Technical Report: NCSA-ISDA09-001
Image Spatial Data Analysis Group
National Center for Supercomputing Applications
1205 W Clark, Urbana, IL 61801

Information Extraction from Scanned Engineering Drawings

Michal Ondrejcek, Jason Kastner, Rob Kooper and Peter Bajcsy
National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL
{mondrejce, jkastner, kooper, pbajcsy}@ncsa.uiuc.edu

December 31, 2009

Abstract

This work is motivated by the need to discover and preserve relationships among engineering drawings and their modern counterparts such as 3D CAD models. We have developed a general prototype system called File2Learn for (1) extracting file system level information using Aperture software, (2) performing content based analyses of 2D engineering drawings by optical character recognition (OCR) techniques and of 3D CAD models by string matching, and (3) discovering and establishing relationships among files by using a semi-automated exploratory framework.

This report focuses only on information extraction from scanned paper copies of engineering drawings. We document the problems related to automation of information extraction, organization and representation of extracted information, as well as information quality control. The automation problem is addressed by scripting the execution of a large volume of OCR operations by using the AutoHotKey scripts with the OCR package by ABBYY. The organization and representation of extracted information are handled by adopting and extending existing ontologies (Engineering Drawing Practices terminology, Friends-of-a-friends (FOAF), Dublin Core ISO/IEC 11179), and using the Resource Description Framework (RDF) triples (subject-predicate-object) for representation. The information quality control is approached by designing an exploratory framework where image areas from engineering drawings and the corresponding OCR string are presented for editing. We have tested the prototype developed system on 784 images of engineering drawings and more than 22 CAD models of the Torpedo Weapon Retriever (TWR 841).

Contents

1 Introduction

The motivation of our work is to design a system for automatic discovery of unknown relationships between the files of engineering drawings and 3D CAD models. The discovery is supported by (1) extraction of metadata from a file system using Aperture JAVA framework [1], (2) automatic optical character recognition (OCR) of the information blocks in 2D engineering drawings using JAVA programming, and AutoHotKeys [2] scripting of the OCR software, (3) design of an ontology for the information extracted by OCR and creation of RDF metadata representation, (4) extraction of metadata about file formats using DROID file format identification scheme against the PRONOM registry [3] and NCSA internal registry about 3D file formats [4], (5) visualization of all metadata to support visual inspection based discovery, and (6) computer-assisted support for discovering relationships.

Our ultimate goal is to discover file relationships automatically given a set of files. The general methodology is illustrated in *Figure 1*.

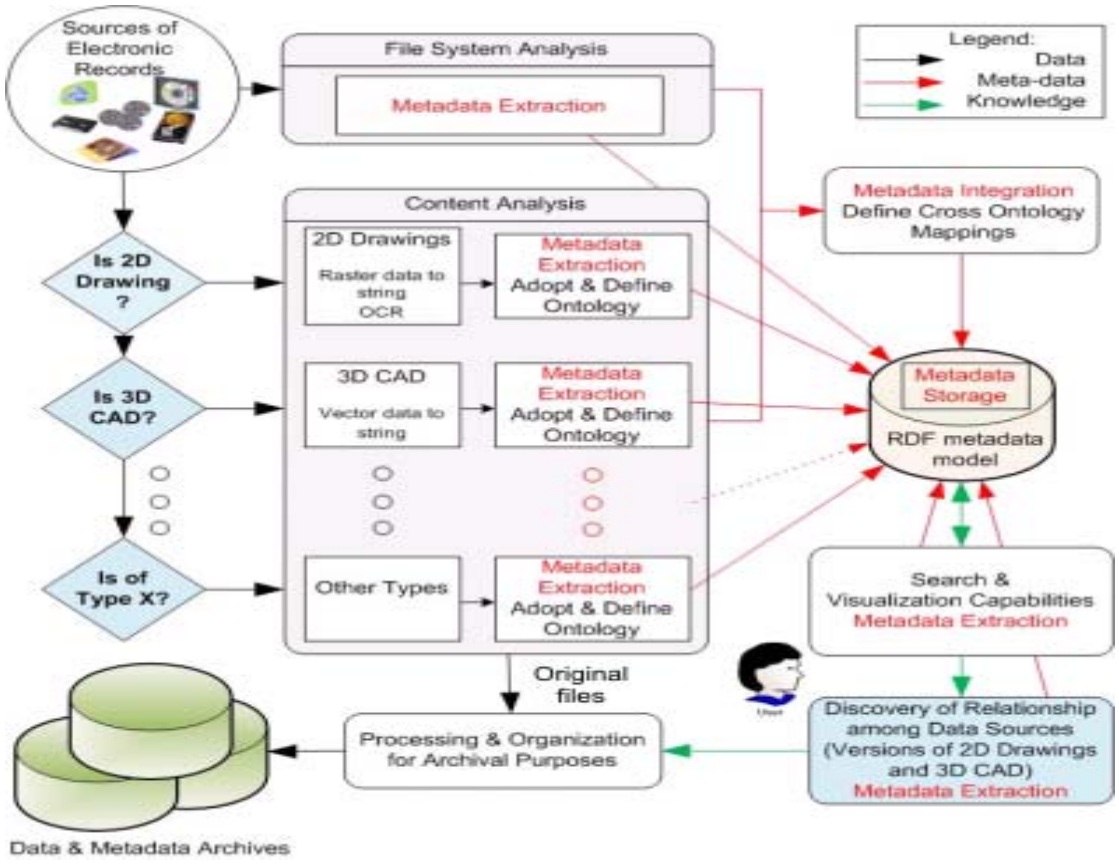


Figure 1. *A methodology for discovering relationships among multiple sources of electronic records.*

At the high level, the methodology consists of (1) file identification and content classification phase (the left column of decisions about the file category), (2) extraction of file system based information (top box), (3) extraction of content based information (the middle box), (4) management of extracted information/ metadata (the boxes with red labels) and (5) services and exploratory interfaces to verification, validation, editing and file relationship discoveries (the bottom right boxes). This methodology feeds the application such as the archival appraisal and management of the files in this case (the bottom center box).

1.1 Specific Focus

In this report, we concentrate only on the problem of information extraction from scanned engineering drawings also denoted as content based information extraction. There are several challenges in automated processing of engineering drawings. For example, the drawings are scanned at low resolution (<300dpi) and hence character recognition is error-prone. The information blocks take on many different shapes or are only partially normalized leading to difficulties in template-based recognition. The lines of information blocks are frequently skewed and/or with gaps due to scanning. Finally, the text in information blocks is a combination of freehand lettering and pre-printed characters without a standard font (uppercase Gothic has been observed as the preferred font). These challenges have to be overcome during automated processing and the processing workflow often includes steps such as (a) information block detection in an engineering drawing, (b) recognition of information block type, (c) localization of sub-fields of an information block, (d) character recognition of the sub-field entries, (e) classification of extracted sub-field entries (taxonomy and ontology), (f) representation and storage of extracted entries referred to as metadata, and (g) visualization of sub-fields and extracted information for the purposes of recognition accuracy evaluation, correction and linking with other information.

In the past, Najman et al. used form-processing method for automating the indexing of title blocks [5]. There have been similar approaches reported in [6, 7] about (a) indexing keyword fields extracted from the title blocks, and (b) automation of the detection of title blocks in the engineering drawing based on location hashing. Several software solutions have been developed in the past for metadata visualization [8, 9] that use the Resource Description Framework (RDF) format for representation [10, 11]. However, these solutions lack capabilities for comparing sub-field images and recognized characters or multiple entries extracted by character recognition. Knowing that the character recognition does not reach 100% accuracy, these solutions also do not support advanced editing, and filtering needed to perform corrections and to support semi-automated decisions about file relationships and document matching.

Several OCR software packages have been tested; TopOCR [12], IRIS ReadIRIS Pro 11 [13], ABBYY FineReader 9.0 [14], OCR engine Tesseract/ OCRopus [15] and Gamera, a toolkit for building document image recognition systems [16]. In order to identify the best OCR solution we used a low resolution (<300 dpi) and a very low resolution (72 dpi) well recognized text images as an input. We have not made any

advanced adjustments of the parameters in these software packages. Our primary goal was to test recognition of the low resolution images. The best package was FineReader by ABBYY with 100% accuracy which was finally purchased and used in this study. The second best package was TopOCR, the free OCR software for smartphones with the accuracy of about 93% of the words being recognized.

Our overall approach to the information extraction problem was to decompose it into sub-problems related to automation of information extraction using one of the OCR software packages, organization and representation of extracted information, and information quality control. The automation problem is addressed by scripting the execution of a large volume of OCR operations using the AutoHotKey scripts with the OCR package by ABBYY. The organization and representation of extracted information are handled by adopting and extending existing ontologies (Engineering Drawing Practices terminology, Friends-of-friends (FOF), Dublin Core ISO/IEC 11179), and using the Resource Description Framework (RDF) triples (subject-predicate-object) for representation [17]. The information quality control is approached by designing an exploratory framework where image areas from engineering drawings and the corresponding OCR string are presented for editing. We have tested the prototype system on 784 images of engineering drawings and more than 22 CAD models of the Torpedo Weapon Retriever (TWR 841).

