Distrib Parallel Databases (2010) 27: 169–210
DOI 10.1007/s10619-009-7059-2

Modeling and querying provenance by extending CIDOC CRM

Maria Theodoridou · Yannis Tzitzikas · Martin Doerr · Yannis Marketakis · Valantis Melessanakis

Published online: 9 January 2010

Abstract This paper elaborates on the problem of modeling provenance for both physical and digital objects. In particular it discusses provenance according to OAIS (ISO 14721:2003) and how it relates with the conceptualization of CIDOC CRM ontology (ISO 21127:2006). Subsequently it introduces an extension of the CIDOC CRM ontology, able to capture the modeling and the query requirements regarding the provenance of digital objects. Over this extension the paper provides a number of indicative examples of modeling provenance in various domains. Subsequently, it introduces a number of indicative provenance query templates, and finally it describes an implementation using Semantic Web technologies.

Keywords Provenance modeling · Querying provenance information · Digital presentation · Semantic web

1 Introduction

Provenance is the origin or the source from which something comes, and the history of subsequent owners (also known in some fields as chain of custody). The term

Communicated by Walid G. Aref and Ouzzani Mourad.

M. Theodoridou · Y. Tzitzikas (✉) · M. Doerr · Y. Marketakis · V. Melessanakis
Institute of Computer Science, FORTH-ICS, Crete, Greece
e-mail: tzitzik@ics.forth.gr

M. Theodoridou
e-mail: maria@ics.forth.gr

M. Doerr
e-mail: martin@ics.forth.gr

Y. Marketakis
e-mail: marketak@ics.forth.gr

V. Melessanakis
e-mail: melesan@ics.forth.gr

is often used in the sense of place and time of manufacture, production or discovery. Comparative techniques, expert opinion, written and verbal records, as well as results of tests, are often used to help establish provenance. The provenance of works of fine art, antiques and antiquities often assumes great importance. Documented evidence of provenance for an object can help to establish that it has not been altered and is not a forgery or a reproduction. Knowledge of provenance can help to assign the work to a known artist and a documented history can be of use in helping to prove ownership. The quality of provenance of an important work of art can make a considerable difference to its selling price in the market; this is affected by the degree of certainty of the provenance, the status of past owners as collectors, and in many cases by the strength of evidence that an object has not been illegally excavated or exported from another country. The provenance of a work of art may be recorded in various forms depending on context or the amount that is known, from a single name to an entry in a full scholarly catalogue several thousand words long.

Scientific research is generally held to be of good provenance when it is documented in detail sufficient to allow reproducibility. As vast amounts of scientific data are produced daily, their management is of prominent importance for e-science. Scientific data cannot be understood without knowledge about the meaning of the data and the ways and circumstances of their creation. Furthermore, in many cases science data are being used in ways not planned by originators. This justifies the need for a comprehensive and extensible modeling approach where provenance/contextual information can be entered/provided/integrated in a gradual manner.

Due to all these reasons the provenance of digital objects has to be properly archived and this is also suggested by the OAIS standard [18] for digital preservation. It should be stressed that the OAIS standard does not propose any particular conceptual model (or formal ontology). To this end, we need conceptual models able to capture the various forms of provenance information that we may have. The availability of such models can enable the exchange and integration of provenance data and can guide the design of provenance services. Moreover, conceptual modeling is important for designing scientific databases and this is orthogonal to the data model of the employed DBMS (which could be relational, semi-structured or graph-based). The contribution of this paper lies in:

- describing an extension of CIDOC CRM [10, 19] appropriate for digital objects that is more rich than the existing proposals, i.e. Open Provenance Model (OPM) [31],
- providing examples of how provenance information can be modeled with this model,
- identifying a number of basic queries that can be used for reasoning about the provenance of digital objects, and
- describing an implementation of the proposed approach using Semantic Web technologies.

We could say that the proposed model is applicable to all e-Science domains (scientific imaging for various purposes, satellite data, medical laboratory tests, physics experiments), as we are not aware of a domain for which this model should not be applicable.

We would like to clarify that the various provenance-related data management techniques and technologies that have been proposed (e.g. see [7] for a brief overview), including methods for archiving and versioning (e.g. [6, 12, 39]), for associating data with metadata in a flexible manner (e.g. [11, 35]), for named graphs (e.g. [8, 13, 41]), certainly tackle several important technical aspects of the problem but do not cover the problem of finding (and deciding on) a modeling approach that allows the integration and exchange of provenance data, a modeling approach that can be systematically specialized (according to the principles of object-orientation).

The remainder of this paper is organized as follows: Sect. 2 discusses provenance from the perspective of both OAIS and CIDOC CRM, while Sect. 3 describes the extension of CIDOC CRM for digital objects. Section 4 describes indicative provenance queries assuming the CIDOC CRM extension. Section 5 provides indicative modeling examples from various domains. Section 6 discusses implementation using Semantic Web technologies, Sect. 7 discusses related work and finally, Sect. 8 concludes the paper.

2 Provenance and OAIS

OAIS is an ISO reference model for Open Archival Information System defined by a recommendation of the Consultative Committee for Space Data Systems. It provides a framework for understanding archival concepts needed for long term digital information preservation and access. In the context of OAIS, provenance describes events that occur during a digital object's life cycle. It documents the history of the content information, i.e. it tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. Examples of provenance information are the principal investigator who recorded the data, and the information concerning its storage, handling, and migration.

The middle column of Table 1 shows examples of OAIS Provenance (as listed in the standard) for various types of content information. The right column comments each row with respect to CIDOC CRM, while a more detailed discussion is given in the subsequent section.

However OAIS does not propose any particular conceptual model or ontology. What is called "OAIS Information Model" (depicted in Fig. 1) is very simplistic and cannot be considered as a conceptual model (it resembles more a requirements diagram). In brief, it states that each digital information object should be associated with Representation Information, i.e. information needed for interpreting the digital object. This may include information about the structure, the semantics of the digital object. Moreover, OAIS suggests that Preservation Description Information (PDI) should contain provenance information documenting the history of the data object (as illustrated in Fig. 2). Again provenance, is a box and no conceptual model is specified.

Table 1 OAIS and CIDOC CRM provenance

Content information type	OAIS provenance	CIDOC CRM provenance
Space science data	Instrument description	Context of observation/experiment
	Processing history	By whom, Derivation chain
	Sensor description	Context of observation/experiment
	Instrument	Context of observation/experiment
	Instrument mode	Context of observation/experiment
	Decommutation map	Context of observation/experiment
	Software interface specification	Context of observation/experiment
Digital library collections	For scanned collections:	For scanned collections:
	Metadata about the digitisation process	Context of digitisation process
	Pointer to master version	Derivation chain
	For born-digital publications:	For born-digital publications:
	Pointer to the digital original	Derivation chain
	Metadata about the preservation process	Context of preservation process
	Pointers to earlier versions of the collection item	Derivation chain
	Change history	Derivation chain
Software package	Revision history	Derivation chain
	License holder	By whom
	Registration	By whom
	Copyright	By whom

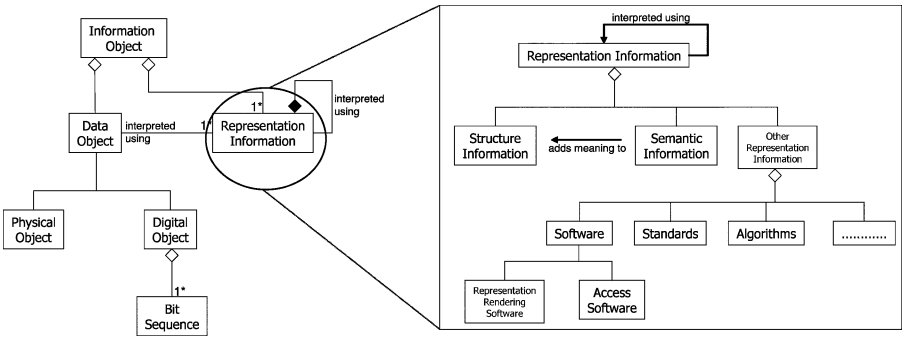


Fig. 1 The information model of OAIS

3 CIDOC CRM extension for digital objects

CIDOC Conceptual Reference Model (ISO 21127) is a core ontology of 80 classes and 132 relations describing the underlying semantics of over a hundred database

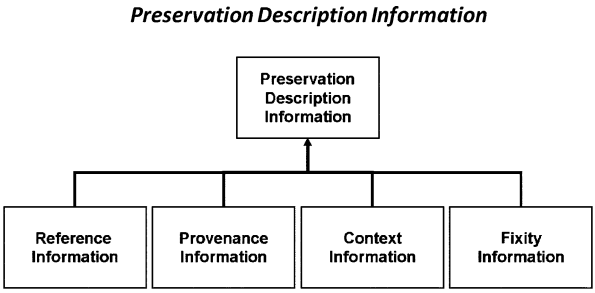


Fig. 2 OAIS PDI preservation description information

schemata and structures from all museum disciplines, archives and libraries. It provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. CIDOC CRM is intended to promote a shared understanding of cultural heritage information by providing a common and extensible semantic framework that any cultural heritage information can be mapped to. CIDOC CRM is the result of long-term interdisciplinary work and agreement. It has been derived by integrating (in a bottom-up manner) hundreds of metadata schemas and is stable (almost no change the last 10 years). We could say that the basic design principles are (a) *empirical bottom-up knowledge engineering* and (b) *object-oriented modeling*. As regards the latter, CIDOC CRM has a rich structure of "intermediate" classes and relations, which apart from being very useful for building query services (enabling queries at various levels of abstraction and granularity), it makes its extension to other domains easier and reduces the risk of over-generalization/specialization. In essence, it is a generic model for recording the "what has happened" in human scale. It can generate huge, meaningful networks of knowledge by a simple abstraction: history as meetings of people, things and information. Figure 3 depicts the main concepts of CIDOC CRM.

Regarding the *modeling methodology*, we have taken as empirical evidence existing data structures from different domains, and analyzed the data structure ele-

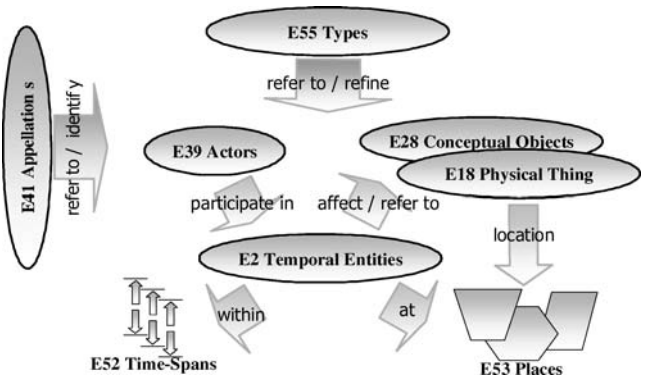


Fig. 3 The main concepts of CIDOC CRM

ments for their underlying common conceptualization necessary to answer questions about the dependency of scientific data on tools, methods and relevant environmental factors of their creation, so that the data quality can be assessed and primary and secondary data can be reused or reprocessed for scientific purposes. The empirical evidence comes from scientific imaging for various purposes, satellite data, medical laboratory tests, physics experiments. In some scientific laboratories, there is not yet an established good practice with respect to complete provenance metadata, or the metadata are highly specific to a particular device. In these cases, our model allows for generalizing and complementing existing metadata creation practices. Top-down approaches, such as OPM [31], suffer from overgeneralization. For instance, due to neglecting the difference between material and immaterial items, OPM cannot describe errors introduced by failures of individual devices, such as dust on a sensor or partial data loss on a DVD.

Regarding the application of CIDOC CRM for scientific data, the idea is that scientific data and metadata can be considered as historical records. Scientific observation and machine-supported processing is initiated on behalf of and controlled by human activity. Things, data, people, times and places are causally related by events. Other relations are either deductions from events or found by observation. In brief, the basic *properties* that we wish to support regarding the extension and application of CIDOC CRM on digital objects are:

- Full *interpretability* of scientific or cultural data with respect to their meaning and quality, in particular all intended and unintended factors possibly influencing the outcome (environmental and hardware effects).
- Ability to *reprocess* primary or half-processed data with different parameters or different algorithms, in particular re-calibration.
- Ability to *trace all dependencies* for digital preservation, such as imminent obsolescence of software to display, process or migrate data. In addition, ability to *clean* reproducible intermediate results, to infer from processing steps *features preserved* between input and output, such as the motif of a digital image under a contrast readjustment ("get all images of this building").
- Ability to search for comparable data sets for integrated evaluation, such as for climate change studies.

Regarding *provenance-related query services*, in the context of CIDOC CRM they can be considered as queries that can take as input an object and answer questions of the form:

- **Context**
  - **by whom (either creator or responsible for creation)**
  - **of observation/experiment**
  - **of digitization**
- **Derivation chain**

The current version of CIDOC CRM (version 5.0.1) can support queries regarding the creator or the responsible for creation of an object ("by whom" type of queries)

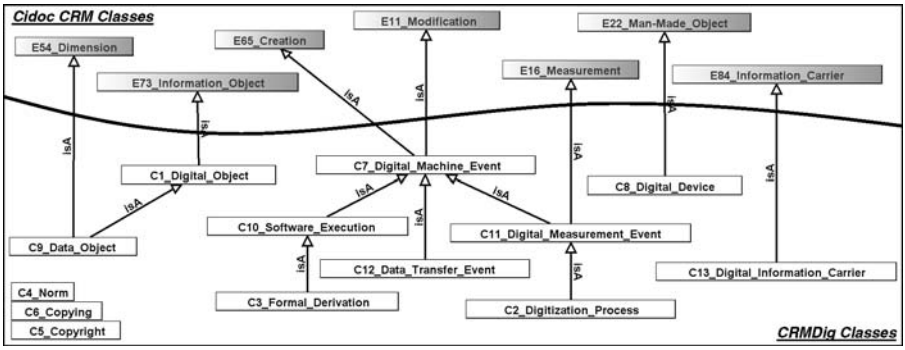


Fig. 4 CIDOC CRM digital ($\texttt{CRM}_{dig}$)

and examples are provided later on. However, the other two types of provenance queries ("context" and "derivation chain" queries) are not directly supported by the current version. For this reason below we describe extensions for capturing such cases. We will refer to this extension with the name $\texttt{CRM}_{dig}$.

