Environmental Modelling & Software 96 (2017) 95–105

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Model provenance tracking and inference for integrated environmental modelling

Mingda Zhang [a], Peng Yue [b, c, *], Zhaoyan Wu [b], Danielle Ziebelin [d], Huayi Wu [a, c], Chenxiao Zhang [a]

[a] *State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, 129 Luoyu Road, Wuhan, Hubei 430079, China*
[b] *School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan, Hubei 430079, China*
[c] *Collaborative Innovation Center of Geospatial Technology, 129 Luoyu Road, Wuhan, Hubei 430079, China*
[d] *University Joseph Fourier in Univ. Grenoble Alpes, CNRS, LIG, F-38000 Grenoble, France*

ARTICLE INFO

Article history:
Received 24 October 2016
Received in revised form
28 June 2017
Accepted 30 June 2017
Available online 7 July 2017

Keywords:
Integrated environmental modelling
OpenMI
Environmental models
Provenance
W3C PROV
Scientific workflow

ABSTRACT

Integrated environmental modeling (IEM) provides a systematic way to couple models for integrated analysis. Coupled models in IEM often exchange data at runtime for time-step based executions. It is a challenge to track which raw observations or intermediate data exchanged at runtime contribute to individual model outputs. Time-step level provenance is needed to audit the trail of model execution or perform diagnosis in case of anomalies. This paper introduces a method to support provenance awareness in IEM. It suggests that individual models should expose necessary interfaces for provenance capturing in IEM environments. The provenance is represented using the W3C PROV model for interoperability. Fine-grained provenance is inferred based on coarse-grained provenance and temporal characteristics of computations of numerical time marching models. The approach is implemented in OpenMI-compliant models. A case study of model provenance tracking and inference on the watershed runoff simulation scenario illustrates the applicability of the approach.

1. Introduction

Environmental models are extensively employed to provide decision support for policy makers. Interdependencies between environmental systems call for an integrated modeling approach to address complex environmental problems (Gaber et al., 2008). For example, the runoff in a catchment depends on climatic forcing inputs such as precipitation and evapotranspiration, and landscape factors such as topography and soil types. Integrated environmental modeling (IEM) is identified as an important discipline to address environmental challenges in a holistic way (Laniak et al., 2013; Voinov and Shugart, 2013). Coupled time marching models in IEM often exchange data at runtime for time-based executions, e.g., a runoff simulation model requests the precipitation at each time step during the simulation. In addition, such models may operate at different temporal resolutions, e.g. daily, weekly, or monthly. The individual inputs at each computational phase are sometimes needed for understanding individual model outputs. Provenance, especially time-step level provenance, could help address such issues by tracing inputs and outputs at each time step. Model provenance focuses on detailed information about the linked models and data flows in generating simulation results. It aids in the reproducibility of scientific results (Tilmes et al., 2013), auditing the trail of model execution, tuning model parameters, and tracing exceptions of model outputs.

Component-based modelling frameworks are widely employed in IEM tools (Granell et al., 2013). A component is a functional unit, which consists of a functional behavior and several interfaces used to interact with other components (Argent, 2004; Gössler and Sifakis, 2005). A component-based modelling framework often includes 1) a software architecture and interfaces to specify component interaction, and 2) methods to create, link, and execute components (Kralisch et al., 2007; Granell et al., 2013). There are various existing frameworks available, such as the Open Modelling Interface (OpenMI) (Gregersen et al., 2007), the Common Component Architecture (Bernholdt et al., 2006), and the Earth System

* Corresponding author. School of Remote Sensing and Information Engineering, Wuhan University, 129 Luoyu Road, Wuhan, Hubei 430079, China.
E-mail address: pyue@whu.edu.cn (P. Yue).

http://dx.doi.org/10.1016/j.envsoft.2017.06.051
1364-8152/© 2017 Elsevier Ltd. All rights reserved.

Modeling Framework (Hill et al., 2004). The OpenMI enables environmental models to exchange data at runtime in a standard way (Moore and Tindall, 2005), which has been a global standard and investigated widely (Laniak et al., 2013). Many existing IEM frameworks, developed originally to meet specific application demands in local communities, are moving forward to accommodate the OpenMI standard to support interoperability (Laniak et al., 2013). It has also recently been adopted and published as an Open Geospatial Consortium (OGC) standard, thereby formally establishing it as a standard protocol for the geospatial community (Vanecek and Moore, 2014). However, to the best of our knowledge, there is no study in the literature that involves provenance tracking in OpenMI-compliant components. Since the OpenMI is being widely accommodated by existing IEM frameworks, supporting provenance tracking for the OpenMI standard could improve the generality of the work and support the wide deployment of the approach.

