. It guides and supports the use and management of Tasmania's land and water resources and protects and promotes its natural, built and cultural assets.

User scenario

CSIRO Tasmanian ICT Centre has developed a Near-Real-Time Water Information System (nrtWIS) for DPIPWE based on the OGC Sensor Web Enablement (OGC-SWE) standards. The nrtWIS produces flow forecasts in near-real-time for the South Esk River catchment in Northeastern Tasmania, Australia. The real information life cycle of the flow forecast production is complex. We use a simplified version (Figure 8) as a running example to demonstrate the idea.

- Firstly the rain gauges, owned by different government agencies (CSIRO, Bureau of Meteorology, and Hydro Tasmania Consulting), sense the phenomena and send rainfall observations every 15 min into the agencies' databases.
- The rainfall observations are published via OGC-SWE SOS. The SOS provides requesting, filtering, and retrieving

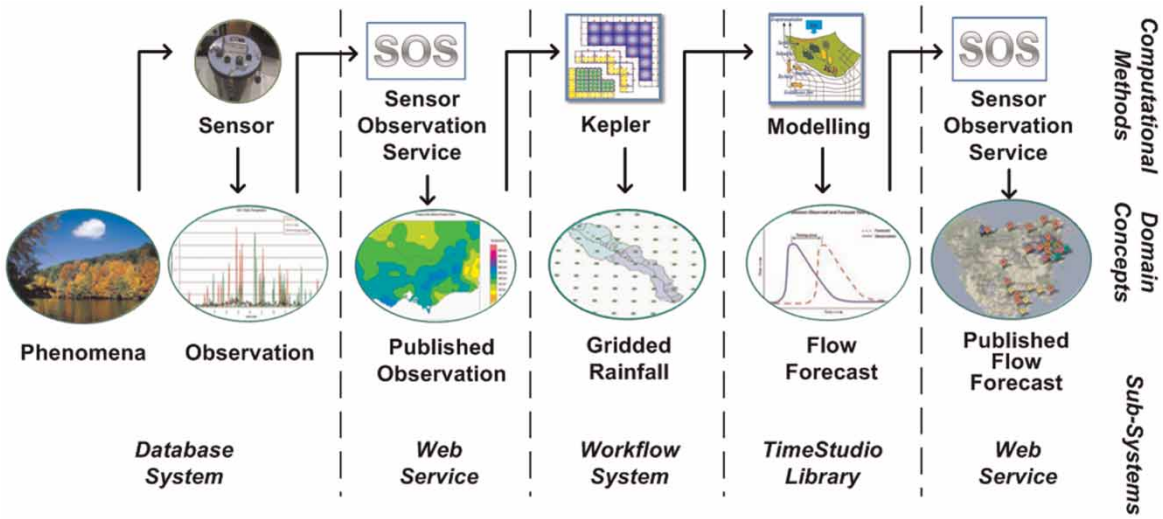

Figure 8 | An example of the simplified information life cycle for flow forecast data generated by nrtWIS.

observations and sensor system information. The interoperability interfaces and metadata encodings are specified to enable integration of heterogeneous sensor information.

- Based on the query received from the forecast model, a Kepler workflow system (Altintas *et al.* 2004), executed by CSIRO, harmonises the time series from the SOSs, checks the quality and validates the errors such as gaps and spikes. It then generates a gridded rainfall surface which is used as an input for the hydrological model.
- The gridded rainfall surface data are received by the HydroTasmania agency as an input for its hydrological model, the Australian Water Balance Model (AWBM). AWBM is calibrated by Hydro Tasmania Consulting and it produces flow forecasts in a 2-hourly interval.
- Finally, the forecasts are published through an SOS.

In Figure 8, we can see the interoperability challenge faced by the Near-Real-Time Water Information System stems from its capturing and querying of knowledge across the heterogeneous systems which do not handle provenance explicitly.

Provenance trace for nrtWIS

Based on the IKnow model, Figure 9 shows a fragment of the provenance trace for the above user scenario. Please note that PML imports the Data Structure Ontology (http://tw.rpi.edu/portal/PML_Data_Structure_OWL_Ontology) (see 'ds' namespace in Figure 9) to encode the 'list' concept.

On the left-hand side, the *NodeSet*s represent the major components involved in the information life cycle to generate the flow forecast: published forecast, forecast, gridded rainfall, published rainfall and rainfall source. Note that the lineage relationships among the components are represented in a backtrack style (dotted line with arrow). For easy presentation, only the published forecast *NodeSet* and rainfall source *NodeSet* are opened for detailed presentation.

As introduced previously, each *NodeSet* captures its conclusion and the justification of the conclusion. For example, in Figure 9, the *iknow:NS_ PublishedForecast* captures the flow forecast result *iknow:Published Forecast_11092010_183026* generated at 18:30:26 on 11 Sept. 2010 through the *iknow:GetObservation* service using the *iknow:ForecastDB* database. The details of forecast (time series) and the *GetObservation* service are described using the domain concept *wml:WaterMonitoringObservation* and the process concept *po:Process*, respectively.

