A CIDOC CRM – compatible metadata model for digital preservation

Panos Constantopoulos and Vicky Dritsou[1]
Information Systems and Databases Laboratory
Athens University of Economics and Business
Patission 76, 10434 Athens, Greece
Tel.: +30-210-8203157, Fax: +30-210-8203134
Email: {panosc | vdritsou}@aueb.gr

Abstract

A number of long-term digital preservation strategies exist, each one appropriate under different circumstances. No matter what strategy is followed, though, the preservation of digital resources requires applying the appropriate metadata elements. Several proposals exist for this purpose, each one indicating a set of the essential metadata. In our work we have studied five such widely known proposals. From this study, we identify a set of elements we consider most important, on the basis of which we propose a "common denominator" metadata model for digital preservation. Furthermore this is taken beyond the level of a mere list of metadata elements by establishing semantic relations between those. In the domain of cultural information, digital resources can be surrogates of non-digital cultural objects and they can also be considered as cultural objects themselves. In the context of a wider effort to ensure interoperability among cultural digital resources we formulate our preservation metadata model so as to be compatible with CIDOC CRM, the standard cultural documentation ontology.

Introduction

Digital assets face two types of perils: physical and technical. Physical perils include various damages of storage media and catastrophic environmental incidents, e.g. fire, flooding, earthquake, etc. Preservation policies to safeguard against such perils include copying and distributing copies in different locations. These are multi-parameter policies with the details of which we are not concerned here. Technical perils include the various kinds of difficulty or inability to access and use data due to the technical evolution of hardware and software. Preservation policies against technical perils employ techniques that fall in nine main classes: migration of digital content, technology emulation, technology preservation, dedication to standards, backward compatibility, encapsulation, permanent identifiers, transformation to non-digital form and digital archaeology [3, 6].

The implementation of preservation policies invariably requires certain information about the digital assets, captured by preservation metadata. In this work we formulate a conceptual model for digital preservation metadata, which (a) abstracts from certain established preservation metadata set, (b) explicitly displays the underlying semantic relations, and (c) is compatible with CIDOC CRM [4], the ISO standard ontology for cultural documentation. The latter is motivated by the fact that, in the cultural domain, preserved digital assets may be surrogates of non-digital objects and/or be cultural objects in their own right.

[1] Vicky Dritsou was in part supported for this work by the General Secretariat for Research and Technology through grant PENED 2003.

Digital preservation metadata sets

Several metadata sets for digital preservation have been proposed. In our work we focus on five well established ones. These are the metadata sets defined by: the Dublin Core [8], the Open Archival Information System (OAIS) [10], the Curl Exemplars Digital Archives (CEDARS) [5], the Pittsburgh Project [1] and the National Library of Australia [9]. Here we only point at certain features of these metadata sets. Interested readers can find a detailed account in [7].

The Dublin Core Metadata Element Set is well documented and easy to apply, comprising only fifteen elements. Being primarily intended for supporting Web-based information access, this metadata set does not convey all the information required by preservation processes. The OAIS reference model includes four basic metadata categories: content information, preservation information, packaging information and descriptive information. All elements are hierarchically organized under these and their total exceeds one hundred different elements including many details. Although the model is documented very well, we believe that this big number of metadata introduces complexity in its application, compounded by the lack of specific data entry instructions. A much similar approach is proposed by the CEDARS project, also featuring high detail and difficulty of application, especially when collections of objects, rather than single files, are concerned. The significance of those detailed metadata sets lies in that they contain, though indiscriminately, all potentially useful elements. The Pittsburgh Project proposal also contains a total of more than a hundred elements, yet it makes a clear distinction between necessary and optional elements and provides clear use instructions. The National Library of Australia proposal contains twenty five basic elements, some of which are further analyzed into sub-elements. What makes this proposal remarkable is the fact that its creators state clearly the essential sub-elements of each object, depending on its type. They analyze six object types: picture, sound, video, text, database, and executives. This is clearly an advantage, yet again no specific guidance for entering the essential data is provided.

