Towards Provenance and Traceability in CRISTAL for HEP

Dr. Jetendr Shamdasani, Andrew Branson, Prof. Richard McClatchey (University of the West Of England, Bristol, UK)

Introduction

CRISTAL is an **object lifecycle oriented, description-driven platform** for modelling processes and data structures as '*descriptions*', our term for model-as-data. It is completely **traceable**, meaning that any change made to the model or the instance data is logged with *when* and *who* made the change, resulting in a complete history (**provenance**). All old versions are preserved and kept available for use. It was originally developed at CERN for CMS ECAL assembly tracking, but has since evolved into a generic provenance gathering tool.

Description-driven Systems

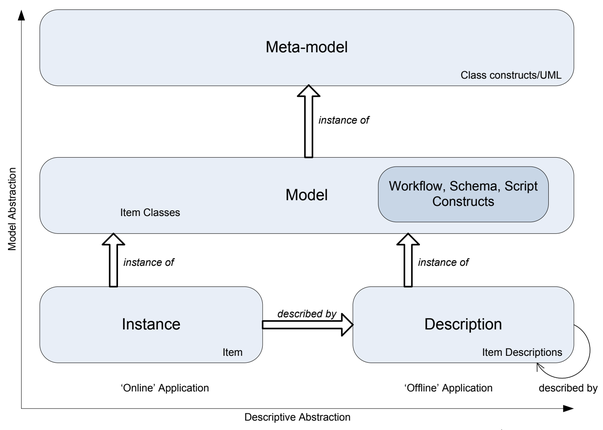


Figure 1: Model vs Descriptive Abstraction in CRISTAL 2[1]

Description-driven systems should rely on configuration data to define their behaviour to the point where *entire applications can be written without compiling additional code*. They are primarily for **data-oriented applications**, as they provide powerful modelling techniques that represent a return to the original principles of object orientation, that modern scripting languages have moved away from.

The CRISTAL Platform

In CRISTAL, objects called **Items** are created which define data structures as **XML Schemas** and application logic as **Scripts**, which are assigned to **Activities** which are then assembled with control flow into **Lifecycles**. These can then be used to create new Items. Crucially, every object in the CRISTAL system is an Item, even the descriptions that define other descriptions.

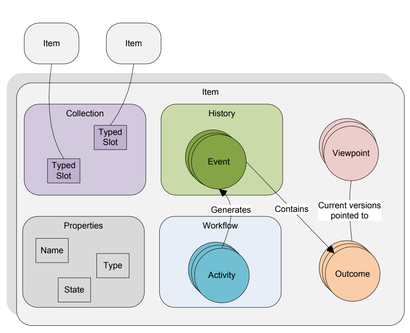


Figure 2: The components of a CRISTAL Item[1]

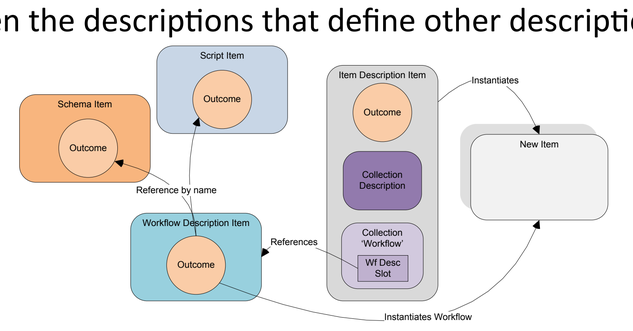
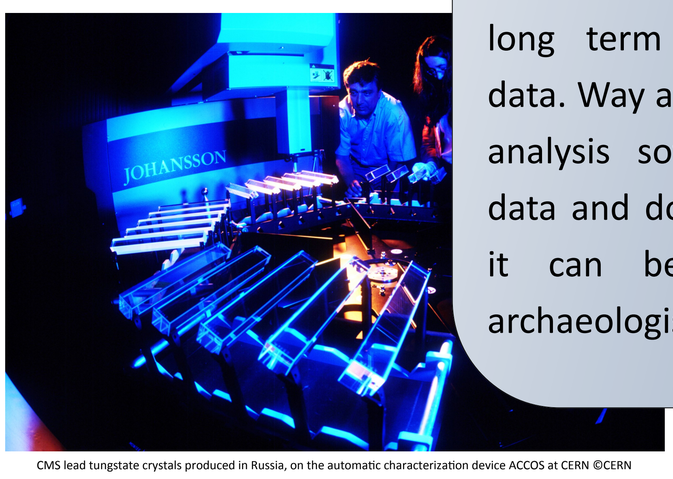


Figure 3: The relationships between Item components and description Items[1]

CRISTAL at CMS ECAL

Originally proposed in 1997 and after a proof-of-concept prototype in 2000, CRISTAL was put into production in 2003 for tracking the assembly of the **CMS ECAL Barrel, Endcaps and Preshower** between different assembly centres. The production lifecycles of the components *changed many times* during production, and CRISTAL seamlessly integrated these changes into a coherent set of data that is now used for calibration.



