View other issues

Contents

Automating the Ingestion and Transformation of Embedded Metadata

by Yannis Tzitzikas and Yannis Marketakis

Can we create automatically and at no cost ontology-based metadata repositories? Work carried out in the context of the CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval) project has been tackling this challenge.

The majority of preservation approaches rely on metadata. However the creation and maintenance of metadata is a laborious task that does not pay off immediately. For this reason there is a need for tools that automate as much as possible the ingestion and management of metadata. Our objective is to bypass the strict (often manual) ingestion process while at the same remaining compatible with it. According to the traditional approach, the ingestion phase starts with assigning identifiers to the objects and then extracting or creating metadata for these objects. These metadata can be expressed using various metadata schemas and formats, and usually the updating or movement of metadata records is prohibited. We want to relax these constraints and automate the process of metadata extraction and ingestion. Automation is crucial for the preservation of emergent systems and structures, like file systems, which are much more complex and dynamic than traditional digital archives.

In general, metadata can be stored either internally, ie in the same file with the data, or externally, ie data and metadata are stored in separate places or systems. The former are called embedded and the latter detached. PreservationScanner (PreScan for short) is a tool that we have developed for automating the extraction, transformation and maintenance of embedded metadata. PreScan is quite similar in spirit to the crawlers of Web search engines (WSE). For the problem at hand, we have to scan file systems, extract the embedded metadata from files of various types and build a metadata repository. In contrast to WSE crawlers, we have to (a) support more advanced extraction services, (b) allow the manual enrichment of metadata, (c) use more expressive frameworks for representing metadata (ie SW languages), (d) associate the extracted metadata with other sources of knowledge (eg registries of format types), and (e) offer rescanning services that do not start from scratch but exploit the previous status of the repository in order to preserve the human-provided metadata.

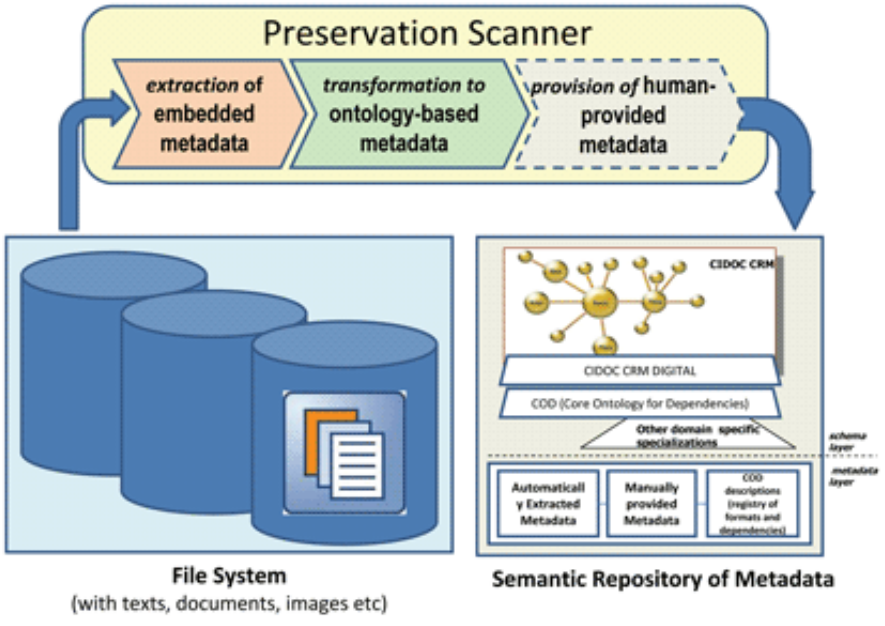

Figure 1: Creating automatically ontology-based metadata repositories.

PreScan starts like an antivirus program by scanning files from a specific folder and continues transitively to its subfolders. For every file it encounters it identifies the file format and extracts the embedded metadata. PreScan has a modular design and can work with several format identifiers and metadata extractors. Currently it uses JHOVE (JSTOR/Harvard Object Validation Environment) for this purpose. PreScan is capable of transforming the extracted metadata into ontological metadata expressed in RDF. Currently the extracted metadata are transformed to descriptions according to CIDOC CRM (CIDOC Conceptual Reference Model) Digital ontology. The user has the option to enrich the metadata of a file by providing additional information.

Periodic scans are supported too. Here PreScan identifies the files that have been renamed or moved to another location by comparing (through hash functions) the contents of files that have vanished (files that existed at the previous scan but are not there now) with new files encountered during the current scan. It suggests these matches to the user who in turn approves the correct ones (this is critical for preserving the human-provided metadata of files that have changed location). PreScan currently recognizes and extracts the embedded metadata from twelve file types (from which we get around 150 attributes in total), and it takes around ten hours to scan, extract and transform the metadata of a hundred thousand files.

Regarding the metadata repository, several (not mutually exclusive) options are supported: (a) all metadata records are stored in a folder specified by the user, (b) each metadata record is stored in the same folder as the scanned file, and (c) the contents of the metadata records are stored in a semantic Web knowledge base (specifically at SWKM (Semantic Web Knowledge Middleware)). The latter choice allows these metadata to be linked with other sources of knowledge (eg from registries). Furthermore, it offers declarative query and update services, which are important for building obsolescence risk detection services, notification services, and services relating to the intelligibility of digital objects.

This work has been done in the context of the CASPAR (Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval) project.

Links:
PreScan: http://www.ics.forth.gr/prescan
CASPAR: http://www.casparpreserves.eu/
Related Publication:
Y. Marketakis, M. Tzanakis and Y. Tzitzikas, "PreScan: Towards Automating the Preservation of Digital Objects", ACM Conference on Management of Emergent Digital EcoSystems, MEDES'2009, Lyon, France, Oct. 2009:

Please contact:
Yannis Tzitzikas
FORTH-ICS, Greece
E-mail: tzitzik@ics.forth.gr

ERCIM News 80

January 2010
Special theme:
Digital Preservation

This issue in pdf
(64 pages; 15 Mb)

Next issue
January 2016
Next special theme:
Life Science
Call for the next issue

Get the latest issue to your desktop