Towards Semantic Provenance in CRISTAL

Jetendr Shamdasani*, Andrew Branson, and Richard McClatchey

CCCS Research Centre, CEMS Faculty, University of the West of England
Coldharbour Lane, Frenchay, Bristol BS16 1QY, UK
firstname.lastname@cern.ch

Abstract. Traceability is an important feature of workflow based systems, and is a key source of provenance data. This paper presents CRISTAL, a mature software platform developed and used at CERN for experiment construction at the LHC. It is entirely workflow based capturing provenance on every aspect of its use from application development to end-user interaction. In this paper we summarize some initial work towards the adaptation of CRISTAL to a more semantic orientation, in particular compliance with the Open Provenance Model.

1 Introduction

Provenance, as in the documentation of the origin or source of something [1], and in particular data and workflow provenance, is an important concern within the area of computer science. This paper describes a stable production level system known as CRISTAL [2] and some early attempts to make its underlying provenance model compliant with semantic web technologies. CRISTAL has been used in many and various projects over the last decade such as neuGRID [3], in the construction of the CMS ECAL at CERN [4], and as the core of the Agilium commercial software suite [5]. It is also currently being employed in the N4U (neuGRID for You) EU FP7 project [6]. A key factor of the N4U project is the requirement for provenance information capture. In N4U clinicians execute analyses on the Grid; provenance information can help them in recreating previous experiments that they or others may have carried out. Also by providing complete traceability of a clinical analysis provenance information can aid in finding areas where a clinicians work may have failed.

The current CRISTAL model is already capable of capturing the information required by the N4U project. However, the core of CRISTAL was developed more than ten years ago, before the semantic web, and so although the data is captured tools are missing to easily communicate that information to other systems. Consequently, we have decided to modernise it by using emerging semantic web based provenance models. This paper presents initial work that has been undertaken at CERN to export the current CRISTAL provenance model to be compliant with semantic provenance vocabularies such as the Open Provenance Model (OPM) [7]. This work is research in progress.

* Corresponding Author

This paper is structured as follows: section 2 presents the CRISTAL platform and how it has been used previously, especially in the context of provenance. Section 3 presents the current CRISTAL provenance model and section 4 describes a preliminary attempt to convert it to the OPM. Finally section 5 presents conclusions with directions for future work.

2 The CRISTAL workbench

CRISTAL is a system designed to provide data management, workflow tracking and change management to an agreed set of user requirements. It is a distributed data and workflow management system which uses a database for its repository, a multi-layered architecture for its component abstraction and dynamic object modelling for the design of its objects and components. These techniques were deemed critical in handling the complexity of data-intensive systems and to provide the flexibility to adapt to the changing production scenarios typical of any research production system. CRISTAL has been based on a so-called *description-driven* approach in which all logic and data structures are described by meta-data, which can be modified and versioned online as the description of the object, component, item or an application changes. A Description-Driven System (DDS) architecture, as advocated previously in [8] is an example of a reflective meta layer architecture (figure 1).

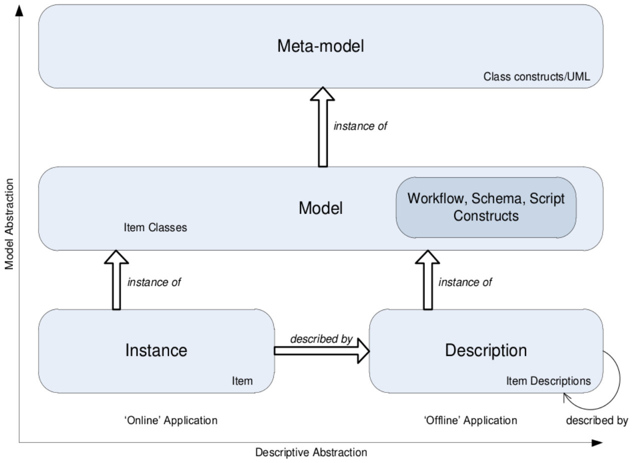

Fig. 1: The CRISTAL meta model

The meta-data along with the instantiated elements of data are stored in the database and the evolution of the design is tracked by versioning the changes in the meta-data over time. Thus DDSs make use of meta-objects to store domain-specific system descriptions that control and manage the lifecycles of domain objects. The separation of descriptions from their instances allows specification and management to evolve independently and asynchronously. This separation is essential in handling the complexity issues facing many web-based computing applications and facilitates interoperability, reusability and system evolution. Separating descriptions from their instantiation allows new versions of defined objects (and in turn their descriptions) to coexist with older versions.