A conceptual model for preservation metadata

Based on our study of the above metadata sets, we identify a set of elements we consider most significant, on the basis of which we propose a parsimonious metadata model for digital preservation. This model explicitly shows the semantic relations between the metadata elements, a feature we consider important for supporting the application of tools and processes. The model also derives from CIDOC CRM so as to enable interoperability with other cultural documentation data and applications. It contains a minimal set of concepts appropriately inter-related and specialized, namely: digital object, digital content, complex object, object identifier, rights, actor, size, title, subject, type, format, language, information carrier, equipment, activity, effect, history. The model is shown in Figure 1, where the name of each entity and property is followed by the code of the corresponding CIDOC CRM item which it is a subclass of. Due to space limitations, here we only briefly present the main elements of the model.

The central concept of the model is Digital Object. This comprises both the content to be preserved, represented by Digital Content, and the relevant metadata, represented by the attributes Identifier, Title, Subject, Size, Natural Language, Type, Format, and Information Carrier. Digital objects can consist of other objects, e.g. an html file containing both text and images. This situation is represented by Complex Object, a subclass of Digital Object. Each digital object admits certain Activities being performed upon it by certain Actors on condition they hold the appropriate Rights. The following Activity Types are distinguished: Creation, Deletion, Modification, Alteration, Read, Copy and Security Enforcement.

The precedence relation between activities is recorded by the attribute Previous. The set of activities that have affected a given digital object is addressed by History. The particular changes that a modification, alteration or security enforcement activity has brought to a digital object are represented by Effect. This element is intended to explicitly record object changes, whether restoration to the previous condition is possible or not. Furthermore the use of Effect is independent of the possible tracking of versions, the latter actually remaining out of the scope of the preservation model. Finally, all activities require Technical Equipment in order to be carried out, Hardware and Software. Technical Equipment is also associated with the Type of digital object.

The preservation conceptual model can be considered as a proposed application ontology derived from CIDOC CRM. The majority of concepts in this model have been defined as specializations of CIDOC CRM concepts. In addition certain independent extensions have been necessary. The has effect property of Activity with range Effect is one presented above. Another is the is formatted in whereby Format is related with the Type (text, image, …) of a digital object.

Conclusion

We have presented a conceptual model for digital preservation metadata, which draws elements from established metadata sets and complies with CIDOC CRM, thus it could be considered as an application ontology within the cultural domain.

We contend it would be interesting to examine in further detail the processes involved in digital preservation and explore the possible differences in modelling requirements that may arise from the fact that these are actually decision and production processes explicitly documented for operational purposes, whereas the historical processes accounted for in the ex post documentation of acquired cultural objects are approached interpretatively.

