



# **Technologies For Appraising and Managing Electronic Records**

**Presented by: Peter Bajcsy**

**-Research Scientist at NCSA**

**-Associate Director of I-CHASS, I3  
Institute**

**-Adjunct Assistant Professor, CS & ECE  
UIUC**

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



# Acknowledgement

- This research was partially supported by a National Archive and Records Administration (NARA) supplement to NSF PACI cooperative agreement CA #SCI-9619019 and NCSA Industrial Partners.
- 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 Archive and Records Administration, or the U.S. government.
- Contributions by: Peter Bajcsy, Kenton McHenry, Rob Kooper, Michal Ondrejcek, Jason Kastner, William McFadden, and Sang-Chul Lee

# Outline

- **Introduction**
- **A discovery of relationships among digital file collections (file2learn)**
- **A comprehensive comparison of contemporary documents (doc2learn)**
- **Automated file format conversions and conversion quality assessment (Polyglot)**
- **Summary**

# Introduction

## Supporting NARA's Strategic Plan

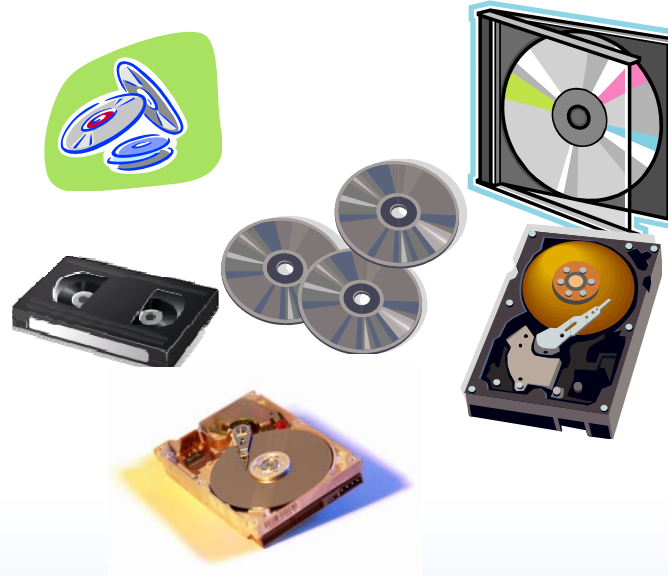
- According to *The Strategic Plan of The National Archives and Records Administration 2006–2016*. “Preserving the Past to Protect the Future”
  - “**Strategic Goal:** We will preserve and process records to ensure access by the public as soon as legally possible”
    - “D. We will improve the efficiency with which we manage our holdings from the time they are scheduled through accessioning, processing, storage, preservation, and public use.”

# To Be Preserved!

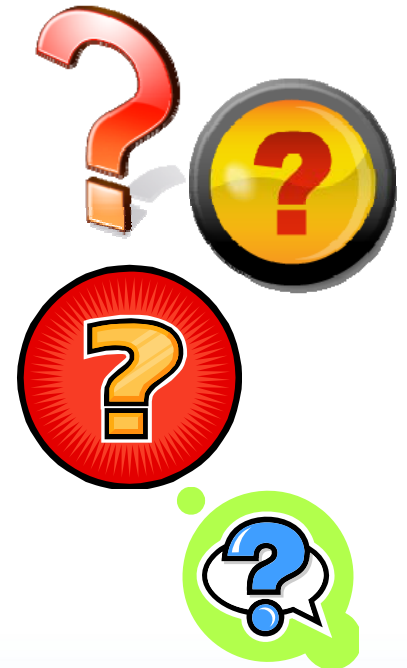


Information  
transfer ?

Digital representation of  
information  
& knowledge



Preservation



AGENCY

ARCHIVES

## Do We Know the Answers?

### Questions During Appraisal of Electronic Records Series

- (1) Given M full DVDs with files, **which files are related?**
- (2) Given N versions of the 'same' file, **which file version(s) should be preserved?**

## Do We Know the Answers?

- (3) Given P file formats, **which file format to use and which conversion software to use** so files would be possible to view in a long run?
  - **How much information is lost during file format conversion?**
- (4) **What is the granularity of information** that one should preserve about a decision process in order to reconstruct it?



# Goal: Design Technologies for Appraising and Managing Electronic Records

- Technologies should address the following problems:
  - (1) a discovery of relationships among digital file collections (file2learn)
  - (2) a comprehensive comparison of contemporary documents (doc2learn)
  - (3) automated file format conversions and conversion quality assessment (Polyglot)

# **A Discovery of Relationships Among Digital File Collections**

# Discovering Relationships Among Files

- How should one establish **relationships among electronic records** coming
  - From disparate sources or
  - From the same source at multiple time instances?
- **Need to Understand the Complexity of the Problem**

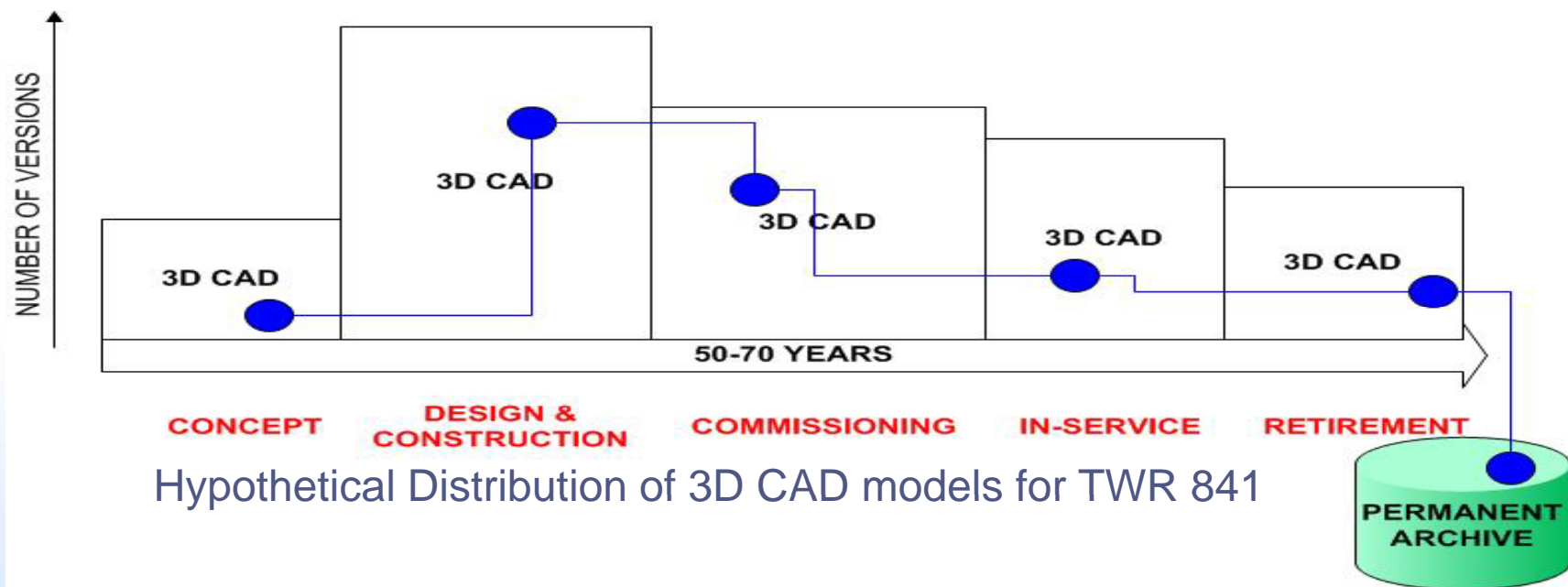
# Discovering Relationships Among Files:

## Components

- Metadata describing electronic records
  - How to extract metadata?
  - How to automate metadata extraction from multiple data types, e.g., 2D drawings and 3D CAD models?
- Storage of metadata
  - What ontology to use to represent the extracted metadata?
  - How to represent and store data and metadata?
- Exploratory and Search Capabilities
  - How to automate discovery of relationships?
  - How to support discovery of relationships between electronic records corresponding to the same physical objects but different multidimensional observations?

# Relationships Among Multiple Data Types

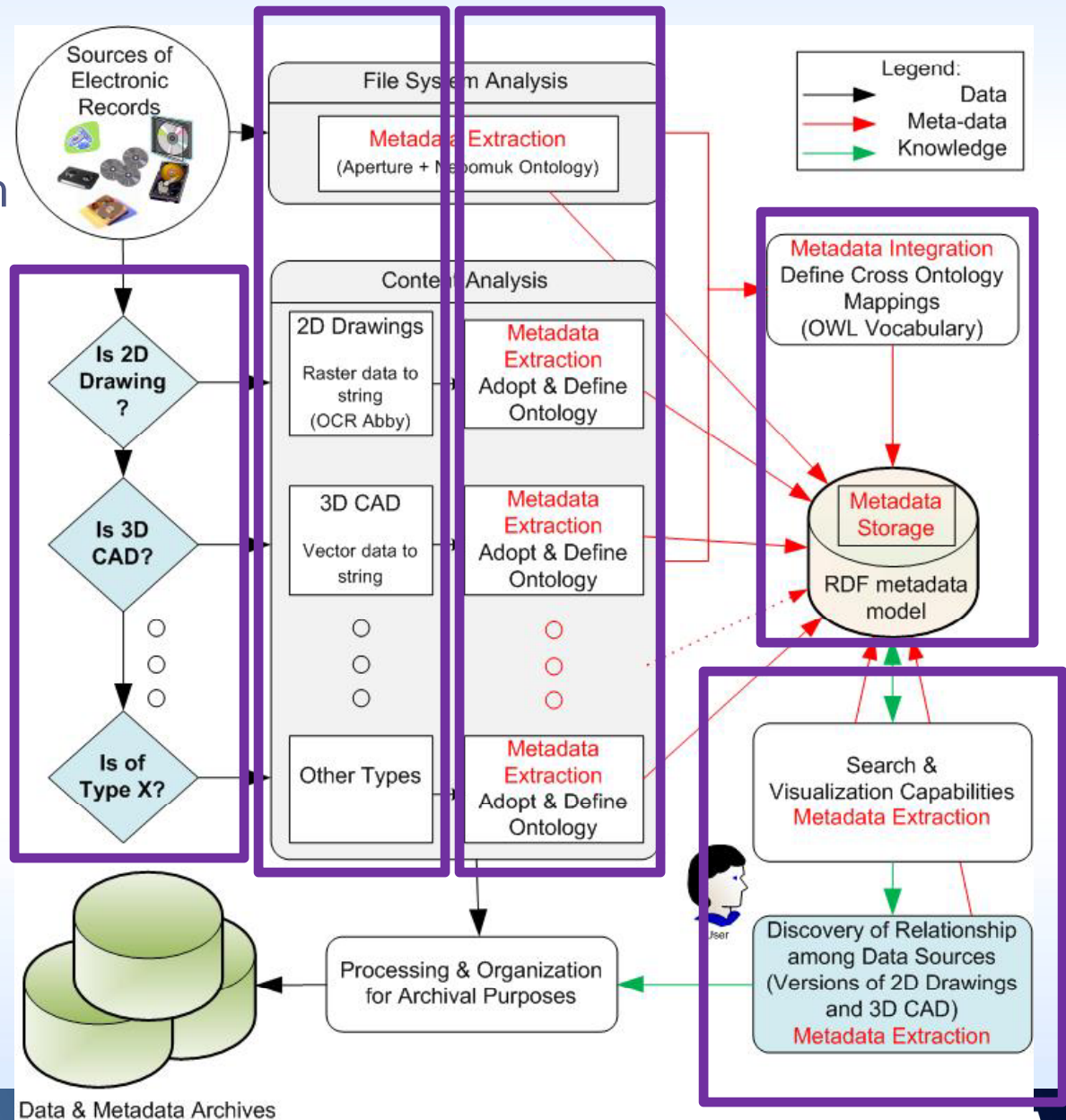
- **Example Data:** Torpedo Weapon Retriever 841
  - 784 existing 2D image drawings and N>22 3D CAD models
- How to establish relationships among the 3D CAD models and 2D image drawings during a product lifecycle?



Hypothetical Distribution of 3D CAD models for TWR 841

# Methodology

- File Identification
- Information Extraction from
  - File System
  - File Content
- Information Organization
  - Taxonomy (classification)
  - Ontology (relationships)
- Information Representation, Integration and Storage
  - XML
  - RDF
- Relationship Discovery

