CRM$_{dig}$: A generic digital provenance model for scientific observation

Martin Doerr, Maria Theodoridou
Institute of Computer Science, FORTH-ICS, Crete, Greece

Abstract

The systematic large-scale production of digital scientific objects, the diversity of the processes involved and the complexity of describing historical relationships among them, imposes the need for an innovative knowledge management system capable to handle all the semantic information in order to monitor, manage and document the origins and derivation of products in a flexible manner. We have implemented CRM$_{dig}$, an extension of the CIDOC-CRM ontology, which is able to capture the modeling and the query requirements regarding the provenance of digital objects for e-science. CRM$_{dig}$ is particularly rich in describing the physical circumstances of scientific observation resulting in digital data.

1 Introduction

Scientific data cannot be understood without knowledge about the meaning of the data and the ways and circumstances of their creation. This knowledge comprises the provenance of the data. According to the W3C Provenance Incubator Group *"Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource."* [11]. Scientific data can be synthetic, such as simulation data, but mostly they are based on results from observation, in particular on measurements by devices creating digital output. Generally, we use metadata to assess meaning (the recorded things, experimental setup, instrument used), relevance (status, conditions of the recording and derived information), quality (calibration, tolerances, measurement errors, processing artifacts and error propagation) and possibilities of improvement and data reprocessing. The key to provenance metadata is the description of processes starting at the level of human activities or actions, which in turn, among others, initiate "machine events" on devices and computers and form a connected graph through the data and things involved in multiple events in roles such as input and output. The relevant context of these actions comprise descriptions of objects, people, places, times which in turn may be related to other things.

The generic provenance models listed by [11], such as OPM [15] or Provenir [8], do not describe the physical context of scientific measurement, the ultimate origin of most scientific data, nor do they foresee to connect scientific data to descriptions of the observed physical items themselves. Therefore, in the context of the European Projects ACGT on cancer research [2], CASPAR on Digital Preservation [3], and 3D-COFORM on large-scale production of 3D Models for scientific and cultural use [1], we have developed a different digital provenance model called CRM$_{dig}$ [6], which describes much more analytically the scientific observation process. It is possible to map OPM to it [19]. The model is implemented in RDFS. In the course of these projects, we have used the model as schema in RDF knowledge bases for cases including cultural digital productions and reproductions, reasoning on digital rights [17], clinical trials and microbiological examinations [12], satellite data [19] and the processes of 3D model generation [14]. For the latter application case, the model has been further extended by specializations and is used in a Repository Infrastructure designed as a workspace for massive 3D model production, built on a SESAME [9] triple store for the metadata and a distributed data object repository [14, 16]. Rather than pipelining data from one tool to another, each intermediate processing step is immediately stored in the repository together with its atomic provenance data maintaining referential integrity between referred entities. This allows for a much more flexible execution of workflows and reliable, exception-free registration of provenance data. In real life, acquisition and processing data with manual interventions and final results form a complex mesh with potentially thousands (!)

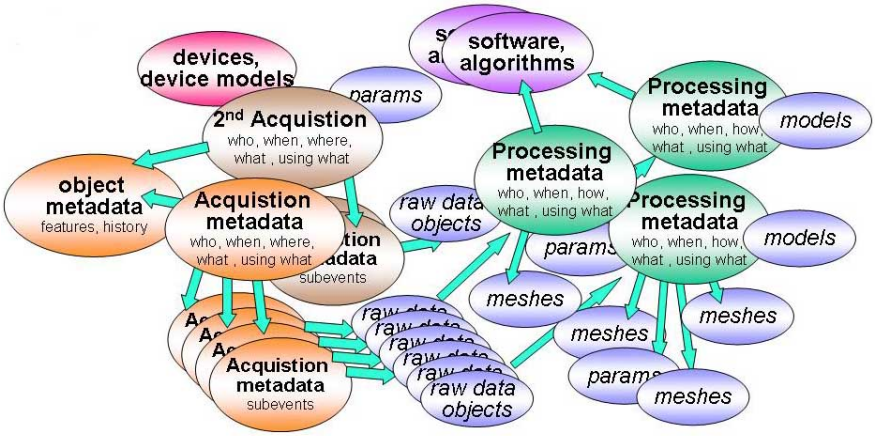

Figure 1: The complex network of metadata

of files, as symbolically represented in Figure 1.

