TOTEM: Trusted Online Technical Environment Metadata - A Long-Term Solution for a Relational Database / RDF Ontologies

Dr Janet Delve
Future Proof Computing Group
School of Creative Technologies
University of Portsmouth
004423 9284 5524
Janet.Delve@port.ac.uk

Dr Leo Konstantelos
Future Proof Computing Group
School of Creative Technologies
University of Portsmouth
004423 9284 5491
Leo.Konstantelos@port.ac.uk

Dr Antonio Ciuffreda
Dr David Anderson
Future Proof Computing Group
School of Creative Technologies
University of Portsmouth
004423 9284 5491
Antonio.Ciuffreda@port.ac.uk

ABSTRACT

For emulation and other preservation actions, metadata is needed to describe the technical environment (operating system, related software libraries, hardware etc.) in which a given file or item of software can be rendered. This paper delineates an enhanced entity attribute relationship model suitable as a basis for a database (relational, object-relational or object-oriented), or for a RDF ontology. This core data model covers the X86, Apple II and Commodore 64 (C64) hardware architectures, as well as games consoles. The model is currently instantiated as a MySQL database with accompanying API, plus a PHP-based browsing system. Data population, and user evaluation are also discussed. The model is extensible over the long term and is compatible with OAIS and PREMIS version 2.

1. INTRODUCTION

A state-of-the art survey [2] for the KEEP project examined in depth the existing technical environment metadata, and found there to be some preparatory work on which to build, but no extant, completed data models / schemas available explicitly for this purpose. PREMIS 2 confirmed this finding, and provided guidelines for technical environment metadata in either database or ontology format [5]. An Enhanced Entity Attribute Relationship conceptual model was thus chosen for KEEP as it provided a basis for either format.

2. DATA MODELS

The core version of the TOTEM Enhanced Entity Relationship Diagram (EERD) is generic, and was created via a bottom-up approach using catalogue data for a PDF file, a multimedia encyclopedia and a console game. The catalogue data held some technical environment information including an initial range of hardware such as 'a multimedia PC' and 'an Apple II'. Three of the publications were on media carriers: CD-ROMs, a 5 ½" floppy disk and a games cartridge. See [3] for full details and EERDs.

A particular feature of this data model is its granularity. For a PDF file, it is vital to know its version, as specific software is required to render a specific PDF version. So a PDF version 1.4 file would run on Adobe Acrobat version 5.x software on a Mac OS version X 10.5.6 (9G55) Operating System (OS) on a MacbookPro 5.1 hardware platform. In contrast, PDF/X files incorporate a Graphics Art Technologies standard and need a software package such as CorelDRAW 11 to render them. Another vital feature of many technical environments is software libraries (for example .DLL files) that the OS might need. Note that further extensions of the EERD will embrace: software patches; system libraries, plug-ins, fonts; plugged hardware devices with their corresponding drivers (together with driver versions) and BIOS revision. This generic model covers a variety of PC architecture together with the software and operating systems that run on them. Computer games running on a PC such as the C64 can be modeled here, but the model for the Games Console in is different in structure to a typical computer, given that the computer chip resides in the console OS. Much thought was given to modeling the plethora of appendages used with games consoles (i.e. joysticks). Concerning the games controllers, generic attributes were defined for a particular version of a game running on a particular games console as being either analogue or digital, and these are then further refined in terms of planes and degrees of movement etc.

Having created these conceptual data models, research was carried out to establish the best way of implementing them. A method embracing linked data was identified in the initial survey [2] and Dublin Core[1] as a good way of achieving semantic interoperability which is achieved through the use of a common RDF format which facilitates the gathering of information via linked data clouds [1]. RDF ontologies would thus appear to be an ideal vehicle for the KEEP technical environment metadata. However, the Planets project[2] had carried out extensive XCL characterization work that included the development of a raft of ontologies [4], but it was pointed out that OWL Protégé had some shortcomings for the general user, and Excel spreadsheets were used instead to house these ontologies:

"A solution may be for non-OWL experts to develop class structures by hierarchically organising relevant concepts in a spreadsheet, and having an OWL expert or software developer

iPRES2011, Nov. 1–4, 2011, Singapore.

[1] http://dublincore.org/metadata-basics/

[2] http://www.planets-project.eu/

develop a script for transforming this spreadsheet into the RDF/OWL language. Such a procedure has been followed manually (not involving scripts) in developing the initial PC ontology, which proved to be a lot faster than building it as an OWL ontology in Protégé, because of the large numbers of classes and individuals involved. This may be an efficient way of developing ontologies in future within the digital preservation community." Collaboration with the University of Cologne (Universität zu Köln) is underway to convert the EERDs into RDFs for the software and hardware classes [8].