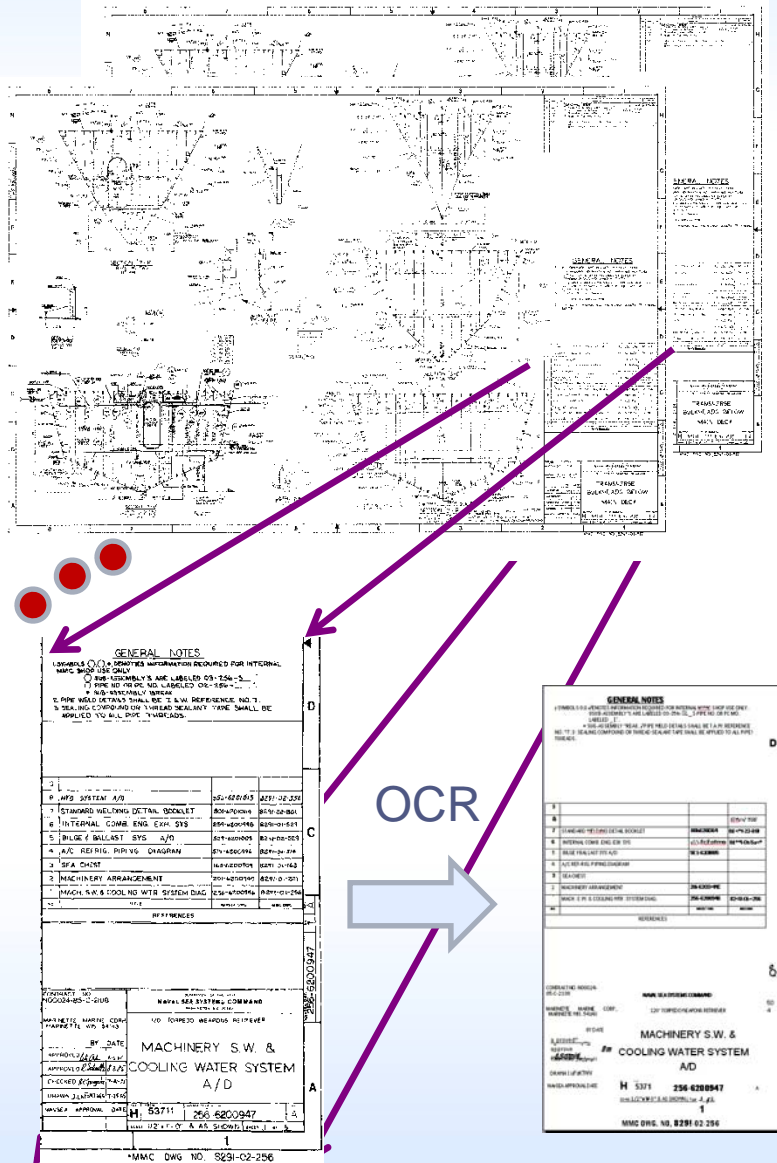


# File Identification and File System Analyses

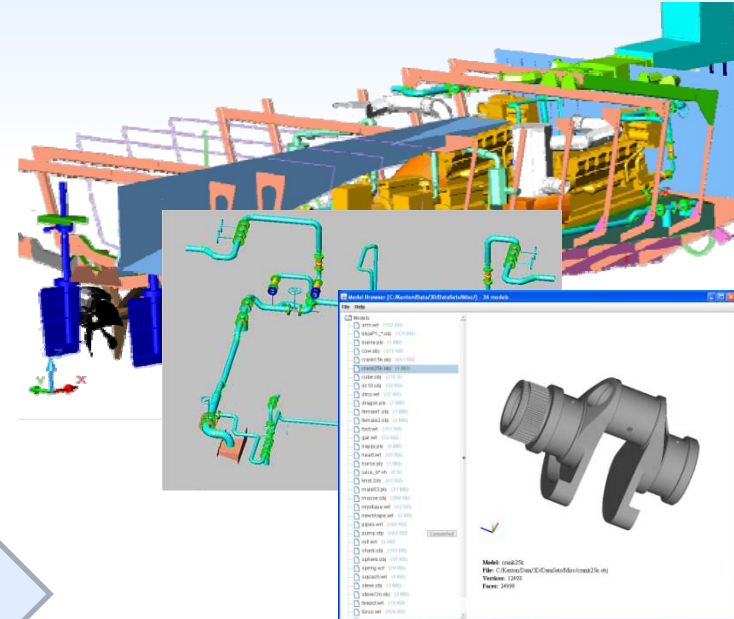
- **File Identification**
  - What is the file format?
  - Is the file format well formed?
- **Approach:** Used DROID built on top of the PRONOM File Registry with additional NCSA support of 3D file identification
- **Metadata extraction about a file system**
  - Where is the file located?
  - What is the file size, time stamp, etc.?
- **Approach:** Use any file system information extraction software, such as **Aperture (cross platform, open source, active development)**, Google desktop, OS specific solutions (e.g., Apple Spotlight, Linux, MS Search)



# Content Analyses: Automation ?



# Relationship Discovery



Part name,  
Author,  
Software,  
Date, ...

## File Descriptors



# Content Analyses: Optical Character Recognition (OCR) of 2D Drawings

16	SHAFTING ARRANGEMENT	243-6200942	01-243
15	BOW THRUSTER ARRANGEMENT	568-6201020	01-568
14	INTERNAL COMBUSTION ENG. EXH. SYSTEM A/D	259-6200940	01-259
13	LONGTL FRAMING & GIRDERS	116-6200994	01-116
12	MOORING & TOWING A/D	582-6201023	01-582
11	STEERING GEAR ARRANGEMENT	561-6201018	01-561
10	SHAFT STRUTS & BARRELS	161-6200903	01-161
9	SEA CHEST	163-6200904	01-163
8	TRANSV. ORDINARY FRAMES	117-6200896	02-117
7	LONGTL BHD. BEL. MAIN DECK	121-6200897	01-121
6	MAIN DECK STRUCTURE	130-6200898	01-130
5	TRANSV BHD. BELOW MAIN DECK	117-6200895	01-117
4	INNER BOTTOM & FIRST PLATFORM	113-6200892	01-113
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

CONTRACT NO N00024-85-C-2108		DEPARTMENT OF THE NAVY NAVAL SEA SYSTEMS COMMAND WASHINGTON DC 20362	
MARINETTE MARINE CORP. MARINETTE WIS 54143		TITLE 120' TOPPEDQ. WEAPONS RETRIEVER	
BY DATE		SHELL EXPANSION	
APPROVED	3/15/85		
CHECKED	3-15-85		
DRAWN	T.B. 3-15-85		
NAVSEA APPROVAL DATE			
H 53711		111-6200891	
SCALE 1/2" = 1'-0" & AS SHOWN		MMC DWG. NO. 78291-01-111	

The main drawing shows a plan view of a weapons retriever structure. It includes various annotations such as 'PLAN VIEW "5-G" BULKHEAD CAP PLATE', 'SECTION "A-A"', 'SECTION "B-B"', and 'SECTION "C-C"'. There are also dimension lines and callouts for specific parts. A 'GENERAL NOTES' section is located on the right side of the drawing. The drawing is framed by a grid with letters A through H and numbers 1 through 7.

Indows Picture and Fax Viewer

Reference Block

Title Block

MMC Block (Marinette Marine Corporation)

# 'Standard' Title Blocks: Organization and Ontology

- Examples of title blocks used on drawings prepared by Naval Construction Battalion and Naval Construction Regiment

CONTRACT NO. N00024-85-C-2108		DEPARTMENT OF THE NAVY NAVAL SEA SYSTEMS COMMAND WASHINGTON D.C. 20342	
MARINETTE MARINE CORP MARINETTE WIS. 54143		120 TORPEDO WEAPONS RETRIEVER	
BY DATE		CONTRACT NO. N00024-85-C-2108	
APPROVED	4-23-85	DEPARTMENT OF THE NAVY NAVAL SEA SYSTEMS COMMAND WASHINGTON D.C. 20342	
APPROVED	4-23-85	120 TORPEDO WEAPONS RETRIEVER	
CHECKED	4-22-85	MARINETTE MARINE CORP MARINETTE WIS. 54143	
DRAWN	4-10-85	BY DATE	
NAVSEA APPROVAL	DATE	MACHINERY S.W. & COOLING WATER SYSTEM A/D	
H	53711	APPROVED	
SCALE 1/2"=1'-0" & AS SHOWN		CHECKED	
		DRAWN	
		NAVSEA APPROVAL	
		DATE	
		H	
		53711	
		256-6200947	
		A	
		SCALE 1/2"=1'-0" & AS SHOWN	
		SHEET 1 OF 5	

\*MMC DWG. NO. 8291-02-256

## TEMPLATES

DSGN		DEPARTMENT OF THE NAVY		WASHINGTON, D.C.	
DR		NAVAL FACILITIES ENGINEERING COMMAND			
CHK					
PROJ LDR					
BR HD					
SPL DES HD					
DIRECTOR					
APPROVED	DATE	SIZE	CODE IDENT NO.	NAVFAC DRAWING NO.	
			80091		
FOR COMMANDER, NAVFAC		SCALE	SPEC	SHEET OF	

MCB - DRAWING NO.		DEPARTMENT OF THE NAVY		NAVAL FACILITIES ENGINEERING COMMAND	
		US NAVAL MOBILE CONSTRUCTION BATTALION			
		FPO SAN FRANCISCO			
		CALIFORNIA, U.S.A.			
DES					
QA					
CHK					
ENGR OFF					
OPERATIONS OFF					
APPROVED	DATE	SIZE	CODE IDENT NO.	NAVFAC DRAWING NO.	
			80091		
COMMANDING OFFICER		SCALE	SPEC	SHEET OF	
SATISFACTORY TO		DATE	CONSTRUCTION CONTR NO.		

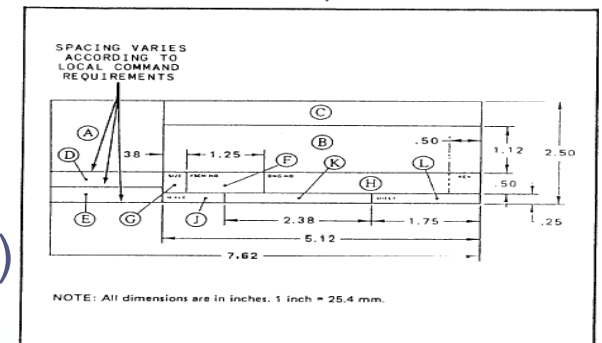
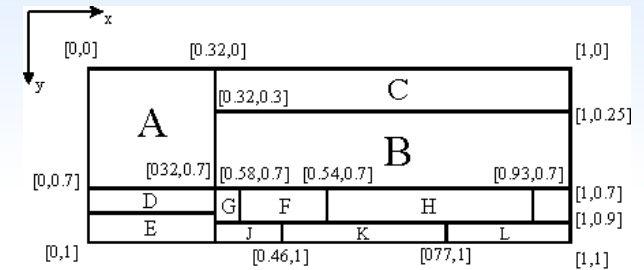
PWO DWG REF		DEPARTMENT OF THE NAVY		NAVAL FACILITIES ENGINEERING COMMAND	
		NAVAL SHIPYARD, LONG BEACH, CALIF			
SPACE SUBDIVIDED TO SUIT PRACTICE OF PWO					
APPROVED	DATE	SIZE	CODE IDENT NO.	NAVFAC DRAWING NO.	
			80091		
OFFICER IN CHARGE		SCALE	SPEC	SHEET OF	

REQUEST NO.		31st NAVAL CONSTRUCTION REGIMENT			
REF. DRG		PORT HUENEME, CALIFORNIA			
DESIGNED BY					
DRAWN BY					
CHECKED BY					
DEPT. HEAD					
SATISFACTORY TO		DATE			
TITLE		PWKS APPROVED		DATE	
31 NCR APPROVED		DATE		31 NCR DWG. NO.	
OPERATIONS OFFICER		SCALE		SHEET OF	

# Title Block: Ontology and Metadata Representation

## *Ontology for sub-fields:*

- A – Record of preparation (<tdrw:recordOfPreparation>),
- B – Drawing title (<tdrw:drawingTitle>),
- C – Preparing Activity <tdrw:preparingActivity>,
- F – Code identification number (<tdrw:FSCMNumber> ),
- G – Drawing size (<tdrw:drawingSize>),
- H – Drawing number (<tdrw:drawingNumber>),
- J – Scale (<tdrw:drawingScale>),
- K – Specification number (<tdrw:drawingNumber>)
- L – Sheet number (<tdrw:sheetNumber>).



## *Resource Description Framework (RDF):*

- Metadata representation: subject – predicate - object

- MMC Blocks

## Inconsistencies

- The list varies in length
- The notation is not standardized

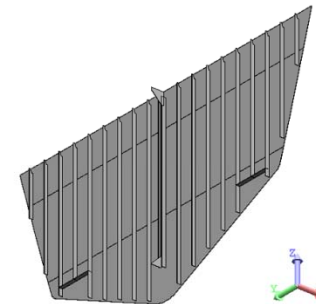


# Summary of OCR Based Analyses

- **Manually encoded block coordinates for 784 files** in PNG (converted from originally LZW compressed TIFF files)
- **Automated OCR and executed OCR on**
  - 700 title blocks,
  - 150 reference blocks,
  - dozen of revision and list of material
  - about 200 additional areas with the drawing numbers (MMC DWG. NO.).
- **Performance benchmarks:**
  - Full OCR of TB, MMC and RF for about 50 image files (105 blocks) took about 6 hours on a quad core machine

# Content Based Extraction from STEP Files

- 3D CAD models in STEP file format are searched for any ASCII strings matching English dictionary and following STEP metadata specification.



Example Metadata for TWR841 ship deck

STEP METADATA SPECIFICATION	EXPECTED STEP METADATA	PARSED STEP METADATA
<b>FILE_DESCRIPTION</b> ( /* description */ ("), /* implementation_level */ '2;1'); <b>FILE_NAME</b> ( /* name */ "  /* time_stamp */ " /* author */ (" /* organization */ (" /* preprocessor_version */ ' /* originating_system */ " /* authorization */ ' ');	<b>FILE_DESCRIPTION</b> (( /* implementation_level */ '2;1'); <b>FILE_NAME</b> ( '120 TORPEDO WEAPONS RETRIEVER, TRANSVERSE BULKHEADS BELOW, MAIN DECK', '04-10-86', ( <b>'LDOBSON'</b> ), ( <b>'NAVAL SEA SYSTEMS COMMAND'</b> ), '', 'IDA-STEP', '');	<b>FILE_DESCRIPTION</b> (( '2;1'); <b>FILE_NAME</b> ( 'D:\\NARA\\Archive_data_samples\\BHD_FR12\\ U2110_BHD12_2007_05_09.stp',  '2007-05-10T13:45:37', ( <b>'rakowpj'</b> ), ('), 'Autodesk Inventor 11', 'Autodesk Inventor 11', ");

# Exploratory Framework – User Interface Overview

The screenshot displays the File2Learn application interface, which is divided into several panes:

- Left Pane (Files):** A list of files with a red box labeled "Filter for Files" at the top. The list includes: Rudder\_Arms\_dwg.jpg, Rudder\_Assembly.stp, Rudder\_Plate.jpg, Rudder\_Plate.stp, Rudder\_Plate\_02.jpg, Rudder\_plate\_dwg.jpg, Rudders\_w\_hull.jpg, Starboard\_Rudder\_Stock.stp, TWR\_aft\_port.jpg, TWR\_aft\_stbd.jpg, Tiller\_Weldment.stp, Tiller\_Weldments.jpg, Tiller\_weldment\_dwg.jpg, Upper\_Rudder\_Arm.stp, and lockfile.lck. A red label "Files" is positioned to the right of this list.
- Central Pane (Graph of Relationships):** A graph titled "Graph of Relationships Between Selected Files". It shows nodes representing files and their relationships. Key nodes include: "file/5f3f99df-043a-4a3e-b013-37e587a3b0da/Tiller\_Weldment.stp", "project.org,2006:/2.0/files/File", "Tiller\_Weldment.stp", "Tiller\_Weldments.jpg", "2009:file/52503b29-e8b4-4ca7-81d1-a27084f6e187/Tiller\_Weldments.jpg", "Positive", and "ISDA\_2D". Relationships are labeled with URIs such as "droid:/IdFile/hasIdentQuality", "http://isda.ncsa.uiuc.edu/2009/isda/hasIsdaClassification", "rdf:type", and "droid:/IdFile/hasFileName". A red label "Files" is also present to the right of the graph.
- Right Pane (Files):** A list of files with a red box labeled "Filter for Files" at the top. The list includes: Rudder\_Arms\_02.jpg, Rudder\_Arms\_dwg.jpg, Rudder\_Assembly.stp, Rudder\_Plate.jpg, Rudder\_Plate.stp, Rudder\_Plate\_02.jpg, Rudder\_plate\_dwg.jpg, Rudders\_w\_hull.jpg, Starboard\_Rudder\_Stock.stp, TWR\_aft\_port.jpg, TWR\_aft\_stbd.jpg, Tiller\_Weldment.stp, Tiller\_Weldments.jpg, Tiller\_weldment\_dwg.jpg, Upper\_Rudder\_Arm.stp, and lockfile.lck. A red label "Files" is positioned to the right of this list.
- Bottom Left Pane (Preview of Selected Data):** A 3D model of a mechanical part, labeled "Preview of Selected Data".
- Bottom Right Pane (Preview of Selected Data):** A 3D model of a mechanical part, labeled "Preview of Selected Data".