3.1 Overview of the extension

Figure 4 depicts an overview of the extensions as visualized by StarLion [38]. CIDOC CRM and $\texttt{CRM}_{dig}$ adopt the following naming conventions:

- **E**XX_Name denote Entities of CIDOC CRM
- **P**XX_Name denote Properties of CIDOC CRM
- **C**XX_Name denote Entities of $\texttt{CRM}_{dig}$
- **S**XX_Name denote Properties of $\texttt{CRM}_{dig}$

The diagram shows the new classes (in *white*) and the directly referred objects from CIDOC CRM (in *gray*).

We have to note that the notion of a digital machine event, digital measurement and formal derivation are very generic, and the essence of e-science. The notion of digitization is specific to certain processes, and assists reasoning about "depicted objects". Similar specializations may be created to reason about other measurement devices.

3.2 Indicative scenario

To introduce the basic concepts of $\texttt{CRM}_{dig}$, we adopt a real world scenario coming from ESA (European Space Agency). The GOME (Global Ozone Monitoring Experiment) dataset, consists of data captured from a sensor on board the ESA ERS-2 (European Remote Sensing) satellite. In general the Satellite (having various properties like name, id) is placed in a particular Orbit (e.g. geosynchronous) and is equipped with a number of Sensors. The captured measurements are sent to a ground earth

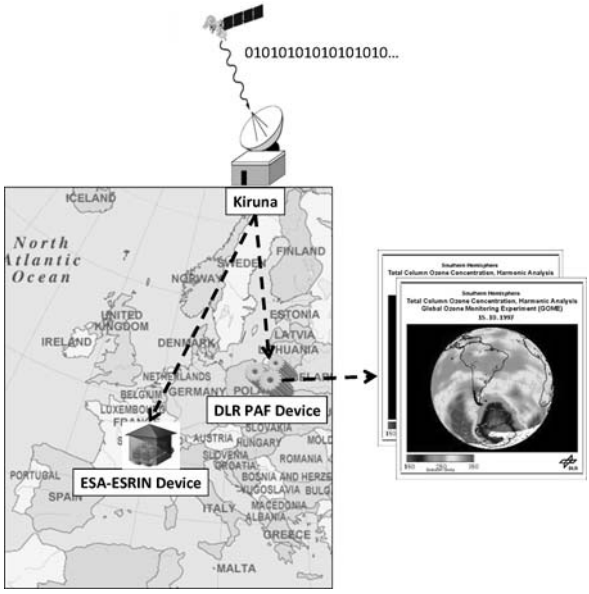


Fig. 5 The trail of GOME data scenario

acquisition station (e.g. at the Kiruna Station), transfered to an Archiving Facility (at ESA-ESRIN) for long term preservation and to a Processing Facility (at DLR—German Aerospace Center) for various data transformations that yield various kinds of Products. Data sets are distinguished according to their processing level to: Level 0 (raw data), Level 1 (radiances/reflectances), Level 2 (geophysical data as trace gas amounts), and Level 3 (a mosaic composed by several level 2 data with interpolation of data values to fill the satellite gaps). Figure 5 illustrates the trail of the GOME data.

Below we describe how this scenario is modeled according to $\texttt{CRM}_{dig}$. Figure 6 shows how data capturing, transmission, processing and archiving events are modelled with $\texttt{CRM}_{dig}$, together with their related products. We adopt a graphical language similar to UML Object Diagrams. The ESA ERS-2 Satellite is modelled as a **C8 Digital Device** whose orbit is recorded in a **E62 String** of **E55 Type** "ORBIT" through the *P3 has note* property and is related through the *P46 is composed of (forms part of)* relationship with Sensors which are also **C8 Digital Device** instances. The data capturing event is modelled as a **C11 Digital Measurement Event** which relates to a Sensor through the *S12 happened on device (was device for)* property and records what it measures through the *S15 measured thing of type (was type of thing measured by)* property. The result of the data capturing event is the creation of a GOME RAW DATA (Level 0) data set, modelled as a **C9 Data Object** and linked to the data capturing event through the *P94 has created (was created by)* property.

The ESA ERS-2 Satellite data transmission to the Kiruna Station is modelled as a **C12 Data Transfer Event**. The data transmission relates to the ESA ERS-2 Satellite through the *S15 has sender (was sender for)* property, to the GOME RAW DATA (Level 0) data set through the *S14 transferred (was transferred by)* property and to

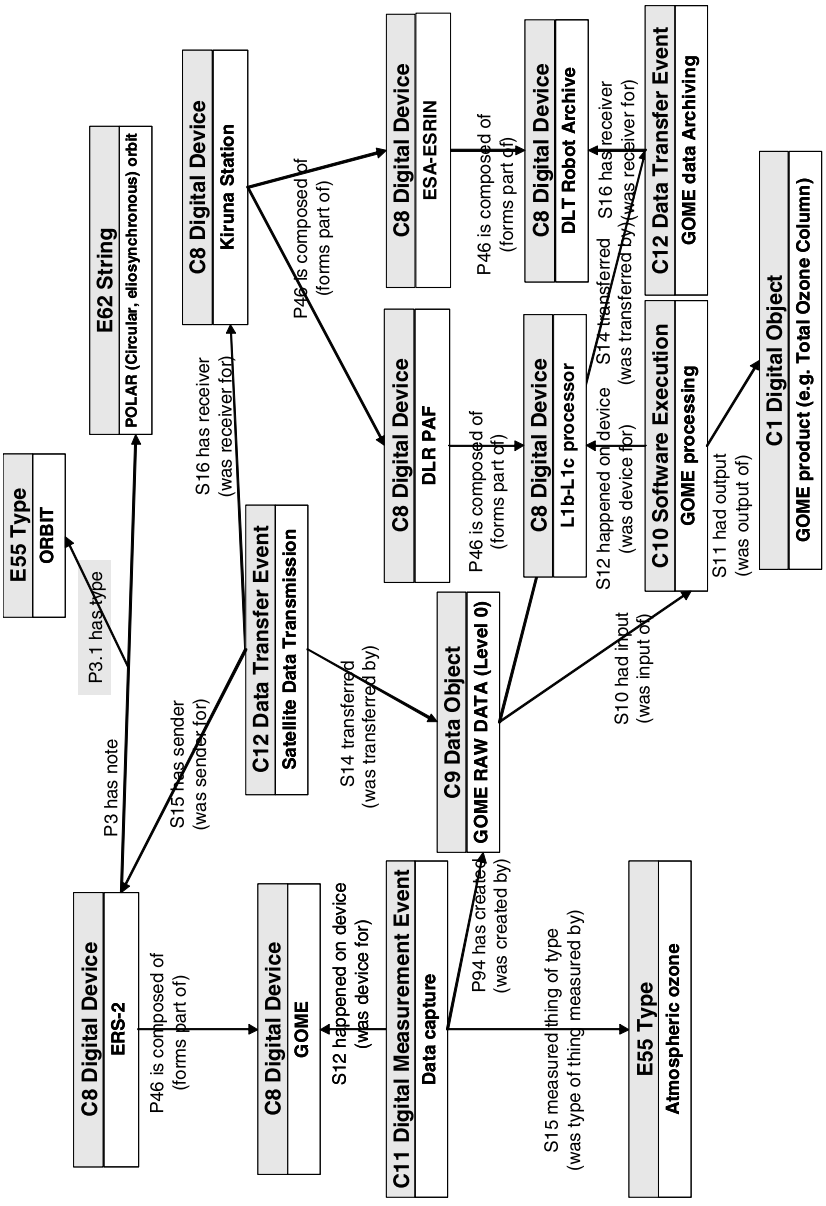


Fig. 6 The trail of GOME data scenario modeled with CRM_{dig}

the Kiruna Station through the *S16 has receiver (was receiver for)* property. The Kiruna Station is modelled as a **C8 Digital Device** which is *P46 composed of (forms part of)* two devices the DLR PAF Device and ESA-ESRIN Device which are also modeled as **C8 Digital Devices**. The DLR PAF is *P46 composed of (forms part of)* the L1b-L1c processor (**C8 Digital Device**) and the ESA-ESRIN is *P46 composed of (forms part of)* the DLT Robot Archive (**C8 Digital Device**) respectively.

GOME processing is modelled as a **C10 Software Execution** which receives as input, through the *S10 had input (was input of)* property, the GOME RAW DATA (Level 0) data set and produces the L0 GOME product (e.g. Total Ozone Column) **C1 Digital Object** (*S11 had output (was output of) property*). GOME data archiving is an event modelled as a **C12 Data Transfer Event** that relates to the DLT Robot Archive through the *S16 has receiver (was receiver for)* property and to the GOME RAW DATA (Level 0) data set through the *S14 transferred (was transferred by)* property.

Figure 7 shows how the transformation of $L0 \rightarrow L1 \rightarrow L2$ products can be modeled using $\texttt{CRM}_{dig}$.

3.3 Detailed description of the new classes

To model context and derivation chain information, we have defined four new specializations of material and immaterial items and six new specializations of events:

- **C1 Digital Object**, which comprises identifiable immaterial items, that can be represented as sets of bit sequences, such as data sets, e-texts, images, audio or video items, software, etc., and are documented as single units. Any aggregation of instances of **C1 Digital Object** into a whole treated as single unit is also regarded as an instance of **C1 Digital Object**. This means that for instance, the content of a DVD, an XML file on it, and an element of this file, are regarded as distinct instances of C1 Digital Object, mutually related by the ***P106 is composed of (forms part of)*** CIDOC CRM property. A **C1 Digital Object** does not depend on a specific physical carrier, and it can exist on one or more carriers simultaneously.
- **C2 Digitization Process**, which comprises events that result in the creation of instances of **C9 Data Object** that represent the appearance and/or form of an instance of **E18 Physical Thing** such as paper documents, statues, buildings, paintings, etc. A particular case is the analogue-to-digital conversion of audiovisual material. This class represents the transition from a material thing to an immaterial representation of it. The characteristic subsequent processing steps on digital objects are regarded as instances of **C3 Formal Derivation**.
- **C3 Formal Derivation**, which comprises events that result in the creation of a **C1 Digital Object** from another one following a deterministic algorithm, such that the resulting instance of digital object shares representative properties with the original object. In other words, this class describes the transition from an immaterial object referred to by property ***S21 used as derivation source (was derivation source for)*** to another immaterial object referred to by property ***S22 created derivative (was derivative created by)*** preserving the representation of some things but in a different form. Characteristic examples are colour corrections, contrast changes and resizing of images.

