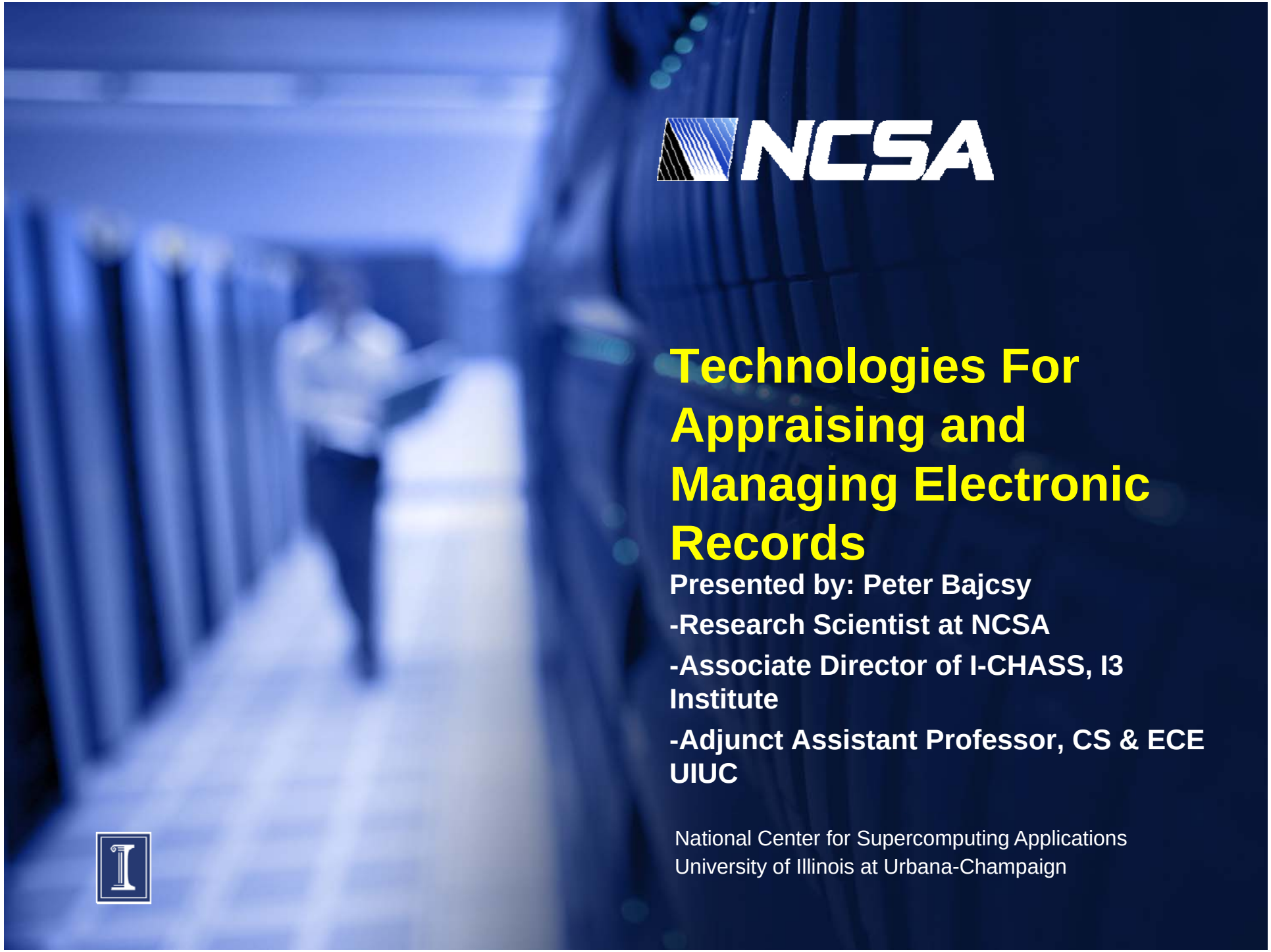


Technologies For Appraising and Managing Electronic Records

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

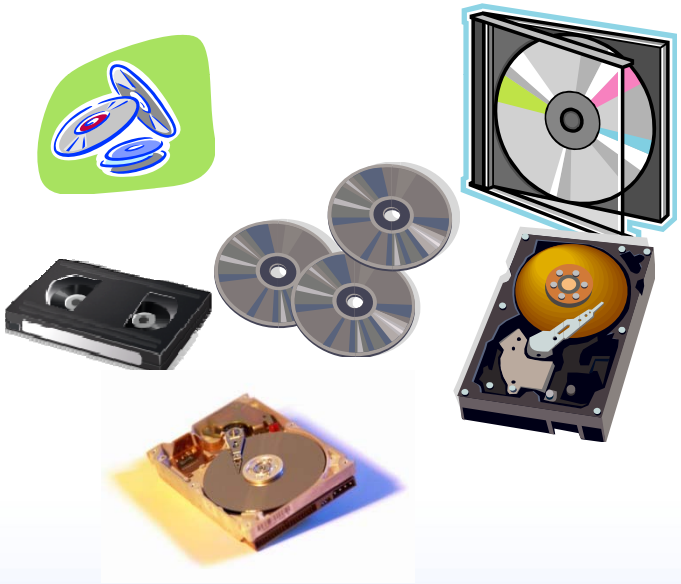
Supporting NARA's Strategic Plan

- 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."***

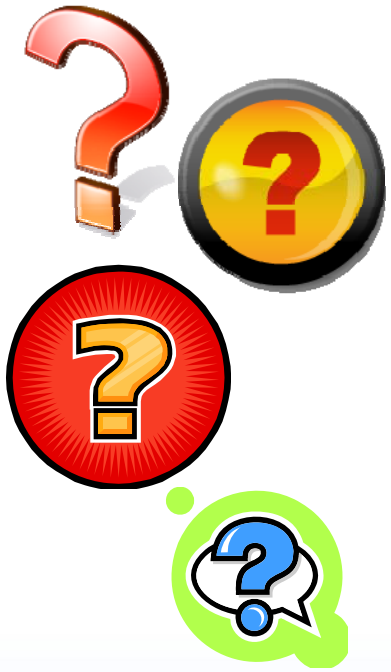
To Be Preserved!

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?

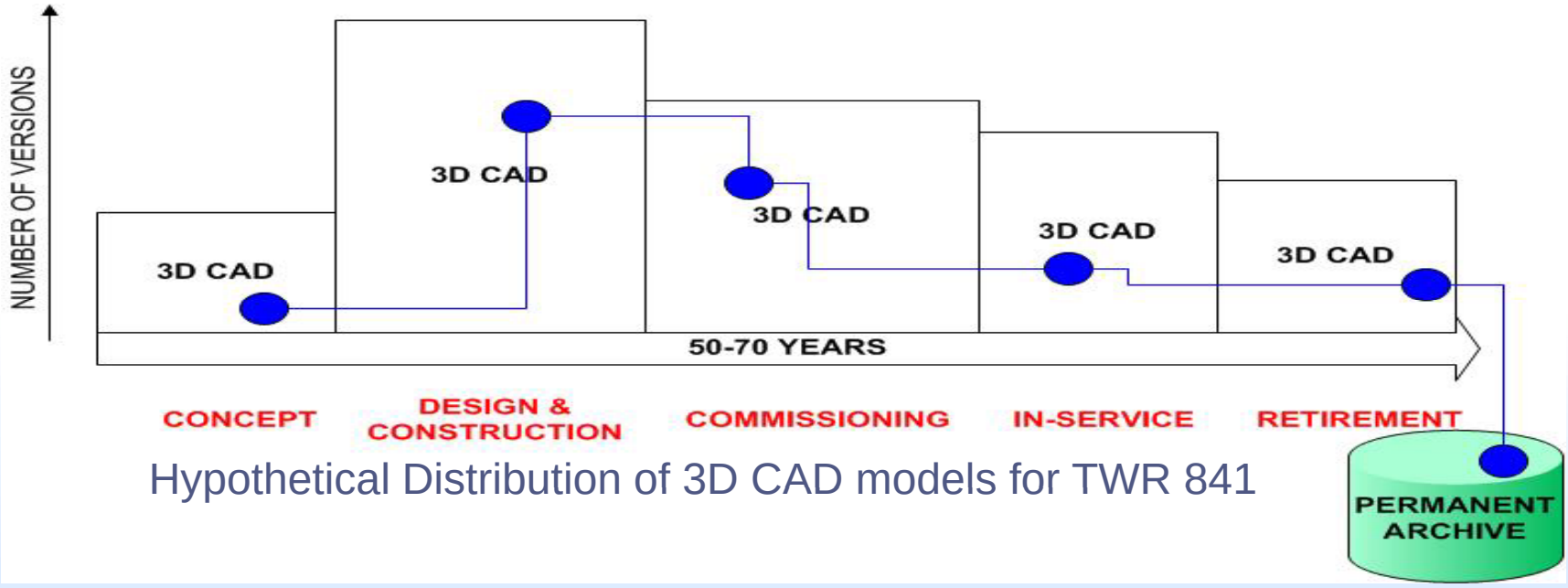
- Need to Understand the Complexity of the Problem