The interface also includes a menu bar (File, Window, Help) and a status bar at the bottom right showing "NaraDemo".

# Exploratory Framework – User Interface Overview

**File2Learn**  
File Window Help

**Additional Import/Export and Preference Options**

type filter text

- Rudder\_Arms\_dwg.jpg
- Rudder\_Assembly.stp
- Rudder\_Plate.jpg
- Rudder\_Plate.stp
- Rudder\_Plate\_02.jpg
- Rudder\_plate\_dwg.jpg
- Rudders\_w\_hull.jpg
- Starboard\_Rudder\_Stock.stp
- TWR\_aft\_port.jpg
- TWR\_aft\_stbd.jpg
- Tiller\_Weldment.stp
- Tiller\_Weldments.jpg
- Tiller\_weldment\_dwg.jpg
- Upper\_Rudder\_Arm.stp
- lockfile.lck

URI :  
<< Link Resources >>

Predicate	Left Resource	Right Resource
http://purl.org/creator	tag://isda.ncsa.uiuc.edu,2...	
http://isda.ncsa		
hasFileName	Tiller_Weldment.stp	Tiller_Weldments.jpg
hasFilePath	C:/NARADemo/RG 428 - G...	C:/NARADemo/RG 428 - Ge
hasIdentQ	Not Identified	Positive
tag		
URI	tag://isda.ncsa.uiuc.edu,2...	tag://isda.ncsa.uiuc.edu,20
http://isda.ncsa		
schema	AUTOMOTIVE_DESIGN { 1...	
http://isda.ncsa		
implemental	2;1	
http://isda.ncsa		
extensionU	Standard for the Exchang...	
hasExtensio	isda/154	
hasIsdaClas	ISDA_3D	ISDA_2D
hasIsdaId	http://kastner-wxp.ncsa....	
http://isda.ncsa		
name	C:\\Documents and Settin...	
originatingS	Autodesk Inventor 11	
preprocess	Autodesk Inventor 11	
timestamp	2007-08-15T07:31:08	
http://isda.ncsa		
hasFormatN		JPEG File Interchange Form
hasFormatV		1.02
hasMimeTyp		image/jpeg
hasPronomI		http://www.nationalarchiv
hasStatus		Positive (Specific Format)
http://www.w3		
22-rdf-synt	tag:tupeloproject.org,200...	tag:tupeloproject.org,2006

Table of Relationships  
Between Selected Files

type filter text

- Rudder\_Arms\_02.jpg
- Rudder\_Arms\_dwg.jpg
- Rudder\_Assembly.stp
- Rudder\_Plate.jpg
- Rudder\_Plate.stp
- Rudder\_Plate\_02.jpg
- Rudder\_plate\_dwg.jpg
- Rudders\_w\_hull.jpg
- Starboard\_Rudder\_Stock.stp
- TWR\_aft\_port.jpg
- TWR\_aft\_stbd.jpg
- Tiller\_Weldment.stp
- Tiller\_Weldments.jpg
- Tiller\_weldment\_dwg.jpg
- Upper\_Rudder\_Arm.stp
- lockfile.lck

Data Viewer

NaraDemo



# Exploratory Framework: Modes of Operations

- **Detection of discrepancies/anomalies in file descriptors**
  - OCR results
    - View 2D drawings and OCR results, and then edit OCR descriptors
  - 3D Model
    - View 3D model and content based extraction, and then edit descriptors
- **Comparison of pairs of files**
  - Pairs of 2D drawings
  - Pairs of 3D models
  - Pairs of (2D drawing, 3D model)
- **Establish file relationships**
  - Insert logical links to relate a pair of files

# Detection of Anomalies in OCR Results

The screenshot displays the File2Learn application interface, which is used for analyzing and visualizing file metadata and content. The interface is divided into several panes:

- Types Pane (Left):** Lists file types and their associated files. A red box highlights the file `6200891-01.png` under the `isda/MMCBlock` category.
- Context View Pane (Center):** Displays a table of metadata for the selected file. A red box highlights the `URI` field, which contains the value `http://isda.ncsa.uiuc.edu/NARA/tdrw/74`. Another red box highlights the `MMCN` field, which contains the value `MMCN DWG. NO. 291-01-111`.
- Types Pane (Right):** Lists file types and their associated files. A red box highlights the file `6200891-01.png` under the `isda/MMCBlock` category.
- Data Viewer Pane (Bottom):** Displays the content of the selected file. A red box highlights the text `NO. 78291-01-111`. A purple box highlights a larger section of the document, which appears to be a technical drawing or diagram.

The application is titled "File2Learn" and includes a menu bar with "File", "Window", and "Help". The status bar at the bottom right indicates "NaraDemo".

# Comparison of Files

Color encoding:

- **Predicates and values match**
- **Predicates match**
- **Predicate occurs only in one file**

File2Learn

File Window Help

Context View

URI :

<< Link Resources >>

Predicate	Left Resource	Right Resource
http://isda.ncsa.uiuc.edu/2009/droi		
hasFilePath	C:/OCR/6201019-02/6201019-02-tbc.png	C:/OCR/6201019-01/6201019-01-tb.png
tag		
URI	http://isda.ncsa.uiuc.edu/NARA/tdrw/97ec3247-eb1f-461a-...	http://isda.ncsa.uiuc.edu/NARA/tdrw/777fc5bd-fffe-4321-9928-be47afc78129/title
http://isda.ncsa.uiuc.edu/2009/isda		
LogicalLink		tag://isda.ncsa.uiuc.edu,2009:file/4b2ab164-44a0-4f8d-b8d5-34cc301a89b4/Upper
http://www.w3.org/1999/02		
22-rdf-syntax-ns#type	http://isda.ncsa.uiuc.edu/NARA/tdrw/TitleBlock	http://isda.ncsa.uiuc.edu/NARA/tdrw/TitleBlock
http://isda.ncsa.uiuc.edu/NARA/tdr		
FSCMNumber	53711	53711
drawingNumber	562-6201019	562-6201019
drawingScale	SCALE 1/2"=1'-0" & AS SHOWN	SCALE 1/2"=1'-0" & AS SHOWN
drawingSize	H	H
drawingTitle		DEPARTMENT OF THE NAVY □ NAVAL SEA SYSTEMS COMMAND □ WASHINGTON OC
fieldInfo		NAVSEA APPROVAL DATE
miniBox	http://isda.ncsa.uiuc.edu/NARA/tdrw/97ec3247-eb1f-461a-...	http://isda.ncsa.uiuc.edu/NARA/tdrw/777fc5bd-fffe-4321-9928-be47afc78129/title
miniBox	http://isda.ncsa.uiuc.edu/NARA/tdrw/97ec3247-eb1f-461a-...	http://isda.ncsa.uiuc.edu/NARA/tdrw/777fc5bd-fffe-4321-9928-be47afc78129/title
miniBox		http://isda.ncsa.uiuc.edu/NARA/tdrw/777fc5bd-fffe-4321-9928-be47afc78129/title
miniBox		http://isda.ncsa.uiuc.edu/NARA/tdrw/777fc5bd-fffe-4321-9928-be47afc78129/title
preparingActivity		120' TORPEDO WEAPONS RETRIEVER □ RUDDER & STEERING GEAR DETAILS
recordOfPreparation		CONTRACT NO N00024 85-C-2108 □ MARINETTE MARINE CORP MARINETTE WIS 5
sheetNumber	SHEET 2 OF 2	SHEET 1 OF 2

Attributes Graph

NaraDemo

# Establish File Relationships

File2Learn

File Window Help

Types

type filter text


- Rudder\_Arms\_dwg.jpg
- Rudder\_Assembly.stp
- Rudder\_Plate.jpg
- Rudder\_Plate.stp
- Rudder\_Plate\_02.jpg
- Rudder\_plate\_dwg.jpg
- Rudders\_w\_hull.jpg
- Starboard\_Rudder\_Stock.stp
- TWR\_aft\_port.jpg
- TWR\_aft\_stbd.jpg
- Tiller\_Weldment.stp
- Tiller\_Weldments.jpg
- Upper\_Rudder\_Arm.stp
- lockfile.lck

Context View

<< Link Resources >>

Predicate	Left Resource	Right Resource
http://isda.ncsa.uiuc.edu/hasFileName	Upper_Rudder_Ar...	
http://isda.ncsa.uiuc.edu/hasFilePath	C:/NARADemo/RG...	C:/OCR/6201019-01/6201019-...
http://isda.ncsa.uiuc.edu/hasIdentQuality	Not identified	
http://purl.org/dc/element/creator	tag://isda.ncsa.uiu...	
tag	URI	tag://isda.ncsa.uiu... http://isda.ncsa.uiuc.edu/NARA...
http://isda.ncsa.uiuc.edu/schema	AUTOMOTIVE_DES...	
http://isda.ncsa.uiuc.edu/implementationLevel	2;1	
http://isda.ncsa.uiuc.edu/extensionUsedBy	Standard for the E...	
http://isda.ncsa.uiuc.edu/hasExtension	isda/154	
http://isda.ncsa.uiuc.edu/hasIsdaClassification	ISDA_3D	
http://isda.ncsa.uiuc.edu/hasIsdaId	http://jastner-wx...	
http://isda.ncsa.uiuc.edu/name	C:\\Documents an...	
http://isda.ncsa.uiuc.edu/originatingSystem	Autodesk Inventor...	
http://isda.ncsa.uiuc.edu/preprocessorVersion	Autodesk Inventor...	
http://isda.ncsa.uiuc.edu/timestamp	2007-08-15T07:29...	
http://www.w3.org/1999/22-rdf-syntax-ns#type	tag:tupeloproject...	http://isda.ncsa.uiuc.edu/NARA...
http://isda.ncsa.uiuc.edu/FSCMNumber		53711
http://isda.ncsa.uiuc.edu/drawingNumber		562-6201019
http://isda.ncsa.uiuc.edu/drawingScale		SCALE 1/2"=1'-0" & AS SHOWN
http://isda.ncsa.uiuc.edu/drawingSize		H
http://isda.ncsa.uiuc.edu/drawingTitle		DEPARTMENT OF THE NAVY □...
http://isda.ncsa.uiuc.edu/fieldInfo		NAVSEA APPROVAL DATE
http://isda.ncsa.uiuc.edu/miniBox		http://isda.ncsa.uiuc.edu/NARA...
http://isda.ncsa.uiuc.edu/miniBox		http://isda.ncsa.uiuc.edu/NARA...

Data Viewer

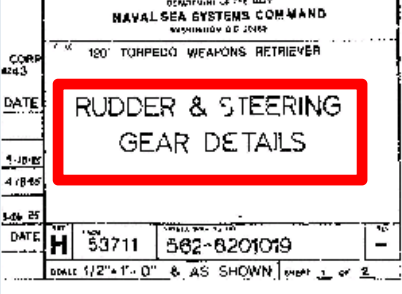


Types

type filter text

- isda//TitleBlock
  - 6200892-02.png
  - 6200895-01.png
  - 6200895-02.png
  - 6201018-01.png
  - 6201018-02.png
  - 6201019-01.png
  - isda//MMCBBlock
    - isda//ReferenceBlock
      - isda/reference/00
      - isda/reference/01
    - isda//TitleBlock
      - isda/actionBox/01
      - isda/actionBox/02
      - isda/actionBox/03
      - isda/actionBox/04
  - 6201019-02.png

Data Viewer



Attributes Graph

NaraDemo

# Establish File Relationships: Logical Link

The screenshot displays the File2Learn application interface, which is used for establishing logical links between files and their metadata. The interface is divided into several panes:

- Types:** A list of file types on the left and a hierarchical tree of metadata types on the right. The left list includes files like `Rudder_Arms_dwg.jpg`, `Rudder_Assembly.stp`, and `lockfile.lck`. The right tree shows a hierarchy starting with `isda//TitleBlock`, including sub-entries like `isda/MMCBlock`, `isda/ReferenceBlock`, and `isda/actionBox/01` through `isda/actionBox/04`.
- Context View:** The central pane showing a graph of relationships. A red box highlights a central node: `tdrw:777fc5bd-fffe-4321-9928-be47afc78129/titleblock`. This node is connected to other nodes via labeled edges:
  - An edge labeled `droid:/IdFile/hasFilePath` connects to the node `C:/OCR/6201019-01/6201019-01-tb.png`.
  - An edge labeled `rdf:type` connects to the node `tdrw:TitleBlock`.
  - An edge labeled `http://isda.ncsa.uiuc.edu/2009/isda/LogicalLink` connects to the node `u,2009:file/4b2ab164-44a0-4f8d-b8d5-34cc301a89b4/Upper_Rudder_Arm.stp`.
  - An edge labeled `rdf:type` connects to the node `t.org,2006:/2.0/files/File`.
  - An edge labeled `droid:/IdFile/hasFilePath` connects to the node `ent of the Navy/Drawings and Engineering Models for TWR 841/2007_08_28/Rudder_data/STEP Files`.
- Data Viewer:** The bottom right pane displays a technical drawing titled "RUDDER & STEERING GEAR DETAILS". The drawing includes a header for "NAVAL SEA SYSTEMS COMMAND" and a table with the following data:

COMP	DATE	DATE	DATE
4143	4/18/05	4/18/05	4/18/05
53711	562-6201019		

# **A Comprehensive Comparison of Contemporary Documents**

# Support of Appraisals by Enabling Comparisons

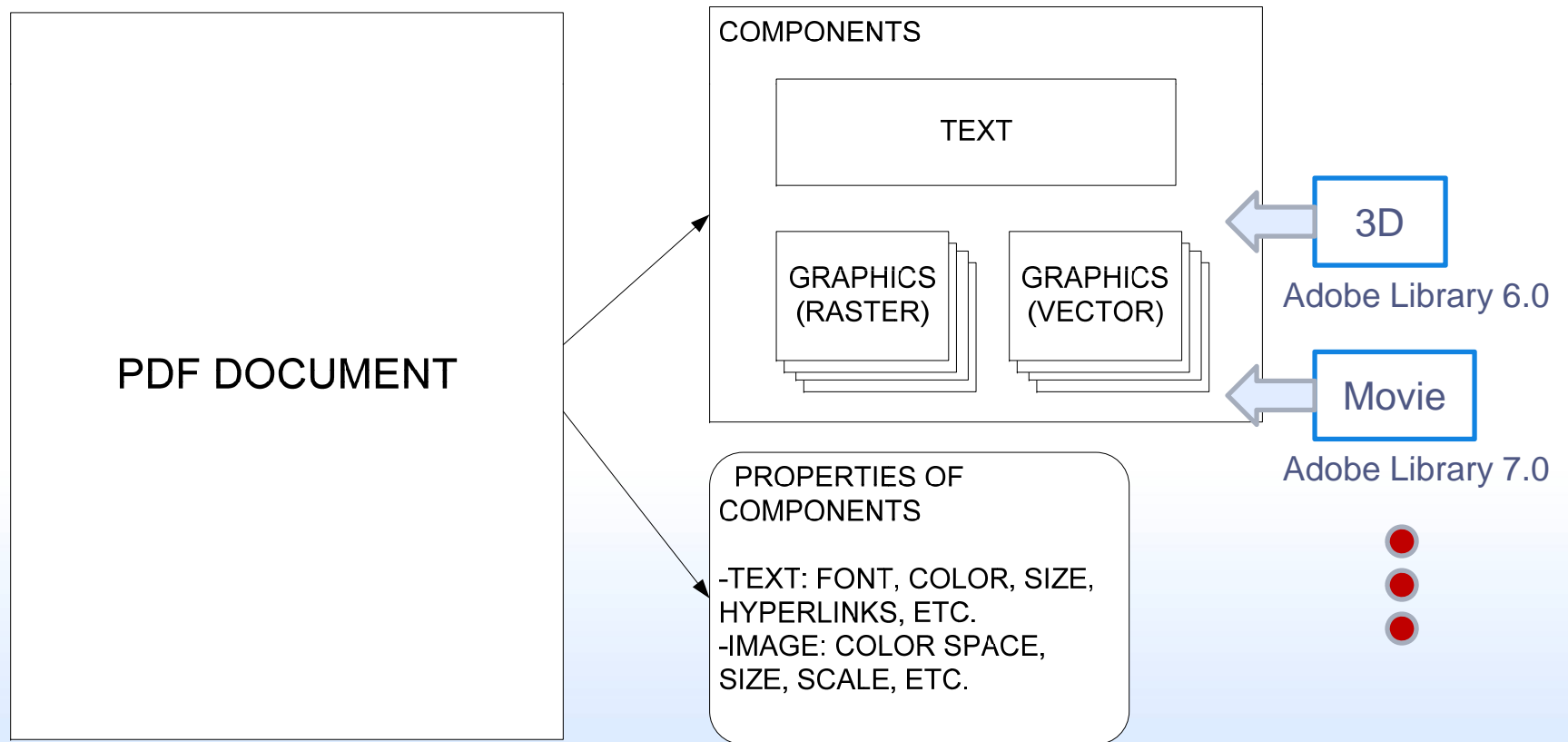
- **How to compare containers with heterogeneous information (text, images, vector graphics, animation, 3D, etc.)?**
  - Methodology
  - Metrics
  - Weighting factors for fusion
- **How to quantify similarities between the same type of information?**
  - Encodings and Representations
  - Metrics
  - Local versus global differences



One of the most important attribute is acc  
attributes to be ~~two~~twenty degrees.

# Example: Adobe Portable Document Format (PDF)

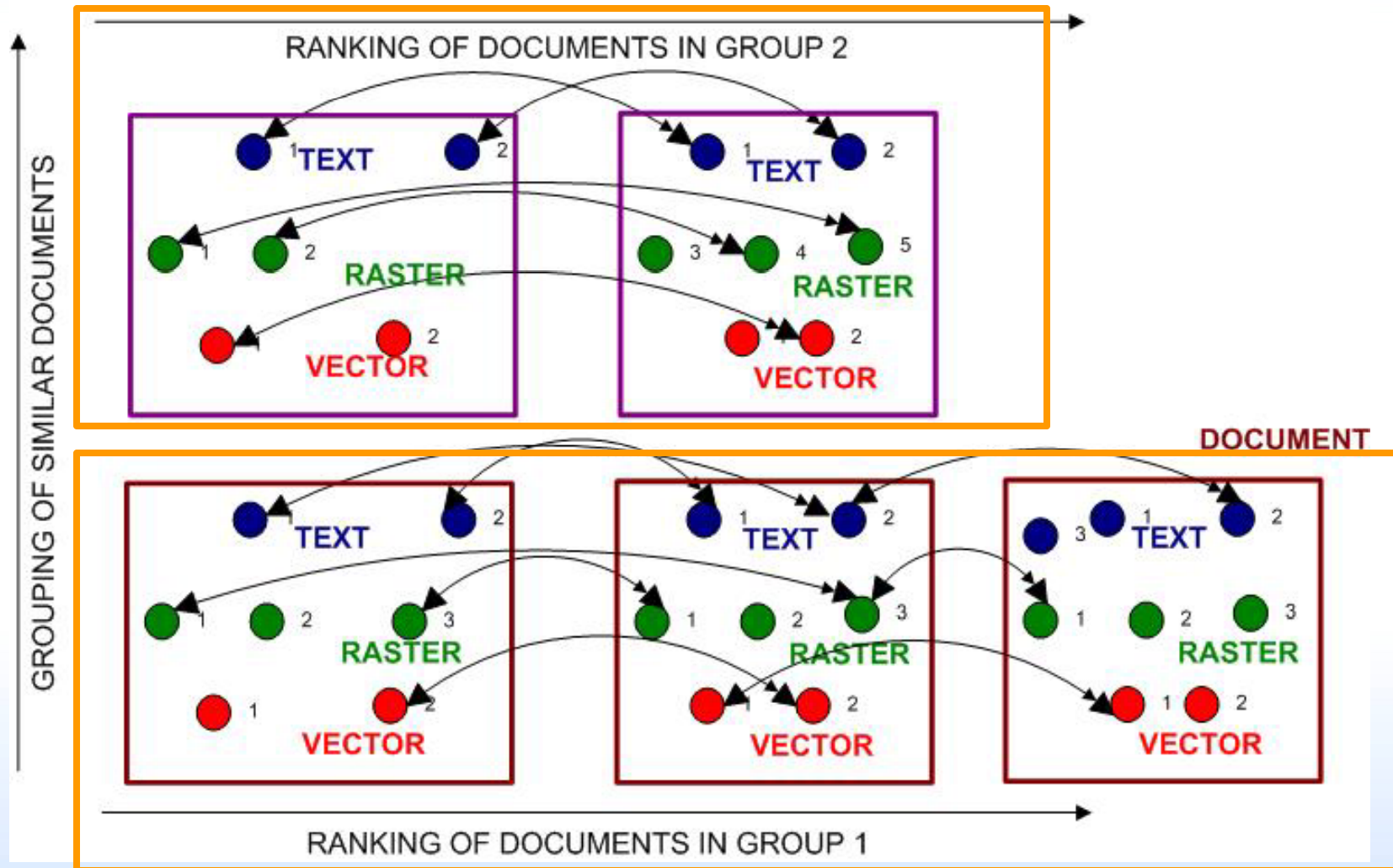
- **Why PDF? - PDF is just an example of a container**
  - Office environment (Adobe PDF, PS, MS Word, HTML, ...)
  - Satellite measurements (HDF, netCDF, ...)





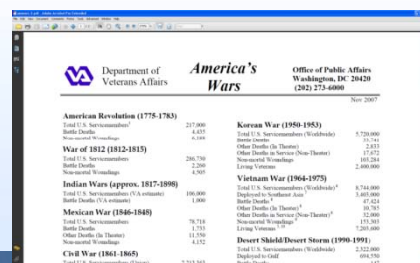
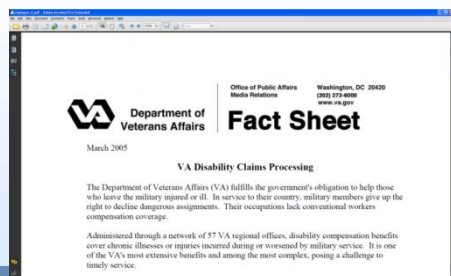
# Comparisons

WANTED: GROUPING, RANKING, AND INTEGRITY VERIFICATION



# Example: Compare Veterans Affairs Fact Sheets in PDF and MS Word file formats

- Test data: 108 files from RG 015 - Records of the Department of Veterans Affairs/Fact Sheets/www1.va.gov/opa/fact/docs.
  - These files are Veterans Affairs Fact Sheets and are stored in both PDF and MS Word file formats (54 MS word and 54 PDF files).
- **Which files have identical content?**
- **Demo: 6 files**
  - **amwars-2.pdf, amwars.pdf**
  - **claimpro-2.pdf, claimpro.pdf**
  - **combrates-2.pdf, combrates.pdf**



Department of Veterans Affairs  
Office of Public Affairs  
Media Relations  
Washington, DC 20420  
(800) 273-8686  
www.va.gov

## Fact Sheet

February 2006

### Disability Compensation — 7000 Rates

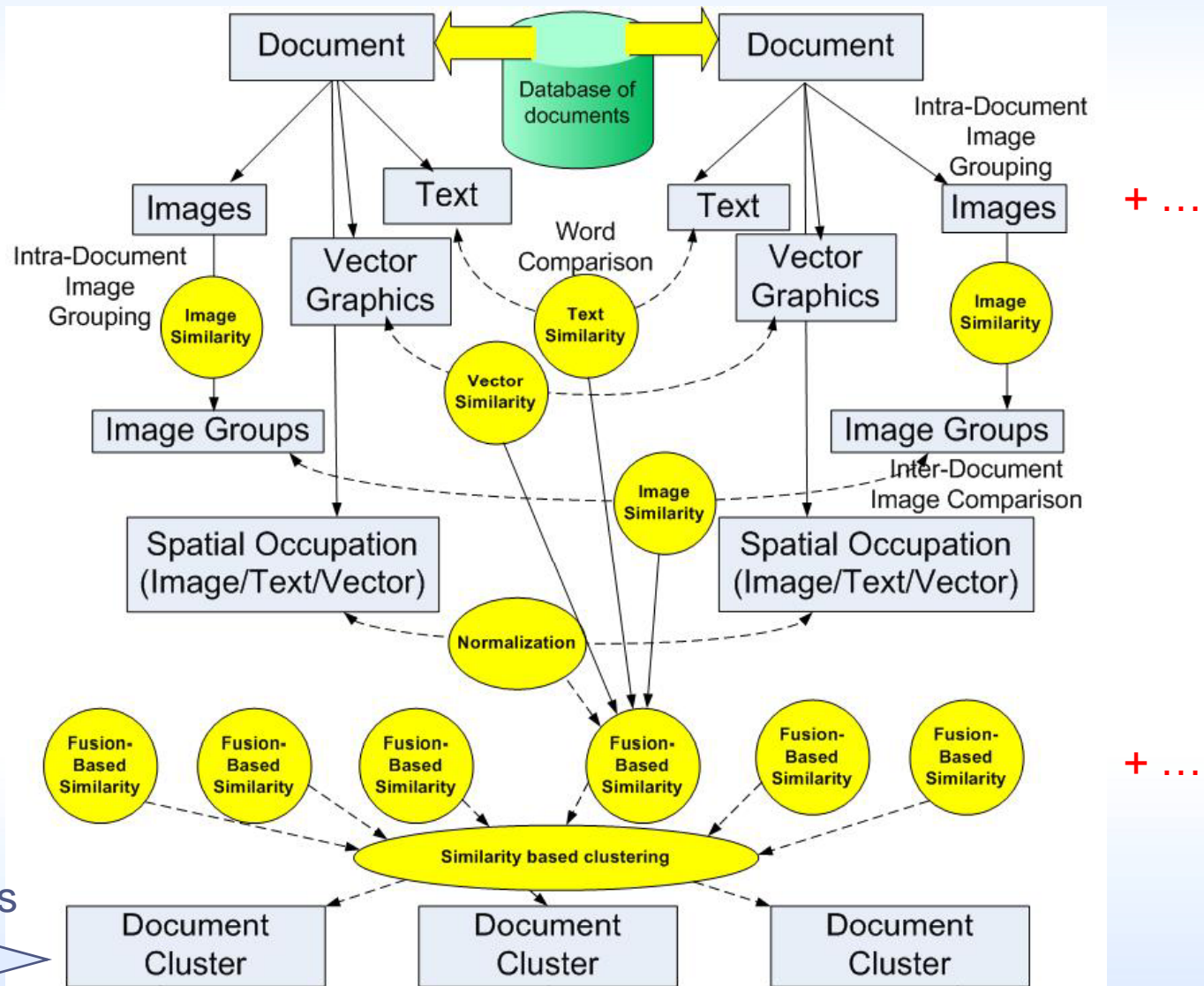
10 percent . . . \$112	60 percent . . . \$873
20 percent . . . 218	70 percent . . . 1,099
30 percent . . . 337	80 percent . . . 1,277
40 percent . . . 485	90 percent . . . 1,436
50 percent . . . 690	100 percent . . . 2,393

# Methodology

Pair-wise  
comparison  
of the same  
digital objects

Comparison of  
multiple and  
heterogeneous  
digital objects

Relationship to  
Permanent Records



# Exploration of Text Components

**Document Analyzer**

Load Launch Document Use Ignore List Compare Group Move to Ignore List Show Comparison Visualization Show Second Window Done Help

Done

Document List

No	Filename	File Size	Num Images
1	amwars-2.pdf		1
2	amwars.pdf		1
3	claimpro-2.pdf	57KB Fri May 22 08:49:59 CDT 2009	3
4	claimpro.pdf	57KB Fri May 22 08:49:59 CDT 2009	3
5	comprates-2.pdf	50KB Fri May 22 08:49:59 CDT 2009	3
6	comprates.pdf	50KB Fri May 22 08:49:59 CDT 2009	3

**Occurrence of words**

**Occurrence of numbers**

**Page List**

Page No	Number of Words
All	393
1	148
2	202
3	162

**All Frequency**

No	Word	Frequency
347	reopened	1
348	14000	1
349	capacity	1
350	starting	1
351	annually	1
352	100	2
353	2393	1
354	signed	1
355	90	1
356	evaluated	1
357	point	1
358	degree	1
359	practicably	1
360	fall	1
361	calculates	1
362	http://www....	1
363	based	2
364	but	2
365	workers	1
366	together	1
367	though	1
368	\$6845	1
369	were	2
370	hearing	1
371	injuries	1
372	regulation	1
373	other	3
374	have	4
375	specific	1
376	one	4
377	state	1
378	beneficial	1
379	\$91	1
380	veteran's	1
381	with	6