Neuroimaging is constantly developing new algorithms and workflows; these may require variations to the provenance data that is collected. At the same time provenance data, to be useful, needs to remain consistent over time, to be traceable, to be queryable and easily accessible and scientists analyses need to be conducted on those data. CRISTAL handles all of this. The reader is directed to previous publications on DDS for further background ([8], [9]). CRISTAL is essentially a provenance tracking system which has previously been used to track the construction of large-scale experiments such as the CMS project [4] at the CERN LHC it has also more recently been used in the neuGRID project and its follow up N4U. It is both a process modelling and provenance capture tool which addresses the harmonisation of processes so that multiple potentially heterogeneous processes can be integrated with each other and have their workflows tracked in the CRISTAL provenance database.

3 Provenance in CRISTAL

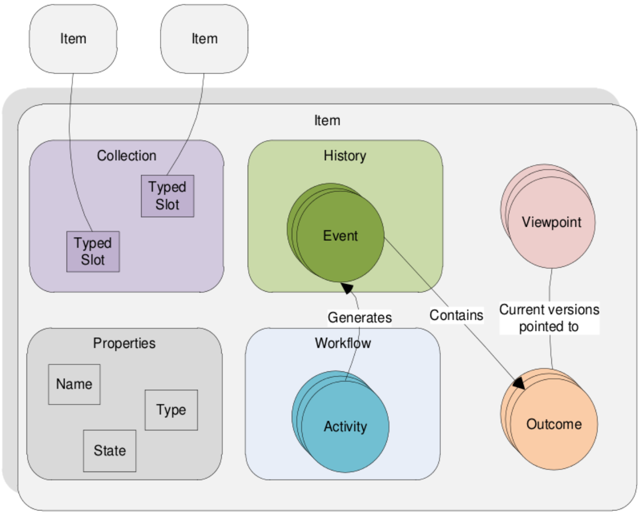

Fig. 2: A representation of the internal CRISTAL model

The current CRISTAL model is shown in figure 2. A collection of all these objects is known as an Item in CRISTAL terminology. An Item contains :

- Workflows i.e. complete layouts of every action that can be performed on that item, connected in a directed graph that enforces the execution order of the constituent activities.
- Activities capture the parameters of each atomic execution step, defining what data is to be supplied and by whom. The execution is performed by agents.
- Agents are either human users or mechanical/ computational agents (via an API), which then generate events.
- Events detail each change of state of an Activity. Completion events generate data, stored as outcomes. From the generation of an Event provenance information is stored.
- Outcomes are XML documents resulting from each execution (i.e. the data from completion Events), for which viewpoints arise.
- Viewpoints refer to particular versions of an Item's Outcome (e.g. the latest version or, in the case of descriptions, a particular version number).
- Properties are name/value pairs that name and type items. Properties also denormalize collected data for more efficient querying, and
- Collections enable items to be linked to each other.

The provenance is captured by Events being generated. For an Agent to write anything within CRISTAL they need to generate an Event. This is done by altering a state in an Activity. The actual provenance information required is application dependent. CRISTAL allows application designers to define their own backend with respect to how they wish to store provenance. Thus a key function of the CRISTAL system is its ability to adapt to changing requirements in terms of provenance storage. The domain of e-science is constantly changing as new workflows, algorithms and research studies are developed. The underlying CRISTAL model allows the system to evolve to handle such challenges whilst retaining provenance information in a consistent and traceable manner. For example in the neuGRID project the CRISTAL provenance service captured:

- Workflow specifications - These were XML based specifications of workflow descriptions which were external to CRISTAL. These were serialised and stored in an relational database.
- Data or inputs supplied to each workflow component - The parameters to each workflow component in the case of neuGrid and N4U these are images, however, they can be any piece of data.
- Annotations added to the workflow and individual workflow components - These consisted of simple name value pairs which allowed uses to store extra information about a workflow.
- Links and dependencies between workflow components - This is a part of the workflow specification.
- Execution errors generated during analysis - A necessary component for any provenance model.
- Output produced by the workflow and each workflow component - In the case of neuGRID and N4U these are images, however, as wit inputs they can be any piece of data that the application requires.

these were all added onto CRISTAL as a relational database model. In the neuGRID project CRISTAL was exposed as a provenance service where information is captured in real time and stored in the current CRISTAL provenance database. However, to enhance CRISTAL we have decided to apply semantic web technologies to the current CRISTAL provenance model. The next section explains some early work on converting the CRISTAL model to be more compliant with the OPM.