CMS lead tungstate crystals produced in Russia, on the automatic characterization device ACCOS at CERN ©CERN

Future CRISTAL Applications for HEP

...for analysis traceability

Building on the success of CRISTAL in the neuroimaging analysis domain, it could bring concurrent algorithm versioning and evolution tracking, advanced traceability and solid reproducibility to HEP analyses.

...as a flexible metadata index

Its unique schemaless database with platform level data validation against schemas defined at runtime mean that CRISTAL can store any XML data you can write a schema for, whenever you like.

This makes CRISTAL ideal for indexing external metadata of file repositories.

...for open data publishing

Most data can be transformed to XML. Although bulky, XML is extremely accessible compared to binary data formats. CRISTAL can publish data and schema this way, and log how that data is being used.

...as a long term data archive

The same reasons that make XML so accessible, make it an ideal long term archive format for data. Way after the detectors and analysis software is gone, the data and documentation describing it can be accessible to data archaeologists of the future!

Commercial Exploitation

CRISTAL is currently commercialised by two external companies. The first company, **Agilium** (Annecy, France), developed a BPM (Business Process Management) solution and has been selling their product in the retail, logistics and manufacturing sectors since 2003. An **EC FP7 IAPP project**, ***CRISTAL-ISE***, is now running to increase the knowledge transfer between Agilium and UWE through secondments, also involving a third company from the Manufacturing Execution Systems world, **Alpha3i** of Rumilly, France.

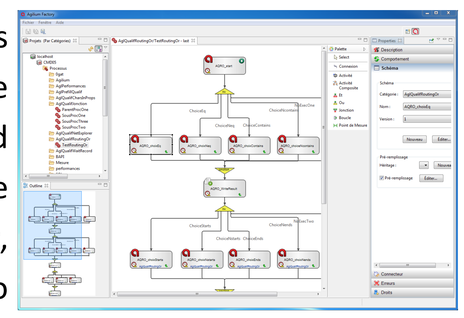


Image: Agilium Factory © Agilium/M1i/Quasar Solutions

In the second company, **Technoledge** (Geneva, Switzerland) it is being applied to fuel cell production lines with a focus on *provenance data capture and management*, therefore demonstrating its maturity as a provenance system. CRISTAL has a commercial license right now, but the next version (3.0, planned for Q1 2014) will be available under an **open source license** (LGPL).

Analysis Provenance and the 'Data Atlas'

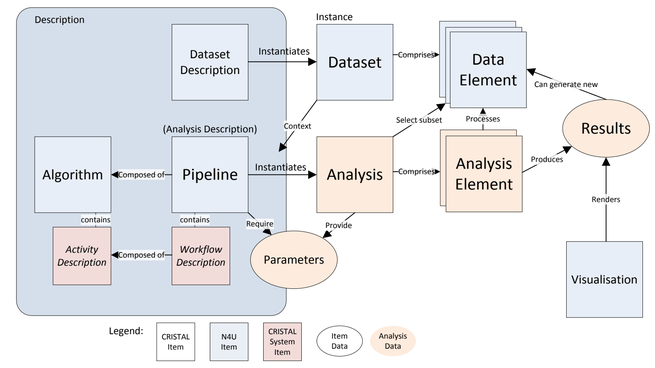


Figure 4: The Item Model for the neuGRID Analysis Service[2]

Since 2007, UWE has been involved in the **neuGRID** consortium, applying CRISTAL to **scientific analysis provenance**. In the first project it monitored grid workflow execution and constructed provenance data from that, but in the second, neuGRID for Users (**N4U**), it is acting as the *full orchestration engine* for executing analysis workflows, sending each job out individually.

CRISTAL's data schema flexibility makes it very useful as a data integration platform. In neuGRID it is used to allow the storage of any metadata that source files supply together in one repository, creating a **'data atlas'** for quickly finding useful subsets. along with analysis results and any user supplied data.

Image: The N4U Project. ©2011 The neuGRID and N4U consortia.

In Summary

This poster is intended to outline the CRISTAL story, its future directions and to suggest possible further uses of it in the HEP community. We are proud of the achievements of CRISTAL so far, and would like to acknowledge the contributions of all those who have helped it get this far.

Contact: ***Jetendr.Shamdasani@cern.ch***
Andrew.Branson@cern.ch
Richard.McClatchey@cern.ch

1. Evolving requirements: Model-driven design to support change. A Branson, R McClatchey, JM Le Goff, J Shamdasani. Under review. Information Systems journal, 2013
2. Deliverable 10.2 of the N4U EC FP7 Project, 2013