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

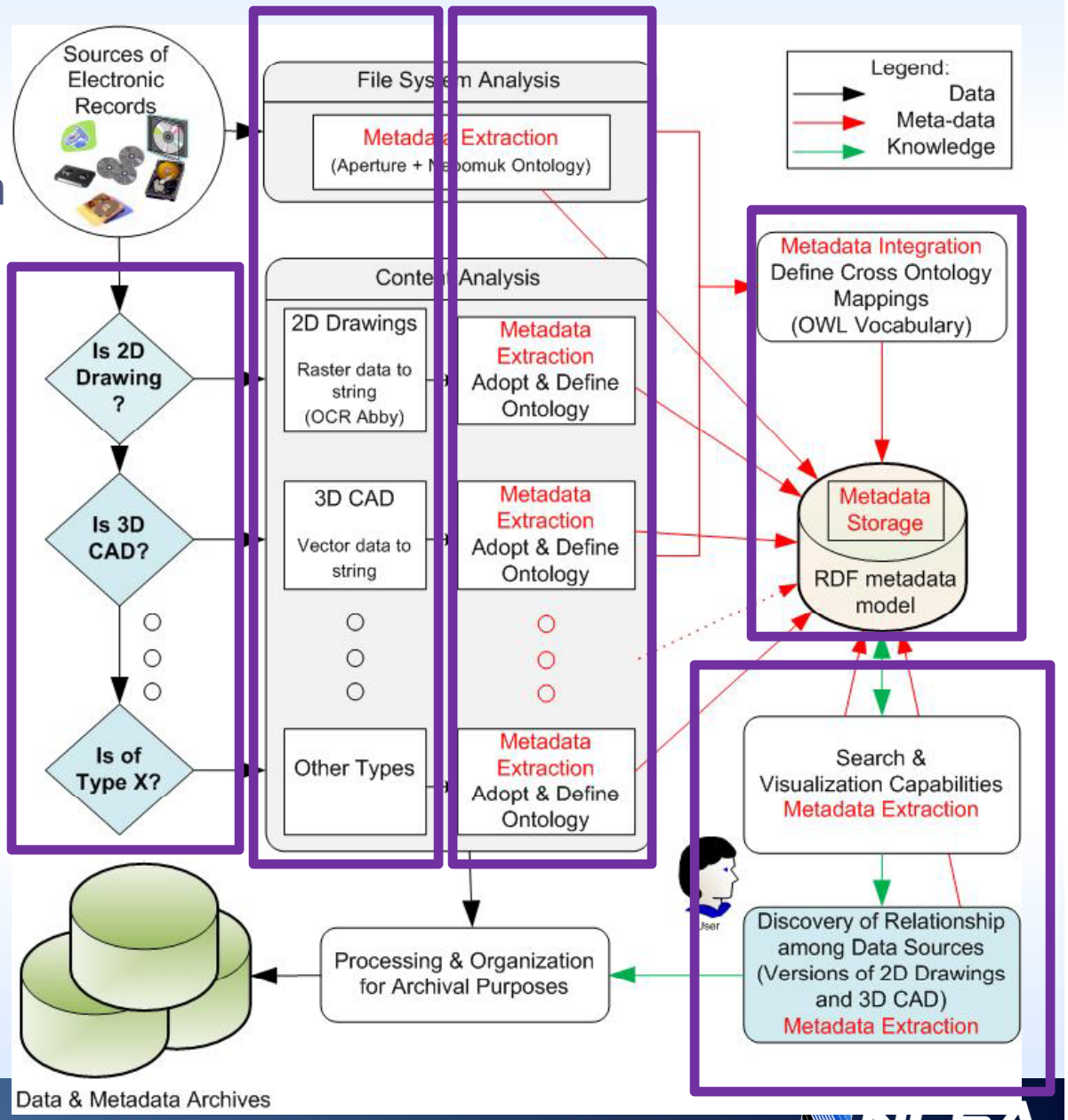
- **Example Data:** Torpedo Weapon Retriever 841
  - 784 existing 2D image drawings and N>22 3D CAD models
- How to establish relationships among the 3D CAD models and 2D image drawings during a product lifecycle?



Hypothetical Distribution of 3D CAD models for TWR 841

Methodology

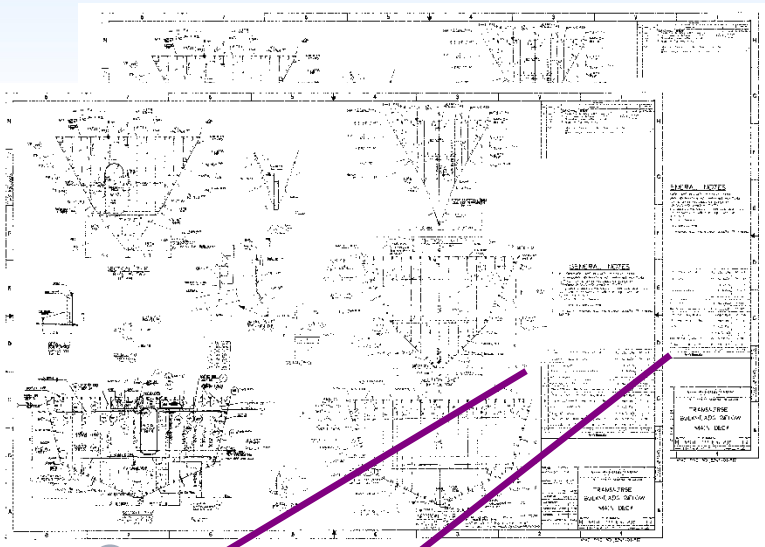
- File Identification
- Information Extraction from
  - File System
  - File Content
- Information Organization
  - Taxonomy (classification)
  - Ontology (relationships)
- Information Representation, Integration and Storage
  - XML
  - RDF
- Relationship Discovery