4 Towards an OPM compliant description of CRISTAL provenance

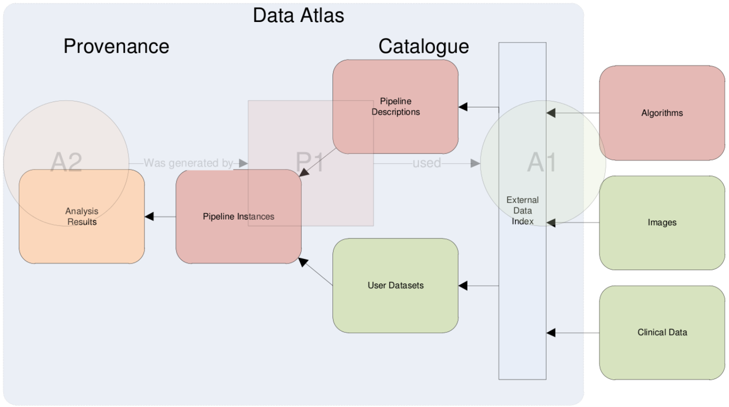

Fig. 3: CRISTAL Provenance in N4U : Green areas represent data, Red areas are processes and Orange areas are outcomes of processes

The OPM is currently a provenance model which is gaining popularity and implementations are becoming available in OWL, RDF and Java. We personally feel that the OPM is a good choice for modernizing the current CRISTAL provenance model since the mapping for the "top layer" of the graphical model fits in well with the current CRISTAL provenance model. This top layer includes Artifact (a physical state such as the result of an action performed), Process (an action that is caused by an artifact and may generate a new artifact) and Agent (an entity which causes the execution of a process). Figure 3 shows the mapping of the tentative N4U CRISTAL Provenance Model onto the OPM.

The colours map onto the conceptual ideas of artifacts (e.g. A1, A2) and processes (e.g. P1) in the OPM. Agents map onto CRISTAL agents since they initiate Processes. Processes from the OPM are equivalent to Workflows and Activities from CRISTAL since they generate Artifacts. Artifacts from the OPM are CRISTAL Events and Outcomes because they are various forms of data. At the time that this initial mapping between the OPM and the CRISTAL provenance model was conceived, there was not an implementation of the OPM available. We are currently in the process of converting the CRISTAL provenance model into the OWL version of the OPM implementation. We believe that once this conversion has been completed it will allow us to demonstrate further how Semantic Web technologies can aid in the provenance of workflows.

5 Conclusion and Future Work

In this research in progress paper we presented CRISTAL which is a system that is able to capture provenance information based on workflows. CRISTAL currently is being used in many different projects and is in the process of being commercialised. An initial mapping of the provenance aspect of CRISTAL to the OPM was shown.

Future work consists of expanding our initial mapping of the OPM to CRISTAL from a simple and preliminary model to using the OWL implementation of the OPM. We believe that this will aid compatibility with other workflow based systems such as Taverna [10] which has an export to OPM option available. As further work we are exploring on how to convert the current CRISTAL model into a full RDF based implementation. This work has already begun and is ongoing. This paper has simply demonstrated how the current provenance aspect of CRISTAL can be made OPM compliant.

References

1. Luc Moreau. The foundations for provenance on the web. *Foundations and Trends in Web Science*, 2(2–3):99–241, 2010.
2. A. Branson et al. Evolving Requirements: Model-Driven Design for Change. *Information Systems*, 2012. Under Final Review.
3. A. Anjum et al. Reusable Services from the neuGRID Project for Grid-Based Health Applications. *Studies in Health Technology and Informatics*, 147:88–99, 2010.
4. The CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3:S08004, 2008.
5. Aglium. http://www.agilium.com. accessed on 03/2012.
6. neuGRID for You (N4U). http://neugrid4you.eu/. accessed on 03/2012.
7. L. Moreau et al. The Open Provenance Model core specification (v1.1). *Future Generation of Computer Systems*, 27(6):743–756, 2011.
8. F Estrella et al. Pattern Reification as the Basis for Description-Driven Systems. *Journal of Software and System Modelling*, 2(2):108–119, 2003.
9. F. Estrella et al. Meta-data Objects as the Basis for System Evolution. In *Proceedings of the Second International Conference on Advances in Web-Age Information Management*, WAIM '01, pages 390–399. Springer-Verlag, 2001.
10. T. Oinn et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, 2004.