1.2 Processing Workflow

We outline the processing workflow for information extraction from 2D images of engineering drawings.

1. Detect the information (title) block in an image and crop the image area.
2. Identify the type of information block template and crop areas for each information sub-field.
3. Apply OCR software to each sub-field and associate it with text strings resulting from OCR.
4. Develop ontology for all sub-fields in information blocks and perform cleansing of the text strings extracted by OCR according to the ontology
5. Represent text strings as metadata RDF triples and store them in a shared repository.
6. Visualize RDF triples in a custom browser called File2Learn that supports verification, validation, metadata editing and viewing of images and 3D CAD Models.

Our current implementation uses (a) the NCSA ImageToLearn library [18] written in Java for image cropping, (b) ABBYY software (a desktop application) for OCR, (c) NCSA Tupelo [19] as a metadata repository, and (d) File2Learn exploratory framework for browsing and quality control of extracted information (images and strings obtained via OCR) [20]. The automation of image cropping is achieved by developing Java code that (a) crops an information (title) block area using Im2Learn library and (b) creates

areas with title block sub-fields. The automation of OCR is accomplished (a) by calling an AutoHotKeys script to launch ABBYY OCR software applied to each area with title block sub-field and (b) by creating RDF triple metadata representation saved into a repository using custom Java code.

2 Information Block Analyses

Navy technical drawings have been analyzed in order to get familiar with the data structure. Appendix A summarizes the types and layouts of the blocks of interest: Title, Reference and MMC (Marinette Marine Corporation) number blocks. *Figure 2* shows the typical NAVY title block with various information fields.

The problem of identifying different types of blocks within the drawing in this stage was solved by manual mapping of the blocks coordinates and by populating a 'master' text file with block coordinates. Due to the fact that the drawings have been scanned at multiple dpi magnifications, the sub-fields in each title block template have to be detected using a relative coordinate system within each template. We have captured the relative position of each sub-field area by manually encoding the left upper corner, and width and height. Appendix B summarizes the selected information blocks.

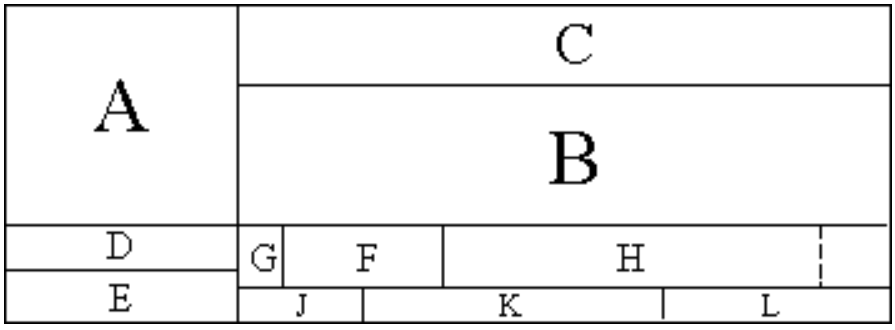


Figure 2: *Title block template used in NAVY showing various information fields: (A) vendor, record of preparation, (B) drawing title, (C) preparing activity, (D)(E) parts are not standardized, (F) Code identification number, (G) drawing size, (H) drawing number, (J) scale, (K) specification number and (L) sheet number.*

We found out that there were discrepancies between the templates described in the Engineering Drawing Practices [21] (denoted as 'Standard') and the blocks found in the drawings of the Torpedo Weapon Retriever (TWR 841) (denoted as 'Real'). The differences in positions were up to ~20% with respect to the template for the same sub-field coordinates.

3 Ontology and Metadata Representation

The sub-fields within a title block are denoted by upper case alphabet. We have designed an ontology describing the sub-fields with in the following selected mapping:

tb - Title block <tdrw:titleBlock>

A0, A01, A02, A03	<tdrw:recordOfPreparation>
B0	<tdrw:preparingActivity>
C0	<tdrw:drawingTitle>
D0101	<tdrw:action>
D0102	<tdrw:actionBy>
D0103	<tdrw:actionDate>

The information in D and sub-fields varies as it is usually reserved for names and drawing activities grouped in boxes (<tdrw:miniBox>, <tdrw:action>, </tdrw:actionBy>, </tdrw:actionDate> etc.). The values enclosed by inequality signs are the predicates of the ontology used for creating metadata RDF triples (subject-predicate-object). The full description of <tdrw> terms is in Appendix C.

One has to be aware of challenges when the extracted information is integrated with metadata obtained by other software packages (e.g., the Aperture software). It is very probable that there would be multiple predicates referring to the same semantic meaning in ontologies (e.g., **creator** and **author**). We have been carefully designing ontologies and namespaces used by the NCSA software in order to follow the already existing standards such as, the Dublin Core Metadata Initiative (DCMI) terms or the friend-of-a-friend [17] (FOAF) namespaces, and the Engineering Drawing Practices terminology. If this problem occurs then it could be corrected in the quality control step by consolidating predicates identified by an end user as described in the next section.

Figure 3. *The user interface to the exploratory framework called file2learn. Each pane is denoted with information presented in the pane.*

4 Quality Control of Extracted Information

We have developed a prototype of a File Relationship Discovery framework called File2Learn for visual inspection based discovery of relationships. This framework is also used for quality control as the OCR step could report unidentified characters or misclassified characters, and the ontologies of multiple software packages might create syntactically different but semantically identical predicates in RDF triples. The user interface of File2Learn is shown in **Figure 3Error! Reference source not found.**. In File2Learn, a user can select two files and preview all connecting information links. We added support for TIFF and STEP files for image preview as well. The File2Learn framework allows a user to view the metadata represented as RDF triples in a graph viewer or in a tabular viewer.

Figure 4. *An example of anomaly detection in OCR results. The left lower pane shows the area of the title block sub-field and the middle pane shows the string extracted using OCR. The String can be edited based on visual inspection of the area image. The sub-field in the left lower pane is associated with the file selected in the upper left pane. The right lower pane shows the entire image by selecting the original file in the upper right pane.*

Figure *4* shows the presentation of an image area in the Title block that has been processed by OCR. The quality control is enabled by providing an interface to visually inspect the image and its corresponding OCR string results, followed by the string editing capabilities.

5 Comparing Extracted Information from Engineering Drawing

In addition to quality control, file2learn can be used for discovering relationships among engineering drawings and 3D CAD models. *Figure 5* shows the pair of 3D file in STEP file format and 2D image file in TIFF file format with three connecting links based on metadata extracted from a file system (files are located in related folders) denoted by a grey ellipse, and one connecting link based on the inserted author (creator) denoted by the red ellipse. *Figure 6* presents the same information in a tabular view with color coded entries. The colors correspond to file descriptors that are equal (blue ~ the same predicate and value), are not equal (red - the same predicate but different value) and are different (green - predicate with a value occurs only in one file as a descriptor).

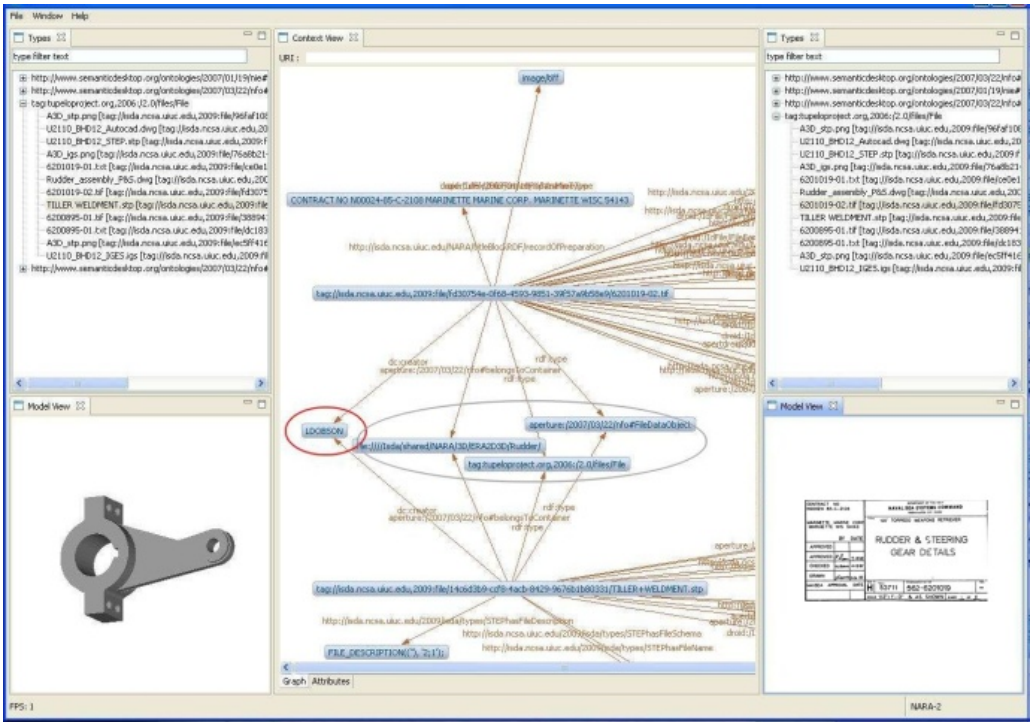


Figure 5. *A File Relationship Discovery framework called File2Learnprovides support for visual inspection based discovery of file relationships. The Tool shows a zoomed-in view of the visualization where the overlapping information in grey ellipse corresponds to the file locations (files belong to folders that are topologically related, e.g., 3D file and 2D file share parent*

folder). The red ellipse corresponds to the overlapping information about authors (creators) extracted from the content of the files.