In order not to reinvent the concepts that describe human activities, physical context and basic notions of scientific observation, we chose CIDOC-CRM (ISO 21127:2006) [4] as the core conceptual schema. Originally, it aids at describing the material past as it is observed or documented by archaeologists, historians and museum experts of all disciplines, including biodiversity and science. It is an event-centric core model of about 80 classes and 130 properties, implemented, among other forms, in RDFS [5] and OWL [18]. Instances of the CIDOC-CRM model can be merged to huge meaningful networks of knowledge about historical facts and contextual relationships [13, 4].
The use of the CRM and extensions of it enables an easy integration of provenance data with descriptions of the observed reality and integrated reasoning.

2 CRM$_{dig}$: a model to support provenance metadata

Consequently we have developed an extension of the CIDOC CRM ontology called CRM$_{dig}$, able to capture the modeling and the query requirements regarding the provenance of digital objects. It is based on events that causally relate physical objects, digital objects, actors, times and places in a similar way that OPM relates processes with artifacts and agents [19, 15]. In addition, being an extension of CIDOC-CRM, it inherits the notion of observation of reality and the physical context of material objects, allowing for descriptions of digital objects participating in actions measuring physical properties and of capturing sensor data in a material environment. CRM$_{dig}$describes and integrates the digital provenance with the physical object that has been measured or even digitized. It also describes the devices that participate in the measurement or digitization and makes it possible to follow the history of individual devices, track factors of possible distortion of results and answer complex queries regarding their status. Such queries are useful in error identification and correction, e.g. faults in photo-shooting devices, drop out of a camera or false input parameters in procedures.

At a high level, events in provenance metadata can be structured by answering the following four main questions about:

- WHO: the persons or organizations playing role in the event,
- WHERE: the place the event occurred ,
- WHEN: the time the event occurred,
- WHAT: the things involved in the event.
- HOW: The kind of process and techniques applied.

Since we support multiple instantiation, an action may simultaneously be of multiple types and thus a new processing class can be introduced specializing the more generic ones. Together with the support of specialized links we are able to describe HOW events and things are related.

The basic classes of CRM$_{dig}$ comprise a hierarchy of event classes and a hierarchy of digital things. The *Digital Machine Event* (Figure 2) is a very generic notion essential for e-science. We regard that any such event happens on behalf of a human Actor responsible for it, and that its products belong to this Actor. Therefore we make it subclass of crm:E65 Creation. It is further specialized into *Digital Measurement Event* (Figure 3) and *Formal Derivation*. A *Digital Measurement Event* is modeled as

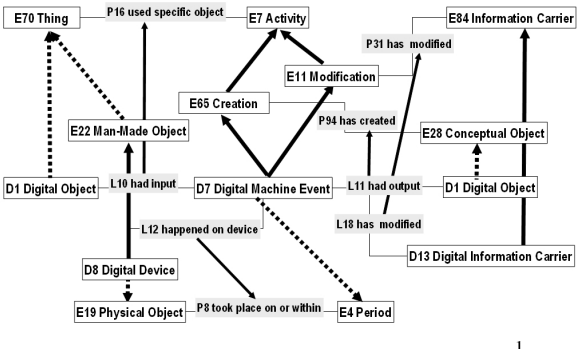

Figure 2: CRM$_{dig}$- Human creation by machine events

a subclass of both *Digital Machine Event* and *Measurement* (a basic CIDOC-CRM class regarding the measuring of physical properties), thus allowing the correlation of the measured object with the device that did the measurement. Being a specialization of ISO21127, it allows for embedding Provenance data in any wider context of social activity and material setting, such as Digitization Processes, taking of sensory data in the ocean, by satellites etc.