Minimum frequency: 0 Set

**Text Frequency**

No	Word	Frequency
0	need	1
1	total	1
2	discussion	1
3	benefit	1
4	maintain	1
5	scar	1
6	higher	3
7	received	2
8	its	1
9	accounting	1
10	over	2
11	veteran	11
12	levels	1
13	spouse	1
14	laws	1
15	year	8
16	bureau	1
17	knee	1
18	these	1
19	turn	1
20	#	3
21	recommends	1
22	arthritis	1
23	under	1
24	number	3
25	residual	1
26	available	1
27		65
28	meet	1
29		4
30	breast	1
31	musculoskel...	1
32	price	2
33	individually	2
34	february	1
35	involv...	1

Minimum frequency: 0 Set

**Integer Frequency**

No	Word	Frequency
0	1436	1
1	12006	4
2	22005	1
3	318	1
4	4485	1
5	5218	1
6	6337	1
7	71099	1
8	820	1
9	914000	1
10	40	3
11	1180	2
12	1260	2
13	13100	2
14	142393	1
15	1590	1
16	161	4
17	1710	2
18	180	2
19	1965000	1
20	30	2
21	775854	1
22	1277	1
23	2370	2
24	788000	1
25	690	1
26	50	1

Minimum frequency: 0 Set

**Float Frequency**

No	Word	Frequency
0	2.6	2
1	2.3	1

Minimum frequency: 0 Set

**TEXT**

61 that  
62 that  
63 the  
64 these  
65 his  
66 to  
67 use

73 where  
74 which  
75 will  
76 with  
77 would

New Remove Load Save

**Compare List**

No	Word	Info
5	are	3:3:3:3:1:1:1:1
6	as	2:2:9:9:12:12
7	based	3:3:1:1:2:2
8	be	1:1:2:2:7:7
9	benefits	1:1:4:4:1:1
10	compensation	1:1:5:5:12:12
11	for	2:2:10:10:17:17
12	from	2:2:6:6:4:4
13	has	1:1:3:3:2:2
14	have	1:1:1:1:4:4
15	in	16:16:20:20:14:14
16	is	4:4:6:6:16:16
17	last	14:14:1:1:1:1
18	new	2:2:1:1:3:3
19	not	4:4:3:3:6:6
20	number	3:3:8:8:3:3
21	of	15:15:42:42:32:32
22	on	6:6:7:7:9:9
23	other	14:14:3:3:3:3
24	public	1:1:1:1:1:1
25	requiring	1:1:1:1:2:2
26	rolls	1:1:1:1:2:2
27	the	7:7:54:54:46:46
28	to	3:3:34:34:15:15
29	total	13:13:1:1:1:1
30	upon	1:1:1:1:1:1
31	va	4:4:19:19:6:6
32	veterans	14:14:7:7:16:16
33	which	1:1:2:2:2:2
34	who	2:2:2:2:1:1

Save

**Department of Veterans Affairs**

**Fact Sheet**

Disability Compensation — 2008 Rates

Percentage	Rate
10 percent	\$112
20 percent	218
30 percent	337
40 percent	485
50 percent	690
60 percent	873
70 percent	1,099
80 percent	1,277
90 percent	1,456
100 percent	2,383

Effective 1/1/07

**Additional Payments for Dependents**

Dependents whose service-connected disabilities are rated at 30 percent or more are entitled to additional allowances for dependents. Depending upon the disability rating of the veteran, monthly allowances for a spouse range from \$61 to \$137 and for a dependent child, \$37 to \$91. Additional amounts are provided for each additional child and there is a higher scale for children in school after age 18.

**Other Major Factors Affecting Payment Levels**

Adjustments to rates are based on a number of factors in addition to dependents. Among factors that can have a significant effect on amounts are:

- \* Veterans with service-connected disabilities may receive compensation at a basic rate as high as \$6,845 per month.
- \* Veterans special monthly compensation rates apply when a veteran experiences loss or loss of use or more limbs, loses use or more of the senses of sight, hearing or speech, or experiences loss of a reproductive organ in its use, or loss of both hands by a double amputation.
- \* Allowances may be made for veterans requiring aids, such as including individuals who need assistance with eating, bathing or dressing other activities of daily living. This adjustment is referred to as "aid and attendance."

# Exploration of Image Components

Document Analyzer

Load Launch Document Use Ignore List Compare Group Move to Ignore List Show Comparison Visualization Show Second Window Done Help

Done

Document List

No	Filename	Size	Date	Num Images
1	amwars-2.pdf	88KB	Fri May 22 08:49:59 CDT 2009	2
2	amwars.pdf	57KB	Fri May 22 08:49:59 CDT 2009	3
3	claimpro-2.pdf	57KB	Fri May 22 08:49:59 CDT 2009	3
4	claimpro.pdf	57KB	Fri May 22 08:49:59 CDT 2009	3
5	comprates-2.pdf	50KB	Fri May 22 08:49:59 CDT 2009	3

List of images Occurrence of colors Preview

Word Frequency Image Color Frequency Vector Graphics Frequency

Image List

No	Filename
1	(numrows=149, numcols= 956, numba...

ColorFrequency

No	Color	Frequency
0	RGB_227_227_227	3436
1	RGB_142_142_142	1274
2	RGB_113_113_113	1271
3	RGB_85_85_85	1369
4	RGB_28_28_28	4637
5	RGB_57_57_57	2449
6	RGB_170_170_170	1891
7	RGB_255_255_255	111261
8	RGB_0_0_0	12907
9	RGB_198_198_198	1949

Preview

Department of Veterans Affairs Fact Sheet

Office of Public Affairs Media Relations Washington, DC 20460 (202) 273-6000 www.va.gov

Minimum frequency: 0 Down Color Bins: 10 Set

Show Color Reduced Image

IMAGES

"Ignore" colors

Department of Veterans Affairs Fact Sheet

February 2006 Disability Compensation — 2006 Rates

10 percent	\$112	40 percent	\$871
20 percent	218	70 percent	1,099
30 percent	337	80 percent	1,277
40 percent	485	90 percent	1,456
50 percent	650	100 percent	2,000

Effective 1/1/07

Additional Payments for Dependents

Dependents whose service-connected disabilities are rated at 30 percent or more are entitled to additional allowances for dependents. Depending upon the disability rating of the veteran, monthly allowances for a spouse range from \$61 to \$117 and for a dependent child, \$17 to \$91. Additional amounts are provided for each additional child and there is a higher scale for children in school after age 18.

Other Major Factors Affecting Payment Levels

Adjustments to rates are based on a number of factors in addition to dependents. Among factors that can have a significant effect on amounts are:

- \* Veterans with service-connected disabilities may receive compensation at a basic rate as high as \$687 per month. Veterans special monthly compensation rates apply when a veteran experiences loss of use of one or more limbs, loses use or sense of the senses of sight, hearing or speech, or experiences loss of a reproductive organ or its use, or loss of breast tissue by a female veteran.
- \* Allowances may be made for veterans requiring aids, such as including individuals who need assistance with eating, bathing or other activities of daily living. This adjustment is referred to as "aid and attendance."

# Exploration of Vector Graphics Components

**Document Analyzer**

☒ Use Ignore List

Done

No	Filename	File Size	File Date	Num Pages	Num Images
1	1amwars-2.pdf		88KB/Fri May 22 08:49:59 CDT 2009	2	1
2	2amwars.pdf				1
3	3claimpro-2.pdf				1
4	4claimpro.pdf				1
5	5comprates-2.pdf				1
6	6comprates.pdf				1

LOADED FILES

VECTORS

Preview

Word Frequency | Image Color Frequency | **Vector Graphics Frequency**

Page List

Page Number	Number of Paths
All	1
0	1
1	1
2	1

Graphics List

Graphics Type	count
1,0,0,0,2,0,0,0,1,56	

Preview

Zoom In Zoom Out

Occurrence of v/h lines

Compare List

No	Word	Info
5	are	3:3:3:3:1:1:1:1
6	as	2:2:9:9:12:12
7	based	3:3:1:1:2:2
8	be	1:1:2:2:7:7
9	benefits	1:1:4:4:1:1
10	compensation	1:1:5:5:12:12
11	for	2:2:10:10:17:17
12	from	2:2:6:6:4:4
13	has	1:1:3:3:2:2
14	have	1:1:1:1:4:4
15	in	16:16:20:20:14:14
16	is	4:4:6:6:16:16
17	last	14:14:1:1:1:1
18	new	2:2:1:1:3:3
19	not	4:4:3:3:6:6
20	number	3:3:8:8:3:3
21	of	15:15:42:42:32:32
22	on	6:6:7:7:9:9
23	other	14:14:3:3:3:3
24	public	1:1:1:1:1:1
25	requiring	1:1:1:1:2:2
26	rolls	1:1:1:1:2:2
27	the	7:7:54:54:46:46
28	to	3:3:34:34:15:15
29	total	13:13:1:1:1:1
30	upon	1:1:1:1:1:1
31	va	4:4:19:19:6:6
32	veterans	14:14:7:7:16:16
33	which	1:1:2:2:2:2
34	who	2:2:2:2:1:1

Save

**Department of Veterans Affairs Fact Sheet**

February 2006

Disability Compensation — 2006 Rates

10 percent	\$112	40 percent	\$873
20 percent	218	70 percent	1,099
30 percent	337	80 percent	1,277
40 percent	485	90 percent	1,456
50 percent	650	100 percent	2,000

Effective 1/1/07

**Additional Payments for Dependents:**

Veterans whose service-connected disabilities are rated at 30 percent or more are entitled to additional allowances for dependents. Depending upon the disability rating of the veteran, monthly allowances for a spouse range from \$65 to \$157 and for a dependent child, \$37 to \$93. Additional amounts are provided for each additional child and there is a higher scale for children in school after age 18.

**Other Major Factors Affecting Payment Levels:**

Adjustments to rates are based on a number of factors in addition to dependents. Among factors that can have a significant effect on payment are:

- Veterans with service service-connected disabilities may receive compensation at a basic rate as high as \$6,847 per month. Veterans special monthly compensation rates apply when a veteran experiences loss of use or loss of sight in one or more limbs, loses use or sight of the senses of sight, hearing or speech, or experiences loss of a reproductive organ or its use, or loss of breast tissue by a female veteran.
- Allowances may be made for veterans requiring aids, such as including individuals who need assistance with eating, bathing or certain other activities of daily living. This adjustment is referred to as "aid and attendance."

More...

# Comprehensive Pair-Wise Comparison of Documents

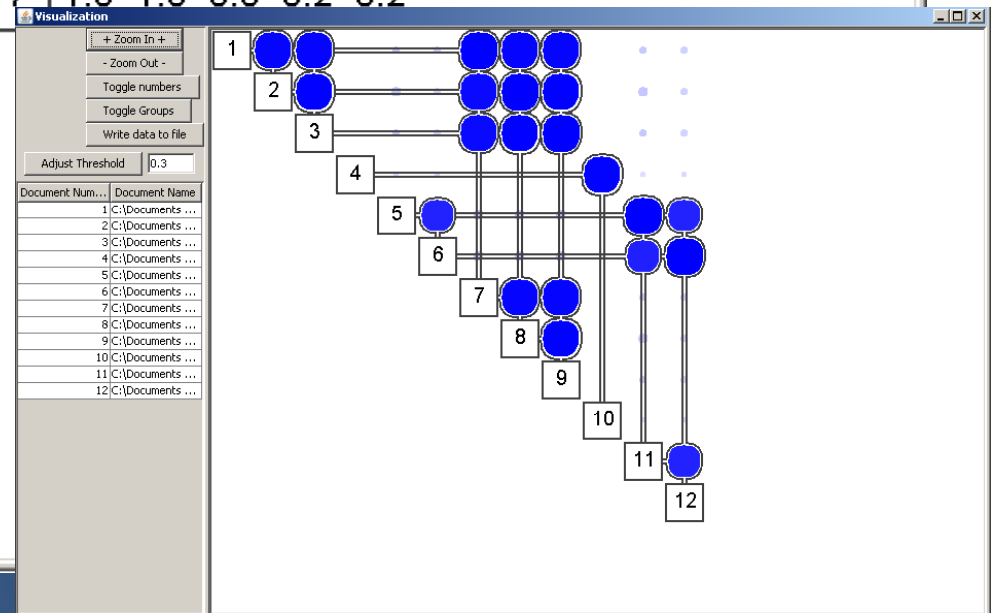
Grouping and Visualization Control

Toggle numbers	
Toggle Groups	
Write data to file	
Adjust Threshold 0.3	
Document Num...	Document Name
1	C:\Documents ...
2	C:\Documents ...
3	C:\Documents ...
4	C:\Documents ...
5	C:\Documents ...
6	C:\Documents ...
7	C:\Documents ...
8	C:\Documents ...
9	C:\Documents ...
10	C:\Documents ...
11	C:\Documents ...