The *iknow:NS_Hydro-Tas_RainfallSource* describes the sensors (e.g. RIMCO_7499 at TowerHill) used to generate the observations which are stored in the HydroTas rainfall database.

The complete provenance trace is much larger than presented in Figure 9. Here we present the main ideas and describe the lineage relationships among the computational methods. It is clear that knowledge of computational

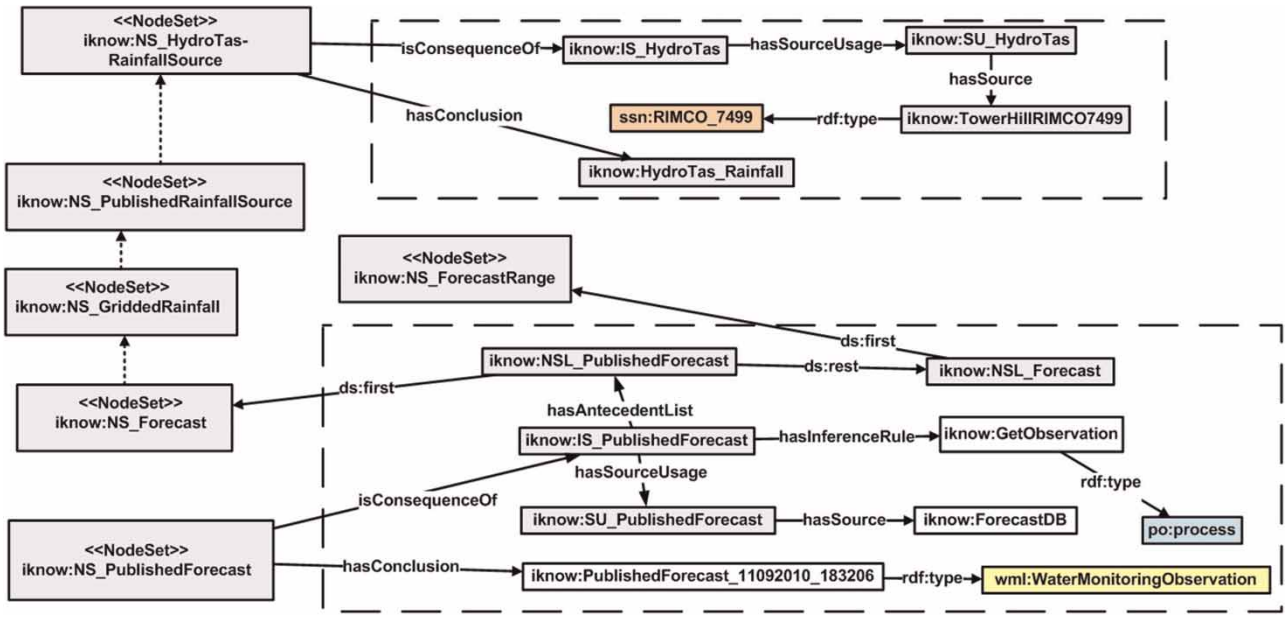

Figure 9 | A fragment of the provenance trace for the near-real-time information system.

methods, the domain concepts and the lineage relationships involved in the above user scenario are all captured based on the designed IKnow model. This enables the cross-domain knowledge (e.g. the accuracy of sensors) to be collected and queried in a unified manner.

Algorithm 1 | Given a forecast, retrieve the simulation model applied

Input:
PublishedForecast_11092010_183026 (the forecast generated at 18:30:26 on 11th Sep., 2010)

Output:
?model (the simulation model that generated the forecast)

Description:

```
1: SELECT ?model
2: WHERE {
3: ?ns pmlj:hasConclusion PublishedForecast_11092010_183026 .
4: ?ns pmlj:isConsequenceOf ?is .
5: ?is pmlj:hasAntecedentList ?nsl .
6: ?nsl ds:first ?ns2 .
7: ?ns2 pmlj:isConsequenceOf ?is2 .
8: ?is2 pmlj:hasInferenceRule ?model .
9: }
```

Since the number of published observations retrieved through a SOS (see Figure 8) may be very large, we decide not to store the retrieved observation result data as part of provenance but rather the SOS query executed by the Kepler

Algorithm 2 | Given a forecast, retrieve the sensor used with its location and accuracy

Input:
PublishedForecast_11092010_183026 (the forecast generated at 18:30:26 on 11th Sep., 2010)

Output:
?sensor, ?location and ?property ?condition (the sensors' locations and accuracies)