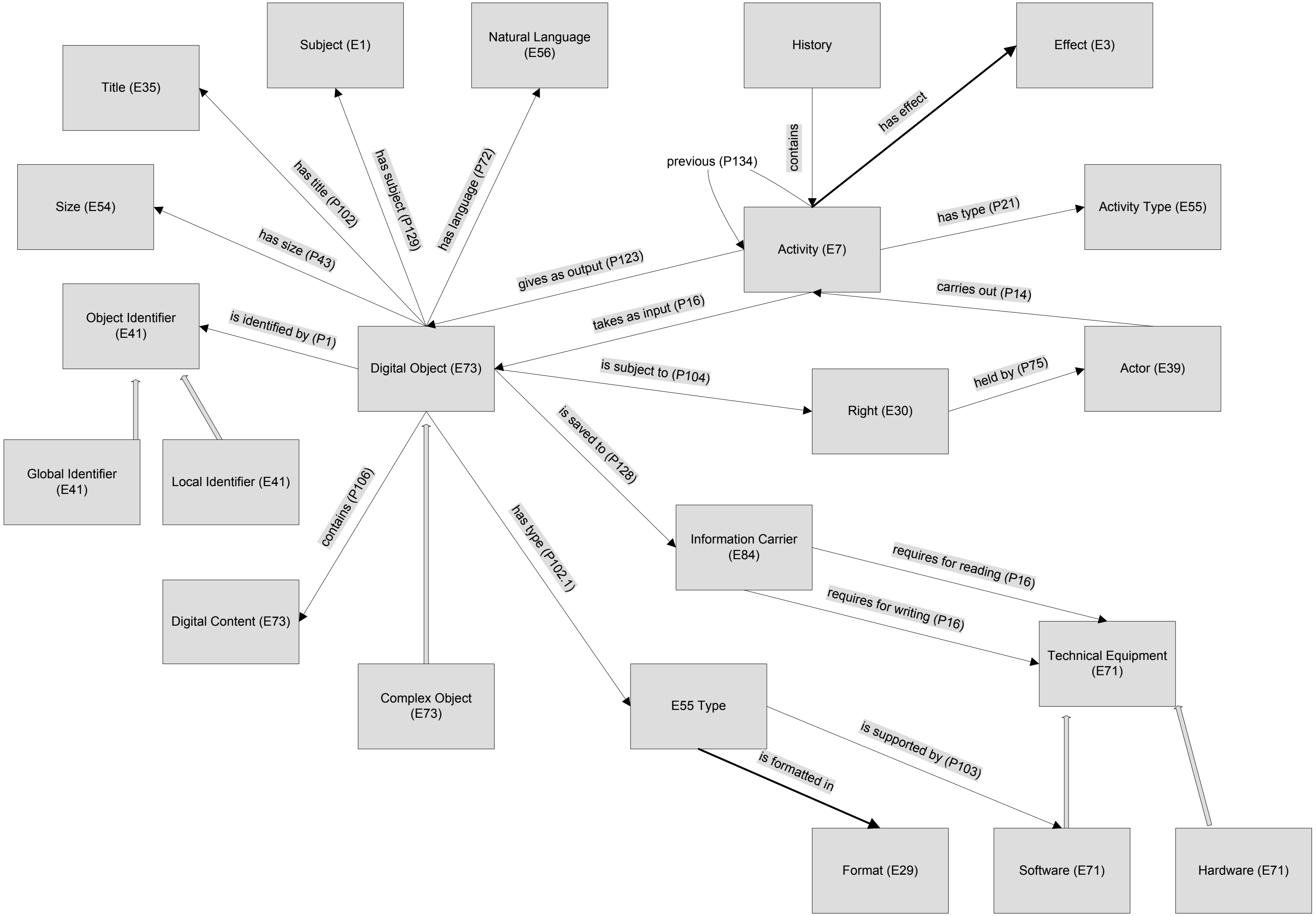

Figure 1: Conceptual model for digital preservation

References

[1] Bearman, D., Sochats, K., “Metadata Requirements for Evidence”, Archives and Museums Informatics, Functional Requirements for Evidence in Recordkeeping: The Pittsburgh Project. Retrieved from: http://www.archimuse.com/papers/nhprc/BACartic.html

[2] Bekiari, C., Constantopoulos, P., Doerr, M. (eds.), “Cultural Documentation and Interoperability Guide” (in Greek). Retrieved from: www.ics.forth.gr/CULTUREstandards

[3] Cornell University Library, Department of Preservation and Collection Maintenance, Digital Preservation Tutorials. “Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems”. Retrieved from: http://www.library.cornell.edu/iris/tutorial/dpm/index.html

[4] Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M., (eds), Definition of the CIDOC Conceptual Reference Model, June 2005. Retrieved from: http://cidoc.ics.forth.gr/official_release_cidoc.html

[5] Day, M. “Metadata for Preservation”. CEDARS Project Document AIW01, Aug. 1998. Retrieved from: www.ukoln.ac.uk/metadata/cedars/AIW01.htm

[6] Digital Preservation Coalition, “Digital Preservation Coalition Handbook”. Retrieved from: www.dpconline.org/graphics/handbook

[7] Dritsou, V., “Digital Content Preservation Metadata” , M.Sc. Thesis, Department of Informatics, Athens University of Economics and Business, Dec. 2004.

[8] Dublin Core Metadata Initiative (DCMI) Usage Board, “DCMI Metadata Terms”, DCMI Recommendation, Sept. 2004. Retrieved from: dublincore.org/documents/2004/09/20/dcmi-terms/

[9] National Library of Australia, “Preservation Metadata for Digital Collections”, Oct. 1999. Retrieved from: www.nla.gov.au/preserve/pmeta.html

[10] Online Computer Library Center (OCLC) / Research Libraries Group (RLG) Working Group on Preservation Metadata, “A metadata framework to support the preservation of digital objects”, June 2002. Retrieved from: www.oclc.org/research/projects/pmwg/pm_framework.pdf