A Methodology for File Relationship Discovery

M. Ondrejcek, J. Kastner, R. Kooper and P. Bajcsy

National Center for Supercomputing Applications of the University of Illinois at Urbana-Champaign

Abstract

We present a framework for file relationship discovery between pairs of 2D engineering drawings, and between 2D engineering drawings and the 3D CAD Models obtained from them. The framework consists of modules for automated file system and content-based metadata extraction, for metadata organization and storage, and for exploratory visual inspection and insertion of discovered relationships between pairs of files. The system level metadata extraction is accomplished by using Aperture software and leads to information about file name, MIME type, and disk location. Additional metadata are obtained by performing file format identification with DROID software based on the PRONOM repository of file formats. Metadata from content based analyses come from 2D engineering drawings by applying Optical Character Recognition (OCR), and from 3D CAD models by keyword based extraction of information. All metadata extracted are represented as RDF triples and stored using Tupelo semantic content management repository. Exploratory visual inspection and insertion of discovered relationships have been enabled by developing a graphic user interface to all metadata acquired and by adding analytical capabilities to support discoveries. We have tested our file relationship discovery framework by processing a test collection of electronic records about the Torpedo Weapons Recovery Vessel (TWR 841) archived by the US National Archive (NARA). This test collection presents a problem of unknown relationships among 784 2D image drawings and 22 CAD models.

Introduction

This work is part of a larger project addressing the problem of finding the relationships and similarities among different files, and their versions in order to support preservation decisions [1]. We focus our effort on discovering relationships among files representing scanned 2D engineering drawings and 3D CAD models obtained from the drawings. We assume that the key information for discovering file relationships is encoded in the normalized blocks of engineering drawings, descriptors in 3D CAD model files, and in the file system used for storing the files. In a digital era, files with relationships are typically placed in similar locations, and contain descriptors in the files with common values. Our framework is based on extraction of metadata from a file system, extraction of metadata about file formats using DROID [2] file format identification scheme against the PRONOM [3] registry and NCSA 3D file registry POLYGLOT [4], an automatic optical character recognition (OCR) of the information blocks in 2D drawings mapped by an ontology for the information extracted. The metadata are stored in RDF triple representation using TUPELO [5], semantic content management repository, and analyzed by the framework to deliver all available information for visual inspection and for establishing file relationship using an exploratory graphic user interface.

Metadata extraction

We have used the Aperture Framework [6] for metadata extraction from the file systems. In order to support the file filtering component of the design, we developed a tool that calls the DROID program. The result from DROID is metadata about each file including the registered PRONOM universal ID. We convert the metadata into RDF triples and store the triples again in a metadata context repository. Several 3D file formats are not supported by PRONOM and DROID returns the unidentified file format flag. Those files are then checked against an internal list of 3D file types we created in the past. For technical drawings a semi-automated extraction tool has been developed with the following algorithm flow: The information blocks are segmented by the JAVA based cropping mechanism using ImageToLearn [7] software classes, each cropped sub-field is analyzed with the OCR ABBYY software, text (and image for hand written signature) sub-fields are stored with corresponding ontological description, text parser creates the final RDF triple representation.

Exploratory framework for relationship discovery

A File Relationship Discovery Tool prototype was developed for visual inspection and search of overlapping information in the metadata repository. For example, if 2D drawings and 3D files belong to folders that are topologically related, or one 2D drawing refers to another through the Reference block data (e.g., by having the same creator), then the overlapping information is presented to the end user for the final decision about the file relationships.

Summary

We have presented a methodology for file relationships discovery from diverse sources and prototyped a specific case of discovering relationships among 2D engineering drawings and 3D CAD models. Overall 784 2D drawings and 22 CAD models have been analyzed and the relationships have been visualized. The system currently works in a semi-automated mode with the coordinates of the information blocks being extracted manually from the original scanned tiff files. In the future, we plan to understand the quality of the initial scans of 2D engineering drawings on the metadata extraction and the implications on the relationship discoveries.

Acknowledgment: The authors would like to thank National Archives and Records Administration (NARA) supplement to NSF PACI cooperative agreement CA #SCI-9619019for funding.

References:

[1] The Strategic Plan of the National Archives and Records Administration (NARA) 2006-2016, *Preserving Past to Protect Future* 2006, URL http://www.archives.gov/about/plans-reports/strategic-plan/

[2] DROID is a software tool to perform automated batch identification of file formats. 2009, URL http://droid.sourceforge.net/wiki/index.php/Introduction

[3] PRONOM is a resource registry (information) about the file formats, software products and other technical components. 2006, See URL http://www.nationalarchives.gov.uk/aboutapps/pronom/puid.htm

[4] K. McHenry, P. Bajcsy; Polyglot, 2009; Polyglot converts 3D file formats from one to another using a set of software packages running on a server at NCSA

[5] J. Futrelle, Harvesting RDF Triples, IPAW’06 International Provenance and Annotation Workshop 2006, 64-72; Tupelo, 2009, http://tupeloproject.ncsa.uiuc.edu/

[6] Aperture, 2008, URL http://sourceforge.net/project/showfiles.php?group_id=150969

[7] R. Kooper, D. Clutter, S.-C. Lee, P. Bajcsy; ImageToLearn, 2006, URL http://isda.ncsa.uiuc.edu/Im2Learn/