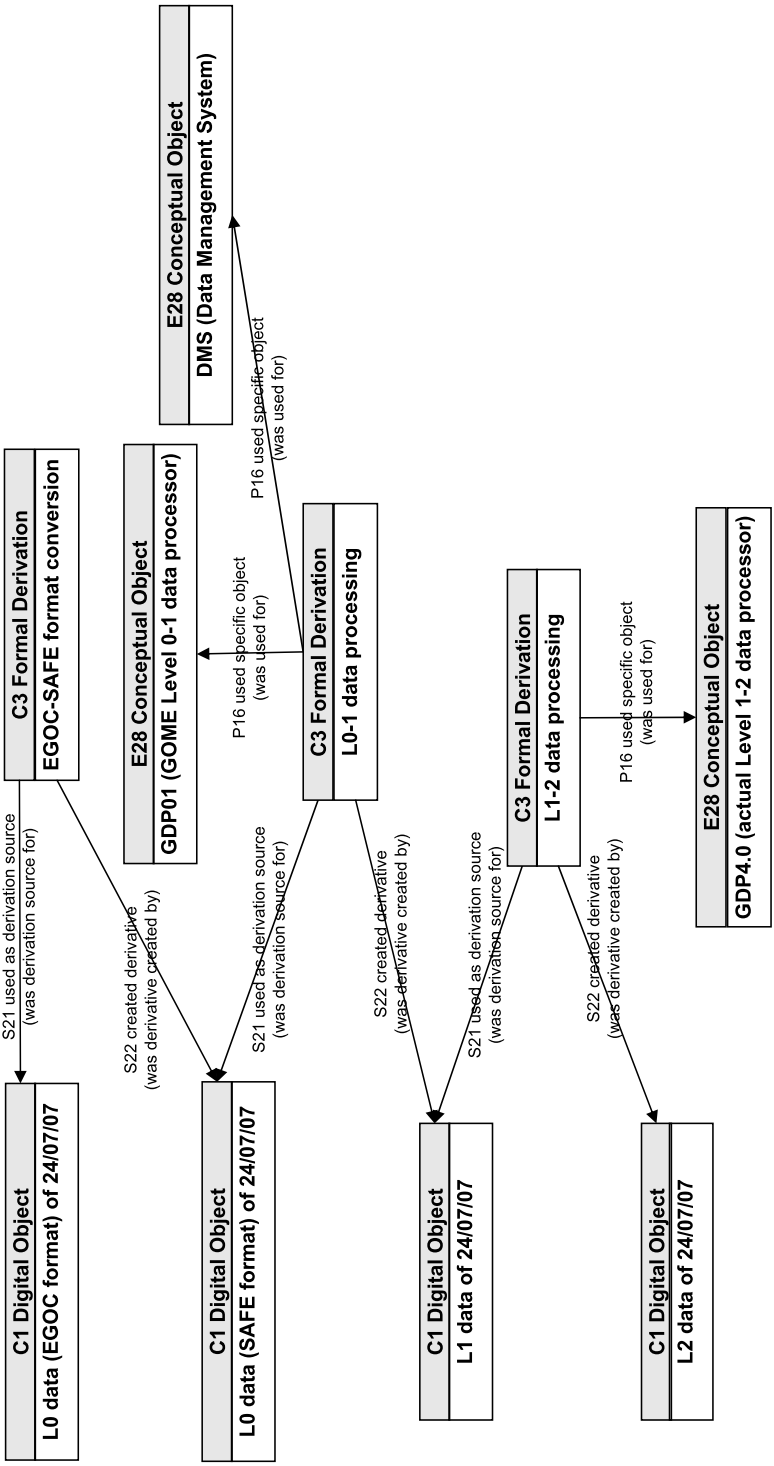


Fig. 7 Modeling the data processing levels of GOME

- **C7 Digital Machine Event**, which comprises events that happen on physical digital devices following a human activity that intentionally caused its immediate or delayed initiation and results in the creation of a new instance of **C1 Digital Object** on behalf of the human actor. The input of a **C7 Digital Machine Event** may be parameter settings and/or data to be processed. Some **C7 Digital Machine Events** may form part of a wider **E65 Creation** event. In this case, all machine output of the partial events is regarded as creation of the overall activity.
- **C8 Digital Device**, which comprises identifiable material items such as computers, scanners, cameras, etc. that have the capability to process or produce instances of **C1 Digital Object**.
- **C9 Data Object**, which comprises instances of **C1 Digital Object** that are the direct result of a digital measurement or a formal derivative of it, containing quantitative properties of some physical things or other constellations of matter.
- **C10 Software Execution**, which comprises events by which a digital device runs a software program or a series of computing operations on a digital object as a single task, which is completely determined by its digital input, the software and the generic properties of the device.
- **C11 Digital Measurement Event**, which comprises actions measuring physical properties using a digital device, that are determined by a systematic procedure and creates an instance of **C9 Data Object**, which is stored on an instance of **C13 Digital Information Carrier**. In contrast to instances of **C10 Software Execution**, environmental factors have an intended influence on the outcome of an instance of **C11 Digital Measurement Event**. Measurement devices may include running distinct software, such as the RAW to JPEG conversion in digital cameras. In this case, the event is regarded as instance of both classes, **C10 Software Execution** and **C11 Digital Measurement Event**.
- **C12 Data Transfer Event**, which comprises events that transfer a digital object from one digital carrier to another. Normally, the digital object remains the same. If in general or by observation the transfer implies or has implied some data corruption, the change of the digital objects may be documented distinguishing input and output rather than instantiating the property ***S14 transferred (was transferred by)***.
- **C13 Digital Information Carrier**, which comprises all instances of **E84 Information Carrier** that are explicitly designed to be used as persistent digital physical carriers of instances of **C1 Digital Object**. A **C13 Digital Information Carrier** may or may not contain information, e.g., an empty diskette.

Below we provide diagrammatic descriptions of various parts of $\texttt{CRM}_{dig}$ and discuss how it can capture various aspects of provenance for digital objects. Regarding graphical notations, simple labeled arrows between classes represent properties. The name of each property is the label of the edge while its domain is the starting class of the edge and accordingly its range is the destination class. The thick arrows represent **IS_A** relations between classes. The dashed thick edges are used to define an **IS_A** relationship transitively. For example in Fig. 9 the class **E22 Man-Made Object** is a (direct) superclass of **C8 Digital Device** while the class **E19 Physical Thing** is an indirect superclass of **C8 Digital Device**.

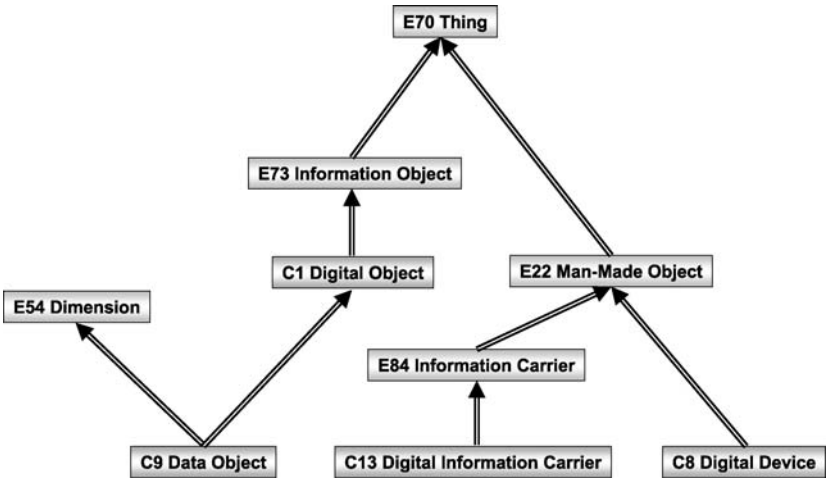


Fig. 8 Basic hierarchy of things. Entities **C1 Digital Object** and **C9 Data Object** model immaterial representations of digital information while entities **C8 Digital Device** and **C13 Digital Information Carrier** model material items

Figure 8 introduces the basic hierarchy of things, Fig. 9 discusses Digital Machine Event, while Fig. 10 focuses on Digital Measurement Event. Figure 11 focuses on the Digitization Process while Fig. 12 focuses on Data Transfer Events. Finally, Fig. 13 focuses on Software Execution, Formal Derivation and Machine Event. Subsequently, Sect. 4 presents specific application examples that use the described classes, while Appendix B contains a detailed description of the classes described in this section following the CIDOC CRM class and property format.

4 Provenance queries over $\texttt{CRM}_{dig}$

Queries regarding provenance, could be based on paths of $\texttt{CRM}_{dig}$. We can identify the following query requirements:

- Get the creator of an object
- Get the earlier versions of an item
- Get the events that changed the custody of an item
- Get the master version of an object
- Get the scanner/resolution of a digital object

Table 2 provides an indicative list of such queries. They can be considered as general purpose query templates that can be refined according to needs. Each template has a name, it takes as input a type-restricted resource (e.g. an instance of **E28 Conceptual Object**), and returns as output a number of typed resources (of course, as in any object oriented system, the type of the actual input/output parameters can be a subtype of the one specified in the template). For each template the path over the semantic graph that should be followed for computing the answer is specified in the form of a sequence consisting of consecutive "edges" of the form:

$$SourceClassName \rightarrow PropertyName \rightarrow TargetClassName$$

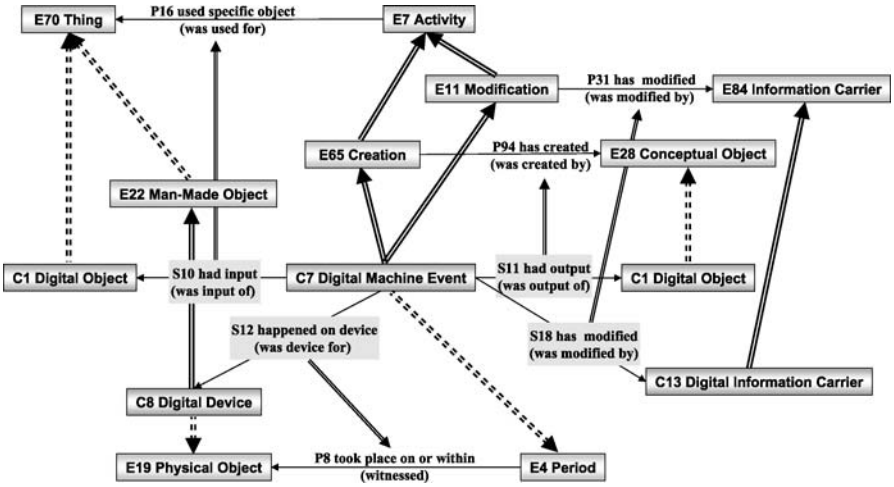


Fig. 9 Machine events. This diagram shows how **C7 Digital Machine Event** relates to **C8 Digital Device**. A **C7 Digital Machine Event** is defined as a subclass of **E65 Creation** and **E11 Modification**. The property ***S10 had input*** is defined as a specialization (in RDF it is called *subproperty*) of ***P16 used specific object (was used for)*** and points to the original digital object. The property ***S11 had output*** is defined as a specialization of ***P94 has created (created by)*** and points to the resulting digital object. The property ***S12 happened on device (was device for)*** which is defined as a specialization of ***P8 took place on or within (witnessed)*** is a pointer to the device used for the machine event. Finally, the property ***S18 has modified (was modified by)*** is a specialization of ***P31 has modified (was modified by)*** and links the **C7 Digital Machine Event** with the **C13 Digital Information Carrier** where the resulting **C1 Digital Object** is stored

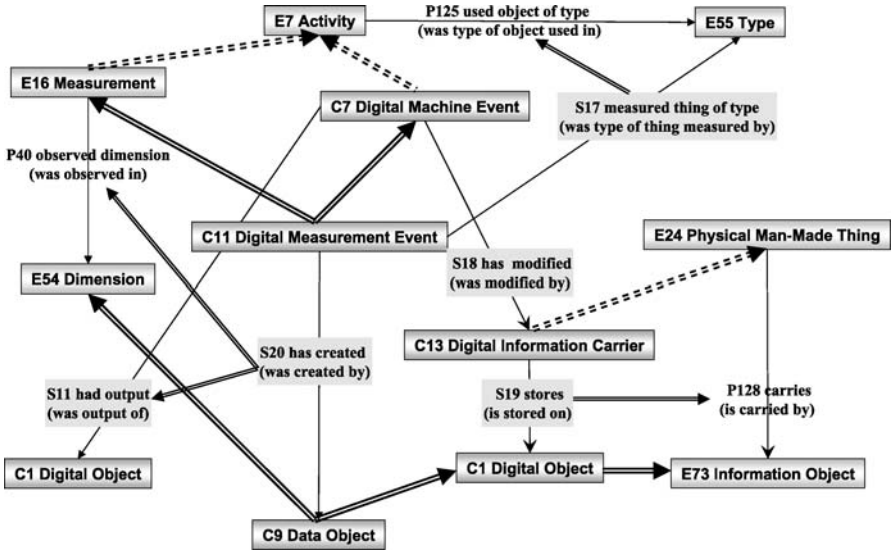


Fig. 10 Digital measurement event. This diagram shows the relationships between **C11 Digital Measurement Event**, **C1 Digital Object**, **C9 Data Object** and **C13 Digital Information Carrier**. **C11 Digital Measurement Event** is defined as a subclass of **C7 Digital Machine Event** and **E16 Measurement**

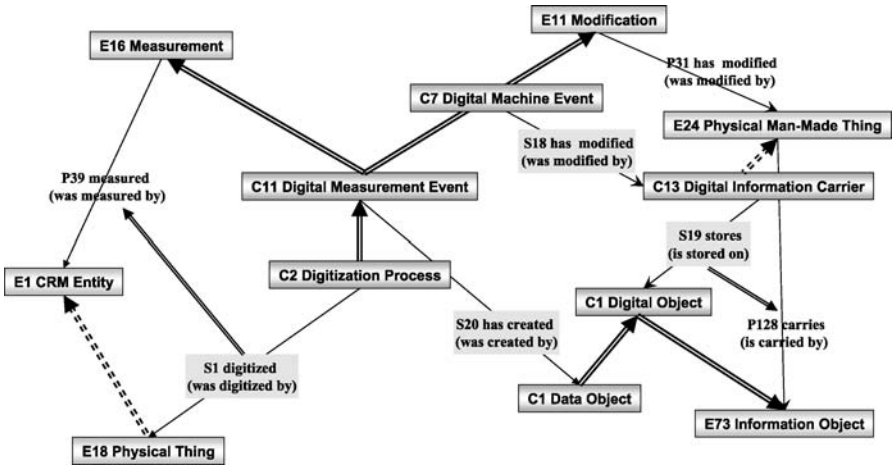


Fig. 11 Digitization process (from a material to an immaterial object). This diagram shows how **C2 Digitization Process**, **C1 Digital Object**, **C9 Data Object** and **C13 Digital Information Carrier** are related. **C2 Digitization Process** is related to class **E18 Physical Thing** through property ***S1 digitized*** which is a specialization of ***P39 measured (was measured by)***. The outcome of a **C2 Digitization Process** is a **C9 Data Object** which can be saved on a digital carrier. An instance of **C2 Digitization Process** represents the transition from an instance of a material thing (**E18 Physical Thing**) to an instance of an immaterial representation of it **C9 Data Object**

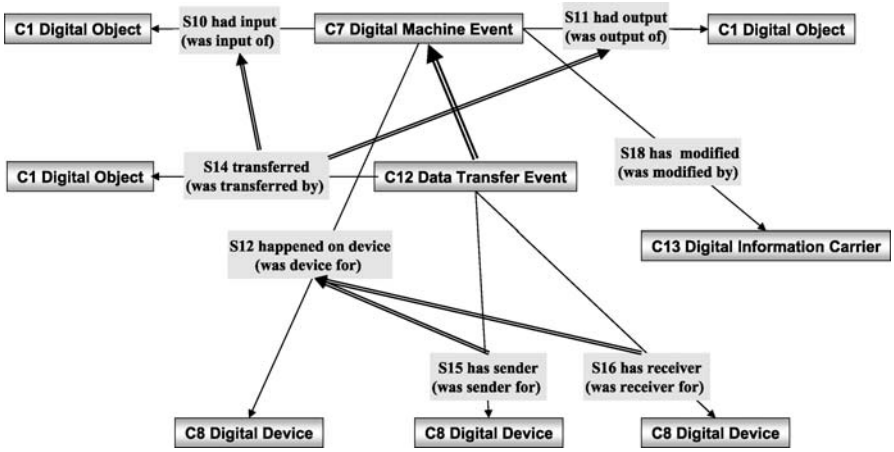


Fig. 12 Data transfer event. Description of the classes **C12 Data Transfer Event**, **C1 Digital Object** and **C8 Digital Device**. A **C12 Data Transfer Event** is related to class **C1 Digital Object** through property ***S14 transferred (was transferred by)*** which is a specialization of ***S10 had input (was input of)*** and ***S11 had output (was output of)***. The properties ***S15 has sender (was sender for)*** and ***S16 has receiver (was receiver for)*** which are specializations of ***S12 happened on device (was device for)*** relate the **C12 Data Transfer Event** with the respective **C8 Digital Devices**