Description:

```
1: SELECT ?Sensor ?Location ?Property ?Condition
2: WHERE {
3: ?ns pmlj:hasConclusion :PublishedForecast_11092010}
4: _183026 .
5: ?ns pmlj:isConsequentOf ?is .
6: ?is pmlj:hasAntecedentList ?nsl .
7: ?nsl ds:first ?forecast_ns .
8: ?forecast_ns pmlj:isConsequentOf ?forecast_is .
9: ?forecast_is pmlj:hasAntecedentList ?forecast_nsl_1 .
10: ?forecast_nsl_1 ds:first ?gridded_ns .
11: ?gridded_ns pmlj:isConsequentOf ?gridded_is .
12: ?gridded_is pmlj:hasAntecedentList ?gridded_nsl .
13: ?gridded_nsl ds:first ?rainfall_ns .
14: ?rainfall_ns pmlj:isConsequentOf ?rainfall_is .
15: ?rainfall_is pmlj:hasAntecedentList ?rainfall_nsl_1 .
16: ?rainfall_nsl_1 ds:first ?hydrotas_db_ns .
17: ?hydrotas_db_ns pmlj:isConsequentOf ?hydrotas_db_is.
18: ?hydrotas_db_is pmlj:hasSourceUsage ?sensing_info .
19: ?sensing_info pmlp:hasSource ?Sensor .
20: ?Sensor DUL:hasLocation ?Location .
21: ?Sensor ssn:hasMeasurementCapability ?cap .
22: ?cap ssn:hasMeasurementProperty ?Property .
23: ?cap ssn:inCondition ?Condition .
24: }
```

workflow. This proved to be very cost-effective. The query assembles the required SOS names based on the pre-defined criteria. Then the Kepler retrieves the related observations from the provided SOSs. Figure 10 shows how to use the PML model to capture the query details, such as query content, query creation time and the query engine used.

By capturing the query issued, the system is able to describe how the rainfall observations can be retrieved at query time. The disadvantage of this approach is that, if the database is updated subsequent to the initial query, then the system may return different observations to those retrieved at the time of the initial workflow execution.

Provenance framework

The harvester services collect discrete knowledge from various log files and web services every 2 h and then the provenance trace is generated. Using the provenance trace, we are able to answer provenance questions. We list two queries as examples to demonstrate how the provenance trace is applied to answer queries using SPARQL.

Query 1: What hydrologic simulation model was applied to generate a particular forecast?

Algorithm 1 demonstrates how the hydrologic model is retrieved using the IKnow model. The returned result is 'AWBM' (Australian Water Balance Model).

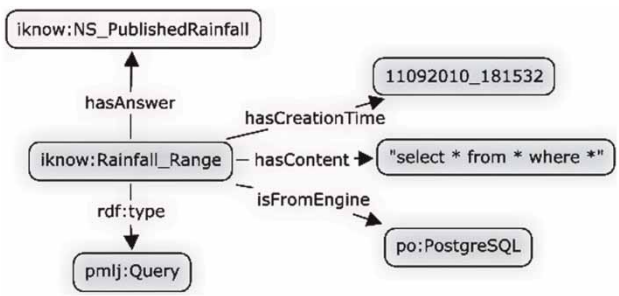

Figure 10 | IKnow query modelling.

Query 2: What sensors are used to generate this forecast and what are their locations and measurement accuracies?

Algorithm 2 demonstrates how SPARQL supports traversing in the IKnow model to answer this query. Lines 3–6 capture the NodeSet *iknow:NS_PublishedForecast* which is linked (line 7) to *iknow:NS_Forecast* (lines 8–10) (see Figure 9). Similarly, *iknow:NS_Grided Rainfall*, *iknow:NS_Published RainfallSouce* and *iknow:NS_HydroTasRainfallSource* are navigated by lines 11–13, lines 14–16 and lines 17–19, respectively. As a result, the queried forecast is linked to the specific sensor that provided the original observations. Lines 20–23 retrieve the sensor capability supported by the Sensor Network Ontology. It is clear that any piece of knowledge could be retrieved as long as the concept is captured in the provenance trace.

Table 1 shows the returned result for query 2. For sensor *RIMCO7499*, its location is at Tower Hill but with two accuracies based on the condition: heavy rainfall or light rainfall. The meaning of *Heavy Rainfall Condition* and *Light Rainfall Condition* could be further queried within the SSN ontology to reveal that they are 250–500 and 0–249 mm/h, respectively.

Figure 11 shows the user interface of the Near-Real-Time Information System. On the left-hand side, the outer blue polygon depicts the South Esk river catchment. Scattered markers within the catchment represent the flow forecast offering at different locations. By clicking on one marker, the short term forecast (red curve) and actual reading till current time (blue curve) is displayed on the right-hand side.