In the sequence, we have modeled further specializations, such as the Digitization Process, which is specific to a certain domain and assists, besides others, reasoning about "depicted objects", and the notions of software execution and their physical and social context. For satellite data, the data transmission events to the ground are important. Similar specializations may be created to reason about other specific classes of measurements according to the needs of an application domain. The complete RDF schema of CRM$_{dig}$ version 2.5 is available at [6].

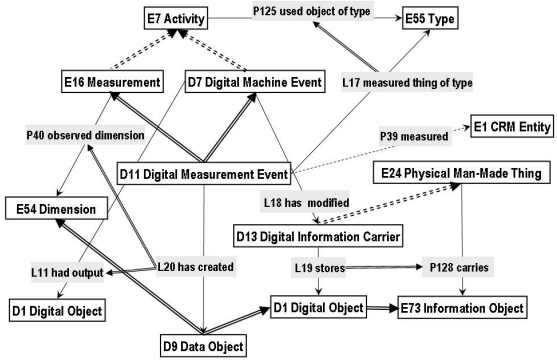

Figure 3: CRM$_{dig}$- Digital Measurement(Activity view)

2.1 Use cases

It is of high importance to record and store semantic information concerning scientific procedures in an integrated semantic network under a common schema which allows for useful inferences on indirect or derived relationships, frequently following deep data paths in the network. If the metadata are consistent and the relationships are well described, chains or data paths are created that correlate events, objects, places, times, people. These paths can be traversed in order to find distant relations relevant for understanding and managing the content. For example, the physical object represented in a given 3D model may not be described in the model's metadata, but may exist in the metadata describing the acquisition process. Figure 4 presents the network of objects and processes for two digitization events and subsequent six processing events related to a specific panel of the V&A Museum. In the context of 3D-COFORM, the panel was digitized with two different techniques and the digitization products were further processed producing finally two 3D models of the object. All relevant data and metadata were stored in the integrated repository RI [1, 14]. Under the assumption, that the subsequent processing events are declared as "subject preserving", which means that the physical object depicted in the derivatives remains the same as the one in the derivation source, it is possible to infer the subject of the model by traversing the following path:

The path connecting the resulting 3D Model with the Physical object:

A.15-1955 3D Model → *was derivative created by* → Formal Derivation
{ → *used as derivation source* → Intermediate Data Object
→ *was derivative created by* → Digitization Process →
(*forms part of*)$^{(0,n)}$→ Digitization Process }$^{(0,n)}$
→ *digitized* → A.15-1955 Physical Thing

If the metadata, that the "A.15-1955 Ivory Panel" was digitized, were physically copied from the acquisition data to each derivative, still they would not capture that all derivatives also "show Jesus Christ". But the museum records significant information regarding the physical object which can be integrated into the RDF knowledge network. For example, we know that "A.15-1955 Ivory Panel" depicts the Ascension and Jesus Christ. Due to this fact, we can answer the question "Find all objects that are about Jesus Christ" exhaustively and the answer set of such a query will consist of the 3D Model and all the intermediate and terminal Data Objects produced during digitization and processing of the A.15-1955 Ivory Panel.

Another issue that can be addressed by queries on the metadata of the processing chain, is tracing device flaws,