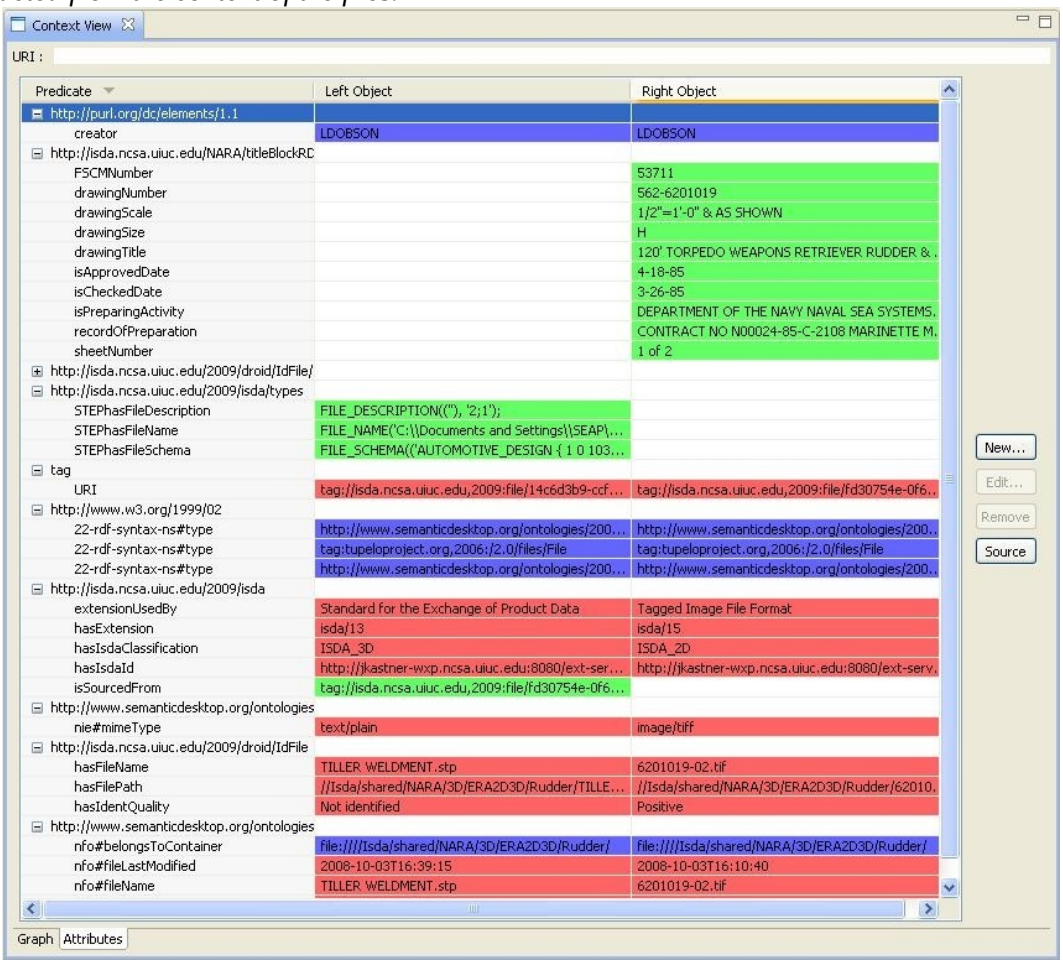


Figure 6. *A tabular presentation of the overlapping information in the File2Learnframework. The colors correspond to file descriptors that are equal (blue ~ the same predicate and value), are not equal (red – the same predicate but different value) and are different (green – predicate with a value occurs only in one file as a descriptor).*

6 Processing Large Numbers of Engineering Drawings

The current workflow contains still a few manual steps in OCR part, such as detecting information blocks in images of engineering drawings, finding the title block template that corresponds to the cropped image area.

We have fully automated (a) metadata file extraction from the file system using Aperture, PRONOM and DROID, (b) image cropping into sub-fields of each title block template, OCR execution, and metadata RDF triple creation and storage. An example of RDF triples extracted from one title block in RDF/XML format is shown in *Figure 7*.

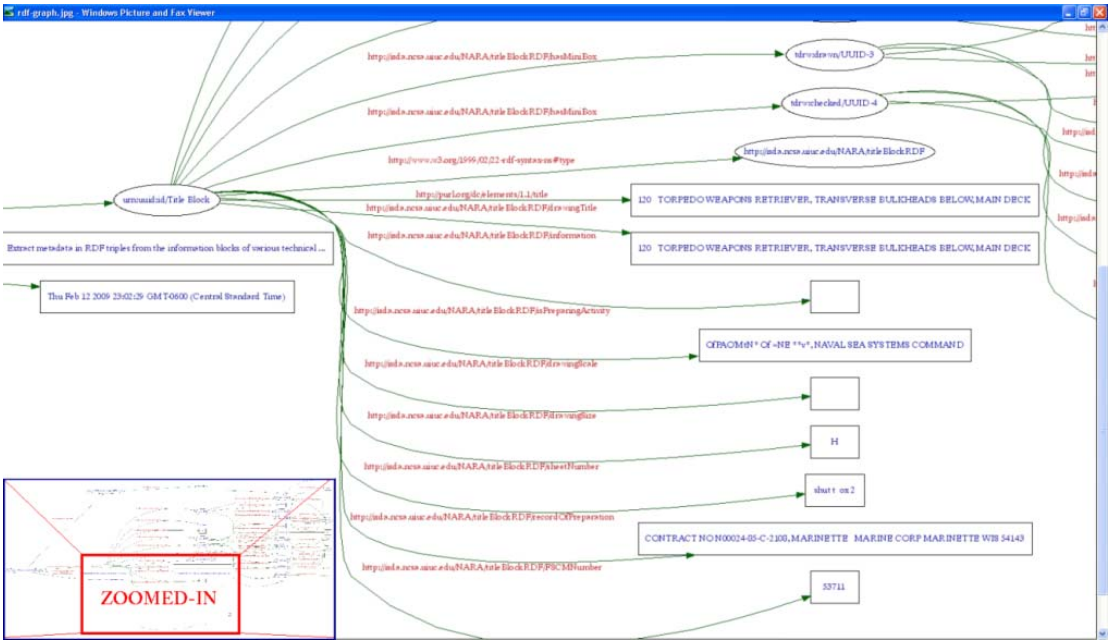


Figure 7. *A graphical representation of the RDF metadata from a title block as displayed in RDF Validator [22] (left lower corner contains the overall graph with the area that has been zoomed-in so that the strings could be readable). There are currently 60 unique triples created from each information (title) block.*

7 Summary

We summarized the overall results and comments below. They are related to the manual coordinated extraction and automation of information extraction from the engineering drawings.

Block coordinates, creation of a master text file:

- Block coordinates obtained manually ~ 50-70 per day depending on the knowledge of the project. The time consuming part is getting familiar with the drawing specifications and 'habits' of the preparers.

Field coordinates, creation of block templates:

- Field coordinates were also obtained manually for block templates. Total number of different blocks (and sizes) in TWR 841 NAVY ship set is about 70 including all variants of reference blocks with different sizes and number of references, all non standardized title blocks, and a few of list of materials and revision blocks. We have created and used about 50 templates. Among them are 40 unique templates only for the reference blocks. It should be possible to derive formula for reference block with n references. Additionally we created few templates for standardized title blocks.

TWR 841 NAVY ship set:

- 784 files in PNG (converted from originally LZW compressed files) with 170 title blocks (mostly H and C), 700 continuation title blocks, 150 reference blocks, about 200 additional areas with the drawing numbers (MMC DWG. NO.)
- Coordinates of two dozen of non standardized title blocks, revision, list of material and other blocks are present in the master text file however they have not been analyzed. In order to analyze them we need more field specifications (coordinates). It is very time consuming since almost all of these blocks have some extra fields and exceptions. Additionally we need more ontological terms especially for list of materials.