There are different forms of models. Numerical models including time marching ones are widely used in the Earth sciences to simulate and evaluate complex physical processes (Oreskes et al., 1994). The behavior of numerical time marching models changes over time according to the numerical time-stepping procedure. Time series data, continuous and ordered sequence of data items, are kinds of inputs to such models. Consequently, computations in such environmental models are divided into different computational phases according to time steps. For example, hourly or daily precipitation is a time-dependent input of catchment runoff simulation models. The amount of runoff generated in a catchment changes over time after model initialization. In this paper, we will present our views of provenance using numerical time marching models as an example.

Provenance can be recorded at coarse-grained or fine-grained levels (Buneman and Tan, 2007; Di et al., 2013a; Henzen et al., 2013). Coarse-grained provenance involves tracking the provenance information associated with whole datasets, while fine-grained provenance focusses on tracking those associated with a data item or part of a resulting larger dataset The latter is a more challenging problem, since more provenance information needs to be captured in this case, which often requires additional work. In the rest of the paper, for clarity the term dataset is used to denote input/output data series in IEM, and a data item refers to a record in the data series. This paper proposes a method for model provenance tracking and inference in IEM. The model provenance at the coarse level is captured from OpenMI-compliant components and aligned with the World Wide Web Consortium (W3C) PROV data model for interoperability. At the time-step level, fine-grained provenance could be inferred based on coarse-grained provenance and temporal characteristics of model computations. The approach is implemented in a workflow-based IEM environment. A case study of model provenance tracking and inference on the watershed runoff simulation scenario illustrates the applicability of the approach.

The remainder of the paper is structured as follows: Section 2 introduces some background concepts and related work. Section 3 presents a motivating scenario for the provenance tracking approach we are proposing. Our proposed method is described in section 4. Section 5 presents our implementation of the proposed provenance tracking method. Our conclusions and suggestions for future work are provided in Section 6.

2. Related work

Provenance is information about entities, activities, and agents involved in producing something (W3C, 2013). For example, the provenance of the runoff product in a catchment may include the information about the particular precipitation and evapotranspiration data as entities, environmental models as activities performing the simulation, and the people or institutions as agents providing the data and models. Provenance could be considered as a part of metadata, such as the lineage element in the International Organization for Standardization (ISO) 19115 Geographic Information—Metadata standard (Di et al., 2013a). .This standard defines a metadata schema for geospatial data, which includes a lineage model. The ISO 19115 lineage model is widely used in the geospatial community to represent geospatial data provenance (Di et al., 2013b; He et al., 2015). In ISO 19115, *LI Lineage* is used to describe provenance information, which consists of sources (*LE_Source*) and process steps (*LE_ProcessStep*). *LE_ProcessStep* describes the detail processing information (*LE_Processing*) used in process steps. *LE_Processing* includes algorithms (*LE_Algorithm*), software, procedures, and runtime parameters. The World Wide Web Consortium (W3C) has published a W3C recommendation for the provenance model, i.e. the PROV Data Model (PROV-DM). Some serializations of the model are also published, including an XML schema (PROV-XML) and the PROV ontology (PROV-O). The specifications provide standard ways to represent the provenance information, thus supporting the interoperable interchange of provenance in distributed environments such as the Web. In the geospatial domain, there is some work to leverage W3C PROV and ISO for representing and sharing geospatial data provenance on the Web. He et al. (2015) presented a service-oriented approach for adding geospatial data provenance into spatial data infrastructures (SDI). It demonstrates how geospatial services could be extended to support provenance tracing on geospatial feature types and feature instances.

Traditionally, provenance has been investigated intensively in the database context, where provenance refers to a set of relational algebra operations yielding database views from tables (Buneman et al., 2001). In the context of scientific workflows, data provenance records the process steps and data dependencies (Chebotko et al., 2011; Davidson and Freire, 2008; Yue et al., 2015a). It has gained considerable attention recently in e-Science, since scientific workflows are widely used in e-Science environments (Davidson and Freire, 2008; Di et al., 2013a; Simmhan et al., 2005; Yue et al., 2015b). Some existing workflow systems have been extended to support provenance capturing and query, such as Kepler (Altintas et al., 2006). Geospatial services in SDI were also extended to share geospatial data provenance in the Web environment (Yue et al., 2011). There are some well-known international workshop series on provenance research, such as the International Provenance and Annotation Workshop (IPAW), and the Workshop on the Theory and Practice of Provenance (TaPP), which are co-located recently in Provenance 2014 and 2016.

Adding provenance in IEM can bring benefits to modelers. The data transfers between models included in the provenance information serve to help audit the trail of model execution, locate errors or exceptions, and assist users in performing error propagation analysis and evaluating the quality of model outputs. Model development ingredients including distributed data and models can be linked through provenance information. This can be accomplished by capturing the dependencies among them using the W3C PROV data model, which facilitates the interoperable provenance interchange and discovery on the Web. Once enriched with ontologies, data and models can be published on the Web of Data, making the discovery of modelling components from disparate development groups on the Web easier and more efficient, by exploiting semantic linkages among components and following the same data model and query language from the Semantic Web (Yuan et al., 2013). For example, the linkages between data (e.g., daily temperature and precipitation) and models (e.g., TOPMODEL and