3. THE TOTEM DATABASE

The conceptual models outlined above have formed the basis for the specification of the TOTEM database Twenty five entities have been modeled, comprising more than 130 elements and their relationships, in a fully normalized structure. Three distinct technical environments are currently supported in the logical data model: the PC architecture, the Commodore 64 architecture and console gaming platforms. It is possible however to represent additional environments (e.g. Apple II or Acorn) in the future, by following the specifications in the conceptual models. The physical model was developed as a MySQL database, accessible as part of KEEP emulation services as well as to general DP users.

Three distinct user roles have been identified: end users, metadata data administrators and metadata database administrators. Database administration is managed via the phpMyAdmin[3] open source tool, currently deployed over an Apache web server. Interaction between the database and end users / data administrators is made via a database application that acts as a front-end and browsing system.

The browsing system, implemented in PHP, allows access to browsing and searching the TOTEM database through a simple interface currently providing three types of search functionality: simple search; advanced search; and compatibility search. In the compatibility search, the user can explore: Software types compatible with a specified file type and version; Software libraries compatible with a specified software type and version; Operating systems compatible with a specified software type and version; and Hardware types compatible with a specified operating system and version.

The greatest challenge **(is)** the process of identifying, populating and maintaining the resource with accurate, pertinent and up-to-date data. TOTEM currently holds PC-related technical metadata, Commodore 64- and Console Game-related metadata. The sheer volume of data – alongside the process of corroborating their accuracy and hence usefulness – clearly indicates that this task cannot be single-handedly undertaken by a sole institution or a sole project. Long-term sustainability of the TOTEM resource and continuing adoption of the model and deriving schema necessitate equally continuing support from user communities. Having identified these caveats, a number of potential solutions are being explored, including collection of data from product documentation and developer blogs, and Crowdsourcing data population by making the database accessible to relevant communities.

The resource will be integrated with the suite of tools provided under the aegis of the Open Planets Foundation (OPF)[4] registry ecosystem. [6] sets out the vision for a new registry for digital preservation, or a "registry ecosystem", which will build on linked data in order to create an interconnected collection of existing (and future) information registries that currently exist in isolation. Although this can be a sustainable solution, the risk of erroneous/contradictory information being inserted, with varying degrees of detail and granularity **still exists**. The OPF registry ecosystem envisages countering this problem by promoting the Crowdsourcing path and by introducing tools that allow institutions to set their own confidence levels on representation information in registries [7]. To conclude, comprehensive user evaluation for TOTEM is almost complete and feedback received so far indicates that this is a useful weapon in the DP armory. Detailed plans for future improvements are already mapped out to make this a robust, versatile, scalable and shareable tool.

[3] http://www.phpmyadmin.net/

[4] http://www.openplanetsfoundation.org/

4. ACKNOWLEDGMENTS

The Keeping Emulation Environments Portable (KEEP) Project is co-financed by the European Union's Seventh Framework Programme for research and technological development (FP7), Grant Agreement number ICT-231954.

5. REFERENCES

[1] Anderson, D., Delve, J., and Pinchbeck, D. 2010. Toward A Workable Emulation-Based Preservation Strategy: Rationale and Technical Metadata. *The New Review of Information Networking* 15, 2 (Nov. 2010), 110-131. DOI=http://dx.doi.org/10.1080/13614576.2010.530132.

[2] Anderson, D., Delve, J., Pinchbeck, D., and Alemu, G. A. 2009. *Preliminary document analyzing and summarizing metadata standards and issues across Europe*. KEEP Technical Report D3.1. URL= http://www.keep-project.eu/ezpub2/index.php?/eng/content/download/4124/20617/file/KEEP_WP3_D3.1.pdf

[3] Delve, J., Ciuffreda, A., and Anderson, D. 2010. *Documents describing meta-data for the specified range of digital objects, as well as requirements and design for the browsing system and user interface of the Emulation Framework.* KEEP Technical Report D3.2.

[4] Montague, L., Nicchiarelli, E., Matthezing, H., Kummer, R., Puhl, J., & Roberts, B. (2010a). *Planets components for the extraction and evaluation of digital object properties*. Planets Technical Report D23B. URL=http:// www.planets-project.eu/docs/reports/Planets_PC3-D23B(DOPWGreport).pdf

[5] PREMIS Working Group, OCLC, & RLG. 2008. *PREMIS data dictionary for preservation metadata Version 2.0.* Washington, DC: Library of Congress. URL=http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

[6] Roberts, B. 2011. *A New Registry for Digital Preservation: Conceptual Overview*. Open Planets Foundation Technical Report version 1.1. URL= http://www.openplanetsfoundation.org/new-registry-digital-preservation-conceptual-overview

[7] Tarrant, D., Hitchcock, S. and Carr, L. 2009. Where the Semantic Web and Web 2.0 meet format risk management: P2 registry. In *Proceedings of the 6th International Conference on Preservation of Digital Objects* (San Fransisco, CA, October 5-6, 2009). CDL, San Fransisco, CA, 187-193. DOI= http://escholarship.org/uc/item/8525r8cn Trust levels are discussed on p17

[8] Thaller, M. 2009. The eXtensible Characterisation Languages - XCL. Verlag Dr. Kovac, Hamburg.