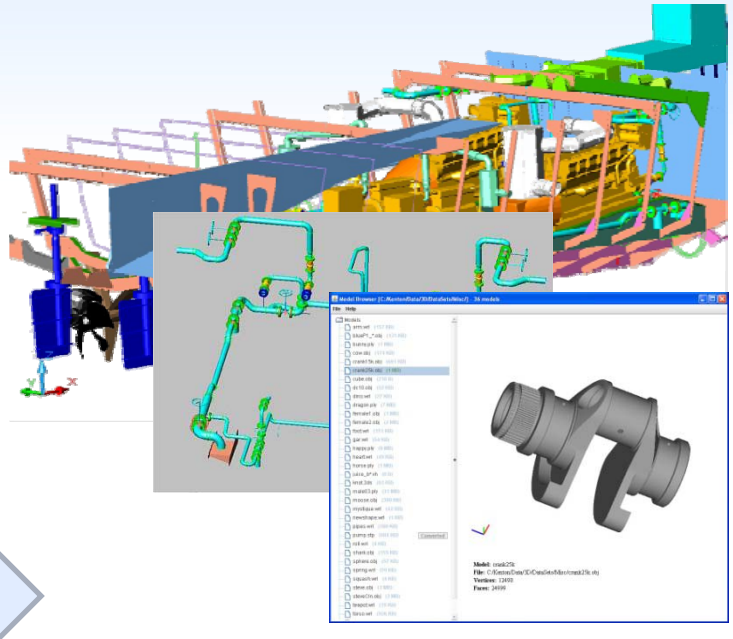


File Identification and File System Analyses

- **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)

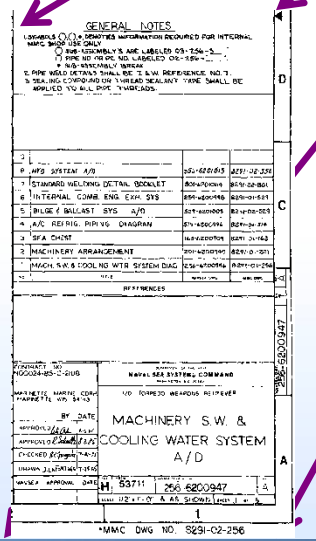
Content Analyses: Automation ?