Document ID

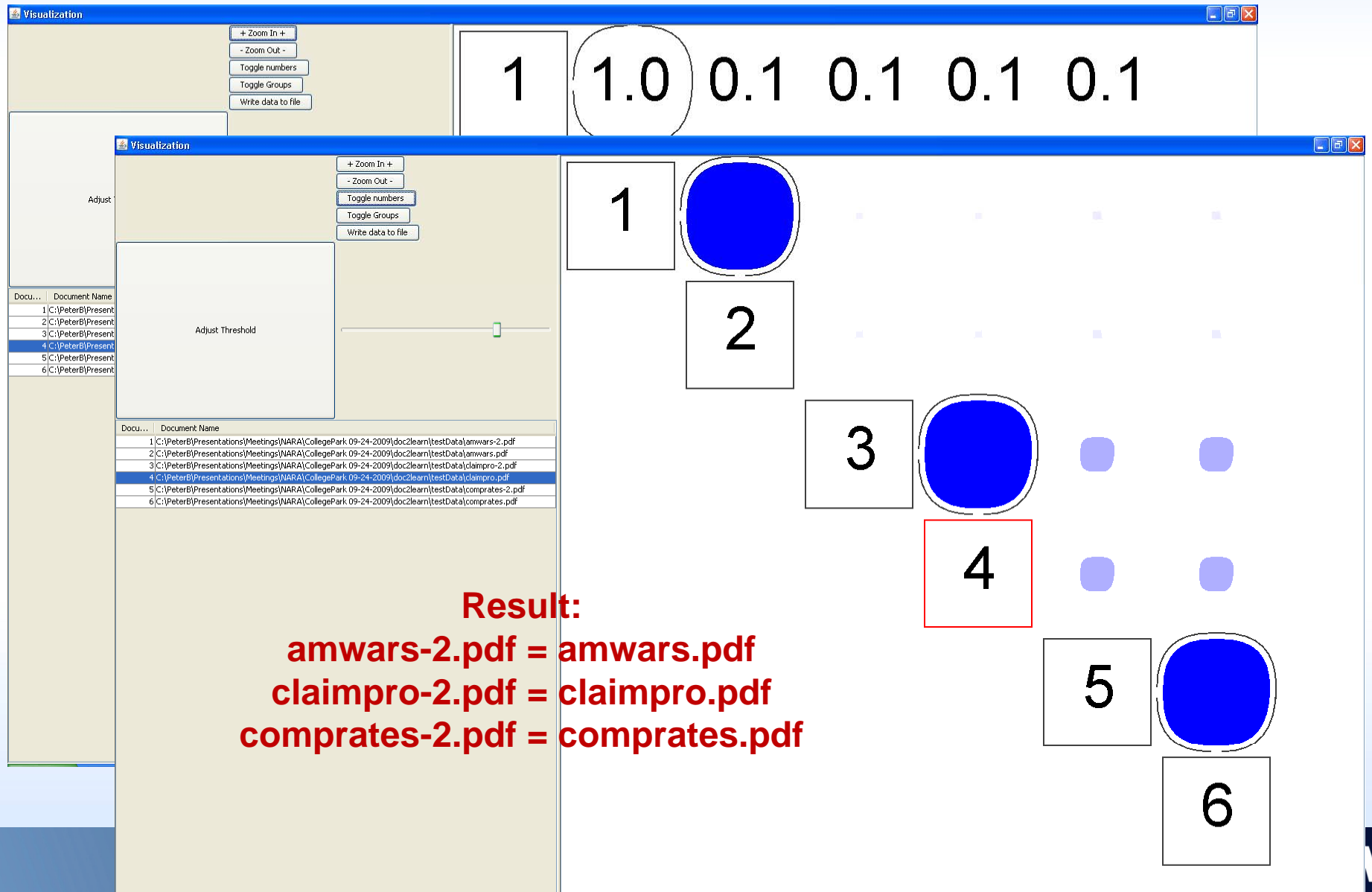
control	0.0	0.2	0.2	1.0	1.0	1.0	0.0	0.2	0.2	
2	1.0	0.0	0.2	0.2	1.0	1.0	1.0	0.0	0.2	0.2
3	0.0	0.2	0.2	1.0	1.0	1.0	0.0	0.2	0.2	0.2
4	0.1	0.2	0.0	0.0	0.0	1.0	0.1	0.2	0.2	0.2
5	0.9	0.3	0.3	0.3	0.1	1.0	0.9	0.9	0.9	0.9
6	0.2	0.2	0.2	0.2	0.2	0.9	1.0	0.9	0.9	0.9
7	1.0	1.0	0.0	0.2	0.2	0.2	0.2	0.2	0.2	0.2

Similarity Values





# Visual Comparison for 6 Test Files



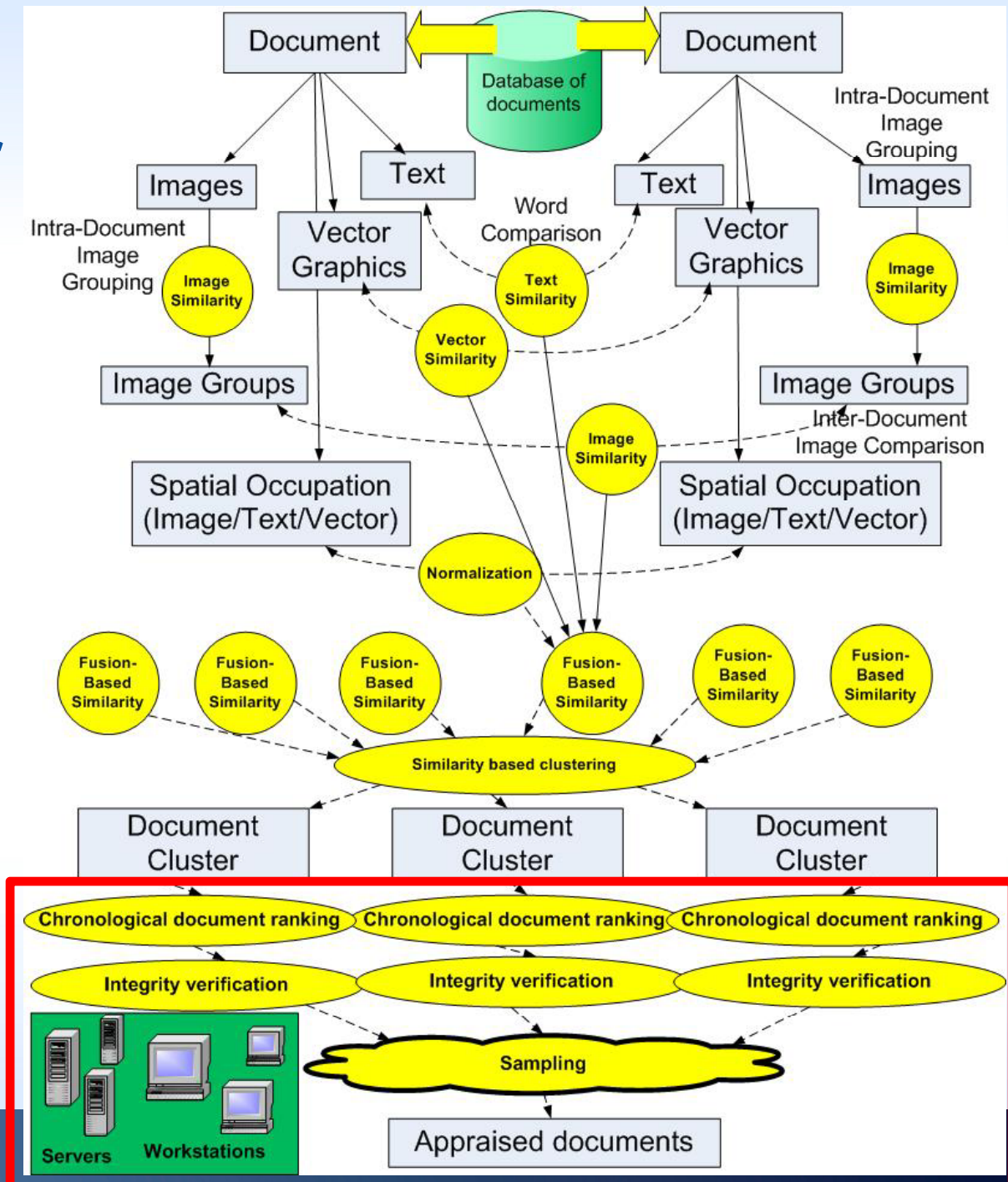


# Computational Requirements for Executing the Methodology

**Yellow indicates computations**

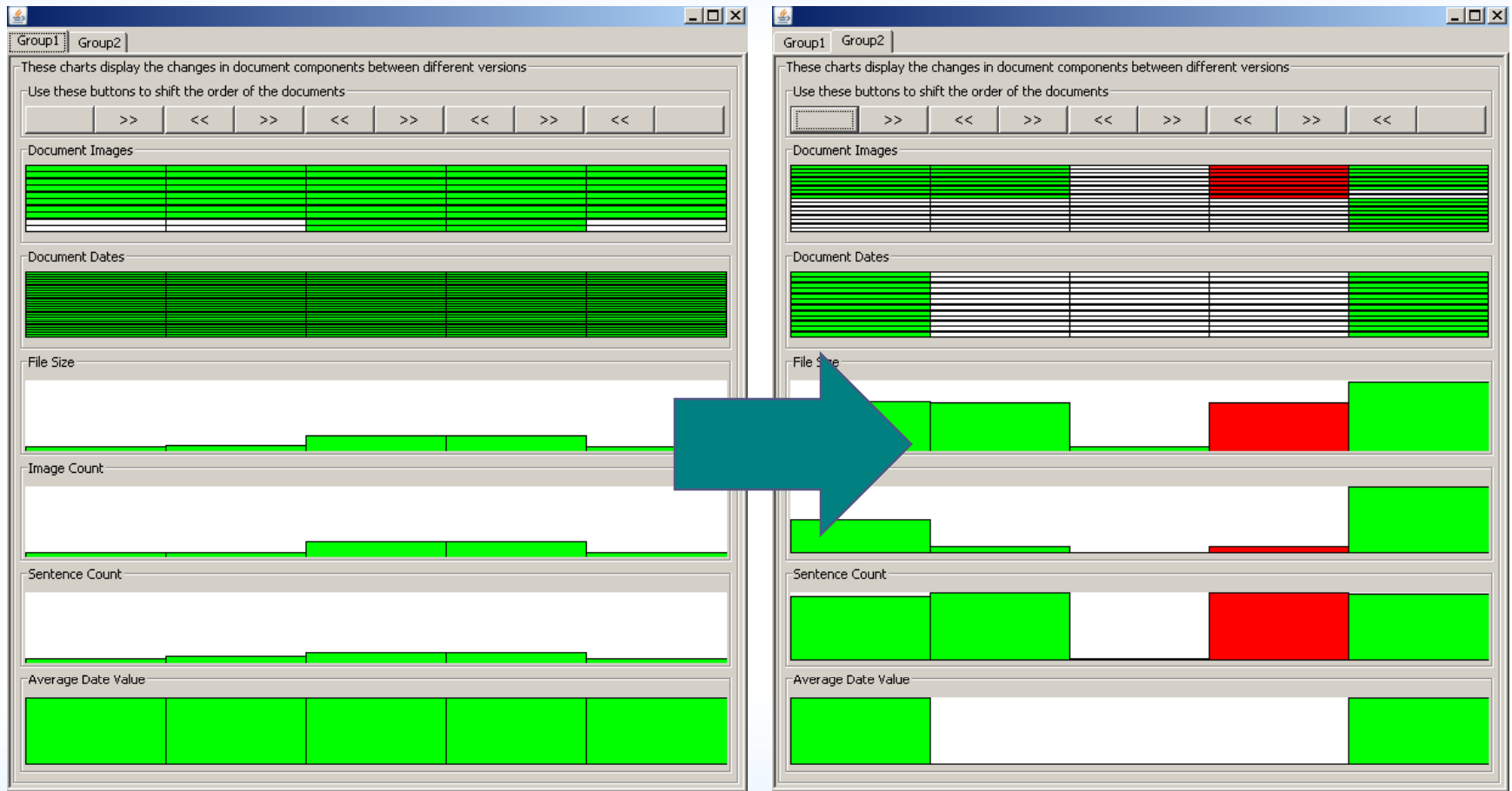
Relationship to  
Permanent Records

Appraisal & Sampling



# Work in progress: Group and Validate Documents

Attributes of documents



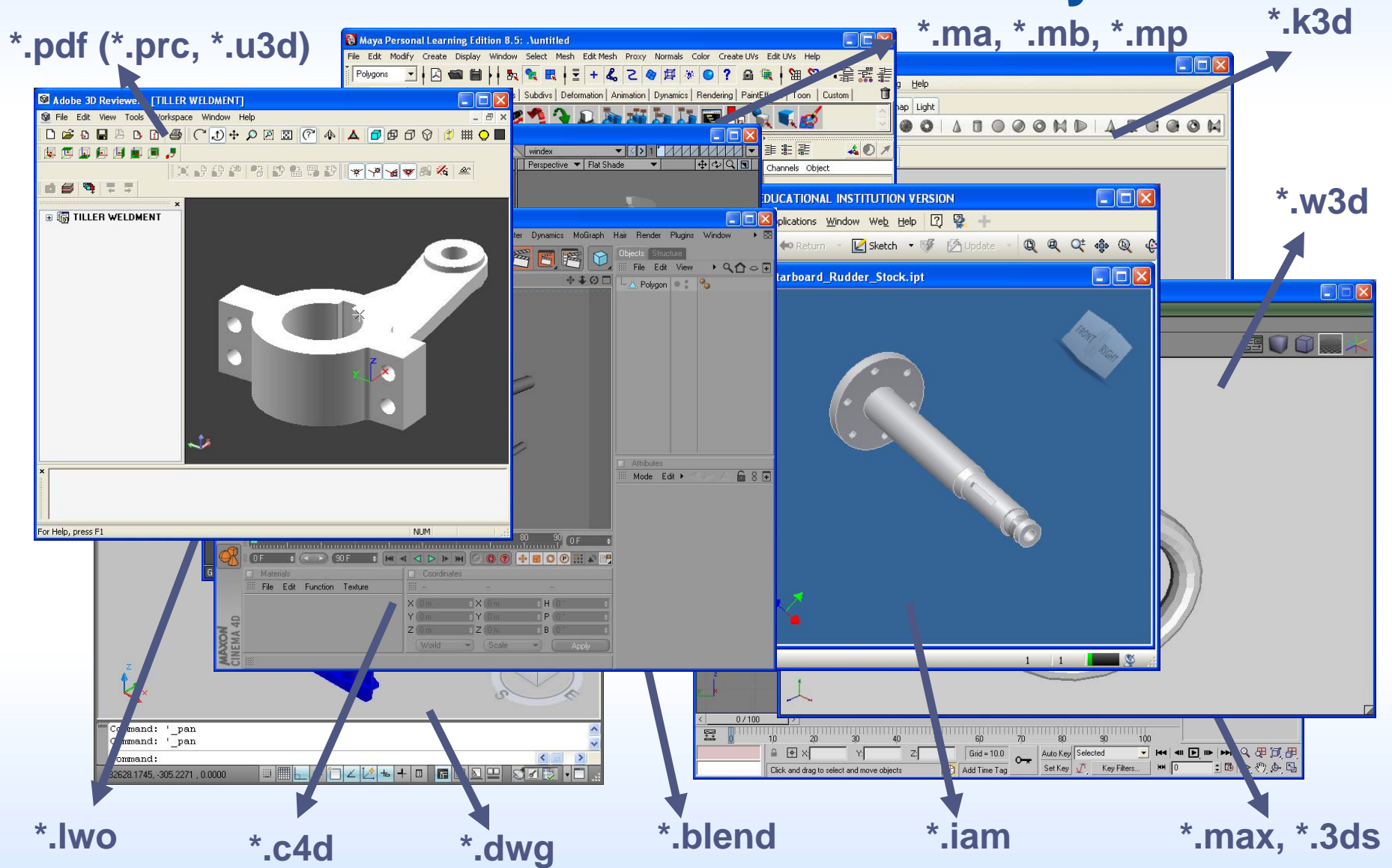
Order of documents

# **Automated File Format Conversions and Conversion Quality Assessment**

# Conversions of Electronic Records

- **Conversions of electronic records are needed** because
  - Visual exploration depends on various software packages
  - Many formats are retired (deprecated) over time
- **How to measure the degree of information preservation when files are converted from format A to format B?**
  - During conversions, information could be lost, added or modified
  - What is the importance of each byte, object, etc. ?
- **How to design a test bed for analyzing the quality of conversion and visualization software?**

# Illustration of 3D File Format Reality



## Our Survey about 3D Content

- **Q: How Many 3D File Formats Exist?**
- A: We have found more than 140 3D file formats. Many are proprietary file formats. Many are extremely complex (1,200 and more pages of specifications).
- **Q: How Many Software Packages Support 3D File Format Import, Export and Display?**
- A: We have documented about 16 software packages. There are many more. Most of them are proprietary/closed source code. Many contain incomplete support of file specifications.