A plug-in function is developed to retrieve the provenance trace associated with each forecast. IWBrowser (http://browser.inference- web.org/iwbrowser/) is used to visualise the provenance trace. IWBrowser provides

Table 1 | Returned result for query 2

Sensor	Location	Property	Condition
esk:TowerhillRIMCO7499	esk:TowerHill	sn:AccuracyPlusMinusThreePercent	esk:RIMCO7499 _HeavyRainlFallCondition
esk:TowerhillRIMCO7499	esk:TowerHill	ssn:AccuracyPlusMinusOnePercent	esk:RIMCO7499 _LightRainfallCondition

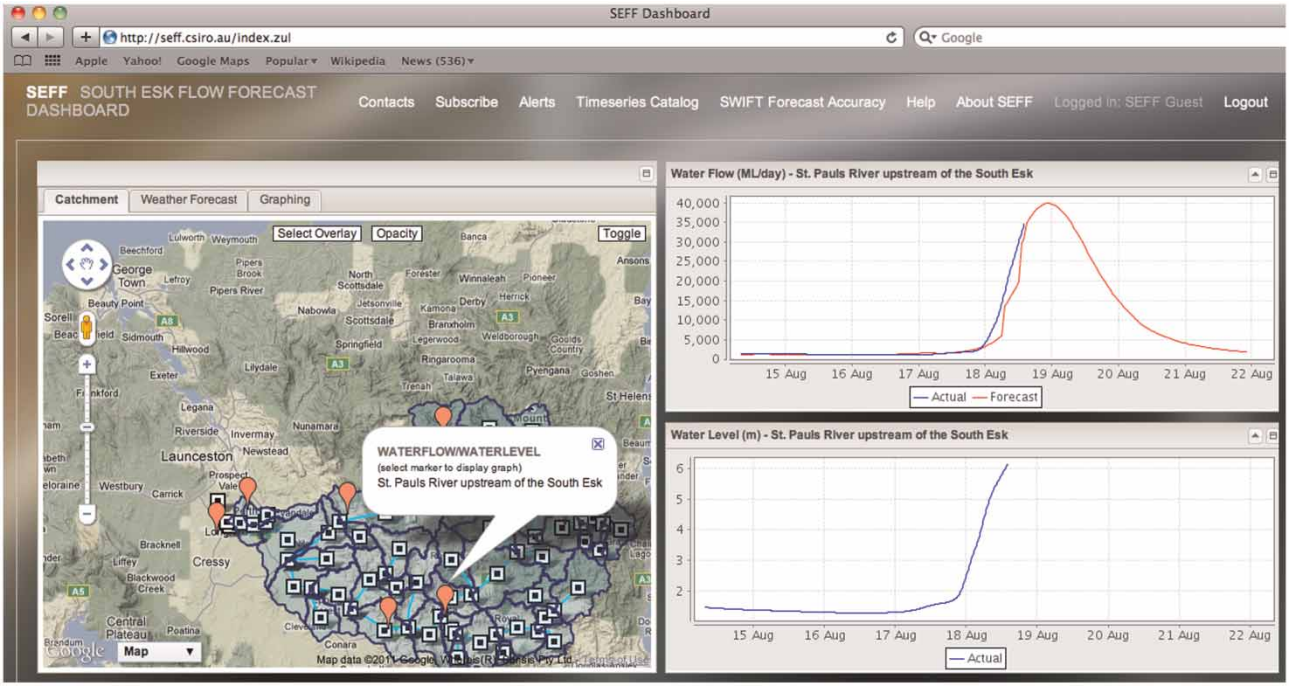

Figure 11 | Interface of the near-real-time information system.

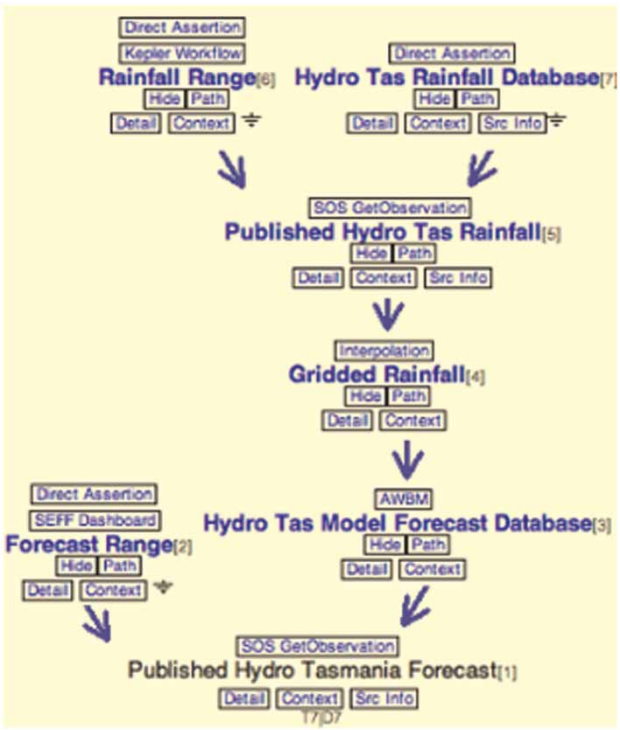

Figure 12 | Provenance trace visualised using inference browser.

graphical rendering capability of a PML trace. Figure 12 shows the simplified provenance trace.

In Figure 12, each component represents a *NodeSet*. Users can click the small box associated with each *NodeSet* to get more details, such as the conclusion, the inference rule or the source used.