Relationship Discovery

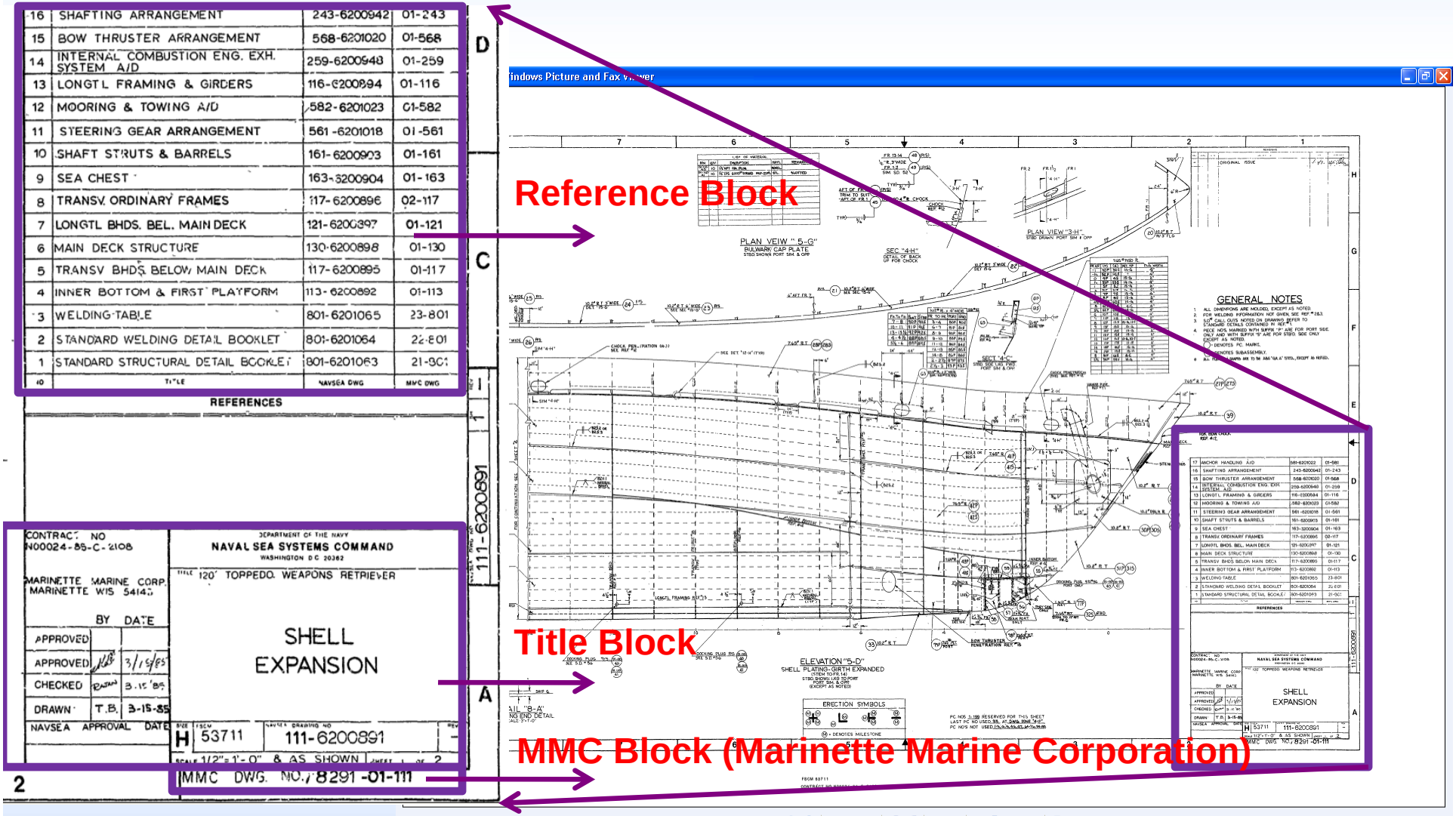
OCR



Part name,
Author,
Software,
Date, …

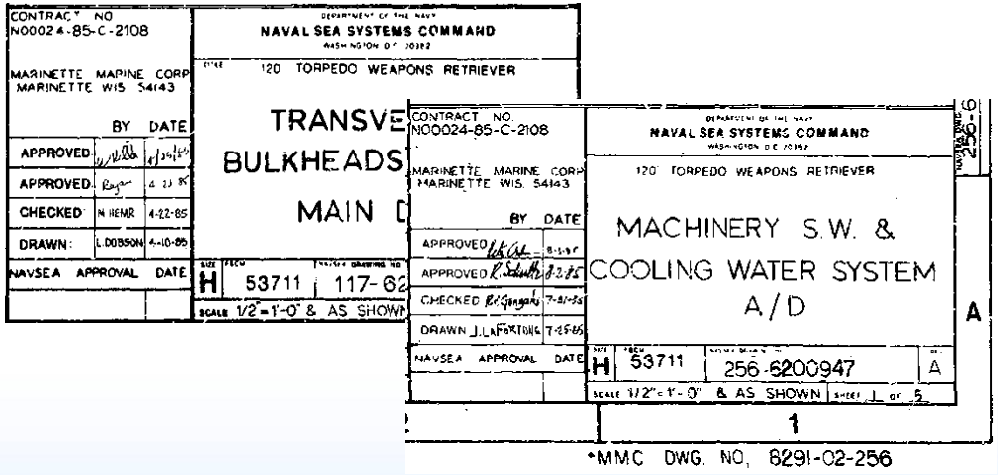
File Descriptors

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

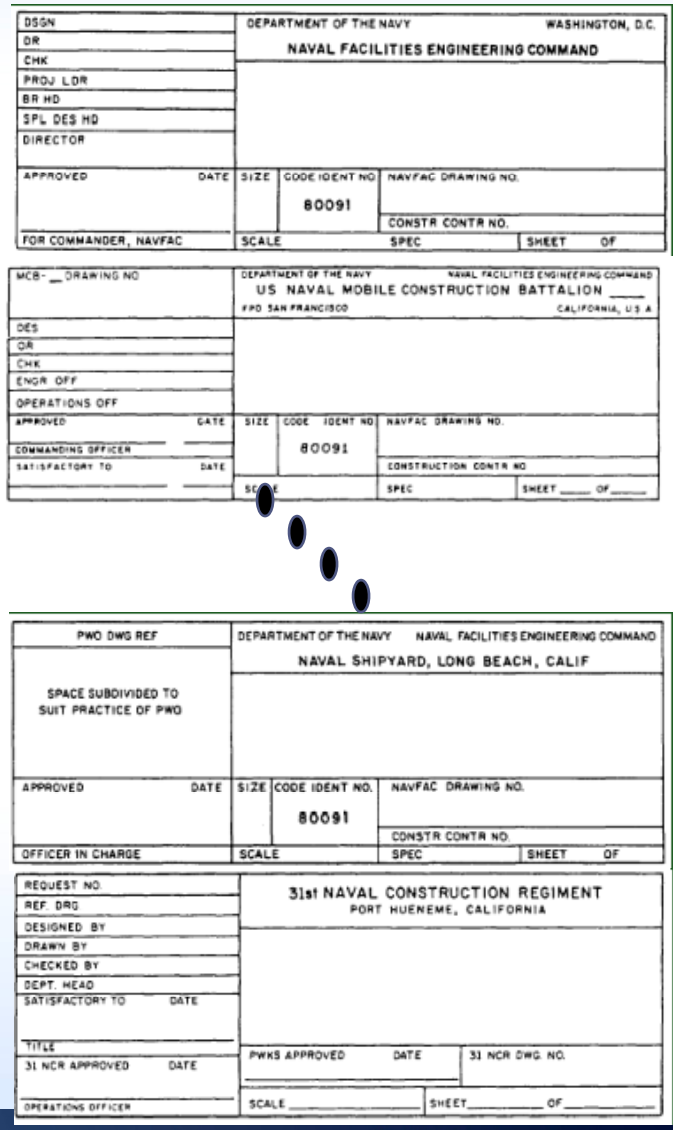


‘Standard’ Title Blocks: Organization and Ontology

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

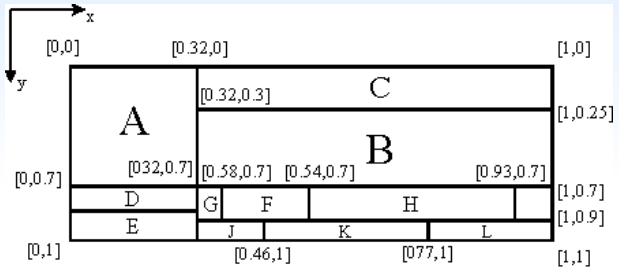


TEMPLATES



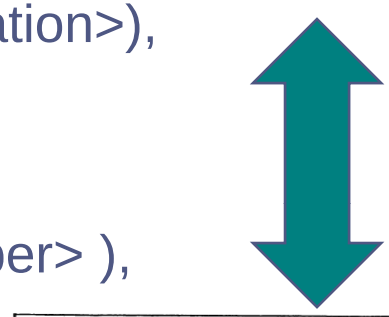
NCSA

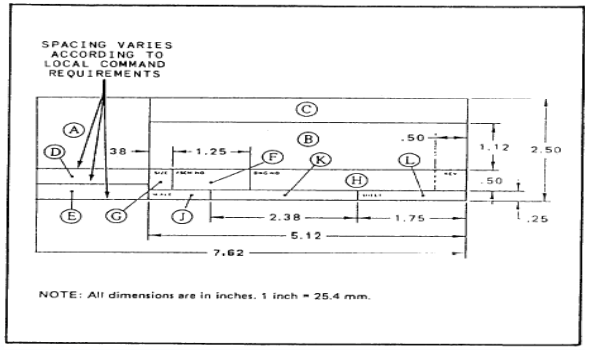
Title Block: Ontology and Metadata Representation



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>).



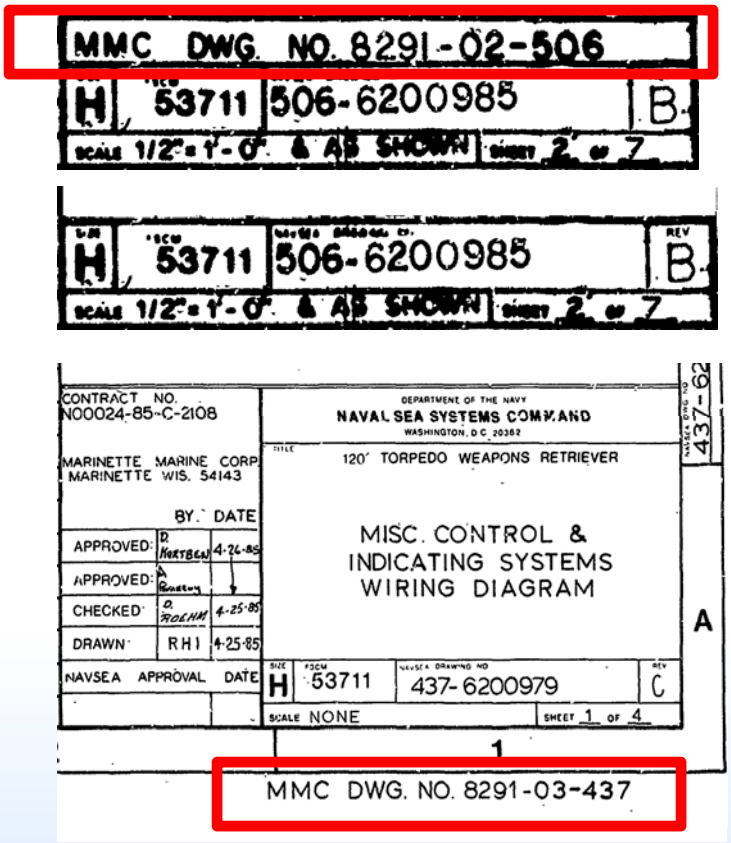


Resource Description Framework (RDF):