OCR:

- Block cropping and field cropping followed by OCR took approximately: 10 drawing files/hour (~20 blocks/hour)
- Total amount of 784 2D drawings in TIFF file format have been identified which include 170 title blocks, 700 continuation title blocks, 150 reference blocks, a dozen of revision and list of material blocks and about 200 additional areas with the drawing numbers. The OCR accuracy drops down to 80% for title blocks characters in certain drawings mainly due to poor original scans quality. The most difficult parts for OCR are the fields with handwritten information such as names and dates and fields with measuring units. The most important fields with drawing numbers and drawing references (digits) are much better resolved. The accuracy of correctly recognized (spelled) words is even lower.

The most time consuming part of the overall process is related to the knowledge of the actual data set. Appendix D shows few exceptions and errors discovered in the NAVY TDR drawings. These are mainly existence of non standard blocks or block fields. The non standard parts make an automated cropping difficult. It is almost certain that any real data set will have to be manually pre-processed and the data exceptions marked.

We presented a prototype for information extraction from scanned engineering drawings which is one part of the system for file relationship discoveries. The prototype has been tested with the data collection of the engineering drawings and 3D CAD Models for the NAVY ship, model TWR 841.

8 Acknowledgement

This research was partially supported by a National Archive and Records Administration (NARA) supplement to NSF PACI cooperative agreement CA #SCI-9619019. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the National Archive and Records Administration, or the U.S. government.

9 Appendix A - NAVY technical drawings

Technical drawings which engineering drawings are part of describe geometric, material and other features of a product required for precise manufacturing. Information blocks and especially the title block are standardized to some extent following the recommendations of American National Standard Engineering and Related Document Practices (ASME Y14/ANSI Y14) [23], International Standard Organization (ISO 128 and 7200) [24] rules. The title block should be included in all sheets and it is located in the lower right corner of the drawing. It should include sufficient information to identify the type of drawing, vendor name and information about the versions, authors and other creators. A title block is divided into several areas (sub-fields) including the drawing title and the drawing number. The drawing number may also contain the sheet number as a substring of the sheet number. Other areas usually contain signatures and approval dates. Most engineering drawings also include a reference block which lists names of other drawings that are related to the system. It is possible to cross link drawings through the title and reference blocks. A drawing might contain other information blocks, such as a list of materials, a list of revisions, and so on.

If the drawing has been born digitally and created by software supporting engineering applications then the title block is pre-defined in the software. *Figure A1* shows an example of the ProfiCAD [25] title block specification which was created according to the ISO 7200 standard.

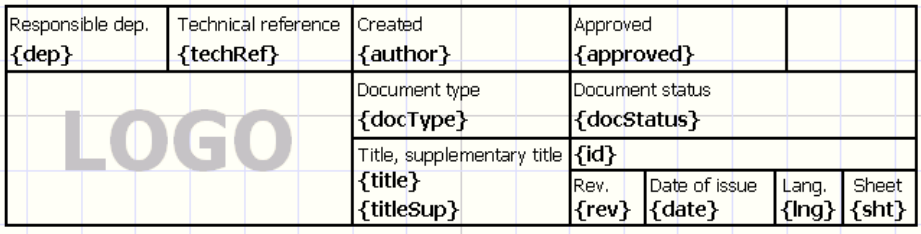


Figure A1. *ProfiCAD specification of the title block according to the ISO 7200 standard.*

In our study we focused on older printed technical drawings described in a DoD-STD-100C specification defined for NAVY [21]. Examples of a title block template with information fields are shown in *Figure A2*. The actual title block is presented in *Figure A4*.

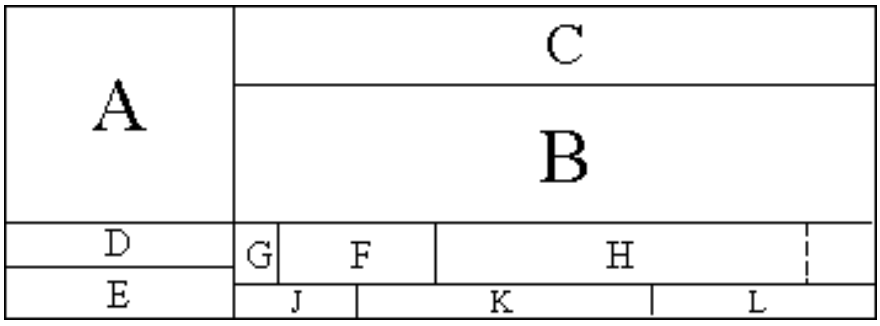


Figure A2. *Title block template used in NAVY drawings with the following information fields: (A) vendor, record of preparation, (B) drawing title, (C) preparing activity, (D)(E) parts are not standardized, (F) Code identification number, (G) drawing size, (H) drawing number, (J) scale, (K) specification number and (L) sheet number.*

The full title block is positioned in the lower right corner of the first sheet of a drawing. The field designation in *Figure A2* varies with each NAVY command or activity. Many fields of this block, especially the A and C sub-fields, are optional in the continuation sheets (the sheets other than the first sheet). Some blocks (D and E) may be extended or absorbed into block A. Other may be split (A and C for example).

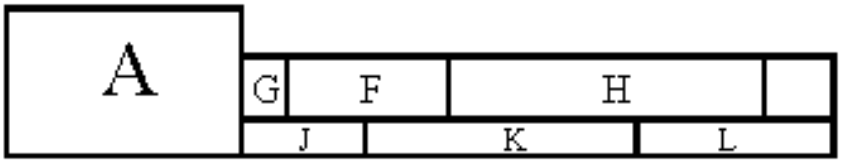
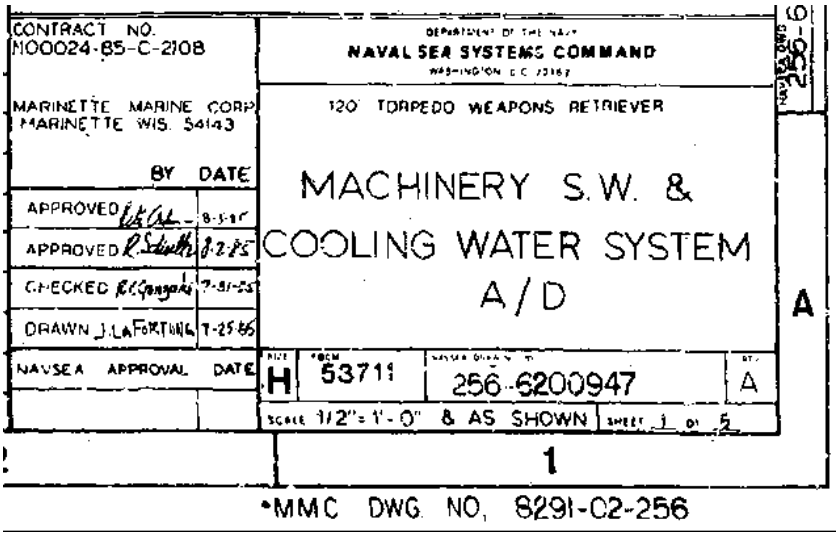
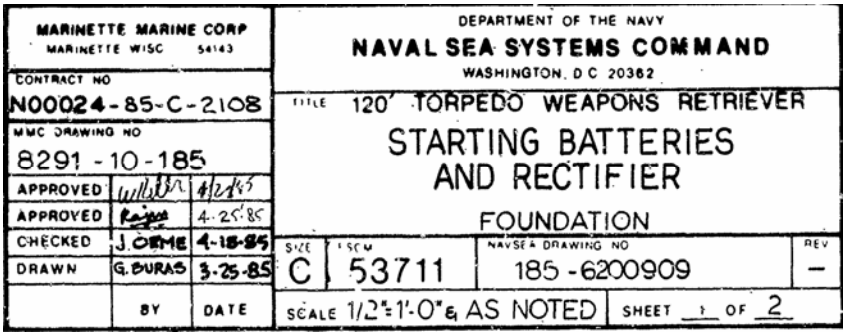


Figure A3. *Continuation title block with similar fields as in Figure A2.*

Each continuation sheet has a smaller version of the title block (called Continuation title block and shown in *Figure A3*) with sub-set of the fields from the main title block.