By developing the knowledge management framework for nrtWIS, we provide a mechanism for DPIPWE water managers to ask questions about what, who, when, where and how the flow forecast was generated. This will help them to make informed decisions.

Our experience shows IWBrowser is a useful tool to visualise provenance traces. However, the visualisation execution time is unsatisfactory (e.g. several minutes) if a provenance trace is complicated. We find that this holds for most of our real-world cases. Furthermore, it is not very intuitive for domain experts who do not understand the structure of the information model. We believe there is considerable opportunity for further research in domain-independent visualisation and navigation of provenance.

User group 2: Australian Bureau of Meterology (http://www.bom.gov.au/)

The Bureau of Meteorology is Australia's national weather, climate and water agency. It provides regular forecasts, warnings, monitoring and advice spanning the Australian region and Antarctic territory.

User scenario

The Bureau of Meteorology is developing a continental-scale modelling system, the Australian Water Resource Accounting and Assessment (AWRA). The system uses observations of different types in a model data assimilation scheme to produce a national water account. It can then be used to produce national water resource assessments. For example, the landscape model (AWRA-L) developed is a variable resolution (250 m–50 km) gridded landscape hydrology model that produces interpretable water balance component estimates. It is running in an experimental form within CSIRO.

The main computations are run within the Hydrological Forecasting and Warning System (Delft-FEWS), a workflow-oriented modelling environment. It provides the capabilities to integrate large datasets, process the data using specialised modules and allow easy integration of existing modelling capacities through open interfaces. A general workflow to produce groundwater storage and evaporation using FEWS is being developed. Figure 13 shows part of the workflow structure.

It is required that a tracking system should be in place to ensure an effective audit trail.

Provenance trace for FEWS

For the FEWS environment, our users developed configuration files to manipulate the data and run the simulation models. Data are stored in file systems. Compared to the provenance trace for nrtWIS, we use this scenario to demonstrate how to manage the complex lineage relationships presented in information life cycles.

In the PML justification module, there is a restriction defined on *pmlj:NodeSet* that its *pmlj:hasConclusion* property can have only one conclusion. However, it is common in scientific workflow that one process can generate multiple outputs.

Figure 13 shows some of the processes involved in the above scenario. The *AWRA-L* process generates 17 outputs which are each processed by the *Import NetCDF* component separately. If we describe *AWRA-L*'s 17 outputs as one conclusion, the *ImportNetCDF* process could not be properly described because each invocation uses different inputs. For example, the *AWRA-L(output$_n$)* is different from the *AWRA-L(output$_{n-1}$)* although they are both generated by the *AWRA-L*.

Algorithm 3 | PML modelling algorithm of multiple outputs

Input:
P_{parent} that generates the output O where $O \supseteq (O_1, O_2, \ldots, O_n)$; P_i ($i \in (1\ldots n)$) is the process that uses O_i ($i \in (1\ldots n)$) as the input; JP_i ($i \in (1\ldots n)$) is the process annotated with 'defined';

Output:
the provenance trace by the IKnow model

Description:
1: Create *NodeSet* P_{parent}_NS;
2: Set $P_{parent)}\_NS$ pmlj:hasConclusion O;
3: for $i = 1 \rightarrow n$
4: Create *NodeSet* JP_i_NS;
5: Set JP_i_NS pmlj:has Conclusion O_i;
6: Set JP_i_NS pmlj:isConsequenceOf JP_i_IS;
7: Set JP_i_IS pmlj:hasInferenceRule JP;
8: Set JP_i_IS pmlj:hasAntecedentList JP_i_NSL;
9: Set JP_i_NSL ds:first P_{parent}_NS;
10: Create *NodeSet* P_i_NS;
11: Set P_i_NS pmlj:isConsequenceOf P_i_IS;
12: Set P_i_IS pmlj:hasInferenceRule P_i;
13: P_i_IS pmlj:hasAntecedentList P_i_NSL;
14: P_i_NSL ds:first JP_i_NS;

A method is developed to handle the above case without changing the PML ontology. We are given a list of processes, $P_1, P_2, \ldots, P_n$, with inputs generated by their common parent process P_{parent}. For the purpose of appropriate knowledge modelling using the IKnow model, intuitively, a process JP_i ($i \in (1\ldots n)$) is needed to extract each sub-result from the multiple outputs and pass to P_i ($i \in (1\ldots n)$), respectively. During the knowledge modelling step, a collection of *JustfiedNodeSet*s are constructed for each 'virtual' JP_i ($i \in (1\ldots n)$). They are defined as an antecedent of *Pi* ($i \in (1\ldots n)$) *NodeSet*. The inference rule is annotated as '*Defined*', meaning that this is not the actual computational method that was executed. Figure 14 depicts the above idea.