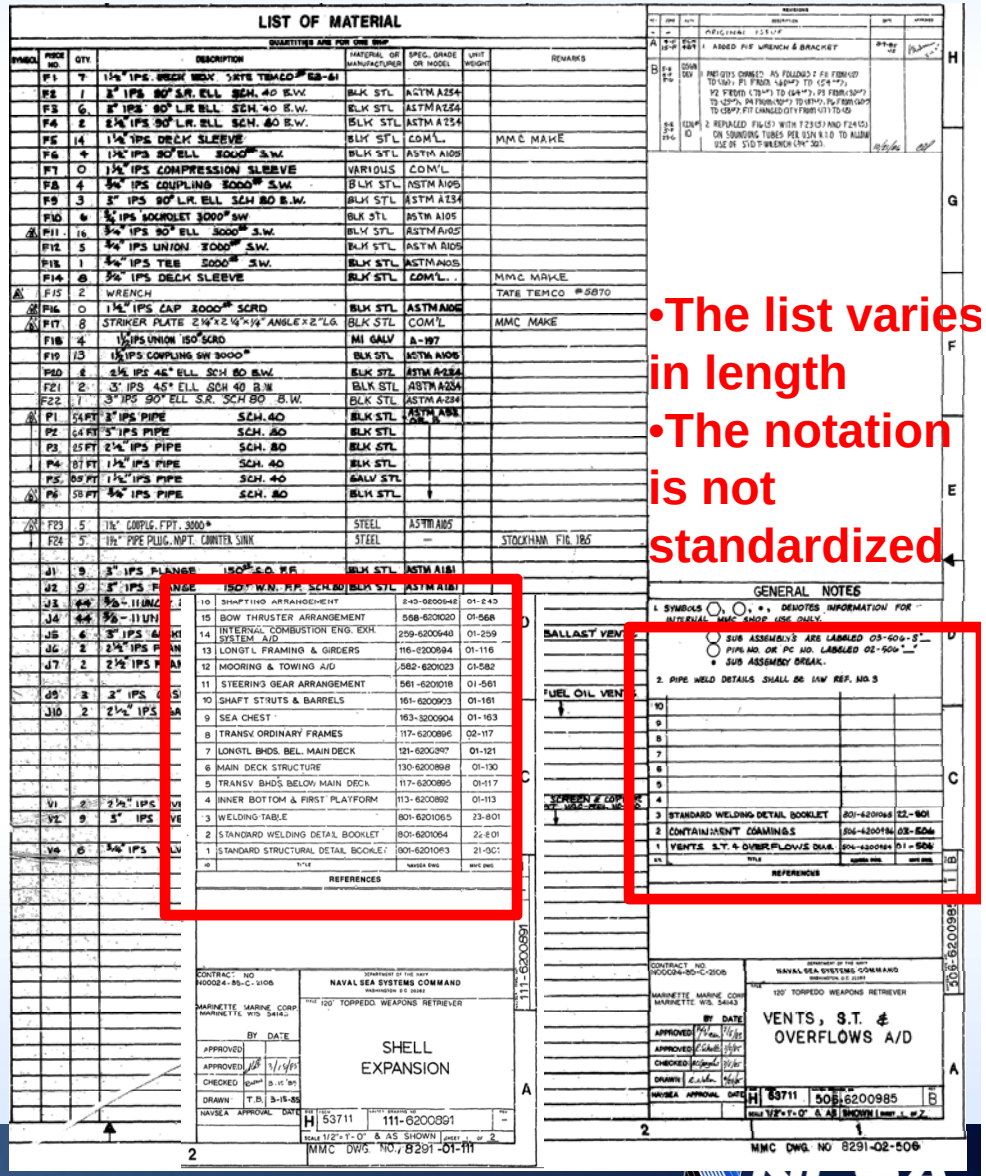
- Metadata representation: subject – predicate - object

MMC and Reference Blocks: Organization

- MMC Blocks



Inconsistencies



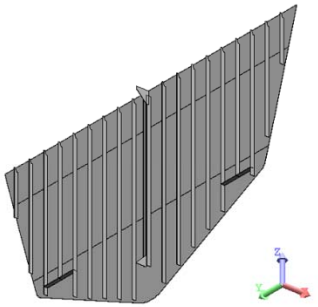
Summary of OCR Based Analyses

- 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

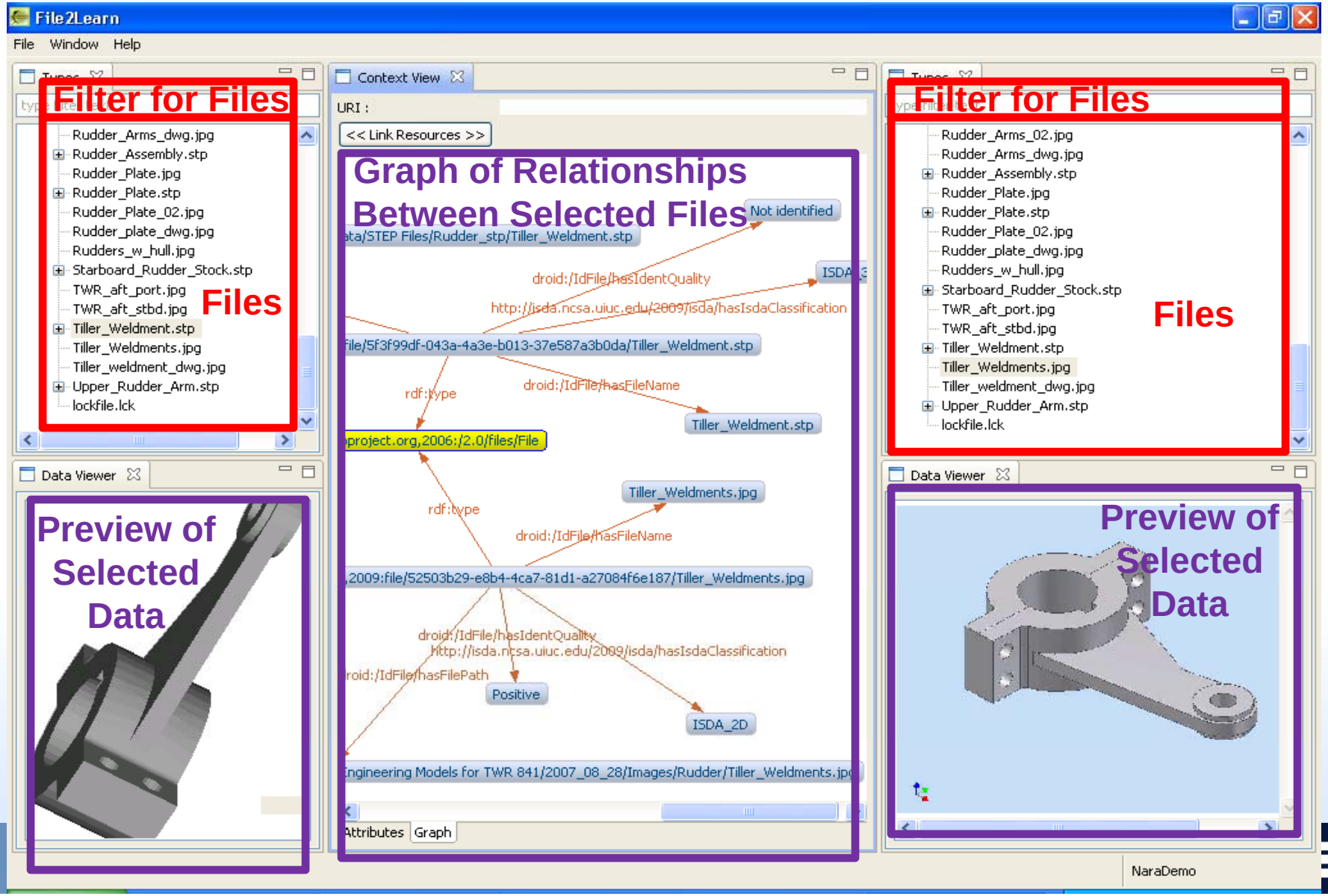
- 3D CAD models in STEP file format are searched for any ASCII strings matching English dictionary and following STEP metadata specification.