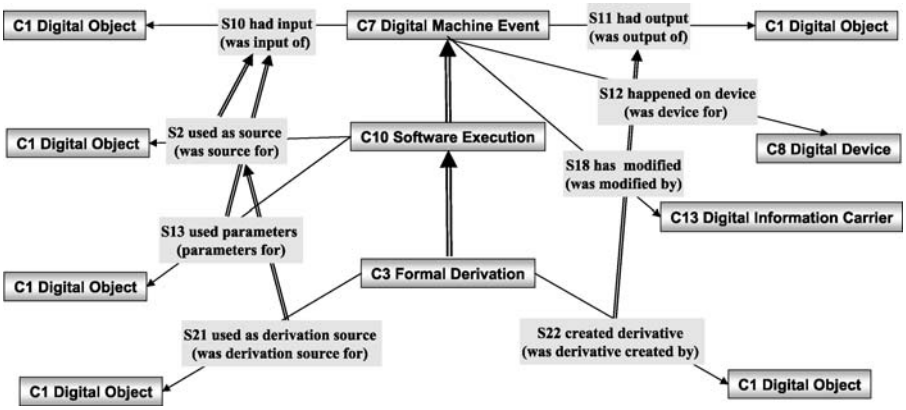


Fig. 13 Software execution, formal derivation, machine event. **C10 Software Execution** and **C3 Formal Derivation** are defined as a subclasses of **C7 Digital Machine Event**. Properties ***S2 used as source (was source for)*** and ***S13 used parameters (parameters for)*** which are defined as specializations of ***S10 had input (was input of)*** are pointers to the input object and the parameters of the **C10 Software Execution**. Properties ***S21 used as derivation source (was derivation source for)*** (which is a specialization of ***S2 used as source (was source for)***) and ***S22 created derivative (was derivative created by)*** which is a specialization of ***S11 had output (was output of)***) point to the input and output objects of the **C3 Formal Derivation** respectively. CIDOC CRM properties ***P32 used general technique (was technique of)*** and ***P33 used specific technique (was used by)*** are used to specify the method, algorithm, software etc. used by the software execution or the formal derivation. Property ***P3 has note*** between **C3 Formal Derivation** and **E62 String** is used to keep information regarding the property list used by the deterministic algorithm that the specific instance of **C3 Formal Derivation** used

Some of these templates are recursive. For instance, consider template number 5:

$$\{ \; E29\_Design\_or\_Procedure \rightarrow P94B\_was\_created\_by \rightarrow E65\_Creation \rightarrow P15F\_was\_influenced\_by \rightarrow E29\_Design\_or\_Procedure \; \} * \; repeat\ until\ P15F\_was\_influenced\_by\ is\ null$$

This query comprises an expression that takes as input an instance of E29 and returns another instance(s) of E29 (those influenced by) and this is continued recursively until there is no other P15F property that could be followed.

Figures 14 and 15 present the CRM_{dig} graphs for Table 2 Query templates 1 and 6 respectively.

5 Modeling provenance in various application domains

This section provides modeling examples from various applications domains. Section 5.1 contains examples from the cultural domain, while Sect. 5.2 gives examples of transformations (conversion and emulation).

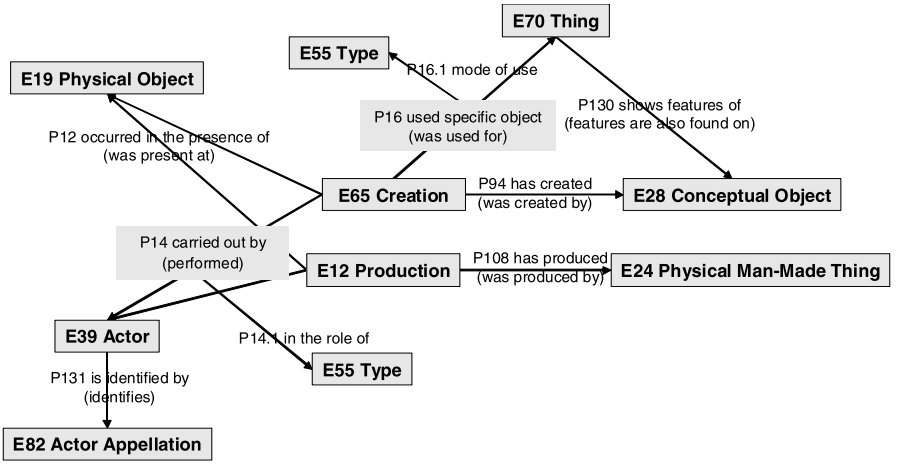


Fig. 14 Sample query 1—find creator/producer

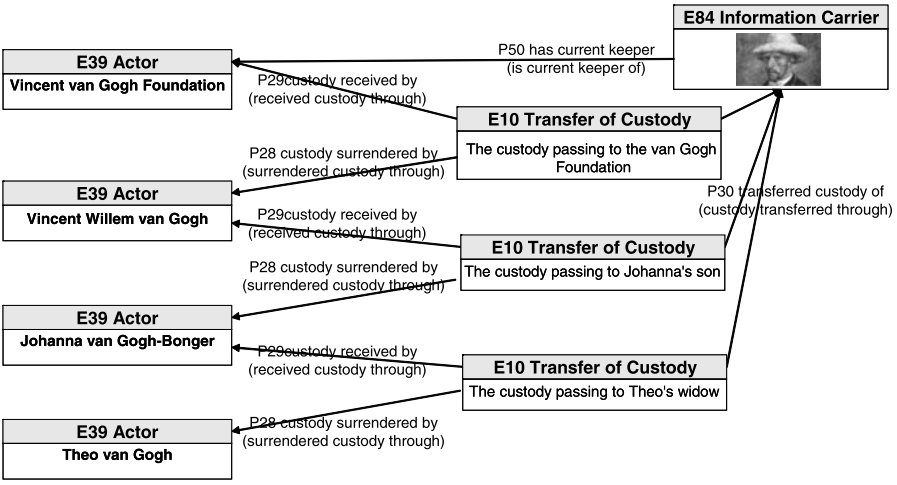


Fig. 15 Sample query 6—change of custody chain

5.1 Provenance in cultural domain

Let's start from the performing arts domain. "Avis de Tempete" is an opera by Georges Aperghis for ensamble and electronics, whose libretto is written by Georges Aperghis and Peter Szendy. Figure 16 shows how this information is modeled, i.e. how different actors can participate with different roles in a creation event, and some related provenance queries.

Philippe Manoury's *Jupiter*, is an opera for flute and live electronics, that was realized at IRCAM and first performed by Pierre-Andre Valade in April 1987. Figure 17

Table 2 Provenance query templates over $\texttt{CRM}_{dig}$

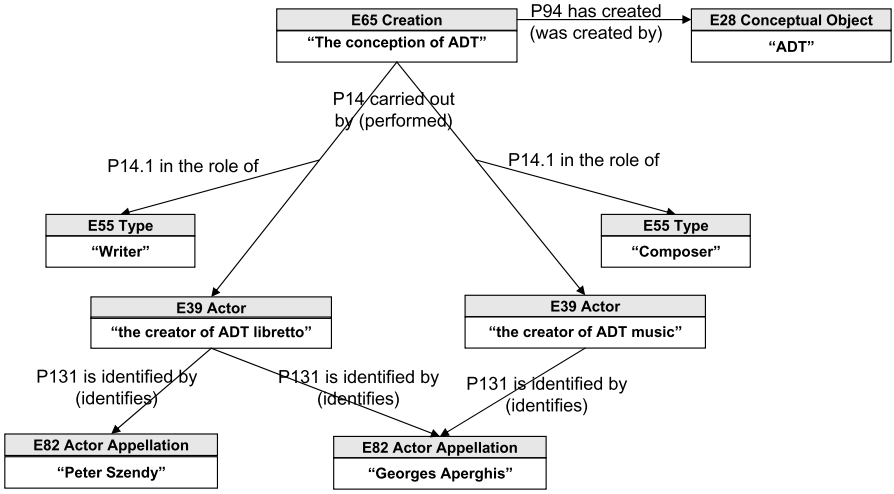
#	Description	Input	Output	Path
1	Get the Creator of a Digital Object	A Digital Object Instance of **E28Conceptual Object**	Instances of **E82 Actor Appellation**	$E28\_Conceptual\_Object \rightarrow P94B\_was\_created\_by \rightarrow E65\_Creation \rightarrow P14F\_carried\_out\_by(P14.1\_in\_the\_role\_of \rightarrow E55\_Type = Developer) \rightarrow E39\_Actor \rightarrow P131F\_is\_identified\_by \rightarrow E82\_Actor\_Appellation$
2	Get the Scanner used to capture a Digital Image	A Digital Image Instance of **C1 Digital Object**	Instance of **C8 Digital Device**	$C1\_Digital\_Object \rightarrow S11B\_was\_output\_of \rightarrow C7\_Digital\_Machine\_Event \rightarrow S12F\_happened\_on\_device \rightarrow C8\_Digital\_Device$
3	Get the Resolution of a Digital Object	A Digital Object Instance of **E73 Information Object** (Digital Image)	Instances of **E60 Number**	$E73\_Information\_Object \rightarrow P39B\_was\_measured\_by \rightarrow C2\_Digitization\_Process \rightarrow P40F\_observed\_dimension \rightarrow E54\_Dimension \rightarrow P90F\_has\_value \rightarrow E60\_Number$
4	Get the Master Version Of a Digital Object	A Digital Object Instance of **E73 Information Object** (Digital Image)	Instance of **E18 Physical Thing**	$E73\_Information\_Object \rightarrow P94B\_was\_created\_by \rightarrow C2\_Digitization\_Process \rightarrow S1F\_digitized \rightarrow E18\_Physical\_Thing$
5	Get Earlier Versions of a Digital Derivative	A Digital Derivative Instance of **E29 Design or Procedure**	List of Instances of **E29 Design or Procedure**	$\{E29\_Design\_or\_Procedure \rightarrow P94B\_was\_created\_by \rightarrow E65\_Creation \rightarrow P15F\_was\_influenced\_by \rightarrow E29\_Design\_or\_Procedure\}*$ *repeat until* $P15F\_was\_influenced\_by$ *is null*
6	Get the custody history of an Object	An Object Instance of **E84 Information Carrier**	List of Instances of **E82 Actor Appellation**	$E84\_Information\_Carrier \rightarrow P50F\_has\_current\_keeper \rightarrow \{E39\_Actor \rightarrow P29B\_received\_custody\_through \rightarrow E10\_Transfer\_of\_Custody \rightarrow P28F\_custody\_surrendered\_by \rightarrow E39\_Actor\}*$ *repeat until* $P29B\_received\_custody\_through$ *is null* $\rightarrow P131F\_is\_identified\_by \rightarrow E82\_Actor\_Appellation$

shows how subsequent performances of the same opera can be modeled and linked, and some related provenance queries.

5.2 Provenance and transformations

5.2.1 Conversion

Here we describe how we can model activities that result in the creation of a digital object from another one, following a deterministic algorithm, as Formal Derivations.



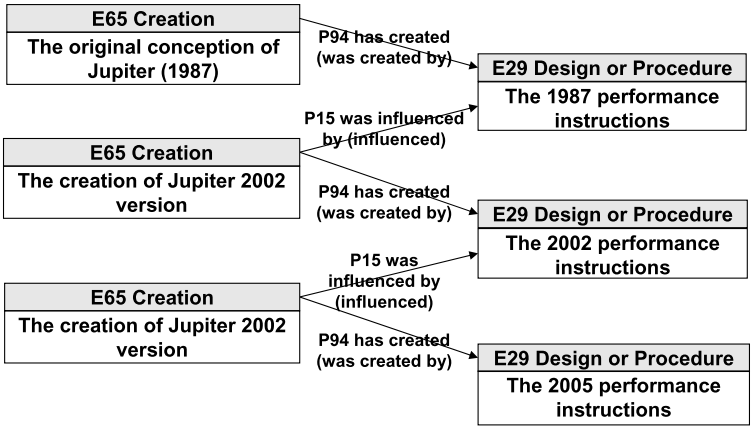
Who wrote the ADT libretto	$E28\_Conceptual\_Object(\text{“}ADT\text{”}) \rightarrow P94B\_was\_created\_by \rightarrow E65\_Creation \rightarrow P14F\_carried\_out\_by(\rightarrow P14.1\_in\_the\_role\_of \rightarrow E55\_Type = Writer) \rightarrow E39\_Actor \rightarrow P131F\_is\_identified\_by \rightarrow E82\_Actor\_Appellation$
Who wrote the ADT music	$E28\_Conceptual\_Object(\text{“}ADT\text{”}) \rightarrow P94B\_was\_created\_by \rightarrow E65\_Creation \rightarrow P14F\_carried\_out\_by(\rightarrow P14.1\_in\_the\_role\_of \rightarrow E55\_Type = Composer) \rightarrow E39\_Actor \rightarrow P131F\_is\_identified\_by \rightarrow E82\_Actor\_Appellation$

Fig. 16 Avis de Tempete: a provenance performing arts domain example

Formal Derivation represents the transition from an immaterial object to another immaterial object. The resulting instance of digital object shares representative properties with the original object and can be mechanically reproduced.

For instance, suppose we have a converter called **JPG2PNG** and consider three photographs **Crete.jpg**, **Crete.png** and **CreteSmall.png**. The latter two derived from the first photograph by using the converter. **CreteSmall.png** has lower resolution. Figure 18 and Figure 19 illustrates how the above scenario can be modeled using CRM_{dig}. In the first case the converter is used to produce **Crete.png** from **Crete.jpg** and then Adobe Photoshop application is used to reduce the resolution of **Crete.png** and produce **CreteSmall.png**. In the second case both **Crete.png** and **CreteSmall.png** are produced from **Crete.jpg** by using the converter with different parameters.

In both cases, the photographs are instances of **C1 Digital Object**. The class **E55 Type** is used to denote the format of each photograph (see Fig. 18) while classes **E54 Dimension** and **E60 Number** are used to model the resolution of each photograph. In general these classes may be used in order to model digital object parameters and their respective values.