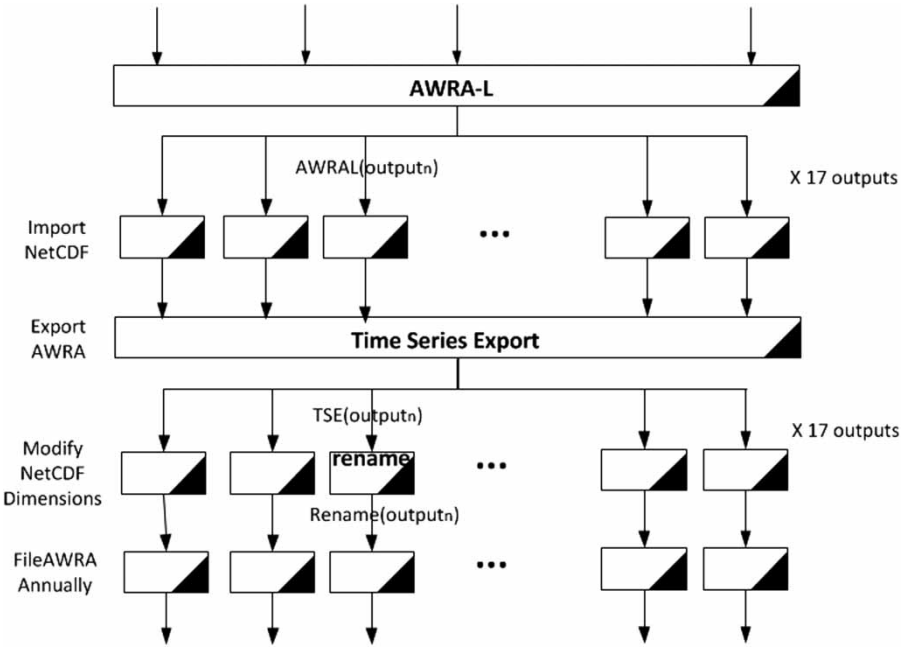

Figure 13 | Part of FEWS workflow.

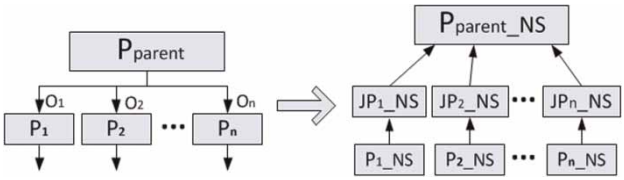

Figure 14 | PML modelling for multiple outputs case.

Algorithm 3 describes the major steps involved to construct *Justified NodeSet* to address the multiple outputs problem. Lines 1–2 set the parent *NodeSet* for P_{parent}. For each child process P_i, lines 4–9 construct the *Justified NodeSet* and the actual *NodeSet* for P_i is constructed between lines 10–14.

In our experience, it is sometimes impossible to track provenance due to the process designed for some particular purposes. In this use case, a process *rename* (see Figure 13) overwrites its input TSE_{output} by its output $Rename_{\text{output}}$. Therefore, TSE_{output} is lost. For PML modelling in this case we set the conclusion of *Time Series Export NodeSet* as 'overwritten'.

Provenance framework

Based on the framework we propose, the harvester services are customised to extract provenance from the log files. The system provides rich search functionality to enable users to search for a particular data and/or computational method.

CONCLUSION

In this paper, we addressed the knowledge management challenge for distributed water information systems. Based on the requirements, we proposed the ontology-based Integrated Knowledge model to describe the provenance semantics captured in the information life cycle of data products. The Semantic Sensor Network Ontology, WaterML, Process Ontology and PML are aligned with each other to represent key domain concepts, computational methods and their relationships in the water domain. The model provides an unambiguous, computer-interpretable form serving as an effective sharing, re-usable and knowledge discovery method for distributed and heterogenous systems. Given the large amount of provenance involved in the information life cycle, a modular-based approach was developed to improve the efficiency and effectiveness of knowledge management.

A generic knowledge management framework was designed to capture provenance generated by distributed heterogeneous systems. It isolates the hydrological

modelling systems from the provenance system and attempts to minimise the impact on the existing heterogeneous systems. This enables scalable, adaptable and domain-agnostic knowledge management.

Based on the designed model and framework, two real-world use cases were examined to verify our approach. The experimental results demonstrate that the IKnow model and the knowledge management framework are practical and extensible and that the provenance queries identified from our user requirements can be answered. The general approach taken in this paper can be applied to encode the domain knowledge of any discipline.

In future, we will investigate how to adopt the Provenance Data Model being developed by the W3C Provenance Working Group to develop the Integrated Knowledge model discussed in this paper. In time, this should enable our framework to leverage additional emerging research for provenance.

ACKNOWLEDGEMENTS

The Tasmanian ICT Centre is jointly funded by the Australian Government through the Intelligent Island Program and CSIRO Water for a Healthy Country Flagship. The Intelligent Island Program is administered by the Tasmanian Department of Economic Development, Tourism and the Arts.