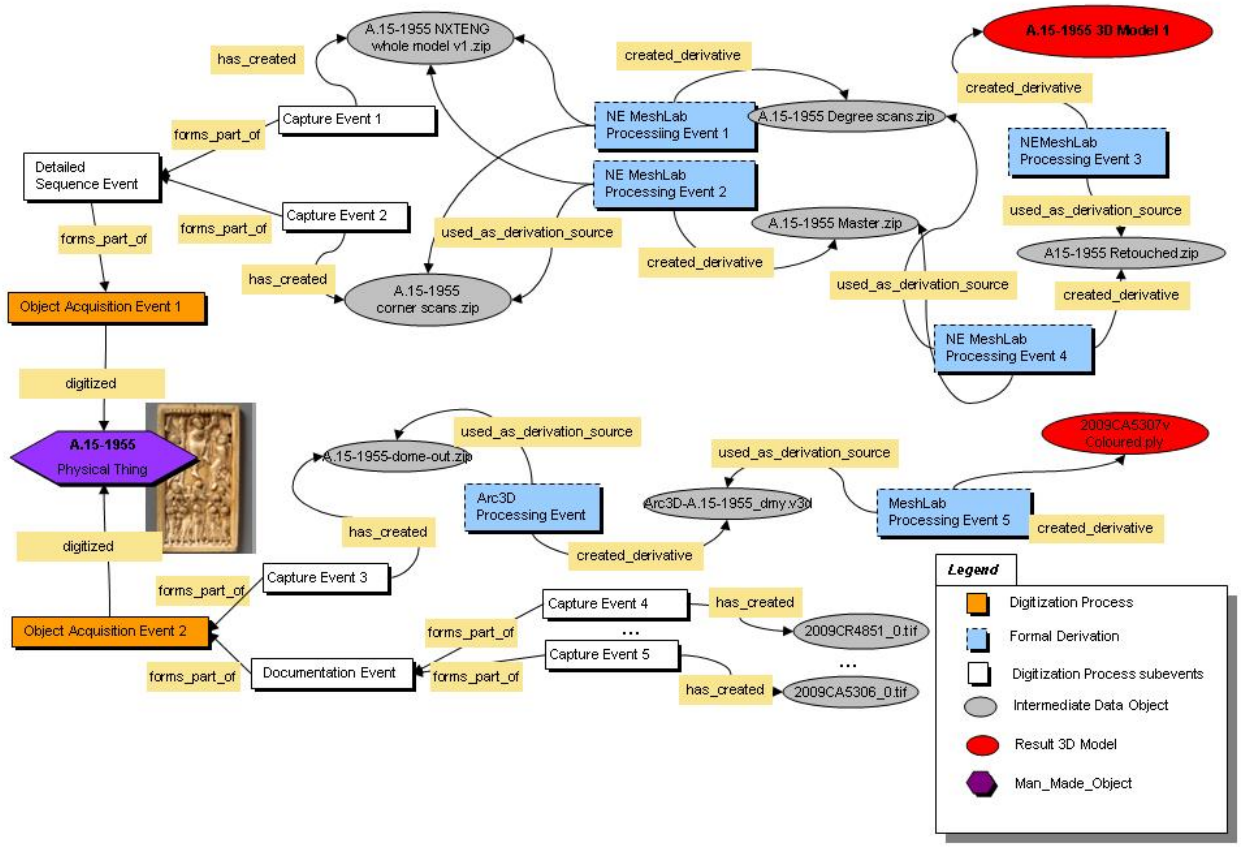

Figure 4: The digitization and processing of a cultural object

e.g. when observing that the same pixels are destroyed in several digital objects that may have been produced by different processes, using the same device. An example that shows the importance of following the history of a device is the Hubble Space Telescope. Within weeks of its launch, the returned images showed that there was a serious problem with the optical system. Nonetheless, during the first three years of the Hubble mission, before the optical corrections, the telescope still carried out a large number of productive observations. The error was well characterized and stable, enabling astronomers to optimize the results obtained using sophisticated image processing techniques such as deconvolution [7]. In December 1993 the First Servicing Mission restored Hubble's Vision. Once Hubble received its corrective "eyeglasses," it began seeing more clearly [10].

3 Conclusions

In this paper we presented CRM$_{dig}$, an extension of the ISO21127 ontology, able to capture the modeling and the query requirements regarding the provenance of digital objects for e-science. The ontology is able to model the physical circumstances of scientific observation resulting in digital data. The ontology is particularly appropriate to describe typical workflows (acquisition, processing, synthesis, presentation) creating a complex semantic network of relationships and to support complex queries which can be resolved by following deep data paths of direct or inferred relationships in the semantic network. Depending on the quality of the required reasoning, more specializations may be introduced. We have verified the model in practice, besides others, by the specialization down to digitization processes. Nevertheless, its contains the constructs at the level of genericity of OPM and other models, and even more, since it is integrated in ISO 21127, which allows for connecting the Provenance view with other parts of reality.