MARINETTE MARINE CORP MARINETTE WISC 54143			DEPARTMENT OF THE NAVY NAVAL SEA SYSTEMS COMMAND WASHINGTON, D C 20362		
CONTRACT NO N00024-85-C-2108			TITLE 120' TORPEDO WEAPONS RETRIEVER STARTING BATTERIES AND RECTIFIER FOUNDATION		
MMC DRAWING NO 8291-10-185					
APPROVED	[illegible]	4/25/85			
APPROVED	[illegible]	4-25-85			
CHECKED	J. OEHME	4-18-85	SIZE C	FSCM 53711	NAVSEA DRAWING NO 185-6200909 REV –
DRAWN	G. BURAS	3-25-85			
	BY	DATE	SCALE 1/2"=1'-0" & AS NOTED		SHEET 1 OF 2

Figure A4. *Example of a title block, size H (top) and size C (bottom) in a NAVY SEA 2D technical drawing files.*

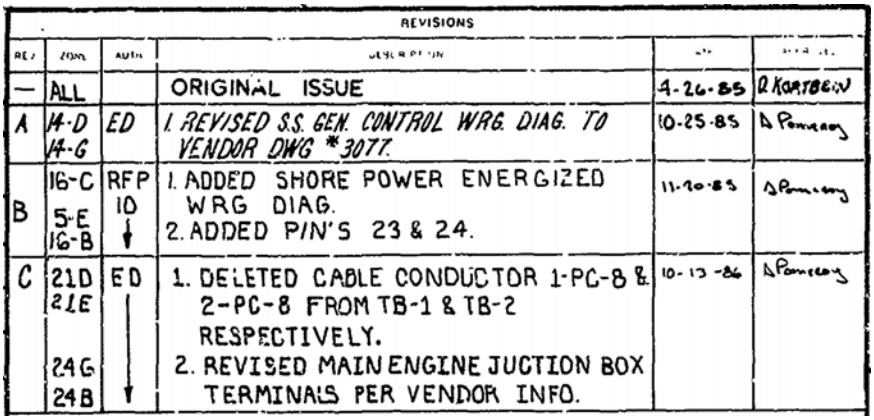
Most engineering drawings will have a reference block (*Figure A5*), which lists other drawings that are related to the entire system or to a component and can be cross-referenced. In the case of NAVY drawings the NAVSEA DWG number is in the form of xxx-yyyyyyy. The MMC (Marinette Marine Corporation) number is in a form of bb-ccc and refers to the drawings from the same sub-project. The second part of the NAVSEA number labeled with y forms also a part of the digital file name. The MMC number forms a part of the full MMC number following the format aaaa-bb-ccc.

NO	TITLE	NAVSEA DWG	MMC DWG
5	LONGITUDINAL FRAMING & GIRDERS	116-6200694	01-116
4	SHELL EXPANSION	111-6200891	01-111
3	WELDING TABLE	801-6201065	23-801
2	STANDARD WELDING DETAIL BOOKLET	801-6201064	22-801
1	STANDARD STRUCTURAL DETAIL BOOKLET	801-6201063	21-801
REFERENCES			

Figure A5. *Example of a Reference block in a NAVY SEA 2D technical drawing. The files can be cross-referenced through the NAVSEA DWG number and the MMC DWG number.*

There are additional information blocks in some technical drawings. For instance, there could be blocks denoted as List of materials, Revision block, General notes and others. Figure A6 shows details of a Revision block. The format and location of these blocks in general is less defined than those for the Title and Reference blocks. The fields are often hand-written with abbreviations and unit specifications.

An example of a complex information part of a particular drawing including the Title and Reference blocks is in *Figure A7*.



REVISIONS

REV	ZONE	AUTH	DESCRIPTION	DATE	APPROVED
—	ALL		ORIGINAL ISSUE	4-26-85	R KORTBEIN
A	14-D 14-G	ED	1. REVISED S.S. GEN. CONTROL WRG. DIAG. TO VENDOR DWG # 3077.	10-25-85	A Pomeroy
B	16-C 5-E 16-B	RFP 10 ↓	1. ADDED SHORE POWER ENERGIZED WRG DIAG. 2. ADDED P/N'S 23 & 24.	11-20-85	A Pomeroy
C	21D 21E 24G 24B	ED ↓	1. DELETED CABLE CONDUCTOR 1-PC-8 & 2-PC-8 FROM TB-1 & TB-2 RESPECTIVELY. 2. REVISED MAIN ENGINE JUCTION BOX TERMINALS PER VENDOR INFO.	10-13-86	A Pomeroy

Figure A6. *Revision block in a NAVY SEA 2D technical drawing.*

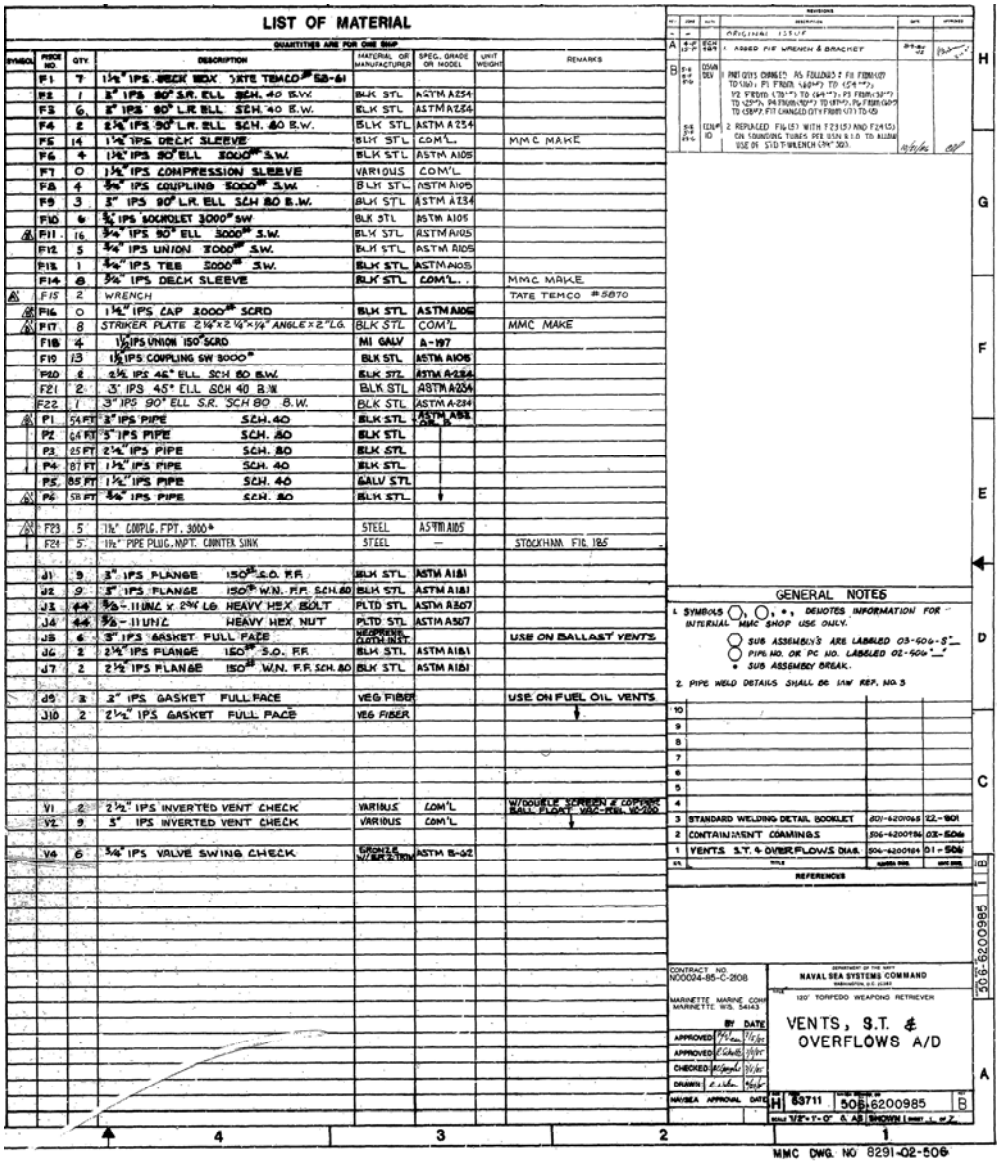


Figure A7. *Example of a complex information part of a drawing with all, Title, Reference block, List of material, General notes and Revision blocks.*

There are three sizes of title blocks: a block used for A-, B-, C-, and G-size drawings, a slightly larger block for D-, E-, F-, H-, J-, and K-size drawings, and a vertical title block. Reference blocks usually follow the title block specifications since they are placed above the title blocks. The same is usually true for the revision block which is placed in the same (title block) column in the upper right corner of the drawing.
It has to be mentioned that almost all specifications are somewhat loose and they depend on the preparation unit and project itself.

10 Appendix B – Text files with block types and coordinates

The problem of identifying different types of blocks within the drawing in this stage was solved by manual mapping of the blocks coordinates. Format of the 'master' text file is shown below. A file name entry is followed by an optional rotation field. This field is either empty or it shows the necessary rotation size plus the degree amount (CCW90 stands for counter clockwise rotation of 90 degrees). Next, a field describes the Title block (TB or TBC for continuation block), size of the block (H) and the upper left and lower right absolute coordinates in pixels with respect to the origin (upper left corner of the drawing). A Reference block is marked RF and its field contains the number of references (0 through 20) and the coordinates in pixels. Other blocks follow with the final field for comments (marked CM).

