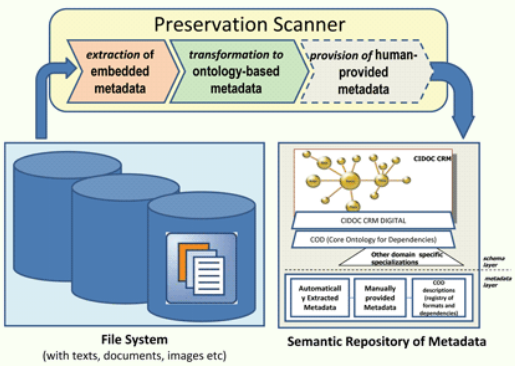
PreScan: A tool offering automatic ingestion and transformation of embedded metadata

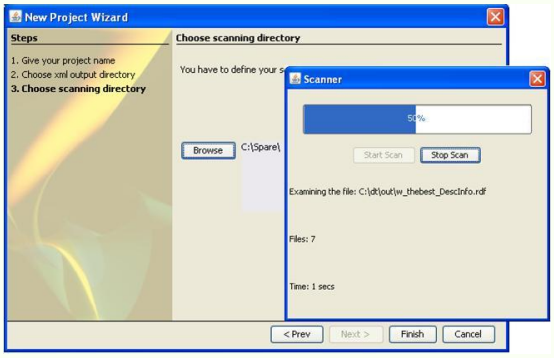
Overview

Digital objects are described in terms of metadata and nowadays there are several approaches that exploit them for different reasons; to assist their discovery, annotation, digital preservation, etc. However the creation and maintenance of metadata is a laborious task that does not pay off immediately. So there is a need for tools that automate as much as possible the ingestion and management of metadata.

PreScan (abbreviation of Preservation Scanner) is a tool that allows the automation of the extraction, the transformation and the maintenance of the embedded metadata of digital objects. It is quite similar in spirit to the crawlers of the web search engines. In contrast to web search engine crawlers, PreScan: (a) focuses on file systems, (b) support more advanced extraction services, (c) allow the manual enrichment of metadata, (d) use more expressive frameworks for representing metadata (i.e. Semantic Web languages), (e) associate the extracted metadata with other sources of knowledge and (f) offer rescanning services that do not start from scratch, but they exploit the previous scans.



Creating automatically ontology-based metadata repositories



Scanning and extracting metadata from the file system

Target Domains

PreScan can be used by digital archives and digital libraries to help archivists in extracting the embedded metadata from large collections objects. It can be used as a tool which is responsible for constructing and maintaining registries of digital objects.

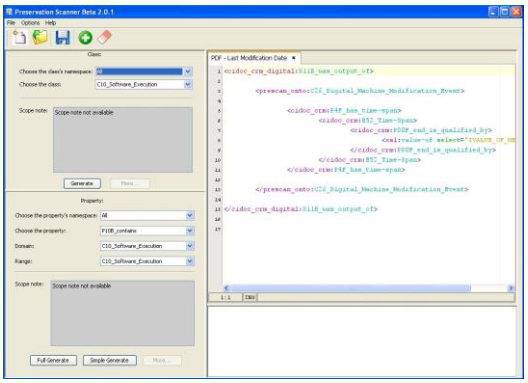
PreScan

www.ics.forth.gr/isl/PreScan/

Description

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). 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 Digital ontology). The user has the option to enrich the metadata of a file by providing additional information.

PreScan supports also periodic scans; it 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.



Updating the metadata representations (in RDF)

Files	Time			
	Exract	MD5	Store	Total
10	3 sec	3 sec	1 sec	9 sec
10^2	27 sec	46 sec	2 sec	77 sec
10^3	11 min	9 min	16 sec	19 min
10^4	95 min	44 min	7 min	145 min
10^5	8.5 hr	48 min	38 min	10 hr

Time performance of PreScan

Additional Information

PreScan was partially supported by the FP6 CASPAR Project and was exploited in FP7 Idea Garden project.

More information is available at the website of PreScan:
http://www.ics.forth.gr/isl/PreScan

PreScan website

Contact details: **Yannis Tzitzikas**
tzitzik@ics.forth.gr
www.ics.forth.gr/isl