Which is the first version of the Jupiter's composition	$\{E29\_Design\_or\_Procedure \rightarrow P94B\_was\_created\_by \rightarrow E65\_Creation \rightarrow P15F\_was\_influenced\_by \rightarrow E29\_Design\_or\_Procedure\} *$ *repeat until* $P15F\_was\_influenced\_by$ *is null*

Fig. 17 Manoury's Jupiter: a provenance performing arts domain example

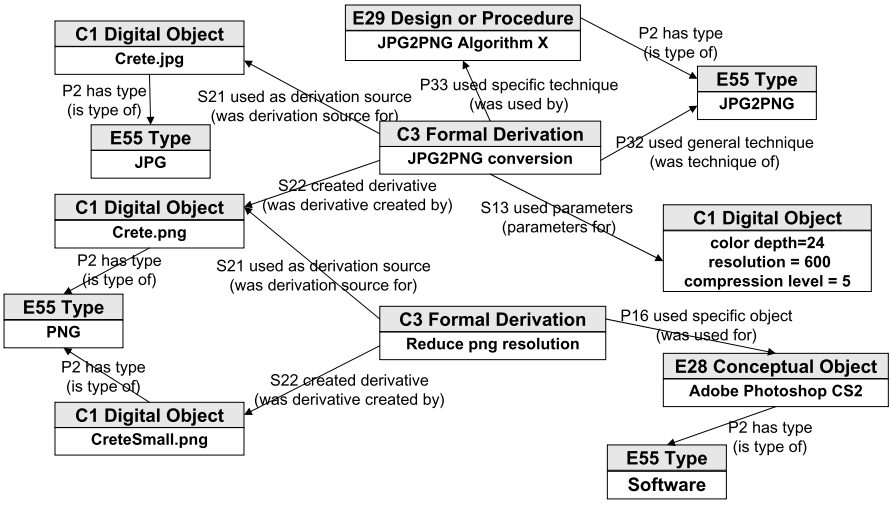


Fig. 18 JPG2PNG converter

The conversion event is an instance of **C3 Formal Derivation** which through the link ***P33 used specific technique (was used by)*** points to the specific algorithm used for the conversion (instance of class **E29 Design or Procedure**). In this example the parameters with which the converter was called are not modeled as separate entities but are implied in the name of the **E29 Design or Procedure** instance. Since **E29**

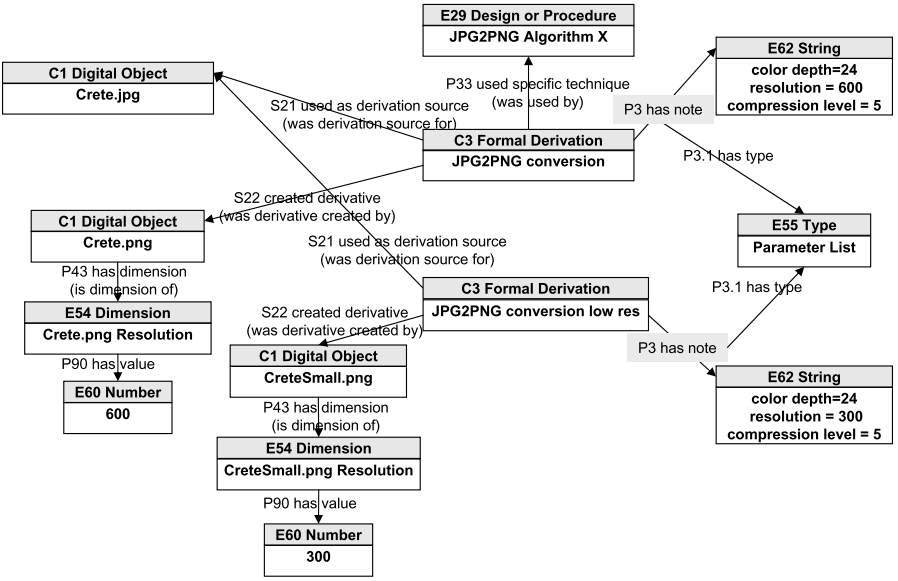


Fig. 19 JPG2PNG converter

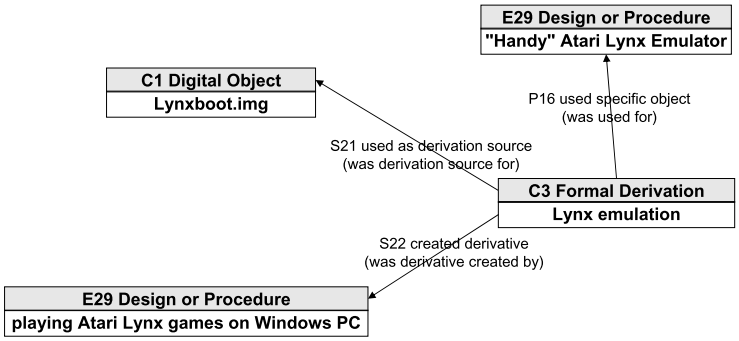


Fig. 20 Modeling emulation

Design or Procedure can be linked with other **E29 Design or Procedure** through ***P69 is associated with***, we can associate the specific call of a converter with the generic converter. **C3 Formal Derivation** is linked to **E55 Type** through ***P32 used general technique (was technique of)*** and denotes the generic type of the conversion. In Fig. 19 we can see how the different parameter list is modeled through the use of property ***P3 has note*** that points to an instance of class **E62 String**.

5.2.2 Emulation

Another example of formal Derivation is *emulation*. Figure 20 shows an example of modeling the Handy emulator. To play Atari Lynx games on a Windows-based PC by

using the Handy emulator the file `LYNXBOOT.IMG` is needed as well as Lynx game ROMs. A specific instance of the Handy emulator is an instance of class **E29 Design or Procedure**. The emulation activity is an instance of **C3 Formal Derivation** which has the property ***S2 used as source*** pointing to the file `LYNXBOOT.IMG` which is an instance of **C1 Digital Object**. The result of the emulation is an instance of **E29 Design or Procedure**. Issues regarding the principles of designing emulators or the reasoning on the properties of emulations, e.g. the Universal Virtual Computer (UVC) [24, 40], go beyond the scope of our work.

6 Implementation using semantic web technologies

There are several possible implementations. Here we describe a graph-based implementation where the ontology structure is directly reflected to the data model of the underlying repository. Specifically, in this section we describe how from an ontology (like CIDOC CRM, FRBRoo and $\texttt{CRM}_{dig}$) one can define a domain-specific schema (in the form of a Semantic Web ontology) and then use it for documenting the objects of interest. The major part of CIDOC CRM can be straightforwardly represented in Semantic Web languages and such artifacts are already available. However, CIDOC CRM Ontology has nine cases of attributes that start from other attributes (instead of starting from classes). This modeling construct is not straightforwardly supported by Semantic Web languages (and systems). However, all these nine attributes aim at capturing type information, therefore they could/should be expressed as elements in the domain specific schemas. This is clarified by the following example.

The left part of Fig. 21 shows a conceptual diagram illustrating a part of an ontology plus an instantiation of it. In particular, there is a property *ab* having domain the class **A** and range the class **B**. There is a property *d* whose domain is the property *ab*, having range the class **C**. The figure shows also an instantiation of this schema,

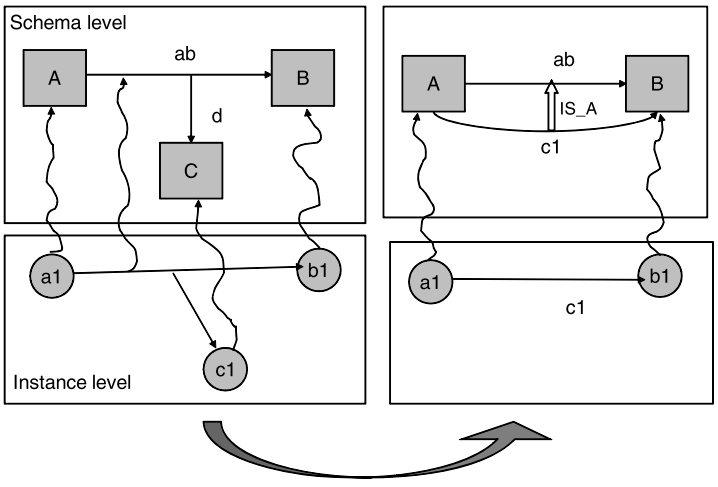


Fig. 21 Properties of properties and SW schemas

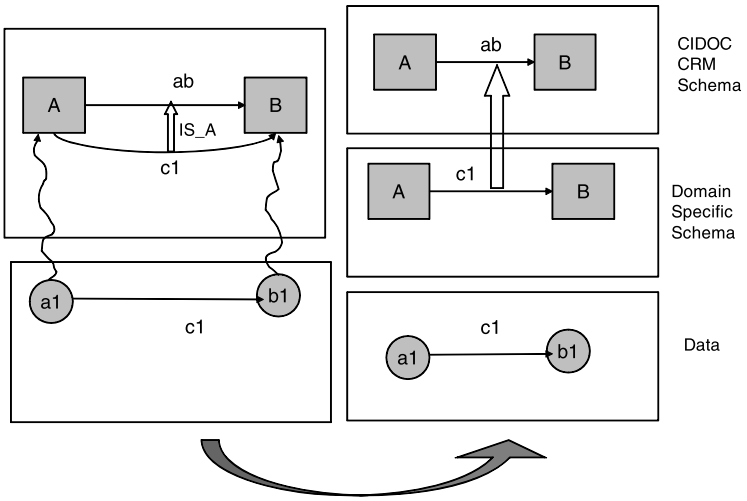


Fig. 22 Partitioning the knowledge into 3 knowledge artifacts

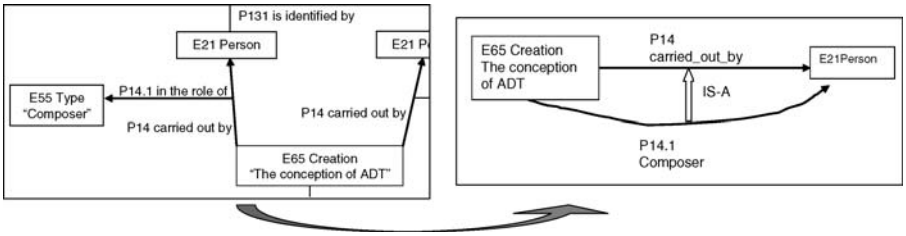


Fig. 23 Example of modeling composer

specifically $a1$ is an instance of class **A**, $b1$ is an instance of class **B**, and $c1$ is an instance of class **C**. The above logical structure should be implemented as the right side of Fig. 21. Specifically, we add at our schema (preferable at the domain specific schema) another property named *ci* which is defined as subproperty (i.e. specialization) of the *ab* property. The definition of $c1$ can be placed at the domain specific schema as in Fig. 22. For example, in the "Avis de tempete" instantiation of CRM-CIDOC we have the *P.14.1 property* "in_the_role_of". We model this by creating a new property named "Composer" as shown in Fig. 23.

The extension of CIDOC CRM for digital objects expressed in RDF/S [4] is given at Appendix A and electronically available at [15]. Moreover, most of the provenance modeling examples of this paper are publicly available in RDF.[1]

[1] http://www.casparpreserves.eu/publications/ontologies/swkmontologies

6.1 Provenance queries in RQL

The declarative languages that have been developed for the Semantic Web (e.g. RQL [20], SPARQL [2]) can be exploited for realizing provenance queries. The expression of some of the query templates in RQL is given and discussed in Table 3 (in Appendix C). The queries assume that there is a resource with identity `&myObject`. Queries for accessing objects based on provenance criteria are also possible and some examples of such templates are given in Table 3.

6.2 Application in CASPAR

The described approach has already been implemented in the context of the (ongoing) project CASPAR [1]. The implementation is over the SWKM (Semantic Web Knowledge Middleware)[2] whose repository, as well as its declarative query and update languages, are based on a relational DBMS (postgreSQL) using a database representation appropriate for RDF/S graphs [37]. The clients are distributed and access the repository through the Web Services provided by SWKM. The declarative update language supported (RUL [26]) is used for updating the repository (e.g. for adding/changing metadata). Apart form this (graph-based) implementation, $\texttt{CRM}_{dig}$ can be exploited for ontology-based integration [21] of relational sources. Figure 24 illustrates some of the CASPAR components. There are graphical components for alleviating the query formulation effort, at least for the provenance queries (e.g. FindingAids[3] is such a component that is based on SWKM, as well as SWKMQuery-

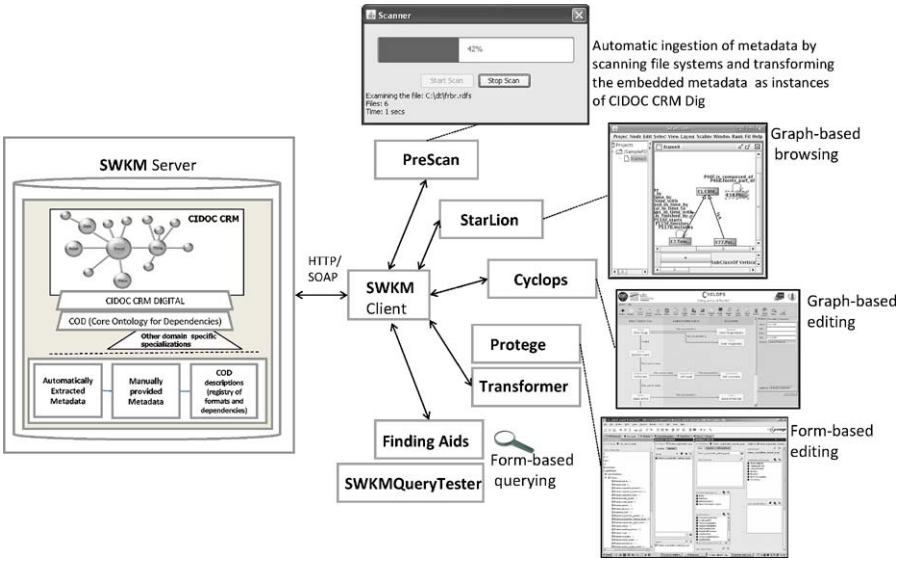


Fig. 24 Implementation in CASPAR

[2] http://athena.ics.forth.gr:9090/SWKM/

[3] http://developers.casparpreserves.eu:8080/CasparGui/

Tester[4]). Apart from issuing provenance queries, one could also explore provenance. For instance, the star-graph views of StarLion[5] can be used for this purpose, or one could adopt the interaction paradigm of *dynamic taxonomies and faceted search* [34] since applications of this interaction paradigm for knowledge bases expressed in RDF are feasible [3, 16, 30, 33]. One difficulty of applying $\texttt{CRM}_{dig}$ to the digital objects of the CASPAR project was the effort required for describing the provenance of existing digital objects. For this purpose we have developed guidelines and small tools that aid the extension of the schema and the creation of instances. For instance, `Transformer`[6] can aid the manual creation of instances using general-purpose ontology editors (like Protege), while `Cyclops`[7] is a specialized graphical editor allowing users to create descriptions according to $\texttt{CRM}_{dig}$. To further automate the ingestion process, we have developed `PreScan` [28],[8] a tool that scans entire file systems, extracts the embedded metadata from each file and transforms them to descriptions according to $\texttt{CRM}_{dig}$.