6200901-01.tif;;TB,H,7074,4891,8529,5689;RF,13,7071,3628,8523,4522;
MMC,7576,5787,8472,5851;CM
6200901-02.tif;;TBC,H,6891,5526,8345,5726;MMC,7367,5498,8341,5566;CM,A empty
6200902-01.tif;;TB,H,7322,4893,8774,5691;RF,13,7316,3554,8764,4525;
MMC,7827,5788,8621,5854;CM
6200902-02.tif;;TBC,H,7239,5532,8697,5736;MMC,7717,5486,8695,5574;CM,A empty
6200903-01.tif;;TB,H,6909,4855,8373,5650;RF,2,6897,3592,8355,4483;
LM,5067,204,6879,1466;MMC,7403,5749,8220,5825;CM
6200903-02.tif;;TBC,H,7048,5543,8500,5749;MMC,7526,5525,8500,5587;CM,A empty
6200904-01.tif;;TB,H,7077,4986,8523,5784;RF,7,7081,3950,8520,4623;
MMC,7551,5880,8389,5942;CM
6200905-01.tif;;TB,H,7446,4871,8890,5654;RF,8,7442,3753,8891,4490;CM, bad

Example of 'master' text file with blocks' coordinates.

Abbreviations of different blocks identified in the full set of TWR 841 NAVY ship drawings (784 scanned images) are summarized in *Table B1* and *Table B2*. All blocks have been included in the master text file. However, only the blocks in Table B1 have been analyzed. *Table B2* represents either odd blocks or not standardized ones.

TB,A; TBC,A	Title block size A; Continuation TB size A
TB,B; TBC,B	Title block size B; Continuation TB size B
TB,C; TBC,C	Title block size C; Continuation TB size C
TB,H; TBC,H	Title block size H; Continuation TB size H
MMC	Marinette Marine Corporation Number
RF,C, 0,1,2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16,17,18,19,20	Reference size C, 0 - 20 references

RF,H, 0,1,2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16,17,18,19,20	Reference size H, 0 - 20 references

Table B1. *Abbreviations and descriptions of main blocks that have been analyzed.*

RV - Revision block	TBC,N956 - no size, not standardized
LM - List of material	TBC,N959 - no size, not standardized
TL - Title logo	TBC,N987 - no size, not standardized
PL - List of parts	TBC,N988 - no size, not standardized
IDX - Index	TB,N989 - no size, not standardized
NM - Name, drawing name or title	TBC,N047 - no size, not standardized
TFA -field A of the B block on top of the block	TBC,N065 - no size, not standardized
GN - General notes	
NA - TB not available, no blocks	

Table B2. *Abbreviations and descriptions of other information blocks.*

Each type of the block analyzed with the OCR package has its own field specifications. Fields coordinates are relative to the block size in order to adjust for different block sizes and scan resolutions. For example, Title block of size H is shown in *Figure B1*. It is similar to the one shown in *Figure A2* of the Appendix A. Each field is marked by a letter and a number (index). More indices denote sub-field of that particular field. Thus H01 and H02 are the two subfields of the *drawing number* H0 fields. For instance, the fields D and E form mini-boxes with related bits of information and hence a ‘matrix’ notation is used (D0102, D0102, D0103).

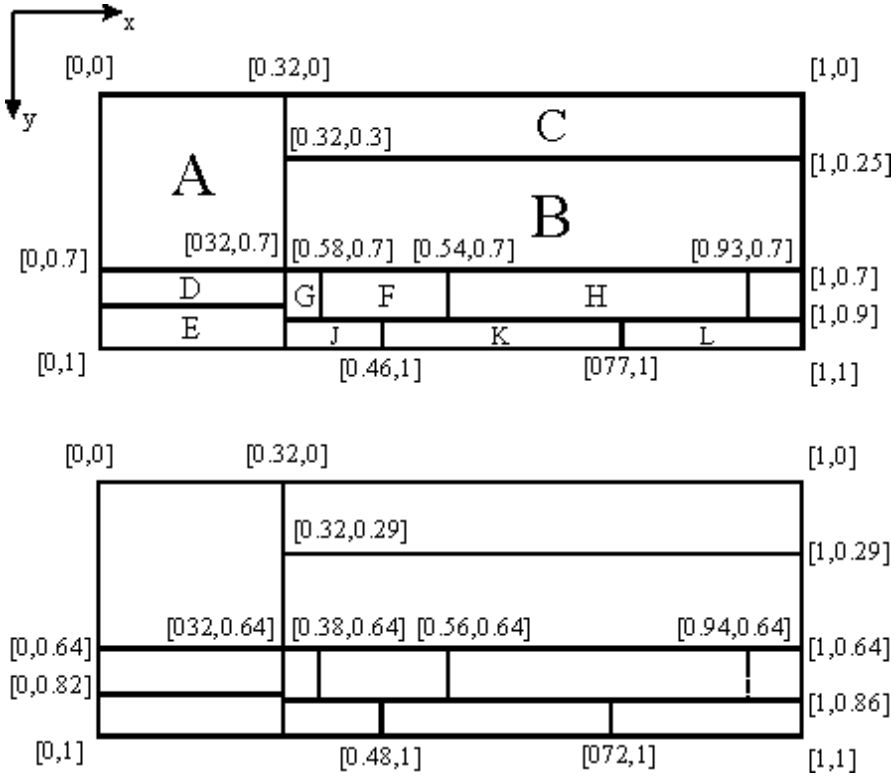


Figure B1: *Top: Relative coordinates of the Title block fields derived from a template following the 'Standard' H size. The field labels follow the description presented in Appendix A, Figure A2. Bottom: Relative coordinates of the Title block fields derived from a 'Real' Title block found in of the NAVY drawings.*

The coordinates of the actual blocks did not always follow the standard ones. In that case the 'Real' block coordinates have been recorded. *Figure B1* shows the normalized coordinates derived from a template block and from a real block. Examples of the detailed field descriptions are in *Table B3*. Creation of this and other text files was necessary for batch cropping of the sub-fields followed by automated Optical Character Recognition (OCR).

%Title Block NAVY REAL (TB) D,E,F,H,J,K	%Reference Block Real R4, 4 fields, C
A0,0,0,0.32,0.42	R0401,0,0,0.08,0.18
B0,0.32,0.17,1,0.8	R0402,0.08,0,0.57,0.18
C0,0.32,0,1,0.17	R0403,0.57,0,0.8,0.18
D0101,0,0.42,0.15,0.52	R0404,0.8,0,1,0.18
D0102,0.15,0.42,0.24,0.52	R0301,0,0.18,0.08,0.36
D0103,0.24,0.42,0.32,0.52	R0302,0.08,0.18,0.57,0.36
D0201,0,0.52,0.15,0.61	R0303,0.57,0.18,0.8,0.36
D0202,0.15,0.52,0.24,0.61	R0304,0.8,0.18,1,0.36
D0203,0.24,0.52,0.32,0.61	R0201,0,0.36,0.08,0.53
D0301,0,0.61,0.15,0.71	R0202,0.08,0.36,0.57,0.53
D0302,0.15,0.61,0.24,0.71	R0203,0.57,0.36,0.8,0.53
D0303,0.24,0.61,0.32,0.71	R0204,0.8,0.36,1,0.53
D0401,0,0.71,0.15,0.8	R0101,0,0.53,0.08,0.70
D0402,0.15,0.71,0.24,0.8	R0102,0.08,0.53,0.57,0.70
D0403,0.24,0.71,0.32,0.8	R0103,0.57,0.53,0.8,0.70
E0,0,0.8,0.32,0.9	R0104,0.8,0.53,1,0.70
D0501,0,0.9,0.24,1	R0001,0,0.70,0.08,0.83
D0503,0.24,0.9,0.32,1	R0002,0.08,0.70,0.57,0.83
F0,0.38,0.8,0.54,0.93	R0003,0.57,0.70,0.8,0.83
G0,0.32,0.8,0.38,0.93	R0004,0.8,0.70,1,0.83
H0,0.54,0.8,1,0.93	RR,0,0.83,1,1
H01,0.54,0.8,0.93,0.93	
H02,0.93,0.8,1,0.93	
J0,0.32,0.93,0.77,1	
L0,0.77,0.93,1,1	

Table B3: *Text files with more detailed field, sub-field and mini-box specifications for a Title block and Reference block with four references*

11 Appendix C – Ontology

Each cropped field has been assigned a letter and a number (index) as mentioned in Appendix B. The notation allows us to map fields with corresponding ontological terms.

Ontology created for the NAVY technical drawings is called <tdrw>. It consists of generic drawing terms such as <tdrw:drawingSize>, <tdrw:information>, <tdrw:FSCMNumber> (federal supply code for manufacturers, and NAVY specific terms such as and <tdrw: MMCNumber> which are type of <tdrw: drawingNumber>.