Example Metadata for TWR841 ship deck

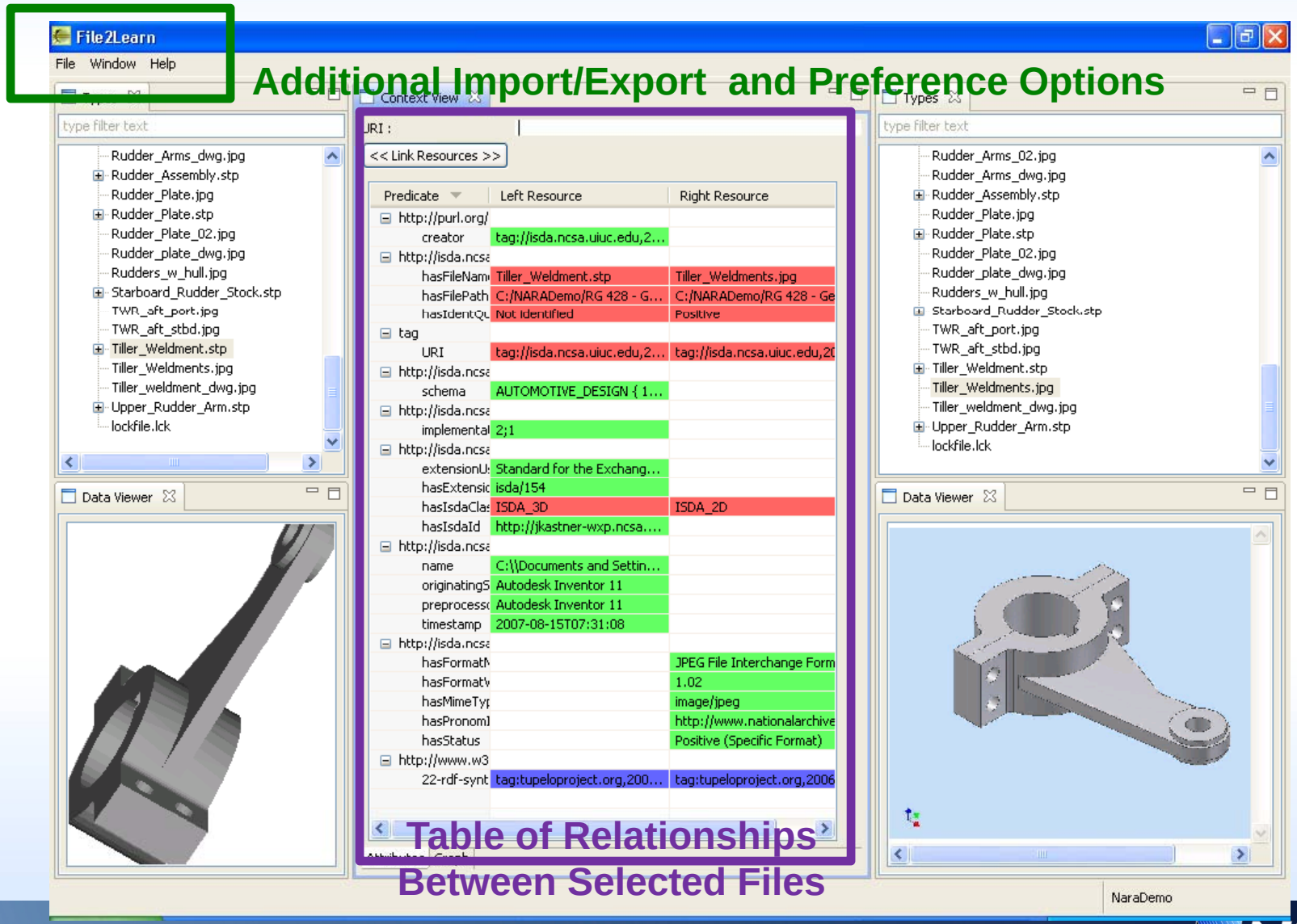


STEP METADATA SPECIFICATION	EXPECTED STEP METADATA	PARSED STEP METADATA
FILE_DESCRIPTION(/* description */ (''), /* implementation_level */ '2;1'); FILE_NAME(/* name */ '', /* time_stamp */'', /* author */ (''), /* organization */ (''), /* preprocessor_version */ ' ', /* originating_system */ '', /*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

Exploratory Framework – User Interface Overview



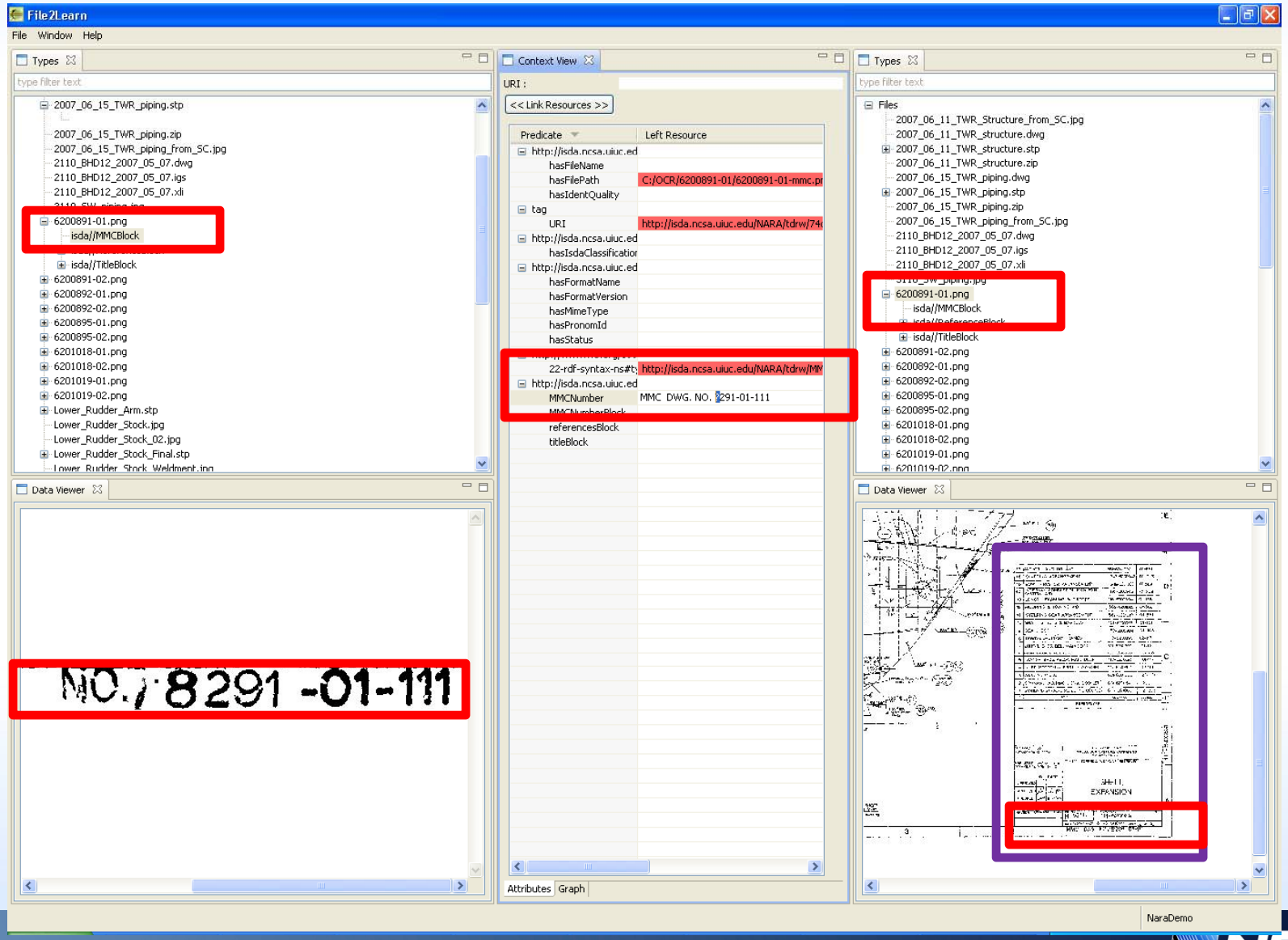
Exploratory Framework – User Interface Overview