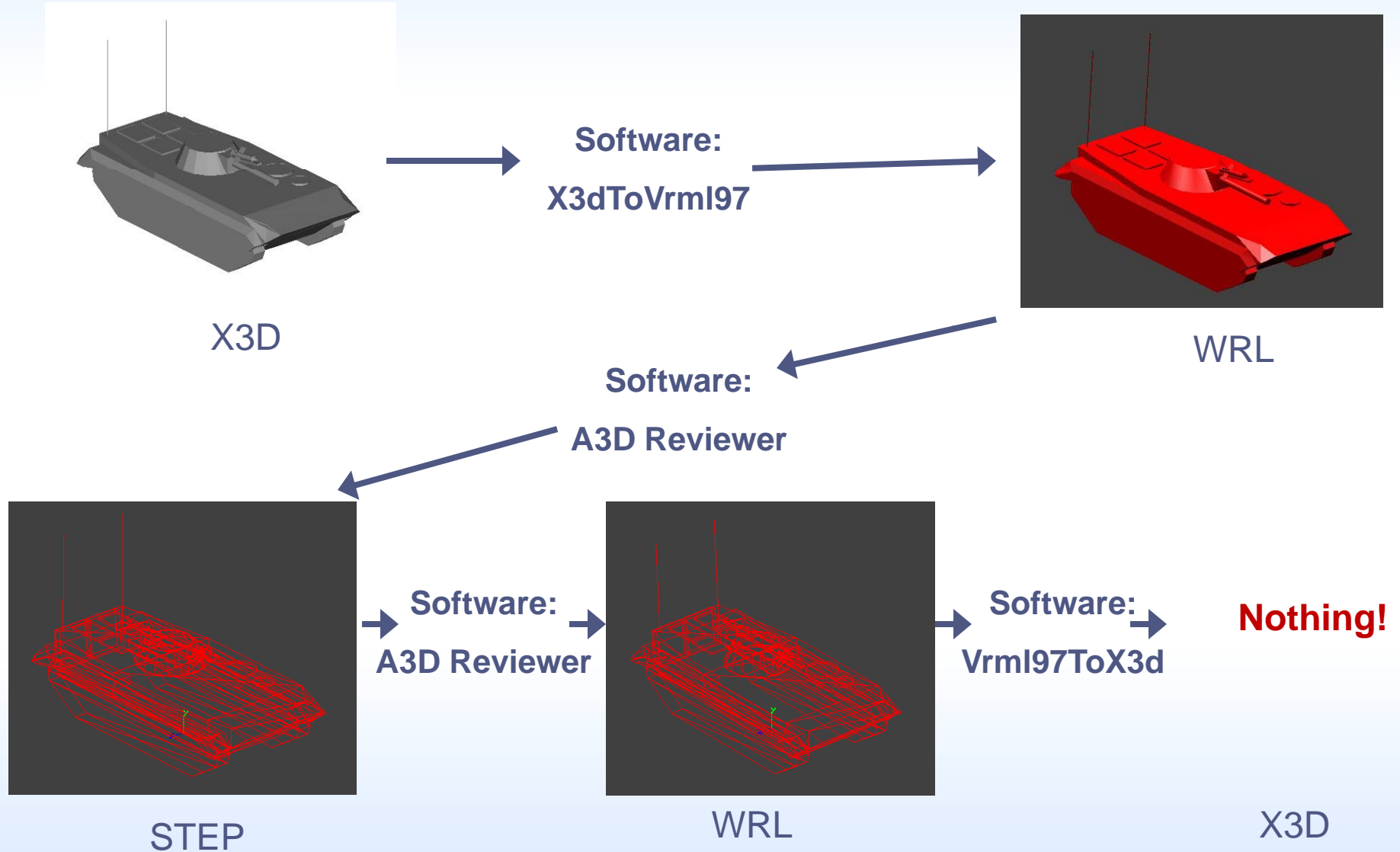
# Examples of 3D Formats and Stored Content

Format	Geometry				Appearance				Scene				Animation
	Faceted	Parametric	CSG	B-Rep	Color	Material	Texture	Bump	Lights	Views	Trans.	Groups	
3ds	✓	✓			✓	✓	✓	✓	✓	✓	✓		
igs	✓	✓	✓	✓	✓						✓	✓	
lwo	✓	✓			✓	✓	✓	✓					
obj	✓	✓			✓	✓	✓	✓				✓	
ply	✓				✓	✓	✓	✓					
stp	✓	✓	✓	✓	✓							✓	
wrl	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓
u3d	✓				✓		✓	✓	✓	✓	✓	✓	✓
x3d	✓	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓

- Some content may be more important than others
  - The relative importance is situation dependent



# Example: Conversion of X3D to STEP to X3D





# Towards a Universal Converter

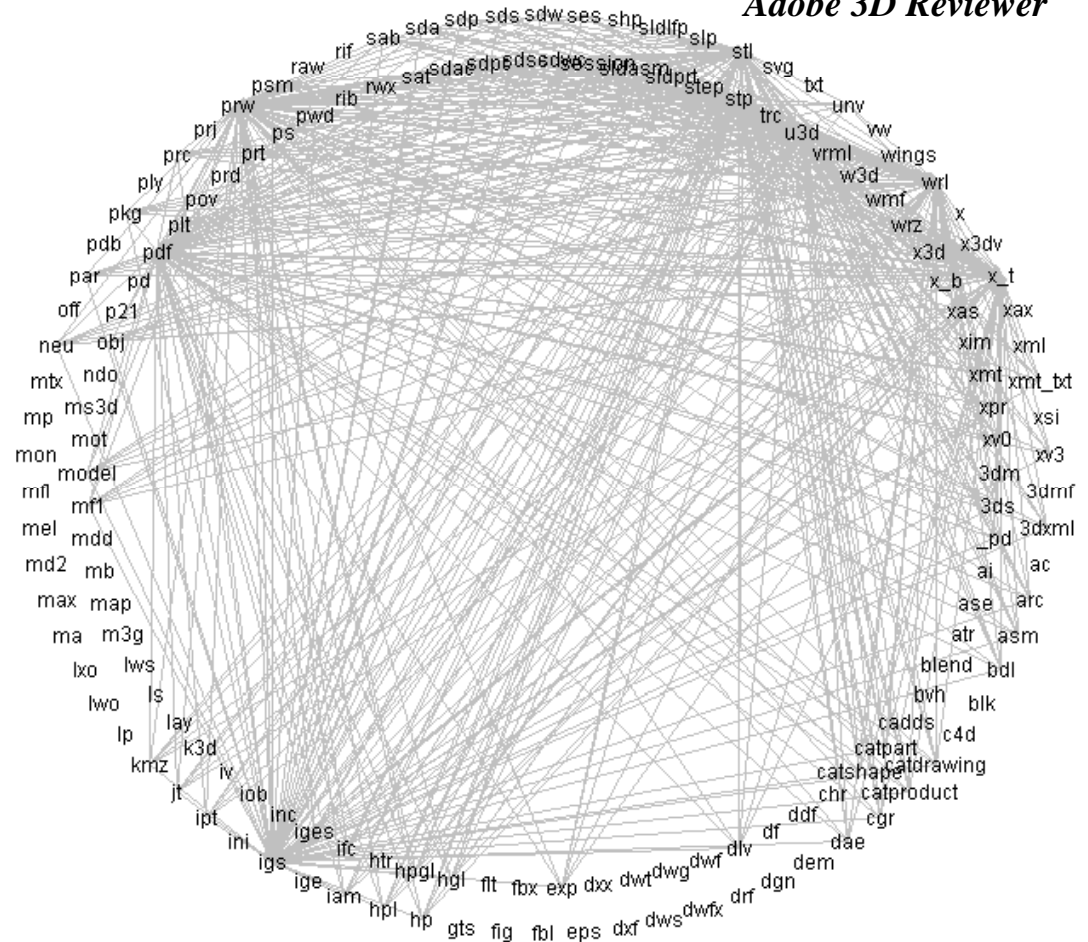
- Use what is available in 3<sup>rd</sup> party software to perform conversions
  - Document **what formats can be opened/imported** by each application
  - Document **what formats can be saved/exported** by each application
  - **Automate the use of each application** and combine their abilities to perform conversions over larger set of formats

# Input/Output Graphs

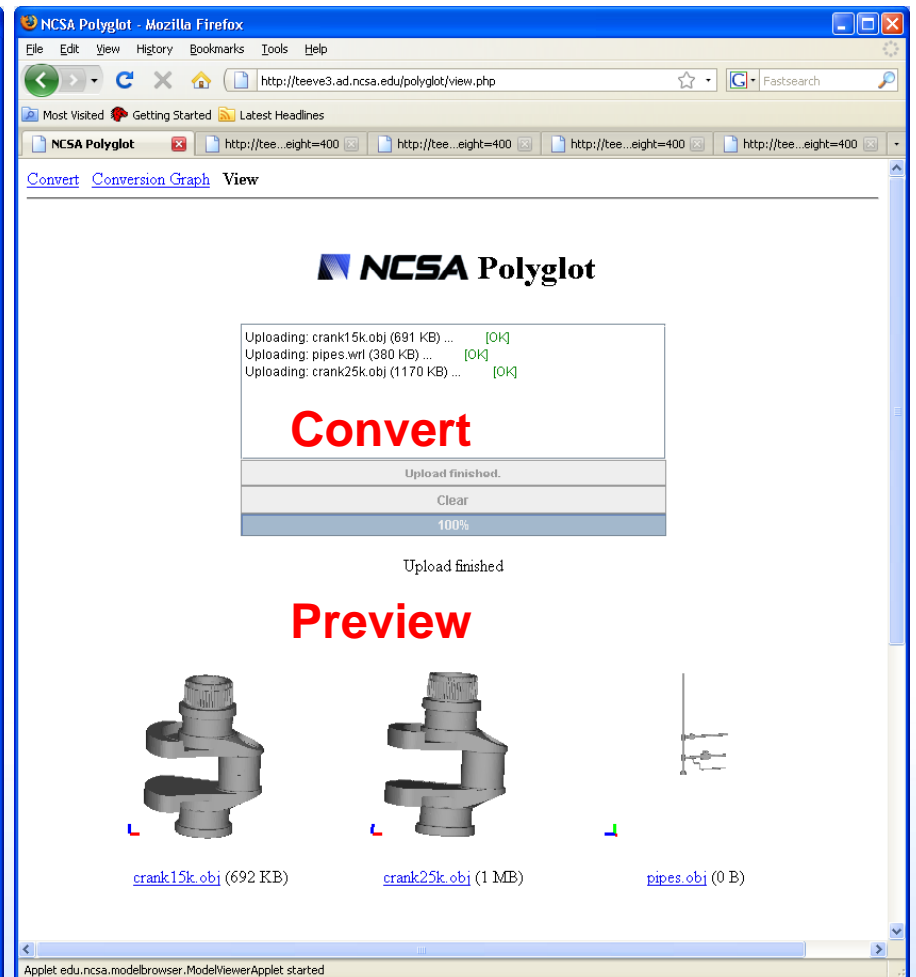
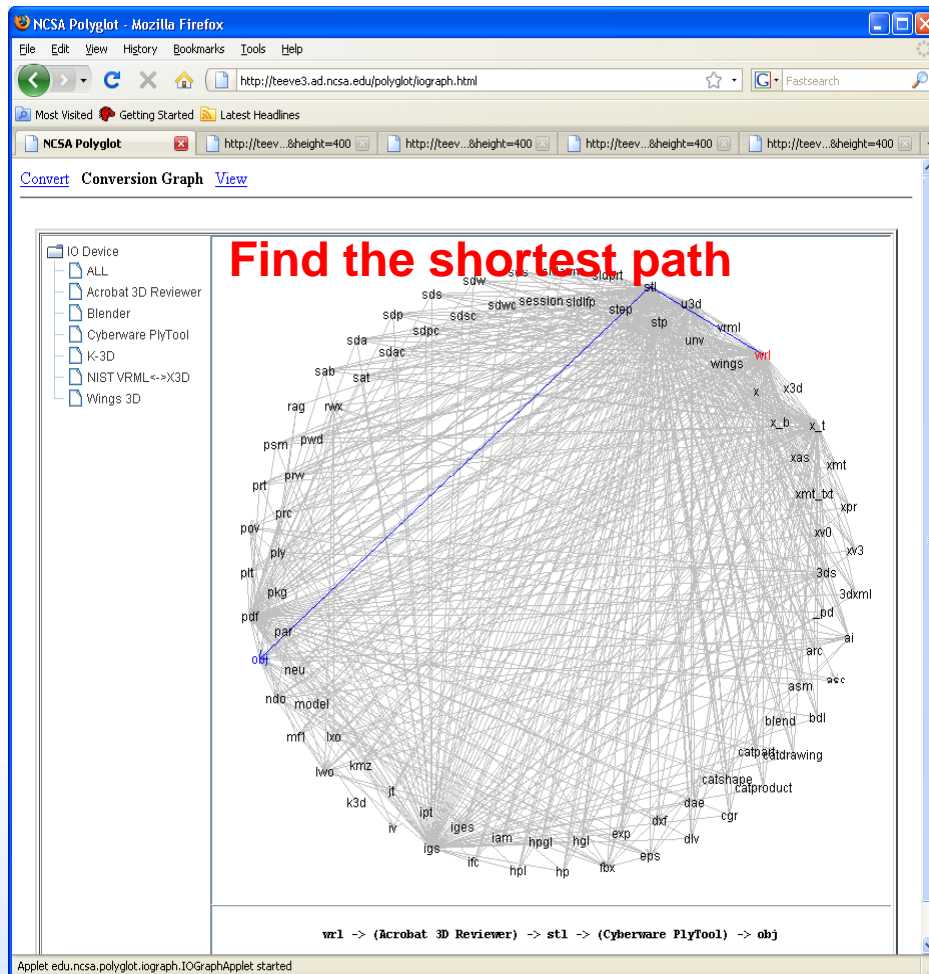
## Adobe 3D Reviewer

Manufacturer	Adobe
Version	9, Trial
URL	<a href="http://www.adobe.com">http://www.adobe.com</a>
Cost	
Opens	3ds, 3dxml, arc, asm, bdl, catdrawing, dae, dlv, exp, hgl, hp, hpgl, hpl, iam, model, neu, obj, pd, par, pdf, pkg, plsda, sdac, sdp, sdpc, sds, sdsc, sdw, sldprt, stl, step, stp, u3d, unv, wrl, vxv0, xv3
Saves	prw
Imports	3ds, 3dxml, arc, asm, bdl, catdrawing, dae, dlv, exp, hgl, hp, hpgl, hpl, iam, model, neu, obj, pd, par, pdf, pkg, plsda, sdac, sdp, sdpc, sds, sdsc, sdw, sldprt, stl, step, stp, u3d, unv, wrl, vxv0, xv3
Exports	igs, pdf, stl, stp, u3d, wrl, x t

## Adobe 3D Reviewer



# Automation of 3D File Format Mapping



# Automation of 3D File Format Conversion

- The I/O-Graph stores the information needed to convert between the formats represented in the graph.
- In order to perform the conversion we must execute the conversion path found.
  - Many high end graphics programs are found on the windows platform
  - Those on other platforms, such as Linux, tend to have windows ports
  - Some are command line driven (usually small converter applications).
  - Many have only GUI interfaces
  - AutoHotKey: a scripting language for the Windows GUI.