Technical drawing <tdrw> ontology:

tb - Title block	<tdrw:titleBlock>
A0, A01, A02, A03	<tdrw:recordOfPreparation>
B0	<tdrw:preparingActivity>
C0	<tdrw:drawingTitle>
D0101	<tdrw:action>
D0102	<tdrw:actionBy>
D0103	<tdrw:actionDate>
D0201	<tdrw:action>
D0202	<tdrw:actionBy>
D0203	<tdrw:actionDate>
D0301	<tdrw:action>
D0302	<tdrw:actionBy>
D0303	<tdrw:actionDate>
D0401	<tdrw:action>
D0402	<tdrw:actionBy>
D0403	<tdrw:actionDate>
D0501	<tdrw:action>
D0502	<tdrw:actionBy>
D0503	<tdrw:actionDate>
E0, E01,E02, E03	<tdrw:fieldInfo>
F0	<tdrw:FSCMNumber>
G0	<tdrw:drawingSize>
H0, H01, H02	<tdrw:drawingNumber>
J0	<tdrw:drawingScale>
K0	<tdrw:specsNumber>
L0	<tdrw:sheetNumber>
mmc - MMC Drawing Number block	<tdrw:MMCNumberBlock>
MMC	<tdrw:MMCNumber>
rf - References block	<tdrw:referenceBlock>
RR	<tdrw:blockInfo>
R0001, R0002, R0003, R0004	<tdrw:fieldInfo>
R0101	<tdrw:refNumber>
R0102	<tdrw:drawingTitle>
R0103	<tdrw:drawingNumber>
R0104	<tdrw:MMCNumber>

R0201	<tdrw:refNumber>
R0202	<tdrw:drawingTitle>
R0203	<tdrw:drawingNumber>
R0204	<tdrw:MMCNumber>
etc	
R2001	<tdrw:refNumber>
R2002	<tdrw:drawingTitle>
R2003	<tdrw:drawingNumber>
R2004	<tdrw:MMCNumber>
	<tdrw:miniBox>

Additional terms, not yet used:

	<tdrw:block>
	<tdrw:drawn>
	<tdrw:drawnBy>
	<tdrw:drawnDate>
	<tdrw:appeoved>
	<tdrw:approvedBy>
	<tdrw:approvedDate>
	<tdrw:checked>
	<tdrw:checkedBy>
	<tdrw:checkedDate>
rv - Revision block	<tdrw:revisionBlock>
	<tdrw:revZone>
lm - List material block	<tdrw:listMatBlock>
	<tdrw:symbol>
	<tdrw:material>
	<tdrw:grade>
	<tdrw:unit>
	<tdrw:weight>
	<tdrw:remarks>
tl - Title logo block	<tdrw:logoBlock>
	<tdrw:logo>
tb - List part block	<tdrw:listPartsBlock>
idx - Index block	<tdrw:indexBlock>
gn - Notes block	<tdrw:notesBlock>

12 Appendix D – Exceptions and errors

In this Appendix non standard blocks or block fields are documented. Non-compliance with templates (standards) has posed challenges on automated cropping. Some of the non-compliance problems can be corrected by manual extraction of the block positions. *Figures D1* and *D2* show two continuation blocks, both with empty A field. The first block follows the specifications for the continuation block more closely than the second one with embedded MMC number field. The MMC number has been marked and cropped as a separate block. The problem is even more complex when other numbers than the MMC number are embedded in the same area as seen in *Figure C3* (NAVSEA DWG. NO instead of the MMC number). The MMC and NAVSEA numbers have been swapped making the correct field mapping almost impossible. The proper extraction here lies in hands of a knowledgeable domain expert.

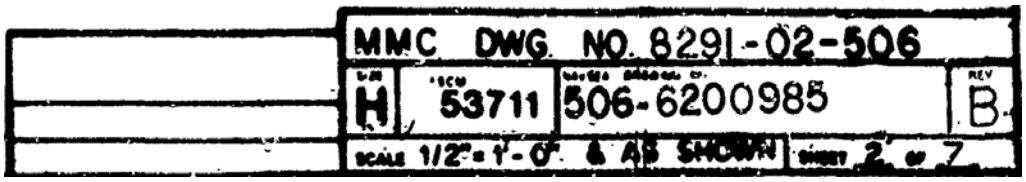


Figure D1. *Continuation title block with embedded MMC number field/block.*

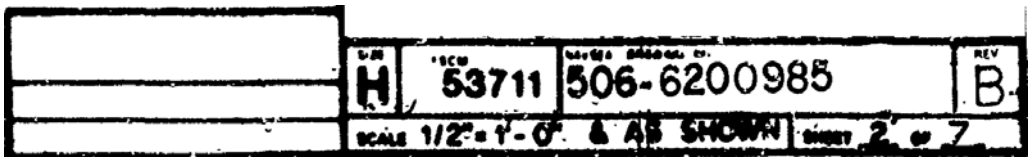


Figure D2. *Continuation title block in a form following a standard template closely.*

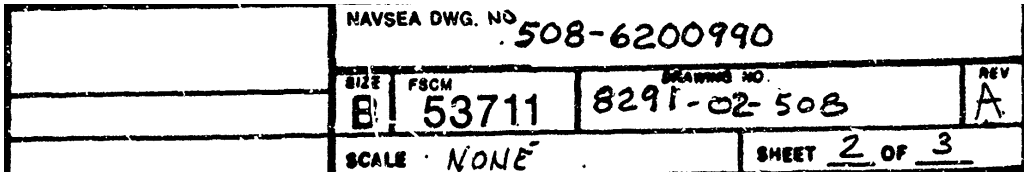


Figure D3. *Continuation title block with swapped MMC and NAVSEA numbers.*

Next example (*Figure D4*) shows the part of the A field positioned above the title block itself. In this case, the coordinate specification for the B field is not compliant with the real block. Thus, the field mapping for the drawing has to be created manually which is very time consuming. Similar problem with field definitions is shown in *Figure D5*. MMC number is embedded in the B field and does not comply with the standard template for the B size title block and its field mapping. Furthermore, the B type block has been used for textual drawings which contain manuals, labels and other additional information. The textual drawings are not in fact drawings per se since they do not need scale and size parameters. We assume that the use of the B type block for annotation and populating the scale allocated field with Not Available (N/A) has been an internal decision of the

creators. However, other textual drawings used different size blocks, custom blocks or even not used blocks at all which indicates the lack of consistency across multiple creators of drawings.

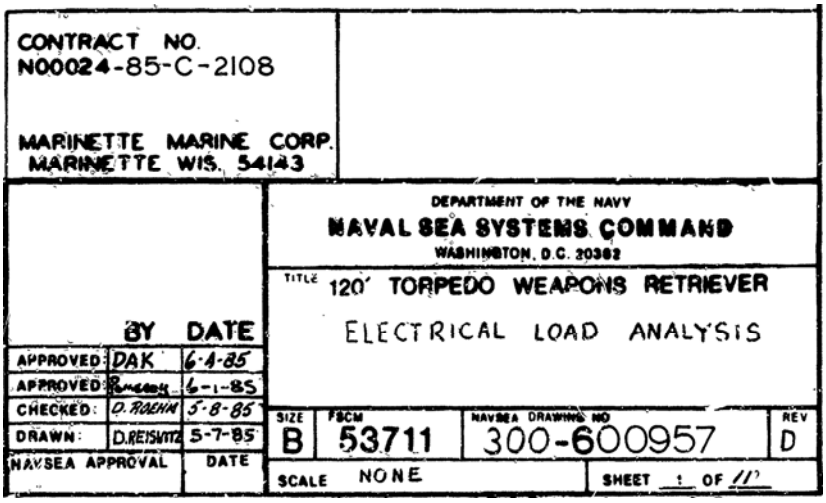


Figure D4. *Title block with an extra field above the field labeled as A before.*

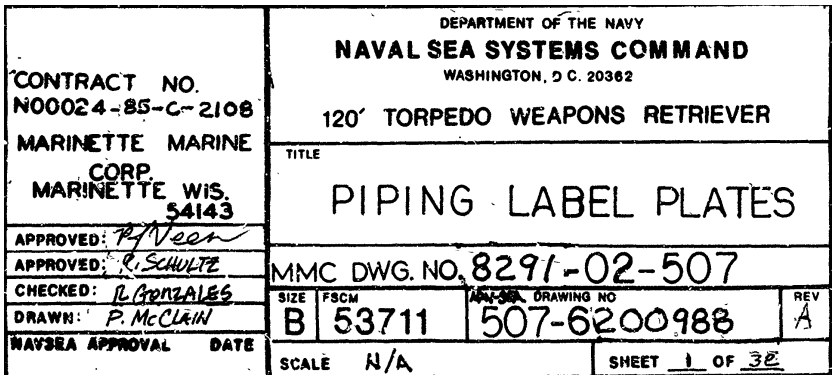


Figure D5. *Title block of the type B in a textual drawing with an embedded MMC number in the B field.*