4 Acknowledgments

We gratefully acknowledge the generous support from the European Commission for the Integrated Projects ACGT (FP6-IST-026996), CASPAR (FP6-IST-033572) and 3D-COFORM (FP7-IST-231809).

References

[1] 3D-COFORM: Tools and expertise for 3D collection formation . http://www.3d-coform.eu.

[2] ACGT - Advancing Clinico Genomic Trials on Cancer. http://www.eu-acgt.org/.

[3] CASPAR - Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval. http://www.casparpreserves.eu.

[4] CIDOC CRM. http://www.cidoc-crm.org/.

[5] CIDOC CRM v5.0.2 Encoded in RDFS. http://www.cidoc-crm.org/rdfs/cidoc-crm.

[6] CRMdig 2.5 Encoded in RDFS. http://www.ics.forth.gr/isl/rdfs/3D-COFORM_CRMdig.rdfs.

[7] Hubble Space Telescope. http://en.wikipedia.org/wiki/Hubble_Space_Telescope.

[8] Provenir - Provenance Management Framework: Provenance Algebra and Materialized View-based Storage. http://knoesis.wright.edu/research/semsci.

[9] Sesame: RDF Schema Querying and Storage. http://www.openrdf.org.

[10] The official Hubble site. http://hubblesite.org/the_telescope/team_hubble/servicing_missions.php.

[11] W3C Provenance Incubator Group. http://www.w3.org/2005/Incubator/prov/wiki/.

[12] BROCHHAUSEN, M., SPEAR, A. D., COCOS, C., WEILER, G., MARTIN, L., ANGUITA, A., STENZHORN, H., DASKALAKI, E., SCHERA, F., SCHWARZ, U., SFAKIANAKIS, S., KIEFER, S., DOERR, M., GRAF, N., AND TSIKNAKIS, M. The ACGT Master Ontology and its applications Towards an ontology-driven cancer research and management system. *Journal of Biomedical Informatics 43*, 6, 859–1044.

[13] DOERR, M. The CIDOC CRM - An Ontological Ap-proach to Semantic Interoperability of Metadata. *AI Magazine 24*, 3 (2003), 75–92.

[14] DOERR, M., TZOMPANAKI, K., THEODORIDOU, M., GEORGIS, C., AXARIDOU, A., AND HAVEMANN, S. A Repository for 3D Model Production and Interpretation in Culture and Beyond. *VAST 2010: 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage, Paris, France. pp. 97-104.* (2010).

[15] MOREAU, L., FREIRE, J., MYERS, J., FUTRELLE, J., AND PAULSON, P. *The Open Provenance Model*. University of Southampton, 2007.

[16] PAN, X., BECKMANN, P., HAVEMANN, S., TZOMPANAKI, K., DOERR, M., AND FELLNER, D. W. A Distributed Object Repository for Cultural Heritage. *VAST 2010: 11th International Symposium on Virtual Reality, Archaeology and Cultural Heritage, Paris, France. pp. 97-104.* (2010).

[17] PRANDONI, C., VALENTINI, M., AND DOERR, M. Formalising a model for digital rights clearance. *ECDL'09 Proceedings of the 13th European conference on Research and advanced technology for digital libraries 5714* (2009), 327–338.

[18] REINHARDT, S. This is a proposal of how to encode CIDOC CRM version 5.0.1 in OWL2, May 2009.

[19] THEODORIDOU, M., TZITZIKAS, Y., DOERR, M., MARKETAKIS, Y., AND MELESSANAKIS, V. Modeling and Querying Provenance by Extending CIDOC CRM. *Distributed and Parallel Databases* (2010).