REFERENCES

Altintas, I., Berkley, C., Jaeger, E., Jones, M., Ludscher, B. & Mock, S. 2004 Kepler: an extensible system for design and execution of scientific workflows. In: *Proc. 16th International Conference on Scientific and Statistical Database Management (SSDBM), Greece*. IEEE Computer Society, Washington, DC, pp. 21–23.

Balazinska, M., Deshpande, A., Franklin, M. J., Gibbons, P. B., Gray, J., Hansen, M., Liebhold, M., Nath, S., Szalay, A. & Tao, V. 2007 Data management in the worldwide sensor web. *IEEE Pervasive Comput.* **6** (2), 30–40.

Barga, R. & Digiampietri, L. 2006 Automatic generation of workflow provenance. In: *Provenance and Annotation of Data. Lecture Notes in Computer Science 4145* (L. Moreau & I. Foster, eds). Springer, Berlin, pp. 1–9.

Bhagwat, D., Chiticariu, L., Tan, W.-C. & Vijayvargiya, G. 2004 An annotation management system for relational databases. In: *Proc. 30th International Conference on Very Large Data Bases (VLDB)*, Toronto, Canada. VLDB Endowment, pp. 900–911.

Bose, R. & Frew, J. 2005 Lineage retrieval for scientific data processing: a survey. *ACM Comput. Surv.* **37**, 1–28.

Buneman, P. & Tan, W.-C. 2007 Provenance in databases. In: *Proc. 2007 ACM SIGMOD International Conference on Management of Data SIGMOD '07*, Beijing, China. ACM, New York, pp. 1171–1173.

Cox, S. (ed.) 2009 *Geographic Information – Metadata – Part 2: Extensions for Imagery and Gridded Data*. Open Geospatial Consortium, Wayland, MA.

Cox, S. (ed.) 2010 *Geographic Information: Observations and Measurements. OGC Abstract Specification Topic 20*. Open Geospatial Consortium, Wayland, MA.

Cui, Y., Widom, J. & Wiener, J.L. 2000 Tracing the lineage of view data in a warehousing environment. *ACM Trans. Database Syst.* **25** (2), 179–227.

Dozier, J. & Frew, J. 2009 Computational provenance in hydrologic science: a snow mapping example. *Phil. Trans. A Math. Phys. Eng. Sci.* **367** (1890), 1021–1033.

Fowler, M. 1998 *Analysis Patterns: Reusable Object Models*. Addison Wesley Longman, Menlo Park, CA.

Freire, J., Koop, D., Santos, E. & Silva, C. T. 2008 Provenance for computational tasks: A survey. *Comput. Sci. Engg.* **10**, 11–21.

Groth, P., Miles, S. & Moreau, L. 2005 PReServ: provenance recording for services. In: *Proc. 4th UK e-Science All Hands Meeting*, Nottingham.

Gruber, T.R. 1993 A translation approach to portable ontology specifications. *Knowl. Acquis.* **5** (2), 199–220.

Hartig, O. & Zhao, J. 2010 Publishing and consuming provenance metadata on the web of linked data. In: *Proc. 3rd Int. Provenance and Annotation Workshop*. Troy, NY.

Kim, J., Deelman, E., Gil, Y., Mehta, G. & Ratnakar, V. 2008 Provenance trails in the Wings-Pegasus system. *Concurr. Comput. Pract. Exper.* **20** (5), 587–597.

de Lange, R.-J. 2010 Provenance Aware Sensor Networks for Real-time Data Analysis. Master thesis, University of Twente.

Ledlie, J., Ng, C., Holland, D. A., Muniswamy-Reddy, K.-K., Braun, U. & Seltzer, M. 2005 Provenance-aware sensor data storage. In: *Proc. Workshop on Networking Meets Databases. IEEE Computer Society*, Washington, DC, p. 1189.

Lefort, L., Henson, C., Taylor, K., Payam,, B., Compton, M., Corcho, O., Castro, G., Graybeal, J., Herzog, A., Janowicz, K., Neuhaus, H., Nikolov, A. & Page, K. 2011 *Semantic Sensor Network XG Final Report*. W3C Incubator Group Report.

Liu, Q., Bai, Q., Terhorst, A. & Shu, Y. 2010a Provenance-aware hydrological sensor web. In: *Proc 9th International Conference on HydroInformatics. Chemical Indutry Press*, Beijing, pp. 1307–1315.

Liu, Y., Futrelle, J., Myers, J., Rodriguez, A. & Kooper, R. 2010b A provenance-aware virtual sensor system using the Open

Provenance Model. In: *Collaborative Technologies and Systems (CTS)* (W. K. McQuay & W. W. Smari, eds). Curran Associates, Red Hook, NY, pp. 330–339.

Martin, D., Burstein, M., McDermott, D., McIlraith, S., Paolucci, M., Sycara, K., McGuinness, D.L., Sirin, E. & Srinivasan, N. 2007 Bringing semantics to web services with OWL-S. *World Wide Web* **10** (3), 243–277.