7 Related work

Below we compare $\texttt{CRM}_{dig}$ with OPM (Open Provenance Model) [31]. The ontology assumed by OPM is minimal. It comprises only 3 classes (*Artifact, Process, Agent*) and five associations among them (*used, wasGeneratedBy, wasControlledBy, wasTriggeredBy, wasDerivedFrom*). It follows that, from the perspective of representation adequacy, we can say that provenance information recorded according to $\texttt{CRM}_{dig}$ can be mapped to an OPM-based view, but not the other way around. In addition, the ontology assumed by OPM does not explicitly model the concept of *Event* a concept that is of prominent importance, not only because events allow tracing the history of an object but also because they enable the integration of several information that concern an object. Without the notion of event and also of physical objects that are carriers (devices) it is not possible for example, to describe adequately the conditions under which a photograph was taken. Nevertheless, we should say that the way OPM treats *Processes* resembles events (however the corresponding ontological structure of OPM is not rich).

In addition, OPM proposes a number of inference rules. Some of these are equivalent to the inferences due to the specialization relationships of CIDOC CRM extension. Some other could be expressed over the CIDOC CRM ontology by adopting an appropriate Rule Language. As an example, [29] describes an extension of the original CIDOC CRM for Interactive Multimedia Performances (IMP) enriched with temporal rules.

Finally, we would like to remark that the proposed approach is orthogonal with the languages that are currently used for digital preservation (such as EAST [36],

[4] Available at http://athena.ics.forth.gr:9090/SWKM/downloads/SWKMQueryTester.rar

[5] http://www.ics.forth.gr/~tzitzik/starlion

[6] Available for download from http://athena.ics.forth.gr:9090/SWKM/mainfiles/transformer.html

[7] http://www.utc.fr/caspar/cyclops_v1

[8] http://www.ics.forth.gr/prescan

DEDSL [22], XDFU [9, 25], SAFE [14]). For instance an archiving package in the SAFE format could contain a description of the provenance according to $\texttt{CRM}_{dig}$ expressed in a RDF/XML or RDF/Trig formatted file). The representation of $\texttt{CRM}_{dig}$ in Semantic Web (SW) languages offers standard formats for exchanging provenance data (i.e. RDF/XML, Trig) while the SW data management tools can be used for storing and declaratively querying/updating such repositories ([2, 20, 26]). Further inference requirements can be defined and exchanged through SWRL [23]. It is worth noting that there are already workflow systems that capture provenance metadata in RDF including Taverna [32], Triana [27] and GridNexus [5], as well as systems for exploring/visualizing provenance trails expressed in RDF (e.g. [17]).

8 Concluding remarks

In this paper we described an extension of the CIDOC CRM ontology (ISO 21127:2006) called $\texttt{CRM}_{dig}$ [15] able to capture the modeling and the query requirements regarding the provenance of digital objects. We discussed the relationship with OAIS (ISO 14721:2003) and provided a number of indicative modeling examples. Finally we described the presentation of this ontology in RDF(/S) and showed how the proposed provenance query templates can be implemented using Semantic Web query languages.

The completeness of the modeling abstractions of CIDOC CRM can be justified from (a) the way it was derived (by integrating hundreds metadata schemas), (b) the fact that is now an ISO standard, and (c) our experience in using it in real applications. The completeness of its extension (i.e. of $\texttt{CRM}_{dig}$) can be justified from our experience in using it for modeling data from CASPAR. Recall that CASPAR aims at preserving data from the cultural domain, the scientific domain and the artistic domain. The proposed model has higher general coverage and deeper specialization than OPM [31] for instance. Further testing is an open ended process for the future, and specialization of such a model is deliberately open ended, but starting with extraordinarily diverse test cases right from the beginning, we have a certain confidence in the completeness.

Acknowledgements This work was partially supported by the EU project CASPAR (FP6-2005-IST-033572). Many thanks to all "CASPARtners" for the fruitful discussions and examples that they provided us, and to Stephen Stead, Paveprime Ltd London, UK, who provided us with the first CIDOC CRM analysis of Digital Provenance data through his work on the photographic process of Cultural Heritage Imaging, San Francisco. This work is continuing in the context of the EU IST IP 3D-COFORM project.

Appendix A: CIDOC CRM extension ($\texttt{CRM}_{dig}$) in RDF/S

```
<?xml version="1.0"?>

<!DOCTYPE rdf:RDF [
    <!ENTITY CIDOC 'http://cidoc.ics.forth.gr/rdfs/caspar/
     cidoc.rdfs#'>
    <!ENTITY CIDOC_DIG 'http://cidoc.ics.forth.gr/rdfs/caspar/
```

```
     cidoc_digital2.3.rdfs#'>
    <!ENTITY rdfs 'http://www.w3.org/2000/01/rdf-schema#'>
    <!ENTITY rdf 'http://www.w3.org/1999/02/
     22-rdf-syntax-ns#'>
]>

<rdf:RDF xml:lang="en"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:CIDOC="http://cidoc.ics.forth.gr/rdfs/caspar/
          cidoc.rdfs#"
        xmlns:CIDOC_DIG="http://cidoc.ics.forth.gr/rdfs/caspar/
          cidoc_digital2.3.rdfs#">

<rdfs:Class rdf:ID="C1_Digital_Object">
    <rdfs:comment></rdfs:comment>
    <rdfs:subClassOf rdf:resource=
     "&CIDOC;E73_Information_Object"/>
</rdfs:Class>
<rdfs:Class rdf:ID="C2_Digitization_Process">
    <rdfs:comment></rdfs:comment>
    <rdfs:subClassOf rdf:resource=
     "#C11_Digital_Measurement_Event"/>
</rdfs:Class>
<rdfs:Class rdf:ID="C3_Formal_Derivation">
    <rdfs:comment></rdfs:comment>
    <rdfs:subClassOf rdf:resource="#C10_Software_Execution"/>
</rdfs:Class>
<rdfs:Class rdf:ID="C4_Norm">
    <rdfs:comment></rdfs:comment>
</rdfs:Class>
<rdfs:Class rdf:ID="C5_Copyright">
    <rdfs:comment></rdfs:comment>
</rdfs:Class>
<rdfs:Class rdf:ID="C6_Copying">
    <rdfs:comment></rdfs:comment>
</rdfs:Class>
<rdfs:Class rdf:ID="C7_Digital_Machine_Event">
    <rdfs:comment></rdfs:comment>
    <rdfs:subClassOf rdf:resource="&CIDOC;E11_Modification"/>
    <rdfs:subClassOf rdf:resource="&CIDOC;E65_Creation"/>
</rdfs:Class>
<rdfs:Class rdf:ID="C8_Digital_Device">
    <rdfs:comment></rdfs:comment>
    <rdfs:subClassOf rdf:resource=
     "&CIDOC;E22_Man-Made_Object"/>
</rdfs:Class>
<rdfs:Class rdf:ID="C9_Data_Object">
    <rdfs:comment></rdfs:comment>
    <rdfs:subClassOf rdf:resource="&CIDOC;E54_Dimension"/>
```

```
    <rdfs:subClassOf rdf:resource="#C1_Digital_Object"/>
</rdfs:Class>
<rdfs:Class rdf:ID="C10_Software_Execution">
    <rdfs:comment></rdfs:comment>
    <rdfs:subClassOf rdf:resource=
     "#C7_Digital_Machine_Event"/>
</rdfs:Class>
<rdfs:Class rdf:ID="C11_Digital_Measurement_Event">
    <rdfs:comment></rdfs:comment>
    <rdfs:subClassOf rdf:resource="&CIDOC;E16_Measurement"/>
    <rdfs:subClassOf rdf:resource=
     "#C7_Digital_Machine_Event"/>
</rdfs:Class>
<rdfs:Class rdf:ID="C12_Data_Transfer_Event">
    <rdfs:comment></rdfs:comment>
    <rdfs:subClassOf rdf:resource=
     "#C7_Digital_Machine_Event"/>
</rdfs:Class>
<rdfs:Class rdf:ID="C13_Digital_Information_Carrier">
    <rdfs:comment></rdfs:comment>
    <rdfs:subClassOf rdf:resource=
     "&CIDOC;E84_Information_Carrier"/>
</rdfs:Class>

<rdf:Property rdf:ID="S1F_digitized">
    <rdfs:domain rdf:resource="#C2_Digitization_Process"/>
    <rdfs:range rdf:resource="&CIDOC;E18_Physical_Thing"/>
    <rdfs:subPropertyOf rdf:resource="&CIDOC;P39F_measured"/>
</rdf:Property>
<rdf:Property rdf:ID="S1B_was_digitized_by">
    <rdfs:domain rdf:resource="&CIDOC;E18_Physical_Thing"/>
    <rdfs:range rdf:resource="#C2_Digitization_Process"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P39B_was_measured_by"/>
</rdf:Property>
<rdf:Property rdf:ID="S2F_used_as_source">
    <rdfs:domain rdf:resource="#C10_Software_Execution"/>
    <rdfs:range rdf:resource="#C1_Digital_Object"/>
    <rdfs:subPropertyOf rdf:resource="#S10F_had_input"/>
</rdf:Property>
<rdf:Property rdf:ID="S2B_was_source_for">
    <rdfs:domain rdf:resource="#C1_Digital_Object"/>
    <rdfs:range rdf:resource="#C10_Software_Execution"/>
    <rdfs:subPropertyOf rdf:resource="#S10B_was_input_of"/>
</rdf:Property>
<rdf:Property rdf:ID="S3F_allows">
    <rdfs:domain rdf:resource="&CIDOC;E30_Right"/>
    <rdfs:range rdf:resource="&CIDOC;E7_Activity"/>
</rdf:Property>
<rdf:Property rdf:ID="S3B_is_allowed_by">
```

```
    <rdfs:domain rdf:resource="&CIDOC;E7_Activity"/>
    <rdfs:range rdf:resource="&CIDOC;E30_Right"/>
</rdf:Property>
<rdf:Property rdf:ID="S4F_violates">
    <rdfs:domain rdf:resource="&CIDOC;E7_Activity"/>
    <rdfs:range rdf:resource="&CIDOC;E30_Right"/>
</rdf:Property>
<rdf:Property rdf:ID="S4B_is_violated_by">
    <rdfs:domain rdf:resource="&CIDOC;E30_Right"/>
    <rdfs:range rdf:resource="&CIDOC;E7_Activity"/>
</rdf:Property>
<rdf:Property rdf:ID="S5F_makes_use_of">
    <rdfs:domain rdf:resource="&CIDOC;E7_Activity"/>
    <rdfs:range rdf:resource="&CIDOC;E30_Right"/>
</rdf:Property>
<rdf:Property rdf:ID="S5B_is_used_by">
    <rdfs:domain rdf:resource="&CIDOC;E30_Right"/>
    <rdfs:range rdf:resource="&CIDOC;E7_Activity"/>
</rdf:Property>
<rdf:Property rdf:ID="S6F_holds">
    <rdfs:domain rdf:resource="&CIDOC;E39_Actor"/>
    <rdfs:range rdf:resource="&CIDOC;E30_Right"/>
</rdf:Property>
<rdf:Property rdf:ID="S6B_is_granted_to">
    <rdfs:domain rdf:resource="&CIDOC;E30_Right"/>
    <rdfs:range rdf:resource="&CIDOC;E39_Actor"/>
</rdf:Property>
<rdf:Property rdf:ID="S8F_copies_to">
    <rdfs:domain rdf:resource="#C6_Copying"/>
    <rdfs:range rdf:resource=
     "&CIDOC;E73_Information_Object"/>
</rdf:Property>
<rdf:Property rdf:ID="S8B_is_created_by">
    <rdfs:domain rdf:resource="&CIDOC;E73_Information_Object"/>
    <rdfs:range rdf:resource="#C6_Copying"/>
</rdf:Property>
<rdf:Property rdf:ID="S9F_has_validity">
    <rdfs:domain rdf:resource="#C4_Norm"/>
    <rdfs:range rdf:resource="&CIDOC;E52_Time-Span"/>
</rdf:Property>
<rdf:Property rdf:ID="S9B_is_validation_period_of">
    <rdfs:domain rdf:resource="&CIDOC;E52_Time-Span"/>
    <rdfs:range rdf:resource="#C4_Norm"/>
</rdf:Property>
<rdf:Property rdf:ID="S10F_had_input">
    <rdfs:domain rdf:resource="#C7_Digital_Machine_Event"/>
    <rdfs:range rdf:resource="#C1_Digital_Object"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P16F_used_specific_object"/>
</rdf:Property>
```

```
<rdf:Property rdf:ID="S10B_was_input_of">
    <rdfs:domain rdf:resource="#C1_Digital_Object"/>
    <rdfs:range rdf:resource="#C7_Digital_Machine_Event"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P16B_was_used_for"/>
</rdf:Property>
<rdf:Property rdf:ID="S11F_had_output">
    <rdfs:domain rdf:resource="#C7_Digital_Machine_Event"/>
    <rdfs:range rdf:resource="#C1_Digital_Object"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P94F_has_created"/>
</rdf:Property>
<rdf:Property rdf:ID="S11B_was_output_of">
    <rdfs:domain rdf:resource="#C1_Digital_Object"/>
    <rdfs:range rdf:resource="#C7_Digital_Machine_Event"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P94B_was_created_by"/>
</rdf:Property>
<rdf:Property rdf:ID="S12F_happened_on_device">
    <rdfs:domain rdf:resource="#C7_Digital_Machine_Event"/>
    <rdfs:range rdf:resource="#C8_Digital_Device"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P8F_took_place_on_or_within"/>
</rdf:Property>
<rdf:Property rdf:ID="S12B_was_device_for">
    <rdfs:domain rdf:resource="#C8_Digital_Device"/>
    <rdfs:range rdf:resource="#C7_Digital_Machine_Event"/>
    <rdfs:subPropertyOf rdf:resource="&CIDOC;P8B_witnessed"/>
</rdf:Property>
<rdf:Property rdf:ID="S13F_used_parameters">
    <rdfs:domain rdf:resource="#C10_Software_Execution"/>
    <rdfs:range rdf:resource="#C1_Digital_Object"/>
    <rdfs:subPropertyOf rdf:resource="#S10F_had_input"/>
</rdf:Property>
<rdf:Property rdf:ID="S13B_parameters_for">
    <rdfs:domain rdf:resource="#C1_Digital_Object"/>
    <rdfs:range rdf:resource="#C10_Software_Execution"/>
    <rdfs:subPropertyOf rdf:resource="#S10B_was_input_of"/>
</rdf:Property>
<rdf:Property rdf:ID="S14F_transferred">
    <rdfs:domain rdf:resource="#C12_Data_Transfer_Event"/>
    <rdfs:range rdf:resource="#C1_Digital_Object"/>
    <rdfs:subPropertyOf rdf:resource="#S10F_had_input"/>
    <rdfs:subPropertyOf rdf:resource="#S11F_had_output"/>
</rdf:Property>
<rdf:Property rdf:ID="S14B_was_transferred_by">
    <rdfs:domain rdf:resource="#C1_Digital_Object"/>
    <rdfs:range rdf:resource="#C12_Data_Transfer_Event"/>
    <rdfs:subPropertyOf rdf:resource="#S10B_was_input_of"/>
    <rdfs:subPropertyOf rdf:resource="#S11B_was_output_of"/>
```

```
</rdf:Property>
<rdf:Property rdf:ID="S15F_has_sender">
    <rdfs:domain rdf:resource="#C12_Data_Transfer_Event"/>
    <rdfs:range rdf:resource="#C8_Digital_Device"/>
    <rdfs:subPropertyOf rdf:resource=
     "#S12F_happened_on_device"/>
</rdf:Property>
<rdf:Property rdf:ID="S15B_was_sender_for">
    <rdfs:domain rdf:resource="#C8_Digital_Device"/>
    <rdfs:range rdf:resource="#C12_Data_Transfer_Event"/>
    <rdfs:subPropertyOf rdf:resource="#S12B_was_device_for"/>
</rdf:Property>
<rdf:Property rdf:ID="S16F_has_receiver">
    <rdfs:domain rdf:resource="#C12_Data_Transfer_Event"/>
    <rdfs:range rdf:resource="#C8_Digital_Device"/>
    <rdfs:subPropertyOf rdf:resource=
     "#S12F_happened_on_device"/>
</rdf:Property>
<rdf:Property rdf:ID="S16B_was_receiver_for">
    <rdfs:domain rdf:resource="#C8_Digital_Device"/>
    <rdfs:range rdf:resource="#C12_Data_Transfer_Event"/>
    <rdfs:subPropertyOf rdf:resource="#S12B_was_device_for"/>
</rdf:Property>
<rdf:Property rdf:ID="S17F_measured_thing_of_type">
    <rdfs:domain rdf:resource=
     "#C11_Digital_Measurement_Event"/>
    <rdfs:range rdf:resource="&CIDOC;E55_Type"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P125F_used_object_of_type"/>
</rdf:Property>
<rdf:Property rdf:ID="S17B_was_type_of_thing_measured_by">
    <rdfs:domain rdf:resource="&CIDOC;E55_Type"/>
    <rdfs:range rdf:resource=
     "#C11_Digital_Measurement_Event"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P125B_was_type_of_object_used_in"/>
</rdf:Property>
<rdf:Property rdf:ID="S18F_has_modified">
    <rdfs:domain rdf:resource="#C7_Digital_Machine_Event"/>
    <rdfs:range rdf:resource=
     "#C13_Digital_Information_Carrier"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P31F_has_modified"/>
</rdf:Property>
<rdf:Property rdf:ID="S18B_was_modified_by">
    <rdfs:domain rdf:resource=
     "#C13_Digital_Information_Carrier"/>
    <rdfs:range rdf:resource="#C7_Digital_Machine_Event"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P31B_was_modified_by"/>
```

```
</rdf:Property>
<rdf:Property rdf:ID="S19F_stores">
    <rdfs:domain rdf:resource=
     "#C13_Digital_Information_Carrier"/>
    <rdfs:range rdf:resource="#C1_Digital_Object"/>
    <rdfs:subPropertyOf rdf:resource="&CIDOC;P128F_carries"/>
</rdf:Property>
<rdf:Property rdf:ID="S19B_is_stored_on">
    <rdfs:domain rdf:resource="#C1_Digital_Object"/>
    <rdfs:range rdf:resource=
     "#C13_Digital_Information_Carrier"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P128B_is_carried_by"/>
</rdf:Property>
<rdf:Property rdf:ID="S20F_has_created">
    <rdfs:domain rdf:resource=
     "#C11_Digital_Measurement_Event"/>
    <rdfs:range rdf:resource="#C9_Data_Object"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P40F_observed_dimension"/>
    <rdfs:subPropertyOf rdf:resource="#S11F_had_output"/>
</rdf:Property>
<rdf:Property rdf:ID="S20B_was_created_by">
    <rdfs:domain rdf:resource="#C9_Data_Object"/>
    <rdfs:range rdf:resource=
     "#C11_Digital_Measurement_Event"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P40B_was_observed_in"/>
    <rdfs:subPropertyOf rdf:resource="#S11B_was_output_of"/>
</rdf:Property>
<rdf:Property rdf:ID="S21F_used_as_derivation_source">
    <rdfs:domain rdf:resource="#C3_Formal_Derivation"/>
    <rdfs:range rdf:resource="#C1_Digital_Object"/>
    <rdfs:subPropertyOf rdf:resource="#S2F_used_as_source"/>
</rdf:Property>
<rdf:Property rdf:ID="S21B_was_derivation_source_for">
    <rdfs:domain rdf:resource="#C1_Digital_Object"/>
    <rdfs:range rdf:resource="#C3_Formal_Derivation"/>
    <rdfs:subPropertyOf rdf:resource="#S2B_was_source_for"/>
</rdf:Property>
<rdf:Property rdf:ID="S22F_created_derivative">
    <rdfs:domain rdf:resource="#C3_Formal_Derivation"/>
    <rdfs:range rdf:resource="#C1_Digital_Object"/>
    <rdfs:subPropertyOf rdf:resource="#S11F_had_output"/>
</rdf:Property>
<rdf:Property rdf:ID="S22B_was_derivative_created_by">
    <rdfs:domain rdf:resource="#C1_Digital_Object"/>
    <rdfs:range rdf:resource="#C3_Formal_Derivation"/>
    <rdfs:subPropertyOf rdf:resource="#S11B_was_output_of"/>
</rdf:Property>
```

```
<rdf:Property rdf:ID="S16_1F_Scanner">
    <rdfs:domain rdf:resource="&CIDOC;E7_Activity"/>
    <rdfs:range rdf:resource="&CIDOC;E70_Thing"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P16F_used_specific_object"/>
</rdf:Property>
<rdf:Property rdf:ID="S14_1F_Developer">
    <rdfs:domain rdf:resource="&CIDOC;E7_Activity"/>
    <rdfs:range rdf:resource="&CIDOC;E39_Actor"/>
    <rdfs:subPropertyOf rdf:resource=
     "&CIDOC;P14F_carried_out_by"/>
</rdf:Property>
</rdf:RDF>
```