# Methodology

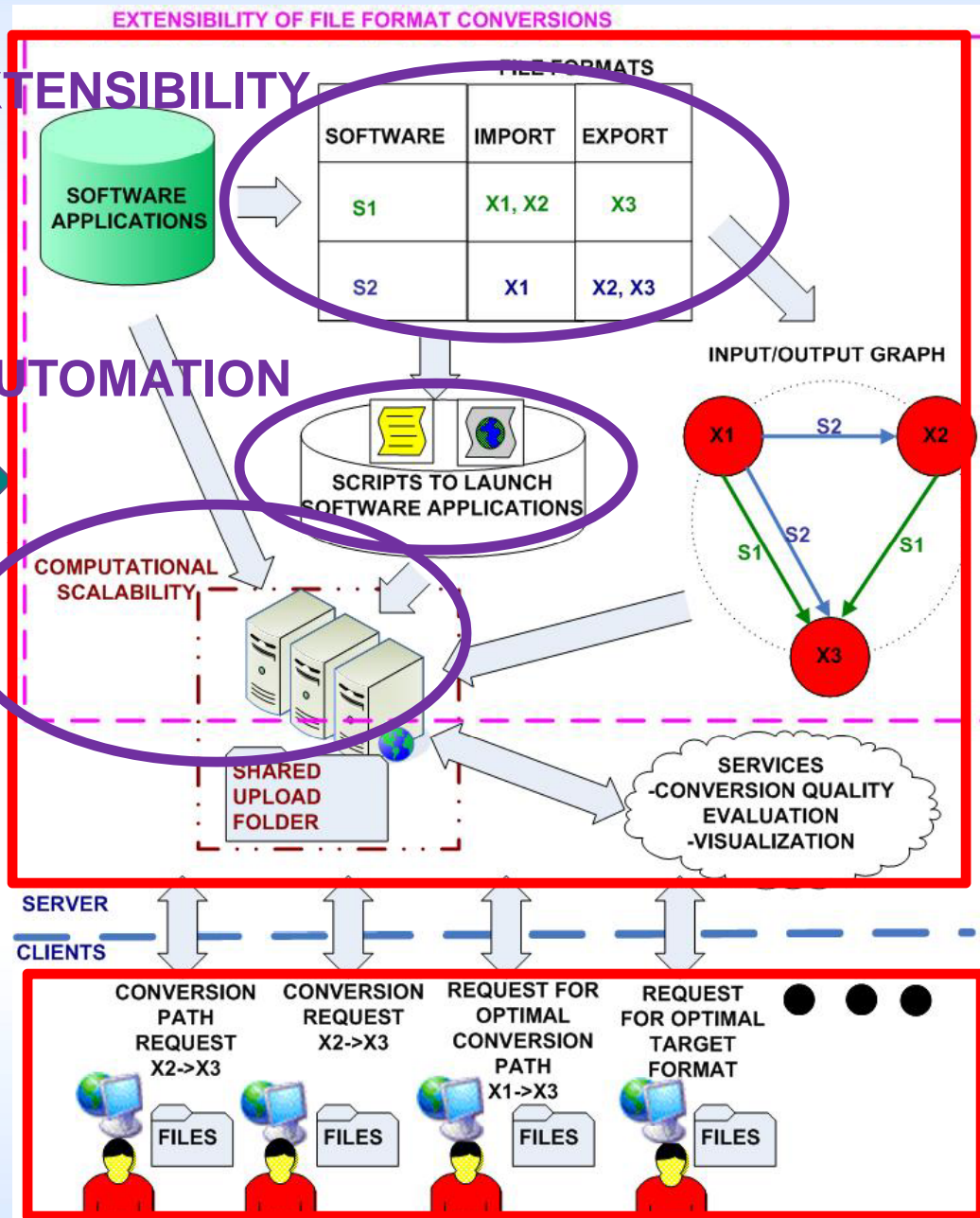
Cloud Computing

COMPUTATIONAL  
SCALABILITY

Services to Archivists

EXTENSIBILITY

AUTOMATION





# NCSA Polyglot – Conversion Services

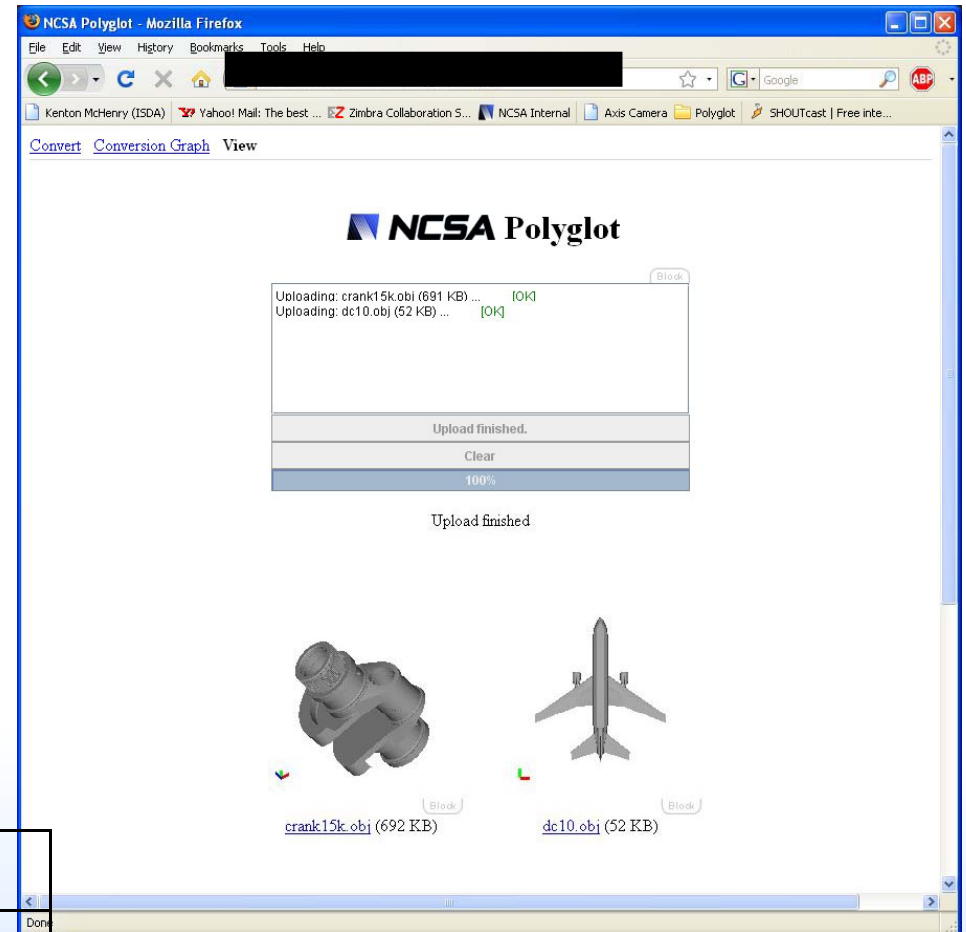
- Web interface: user can drag and drop files into upload area for conversion

- Java interface:

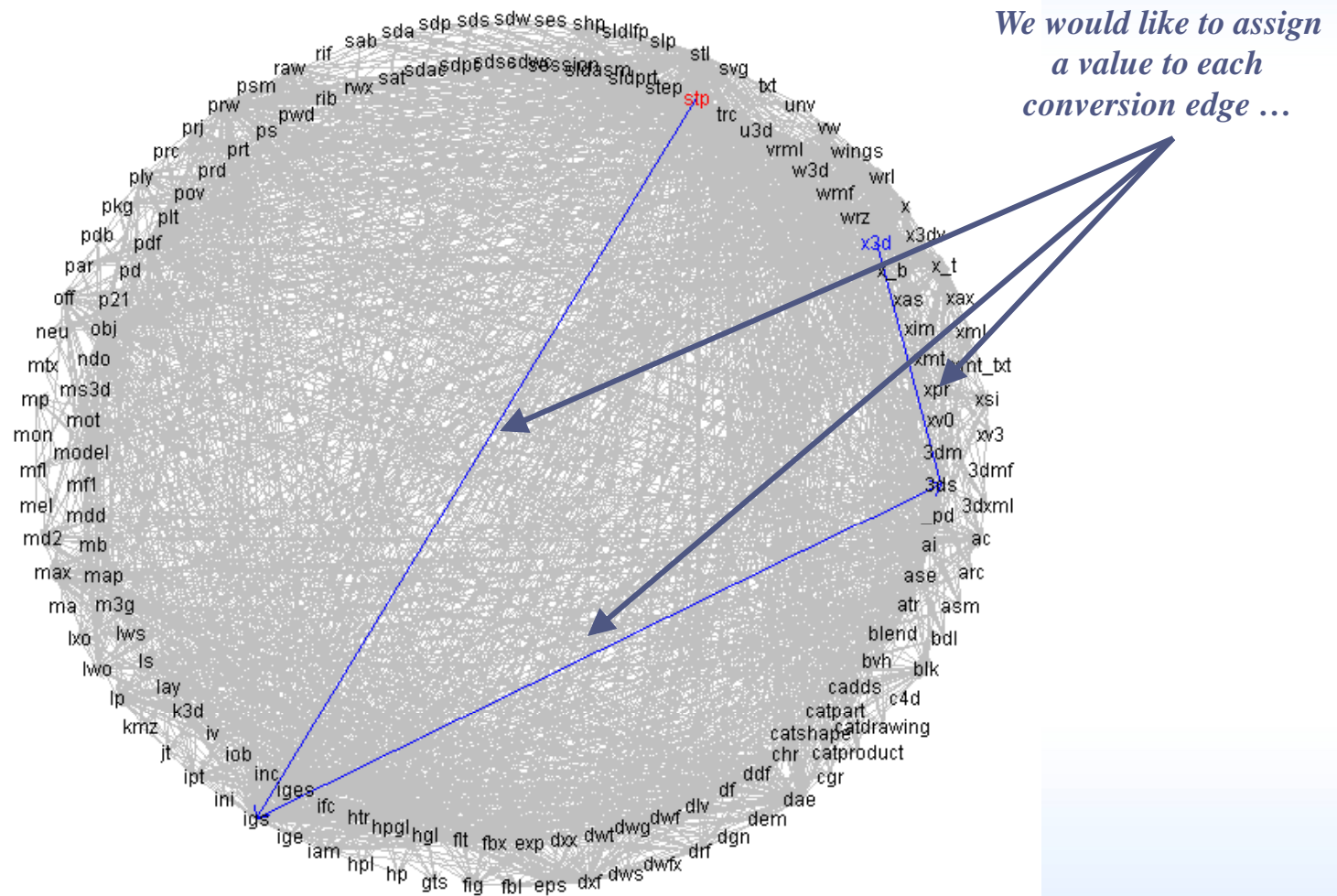
```
PolyglotRequest pgr;  
pgr = new PolyglotRequest("http://???", "obj");  
pgr.convertFile("file.wrl", ".");
```

- Scalability Test

Number of PCs	One PC	Two PCs
Processing Time	33 minutes 6 seconds	16 minutes 40 seconds



# NCSA Polyglot – Data Loss Measurement Services



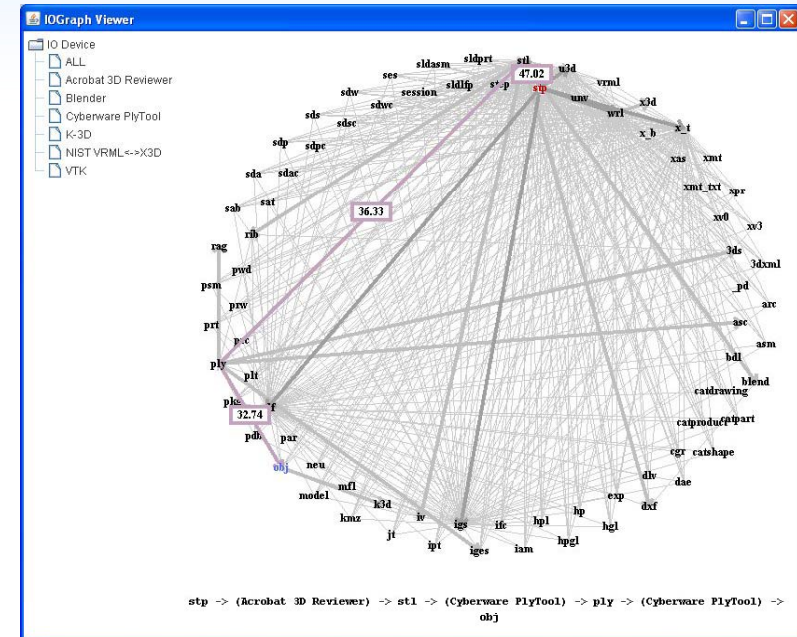
stp -> (Adobe 3D Reviewer) -> igs -> (3ds Max) -> 3ds -> (Blender) -> x3d



# Geometry Based Content Retention

- Several metrics
- Data driven assignment

- **Example results**



Metric\Result	Single Optimal Conversion				‘Best’ File Format	
	Software	From	To	Information Retention	Format	Information Retention
Light Fields	Adobe 3D Reviewer	.pdf	.stp	61.67	.stp	40.73
Spin Images	Adobe 3D Reviewer	.obj	.pdf	59.07	.stl	34.89

# Summary

- Technologies for appraisal of electronic records should assist archivists
  - They are designed to support decisions and data explorations by automating appraisal tasks
  - The software for doc2learn and Polyglot is available for downloading at <http://isda.ncsa.uiuc.edu/download/>
  - File2learn software – the work is still in progress
  - Feedback is very welcome
- 
- Questions: Peter Bajcsy – pbajcsy@ncsa.uiuc.edu

# Demo exercise

- Step 1: Check the path exists between wrl and pdf
- Step 2: drag and drop heart.wrl; select target to be pdf, click upload
- Step 3: download to desktop and open in Adobe PDF Viewer