McGuinness, D. L., Ding, L., Pinheiro da Silva, P. & Chang, C. 2007 PML 2: a modular explanation interlingua. In: *Proc. AAAI 2007 Workshop on Explanation-Aware Computing. AAAI Press*, Palo Alto, CA, pp. 49–55.

Moreau, L. 2010 The foundations for provenance on the web. *Found. Trends Web Sci.* **2** (2), 99–241.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E. & Van den Bussche, J. 2011 The Open Provenance Model core specification (v1.1). *Future Gen. Comput. Syst.* **27** (6), 743–756.

Muniswamy-Reddy, K.-K., Holland, D. A., Braun, U. & Seltzer, M. 2006 Provenance-aware storage systems. In: *Proc. Annual Conference on USENIX '06*, Boston, MA. USENIX Association, Berkeley, CA, pp. 43–56.

Novak, J. D. & Canas, A. J. 2008 *The Theory Underlying Concept Maps and How to Construct Them*. Technical Report IHMC CmapTools. Florida Institute for Human and Machine Cognition.

Park, U. & Heidemann, J. 2008 Provenance in sensornet republishing. In: *Proc. 2nd International Provenance and Annotation Workshop*. Springer, Berlin, pp. 208–292.

Patni, H., Sahoo, S. S., Henson, C. & Sheth, A. 2010 Provenance aware linked sensor data. In: *2nd Workshop on Trust and Privacy on the Social and Semantic Web*. Heraklion, Greece, May 31.

Sahoo, S. S. & Sheth, A. 2009 Provenir ontology: towards a framework for escience provenance management. In: *Microsoft eScience Workshop*. Microsoft Research, Redmond, WA.

Sahoo, S. S., Barga, R. S., Goldstein, J. & Sheth, A. 2008 *Provenance Algebra and Materialized View-based Provenance Management. MSR-TR- 2008–170*. Microsoft Research, Redmond, WA.

Scheidegger, C., Koop, D., Santos, E., Vo, H., Callahan, S., Freire, J. & Silva, C. 2008 Tackling the provenance challenge one layer at a time. *Concurr. Comput. Pract. Exper.* **20** (5), 473–483.

Shu, Y., Taylor, K., Hapuarachchi, P. & Peters, C. 2012 Modelling provenance in hydrologic science: a case study on streamflow forecasting. *J. Hydroinf.* (http://www.iwaponline.com/jh/up/jh2012134.htm).

Simmhan, Y. L., Plale, B. & Gannon, D. 2005a A survey of data provenance in e-science. *SIGMOD Rec.* **34** (3), 31–36.

Simmhan, Y. L., Plale, B. & Gannon, D. 2005b *A Survey of Data Provenance Techniques*. Technical report. ACM, New York.

Simmhan, Y. L., Plale, B. & Gannon, D. 2008 Karma2: provenance management for data driven workflows. *Int. J. Web Serv. Res.* **5**, 1–22.

Tan, W. C. 2007 Provenance in databases: past, current, and future. *IEEE Data Eng. Bull.* **30** (4), 3–12.

Tarboton, D. G., Maidment, D. R., Zaslavsky, I., Ames, D. P., Goodall, J. & Horsburgh, J. S. 2010 *CUAHSI Hydrologic Information System 2010 Status Report*. Consortium of Universities for the Advancement of Hydrologic Science, Washington, DC.

Taylor, P., Walker, G., Valentine, D. & Cox, S. 2010 WaterML2.0: harmonising standards for water observation data. *Geophys. Res. Abstr.* **12**.

Vahdat, A. & Anderson, T. 1998 Transparent result caching. In: *Proc. 1998 USENIX Technical Conference*, New Orleans, LA. USENIX Association, Berkeley, CA.

W3C Provenance Incubator Group 2010 *Provenance XG Final Report*. Available from: http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/.

W3C Provenance Working Group 2012 *The PROV Data Model and Abstract Syntax Notation*. Available from: http://www.w3.org/TR/2012/WD-prov-dm-20120202/.

W3C 2004 *OWL Web Ontology Language Overview*. Available from: http://www.w3.org/TR/owl-features/.

W3C 2004 *RDF Vocabulary Description Language 1.0: RDF Schema*. Available from: http://www.w3.org/TR/rdf-schema/.

Yue, P., Gong, J. & Di, L. 2010 Augmenting geospatial data provenance through metadata tracking in geospatial service chaining. *Comput. Geosci.* **36** (3), 270–281.

Zednik, S., Fox, P., McGuinness, D., Pinheiro da Silva, P. & Chang, C. 2009 Semantic provenance for science data products: application to image data processing. In: *Proc. 1st International Workshop on the role of Semantic Web in Provenance Management*, Washington, DC, pp. 1–7.

Zhao, J., Goble, C., Greenwood, M., Wroe, C. & Stevens, R. 2003 Annotating, linking and browsing provenance logs for e-science. In: *Proc. Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data. CEUR Workshop Proceedings*, Aachen, pp. 158–176.

First received 26 October 2011; accepted in revised form 26 April 2012. Available online 26 July 2012