Appendix B: CIDOC CRM extensions

C1 Digital Object

SubClassOf: E73 Information Object

SuperClassOf: C9 Data Object

Scope note: This class comprises identifiable immaterial items that can be represented as sets of bit sequences, such as data sets, e-texts, images, audio or video items, software, etc., and are documented as single units. Any aggregation of instances of C1 Digital Object into a whole treated as single unit is also regarded as an instance of C1 Digital Object. This means that for instance, the content of a DVD, an XML file on it, and an element of this file, are regarded as distinct instances of C1 Digital Object, mutually related by the P106 is composed of (forms part of) property.
A C1 Digital Object does not depend on a specific physical carrier, and it can exist on one or more carriers simultaneously.

Examples

- image BM000038850.JPG from the Clayton Herbarium in London
- texas_flood_21-28june07_graph.gif

Properties:

C2 Digitization Process

SubClassOf: C11 Digital Measurement Event

Scope note: This class comprises events that result in the creation of instances of C9 Data Object that represent the appearance and/or form of an instance of E18 Physical Thing such as paper documents, statues, buildings, paintings, etc. A particular case is the analogue-to-digital conversion of audio-visual material
This class represents the transition from a material thing to an immaterial representation of it. The characteristic subsequent processing steps on digital objects are regarded as instances of C3 Formal Derivation.

Examples
- the scanning of the performance handbook of Avis de Tempête
- the digital photographing of van Gogh's self portrait
- the audio visual recording of the 17-11-2004 performance of ADT at the Opéra de Lille

Properties: S1 digitized (was digitized by): E18 Physical Thing

C3 Formal Derivation

SubClassOf: C10 Software Execution

Scope note: This class comprises events that result in the creation of a C1 Digital Object from another one following a deterministic algorithm, such that the resulting instance of digital object shares representative properties with the original object. In other words, this class describes the transition from an immaterial object referred to by property *S21 used as derivation source (was derivation source for)* to another immaterial object referred to by property *S22 created derivative (was derivative created by)* preserving the representation of some things but in a different form. Characteristic examples are colour corrections, contrast changes and resizing of images.

Examples
- the reduction of the resolution of image BM000038850.JPG from the Clayton Herbarium in London to 300dpi

Properties: S21 used as derivation source (was derivation source for): C1 Digital Object
S22 created derivative (was derivative created by): C1 Digital Object

C7 Digital Machine Event

SubClassOf: E65 Creation
E11 Modification

Scope note: This class comprises events that happen on physical digital devices following a human activity that intentionally caused its immediate or delayed initiation and results in the creation of a new instance of C1 Digital Object on behalf of the human actor. The input of a C7 Digital Machine Event may be parameter settings and/or data to be processed. Some C7 Digital Machine Events may form part of a wider E65 Creation event. In this case, all machine output of the partial events is regarded as creation of the overall activity.

Examples

- the scanning with ISL's EPSON of the performance handbook of Avis de Tempête
- the digital photographing of van Gogh's self portrait with OLYMPUS FE-3010
- resizing image BM000038850.JPG with Adobe Photoshop

Properties: S10 had input (was input of): C1 Digital Object
S11 had output (was output of): C1 Digital Object
S12 happened on device (was device for): C8 Digital Device
S18 has modified (was modified by): C13 Digital Information Carrier

C8 Digital Device

SubClassOf: E22 Man-Made Object

Scope note: This class comprises identifiable material items such as computers, scanners, cameras, etc. that have the capability to process or produce instances of C1 Digital Object.

Examples

- Doerr's Olympus FE-3010
- ISL's EPSON digital scanner

C9 Data Object

SubClassOf: C1 Digital Object
E54 Dimension

Scope note: This class comprises instances of C1 Digital Object that are the direct result of a digital measurement or a formal derivative of it, containing quantitative properties of some physical things or other constellations of matter.

Examples

- GOME RAW L0 data (EGOC format) of 24/07/07
- The monymusk reliquary OBJECT/Kestrel_3DH_KE_001c

C10 Software Execution

SubClassOf: C7 Digital Machine Event

SuperClassOf: C3 Formal Derivation

Scope note: This class comprises events by which a digital device runs a software program or a series of computing operations on a digital object as a single task, which is completely determined by its digital input, the software and the generic properties of the device.

Examples

- The GOME L1B → 1C Processing

Properties: S2 used as source (was source for): C1 Digital Object
S13 used parameters (parameters for): C1 Digital Object

C11 Digital Measurement Event

SubClassOf: E16 Measurement
C7 Digital Machine Event

SuperClassOf: C2 Digitization Process

Scope note: This class comprises actions measuring physical properties using a digital device, that are determined by a systematic procedure and creates an instance of C9 Data Object, which is stored on an instance of C13 Digital Information Carrier. In contrast to instances of C10 Software Execution, environmental factors have an intended influence on the outcome of an instance of C11 Digital Measurement Event. Measurement devices may include running distinct software, such as the RAW to JPEG conversion in digital cameras. In this case, the event is regarded as instance of both classes, C10 Software Execution and C11 Digital Measurement Event.

Examples

– the atmospheric ozone data capture with GOME on 27 July 2007

Properties: S17 measured thing of type (was type of thing measured by): E55 type
S20 has created (was created by)

C12 Data Transfer Event

SubClassOf: C7 Digital Machine Event

Scope note: This class comprises events that transfer a digital object from one digital carrier to another. Normally, the digital object remains the same. If in general or by observation the transfer implies or has implied some data corruption, the change of the digital objects may be documented distinguishing input and output rather than instantiating the property *S14 transferred (was transferred by).*

Examples

– the GOME raw satellite data (level 0) transmission from ERS-2 to the Kiruna Station on 27 July 2007

Properties: S14 transferred (was transferred by): C1 Digital Object
S15 has sender (was sender for): C8 Digital Device
S16 has receiver (was receiver for): C8 Digital Device

C13 Digital Information Carrier

SubClassOf: E84 Information Carrier

Scope note: This class comprises all instances of E84 Information Carrier that are explicitly designed to be used as persistent digital physical carriers of instances of C1 Digital Object. A C13 Digital Information Carrier may or may not contain information, e.g., an empty diskette.