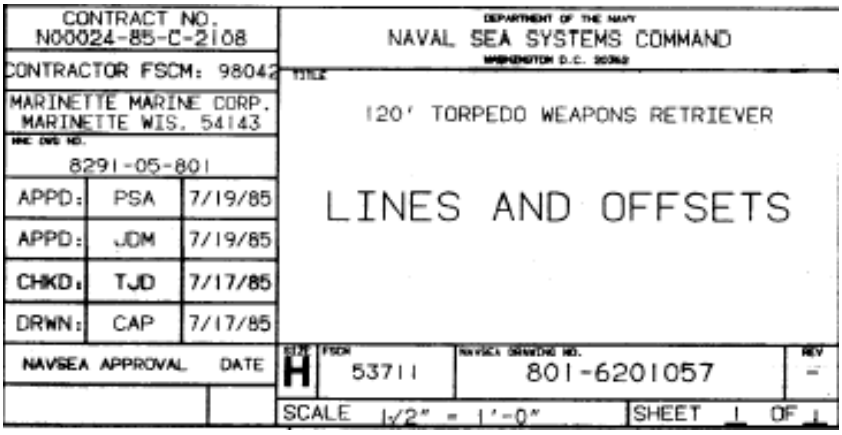
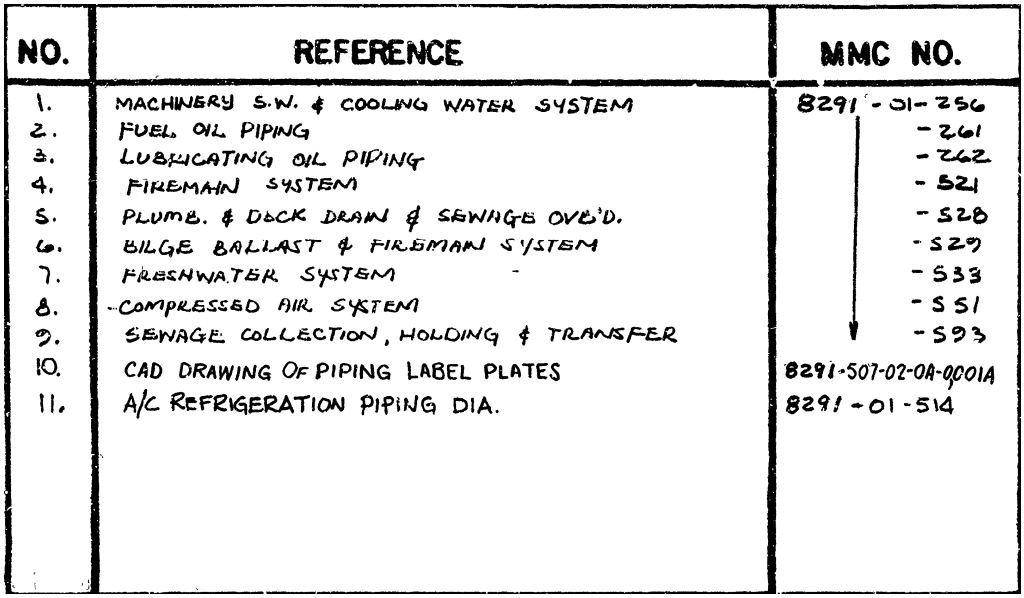


Figure D6. *Title block of the type H containing additional FSCM and MMC numbers in the A field.*

Figure D6 shows extra numbers in the A field. Although permitted by the specifications, it adds additional complexity to the automated extraction process. For example, in this case, the contractor number (FSCM = federal supply code for manufacturers) has been embedded in the A field in addition to the MMC number. Again this type of block needs additional field description file for automated information extraction.

References in reference blocks are defined from the bottom up with increasing number of fields as seen in *Figure A5*. In some cases, a reference block has different structure shown in *Figure D7*. For instance, the structure can be described by items that are not separated by lines, missing drawing number, and inconsistent formatting of the MMC numbers.



NO.	REFERENCE	MMC NO.
1.	MACHINERY S.W. & COOLING WATER SYSTEM	8291 - 01 - 256
2.	FUEL OIL PIPING	- 261
3.	LUBRICATING OIL PIPING	- 262
4.	FIREMAIN SYSTEM	- 521
5.	PLUMB. & DECK DRAIN & SEWAGE OVB'D.	- 528
6.	BILGE BALLAST & FIREMAIN SYSTEM	- 529
7.	FRESHWATER SYSTEM	- 533
8.	COMPRESSED AIR SYSTEM	- 551
9.	SEWAGE COLLECTION, HOLDING & TRANSFER	- 593
10.	CAD DRAWING OF PIPING LABEL PLATES	8291-507-02-0A-Q001A
11.	A/C REFRIGERATION PIPING DIA.	8291 - 01 - 514

Figure D7. *Reference block with a different layout compared to the most common type (see Figure A5).*

In order to extract information automatically from reference blocks would require solving handwritten character and word recognition problems and resolving typographical and printing inconsistencies (e.g., missing dots - MMC vs. M.M.C.). These problems could be handled by post-processing routines.

13 References

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

2. AutoHotkey is an open source scripting language for Windows, 2009, URL: http://www.autohotkey.com/

3. DROID is a software tool to perform automated batch identification of file formats. 2009, URL: http://droid.sourceforge.net/wiki/index.php/Introduction; PRONOM is a resource registry (information) about the file formats, software products and other technical components. 2006, URL:http://www.nationalarchives.gov.uk/aboutapps/pronom/puid.htm

4. K. McHenry and P. Bajcsy, "3D Data Analysis," WVU/NETL/ERA Workshop on Digital Preservation of Complex Engineering Data, April 21-22, 2009, Morgantown, WV

5. L. Najman, O. Gibot, and S. Berche. *Indexing Technical Drawings Using Title Block Structure Recognition*. in *Proc. 6th International Conference on Document Analysis and Recognition (ICDAR'01)*. 2001. Seattle (USA).

6. T.F. Syeda-Mahmood. Extracting Indexing Keywords from Image Structures in Engineering Drawings. in Proc. 5th International Conference on Document Analysis and Recognition (ICDAR'99). 1999. Bangalore, India.

7. T.F. Syeda-Mahmood. Locating Indexing Structures in Engineering Drawing Databases Using Location Hashing. in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'99). 1999. Fort Collins, CO, USA.

8. RDF Gravity, 2004, URL: http://semweb.salzburgresearch.at/apps/rdf-gravity/index.html.

9. Welkin. graph-based RDF visualize. 2008, URL: Available from: http://simile.mit.edu/welkin/

10. RDF semantics. 2004, URL: Available from: http://www.w3.org/TR/rdf-mt/

11. D. Alemang and J. Hendler, eds. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL. 2009, Morgan Kaufmann: San Francisco

12. TopSoft, , a free OCR Software for cameras or smartphones by TopSoft, Ltd.,2009, URL: http://www.topocr.com/

13. Readiris OCR Software demo version, 2009, URL: http://www.irislink.com/

14. ABBYY FineReader 9.0, 2009, URL: http://finereader.abbyy.com/

15. Tesseract/ OCRopus is an OCR engine developed at the HP Labs between 1985 and 1995. In 2005, HP made Tesseract open source (Apache License). URL: http://sites.google.com/site/ocropus/install-02

16. Gamera, a toolkit for building document image recognition systems. We tested training phase of the Gamera framework only. Gamera, 2009, URL: http://gamera.informatik.hsnr.de/

17. RDF vocabulary, rdf: URL, http://www.w3.org/1999/02/22-rdf-syntax-ns#, Dublin Core Metadata Element Set, dc:, URL, http://xmlns.com/foaf/0.1/, Friend of a Friend project, foaf: URL, http://purl.org/dc/elements/1.1/

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

19. J. Futrelle, Harvesting RDF Triples, IPAW'06 International Provenance and Annotation Workshop 2006, 64; Tupelo, 2009, URL: http://tupeloproject.ncsa.uiuc.edu/

20. M. Ondrejcek, J. Kastner and P. Bajcsy, "A Methodology for File Relationship Discovery," the 5th International IEEE eScience conference, Oxford, UK, Dec 9-11. 2009.

21. DoD-STD-100C, *Engineering Drawing Practices,* MIL-HDBK-1006/*1, Policy and Procedures for Project Drawing and Specification Preparation*. 1978, replaced by DOD-STD-00100D, see for example URL http://www.everyspec.com/DoD/DoD-STD/; see also URL, http://www.tpub.com/engbas/3-12.htm

22. RDFValidator, W3C RDF Validation Service, 009, URL http://www.w3.org/RDF/Validator

23. ASME Y14/ANSI Y14, *American National Standard Engineering and Related Document Practices*, 2000, URL http://webstore.ansi.org/

24. ISO 128, *Technical drawings - General principles of presentation*, 2003; ISO 7200, *Technical product documentation -- Data fields in title blocks and document headers*, 2004, International Standard Organization (ISO), URL http://www.iso.org/iso/home.htm

25. ProfiCAD, 2009, URL: http://www.proficad.com/index.html