Examples

– the computer disk at ICS-FORTH that stores the canonical Definition of the CIDOC CRM

Properties: S19 stores (is stored on): C1 Digital Object

Appendix C: Provenance queries in RQL

Table 3 Provenance query templates in RQL

#	Description	Query template in RQL
1	Get the Creator of a Digital Object	*Input:* A Digital Object Instance of E28_Conceptual_Object *Output:* Instances of E82_Actor_Appellation *Description:* $E28\_Conceptual\_Object \rightarrow P94B\_was\_created\_by \rightarrow E65\_Creation \rightarrow P14F\_carried\_out\_by(\rightarrow P14.1\_in\_the\_role\_of \rightarrow E55\_Type = Developer) \rightarrow E39\_Actor \rightarrow P131F\_is\_identified\_by \rightarrow E82\_Actor\_Appellation$

	RQL	`select X5 from`
		`{X1;E28_Conceptual_Object}P94B_was_created_by`
		`{X2;E65_Creation},`
		`{X2;E65_Creation}P14F_carried_out_by{X3;E39_Actor},`
		`{X2;E65_Creation}S14_1F_Developer{X3;E39_Actor},`
		`{X3;E39_Actor}P131F_is_identified_by`
		`{X5;E82_Actor_Appellation}`
		`where X1 like "*myObj"`
		Note: Instead of $S14\_1F\_Developer$ one could use one of the following:
		{PY1_composer, PY3_commissioner, PY2_writer }
		All of them are subproperties of P14F_carried_out_by (and have the same domain with the property P14_1F_Developer).

2	Get the Scanner used to capture a Digital Image	*Input:* A Digital Image Instance of C1_Digital_Object *Output:* Instances of C8_Digital_Device *Description:* $C1\_Digital\_Object \rightarrow S11B\_was\_output\_of \rightarrow C7\_Digital\_Machine\_Event \rightarrow S12F\_happened\_on\_device \rightarrow C8\_Digital\_Device$

	RQL	`select X3 from`
		`{X1;C1_Digital_Object}S11B_was_output_of`
		`{X2;C7_Digital_Machine_Event},`
		`{X2;C7_Digital_Machine_Event}S12F_happened_on_device`
		`{X3;C8_Digital_Device}`

3	Get the Resolution of a Digital Object	*Input:* A Digital Object Instance of E73_Information_Object (Digital Image) *Output:* Instances of E60_Number *Description:* $E73\_Information\_Object \rightarrow P39B\_was\_measured\_by \rightarrow C2\_Digitization\_Process \rightarrow P40F\_observed\_dimension \rightarrow E54\_Dimension \rightarrow P90F\_has\_value$

	RQL	`select X4 from`
		`{X1;E73_Information_Object}P39B_was_measured_by`
		`{X2;C2_Digitization_Process},`
		`{X2;C2_Digitization_Process}P40F_observed_dimension`
		`{X3;E54_Dimension},`
		`{X3;E54_Dimension}P90F_has_value{X4;Literal}`

Table 3 (*Continued*)

#	Description	Query template in RQL
4	Get the Master Version Of a Digital Object	*Input:* A Digital Object Instance of E73_Information_Object (Digital Image) *Output:* Instance of E18_Physical_Thing *Description:* $E73\_Information\_Object \rightarrow P94B\_was\_created\_by \rightarrow C2\_Digitization\_Process \rightarrow S1F\_digitized \rightarrow E18\_Physical\_Thing$

	RQL	`select X3 from`
		`{X1;E73_Information_Object}P94B_was_created_by`
		`{X2;C2_Digitization_Process},`
		`{X2;C2_Digitization_Process}S1F_digitized`
		`{X3;E18_Physical_Thing}`
		`where X1 like "*myObj"`

5	Get Earlier Versions of a Digital Derivative	*Input:* A Digital Derivative Instance of E29_Design_or_Procedure *Output:* List of Instances of E29_Design_or_Procedure *Description:* $\{E29\_Design\_or\_Procedure \rightarrow P94B\_was\_created\_by \rightarrow E65\_Creation \rightarrow P15F\_was\_influenced\_by \rightarrow E29\_Design\_or\_Procedure$ }* repeat until $P15F\_was\_influenced\_by$ is null

	RQL	`select X3 from`
		`{X1;E29_Design_or_Procedure}P94B_was_created_by`
		`{X2;E65_Creation},`
		`{X2;E65_Creation}P15F_was_influenced_by`
		`{X3;E29_Design_or_Procedure}`
		`where X1 like "*myObj"`
		Note:
		The above query returns the immediate earlier version(s)f &myobj. To get transitively all earlier version(s), we have to apply the same query again with only difference that instead of "where X1='&myobj"' we should write "where X1 In Z" where Z is the result of the previous query. To get all earlier versions we continue in this way until we get an empty result.

6A	Get the owner of an object	*Input:* A physical object *Output:* The actor that currently owns that thing

	RQL	`select X2 from`
		`{X1;E84_Information_Carrier}P50F_has_current_keeper`
		`{X2;E39_Actor}`
		`where X1 like "*myObj"`

6B	Get the previous owner of an object	*Input:* An actor &actor1 and a physical object &object1 *Output:* The actor that owned &object1 just before &actor1

	RQL	`select X4 from`
		`{X1;E84_Information_Carrier}P50F_has_current_keeper`
		`{X2;E39_Actor},`
		`{X2;E39_Actor}P29B_received_custody_through`
		`{X3;E10_Transfer_of_Custody},`
		`{X3;E10_Transfer_of_Custody}P28F_custody_surrendered_by`
		`{X4;E39_Actor}`
		`where X1 like "*object1" and X2 like "*actor1"`

Table 3 (*Continued*)

#	Description	Query template in RQL
7	Find all png images derived from a tool whose name is JPG2PNG	*Input:* Instance of C1_Digital_Object *Output:* Instance of C1_Digital_Object *Description:* *C1_Digital_Object* → *P94B_was_created_by* → *C3_formal_derivation* → *S2F_used_as_source* → *C1_Digital_Object* RQL `select X3 from {X1;C1_Digital_Object}P94B_was_created_by {X2;C3_Formal_Derivation}, {X2;C3_Formal_Derivation}S2F_used_as_source {X3;C1_Digital_Object} where X1 like "*JPG2PNG"`
8	Find all L1 data products created from DMS tool	*Input:* Instance of C1_Digital_Object *Output:* Instance of C1_Digital_Object *Description:* *C1_Digital_Object* → *P94B_was_created_by* → *C3_formal_derivation* RQL `select X1 from {X1;C1_Digital_Object}P94B_was_created_by {X2;C3_Formal_Derivation} where X1 like "*L1" and x2 like "*DMS_tool"`

References

1. CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval), FP6-2005-IST-033572. http://www.casparpreserves.eu/
2. SPARQL Query Language for RDF. W3C Candidate Recommendation, 6 April 2006. http://www.w3.org/TR/rdf-sparql-query/
3. Allard, P., Ferre', S.: Dynamic taxonomies for the semantic web. In: Proccedings of FIND'2008 (at DEXA'08), Turin, Italy, September 2008
4. Brickley, D., Guha, R.V.: Resource description framework (RDF) schema specification: proposed recommendation. W3C, March 1999. http://www.w3.org/TR/1999/PR-rdf-schema-19990303
5. Brown, J.L., Ferner, C.S., Hudson, T.C., Stapleton, A.E., Vetter, R.J., Carland, T., Martin, A., Martin, J., Rawls, A., Shipman, W.J., et al.: Gridnexus: a grid services scientific workflow system. Int. J. Comput. Inf. Sci. **6**(2), 77–82 (2005)
6. Buneman, P., Khanna, S., Tajima, K., Tan, W.C.: Archiving scientific data. ACM Trans. Database Syst. **29**(1), 2–42 (2004)
7. Buneman, P., Tan, W.C.: Provenance in databases. In: Proccedings of the 2007 ACM SIGMOD International Conference on Management of Data, pp. 1171–1173. ACM, New York (2007)
8. Carroll, J.J., Bizer, C., Hayes, P., Stickler, P.: Named graphs, provenance and trust. In: Proccedings of the 14th International Conference on World Wide Web, pp. 613–622. ACM, New York (2005)
9. XFDU development site. http://sindbad.gsfc.nasa.gov/xfdu
10. Doerr, M., Crofts, N.: Electronic communication on diverse data—the role of an object-oriented CIDOC reference model. In: Proccedings of CIDOC'98, Melbourne, October 1998. http://www.ics.forth.gr/proj/isst/Publications/Conference_Proc.html
11. Eltabakh, M.Y., Aref, W.G., Elmagarmid, A.K., Ouzzani, M., Laura-Silva, Y.: Supporting annotations on relations. In: 12th International Conference on Extending Database Technology (EDBT 2009), Saint-Petersburg, Russia, March 2009
12. Factor, M., Henis, E., Naor, D., Rabinovici-Cohen, S., Reshef, P., Ronen, S., Michetti, G., Guercio, M.: Authenticity and provenance in long term digital preservation: modeling and implementation

in preservation aware storage. In: Proccedings of the USENIX First Workshop on the Theory and Practice of Provenance (TaPP), San Francisco, USA, February 2009
13. Flouris, G., Fundulaki, I., Pediaditis, P., Theoharis, Y., Christophides, V.: Coloring rdf triples to capture provenance. In: Proccedings of the 8th International Semantic Web Conference (ISWC'09), October 2009
14. SAFE (Standard Archive Format for Europe). http://earth.esa.int/safe/
15. FORTH-ICS/ISL. The CIDOC conceptual reference model for digital objects (2008). http://cidoc.ics.forth.gr/rdfs/caspar/cidoc_digital2.3.rdfs
16. Hildebrand, M., van Ossenbruggen, J., Hardman, L.: /facet: a browser for heterogeneous semantic web repositories. In: Lecture Notes in Computer Science, vol. 4273, p. 272. Springer, Berlin (2006)
17. Hunter, J., Cheung, K.: Provenance explorer-a graphical interface for constructing scientific publication packages from provenance trails. Int. J. Digit. Libr. **7**(1), 99–107 (2007)
18. International organization for standardization: OAIS: open archival information system—reference model (2003). Ref. No. ISO 14721:2003
19. International organization for standardization: The CIDOC conceptual reference model (2006). Ref. No. ISO 21127:2006. http://cidoc.ics.forth.gr/
20. Karvounarakis, G., Christophides, V., Plexousakis, D.: RQL: a declarative query language for RDF. In: Eleventh International World Wide Web Conference (WWW), Hawaii, USA, May 2002
21. Kondylakis, H., Analyti, A., Plexousakis, D.: Quete: ontology-based query system for distributed sources. In: Advances in Databases and Information Systems (ADBIS 2007), pp. 359–375. Springer 2007
22. DEDSL Language (Data Entity Dictionary Specification Language). http://east.cnes.fr/english/page_dedsl.html
23. SWRL (Semantic Web Rule Language). http://www.w3.org/submission/swrl/ (2004)
24. Lorie, R.A.: Long term preservation of digital information. In: Proccedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 346–352 (2001)
25. Lucas, A.: XFDU packaging contribution to an implementation of the OAIS reference model. In: Proccedings of the International Conference PV'2007 (Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data), Edinburgh, November 2005
26. Magiridou, M., Sahtouris, S., Christophides, V., Koubarakis, M.: RUL: a declarative update language for RDF. In: Proccedings of the 4th International Conference on the Semantic Web (ISWC-2005), Galway, Ireland, November 2005
27. Majithia, S., Shields, M.S., Taylor, I.J., Wang, I.: Triana: a graphical web service composition and execution toolkit. In: Proccedings of the IEEE International Conference on Web Services (ICWS'04), San Diego, California, USA, July 2004
28. Marketakis, Y., Tzanakis, M., Tzitzikas, Y.: PreScan: towards automating the preservation of digital objects. In: Proccedings of the International Conference on Management of Emergent Digital Ecosystems (MEDES'09), Lyon, France, October 2009
29. Mikroyannidis, A., Bee, O., Ng, K., Giaretta, D.: Ontology-based temporal modelling of provenance information. In: Proccedings of Electrotechnical Conference, MELECON 2008. The 14th IEEE Mediterranean, Tenerife, Spain, May 2008, pp. 176–181
30. Mäkelä, E., Hyvönen, E., Saarela, S.: Ontogator—a semantic view-based search engine service for web applications. In: International Semantic Web Conference. Lecture Notes in Computer Science, vol. 4273. Springer, Berlin (2006)
31. Moreau, L., Freire, J., Myers, J., Futrelle, J., Paulson, P.: The open provenance model. University of Southampton (2007)
32. Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A., et al.: Taverna: a tool for the composition and enactment of bioinformatics workflows (2004)
33. Oren, E., Delbru, R., Decker, S.: Extending faceted navigation for RDF data. In: Lecture Notes in Computer Science, vol. 4273, p. 559. Springer, Berlin (2006)
34. Sacco, G.M., Tzitzikas, Y. (eds.): Dynamic Taxonomies and Faceted Search: Theory, Practise and Experience. Springer, Berlin (2009)
35. Srivastava, D., Velegrakis, Y.: Using queries to associate metadata with data. In ICDE, pp. 1451–1453 (2007)
36. EAST Language (Enhanced Ada Subse T). http://east.cnes.fr/english/page_east.html
37. Theoharis, Y., Christophides, V., Karvounarakis, G.G.: Benchmarking database representations of RDF/S stores. In: Proccedings of the 4th International Semantic Web Conference (ISWC'05). Springer, Berlin (2005)

38. Tzitzikas, Y., Kotzinos, D., Theoharis, Y.: On ranking RDF schema elements (and its application in visualization). J. Univers. Comput. Sci. **13**(12), 1854–1880 (2007)
39. Tzitzikas, Y., Theoharis, Y., Andreou, D.: On storage policies for semantic web repositories that support versioning. In: Proccedings of the 5th European Semantic Web Conference (ESWC'08), Tenerife, Spain, June 2008, pp. 705–719. Springer, Berlin (2008)
40. van der Hoeven, J.R., van Diessen, R.J., van der Meer, K.: Development of a universal virtual computer (UVC) for long-term preservation of digital objects. J. Inf. Sci. **31**(3), 196 (2005)
41. Watkins, E.R., Nicole, D.A.: Named graphs as a mechanism for reasoning about provenance. In: Proccedings of the 8th Asia-Pacific Web Conf. (APWeb'2006), Harbin, China, pp. 943–948 (2006)