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

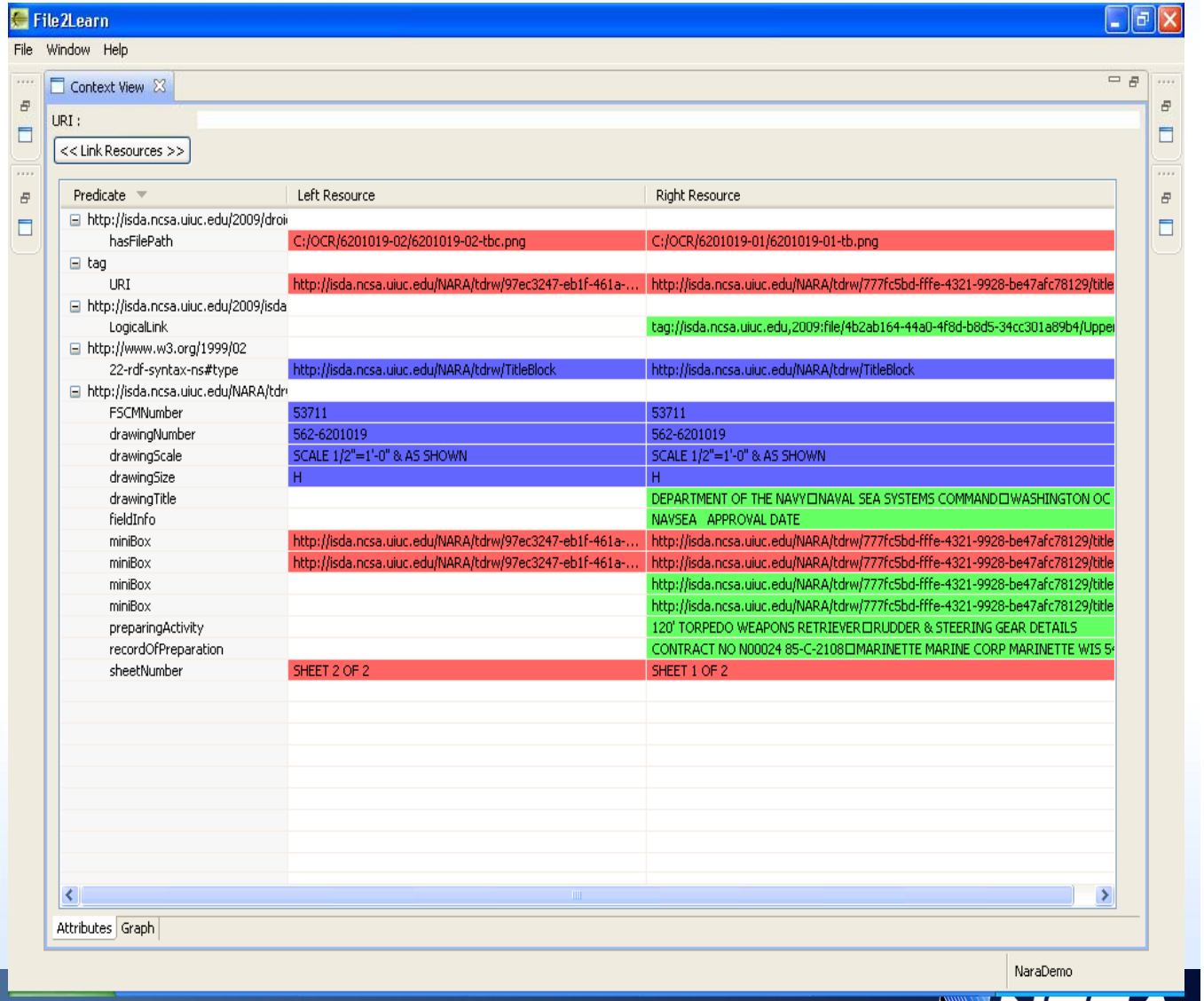
Detection of Anomalies in OCR Results



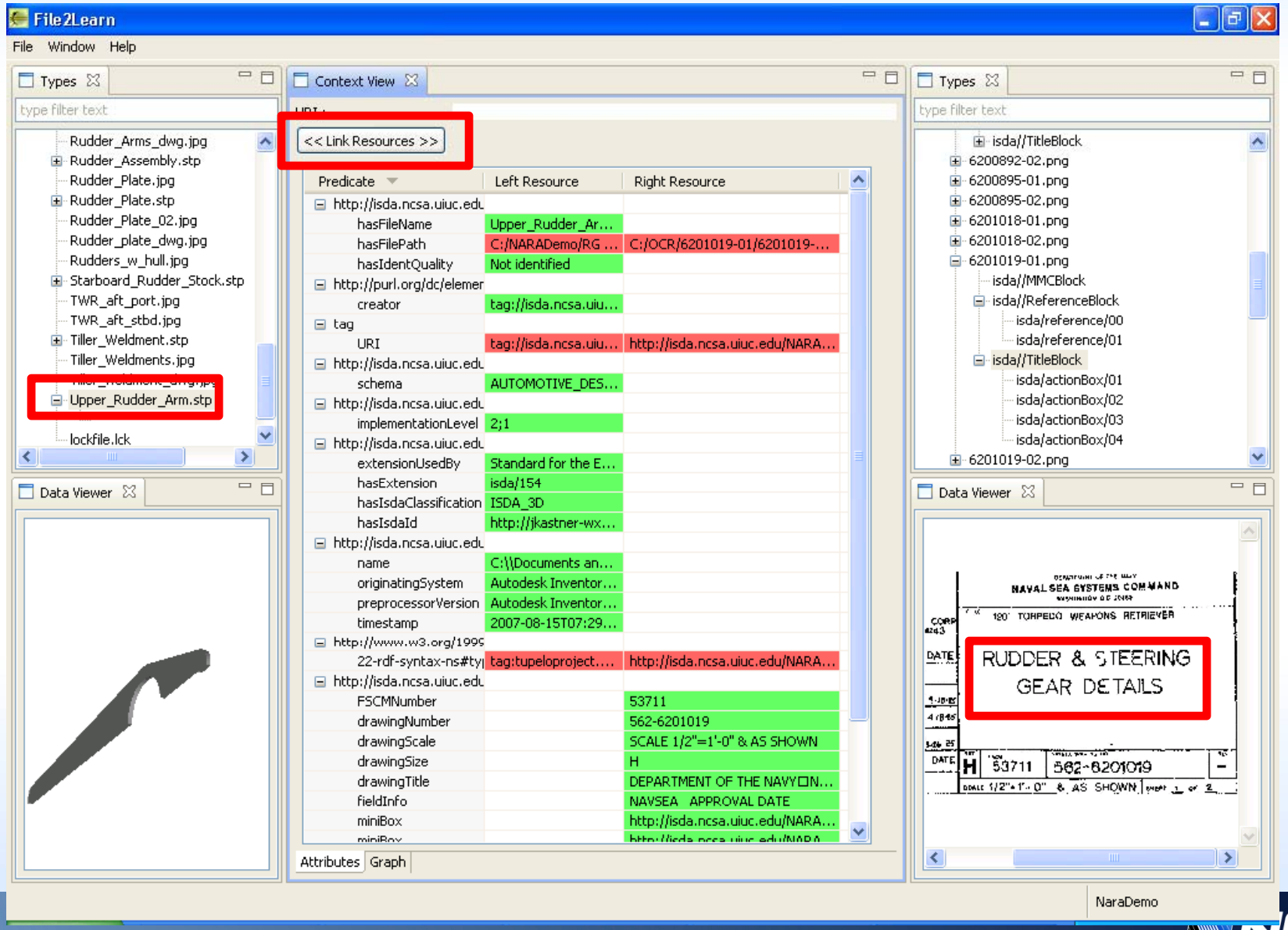
Comparison of Files

Color encoding:

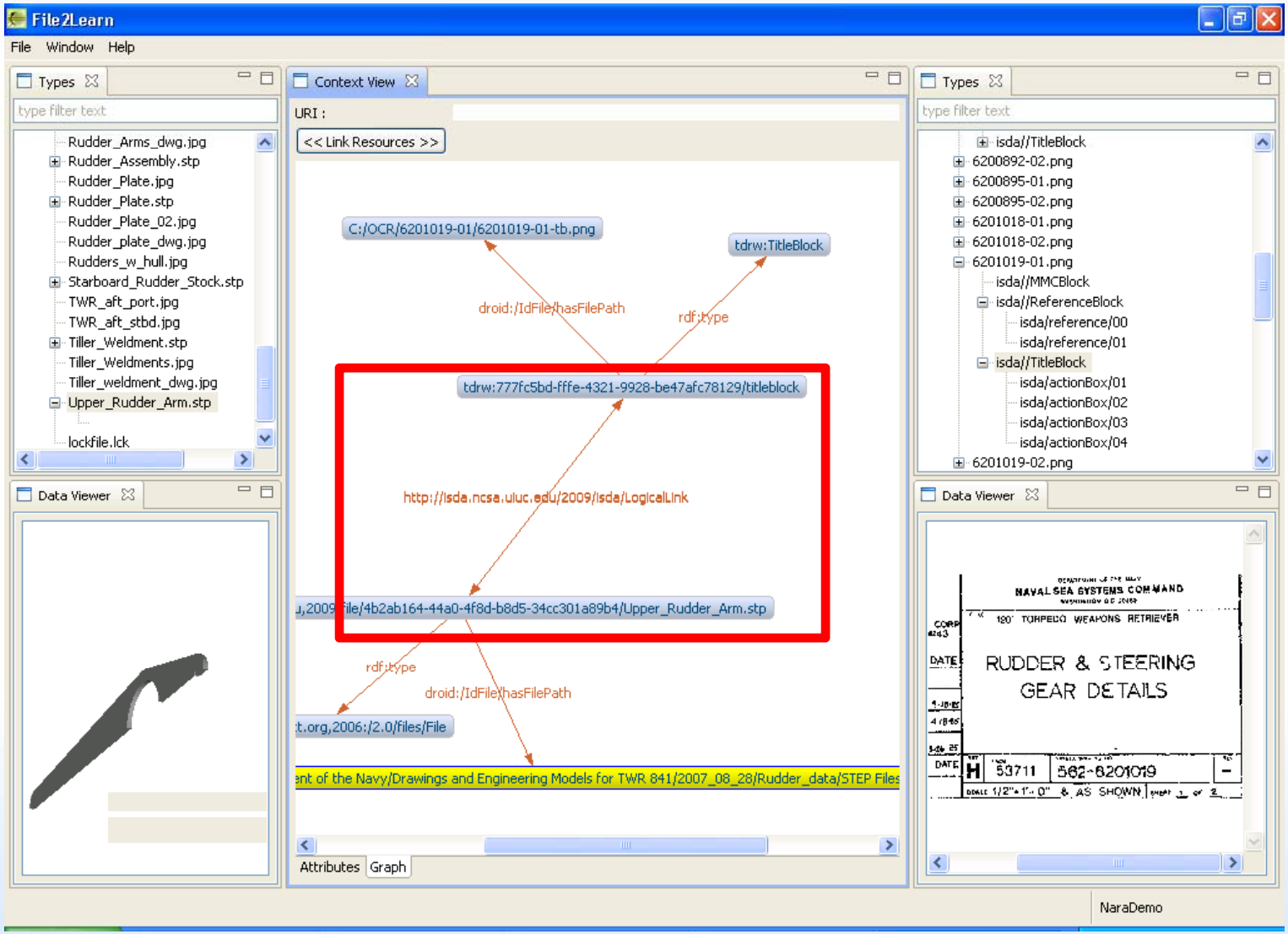
- **Predicates and values match**
- **Predicates match**
- **Predicate occurs only in one file**



Establish File Relationships



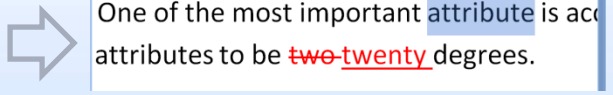
Establish File Relationships: Logical Link



A Comprehensive Comparison of Contemporary Documents

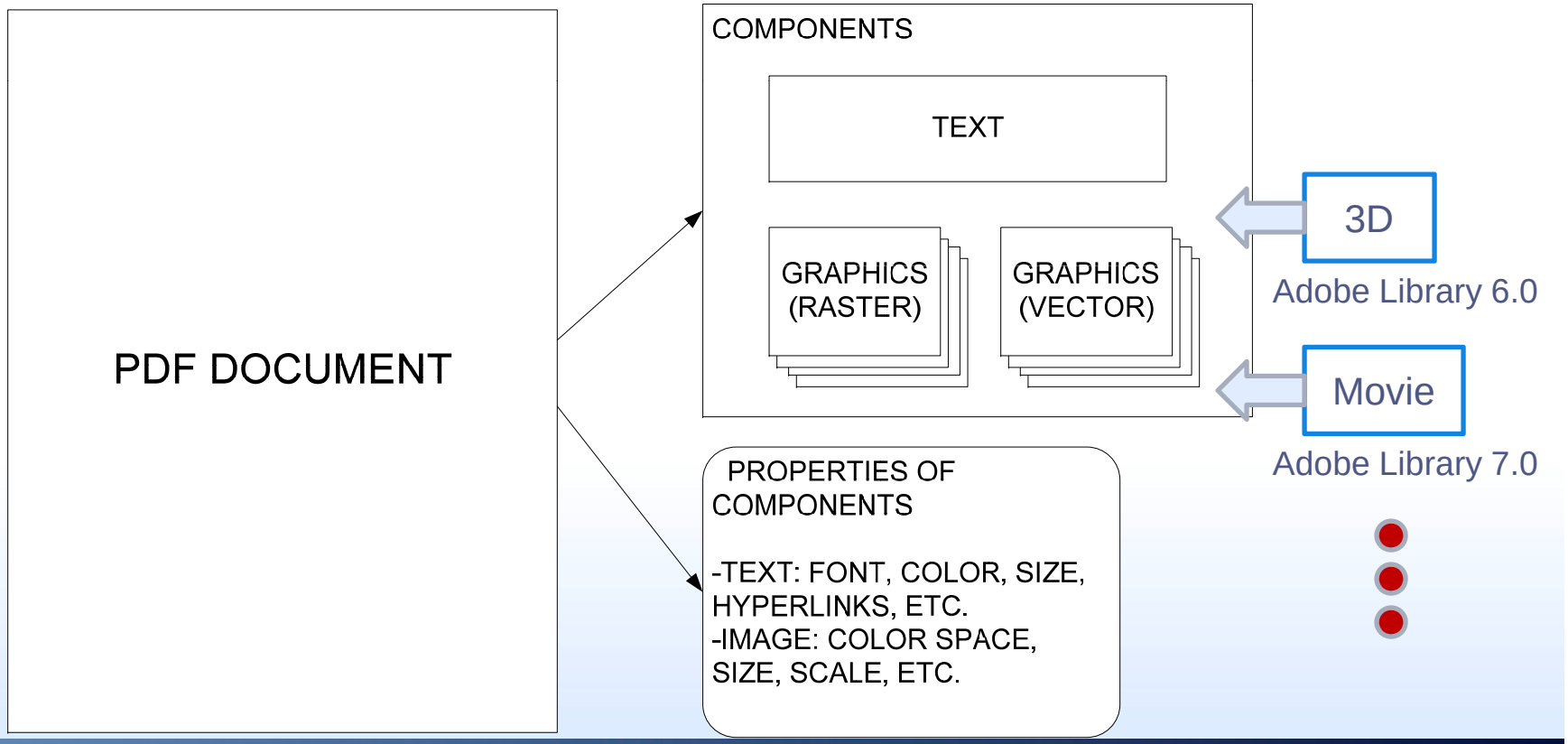
Support of Appraisals by Enabling Comparisons

- **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

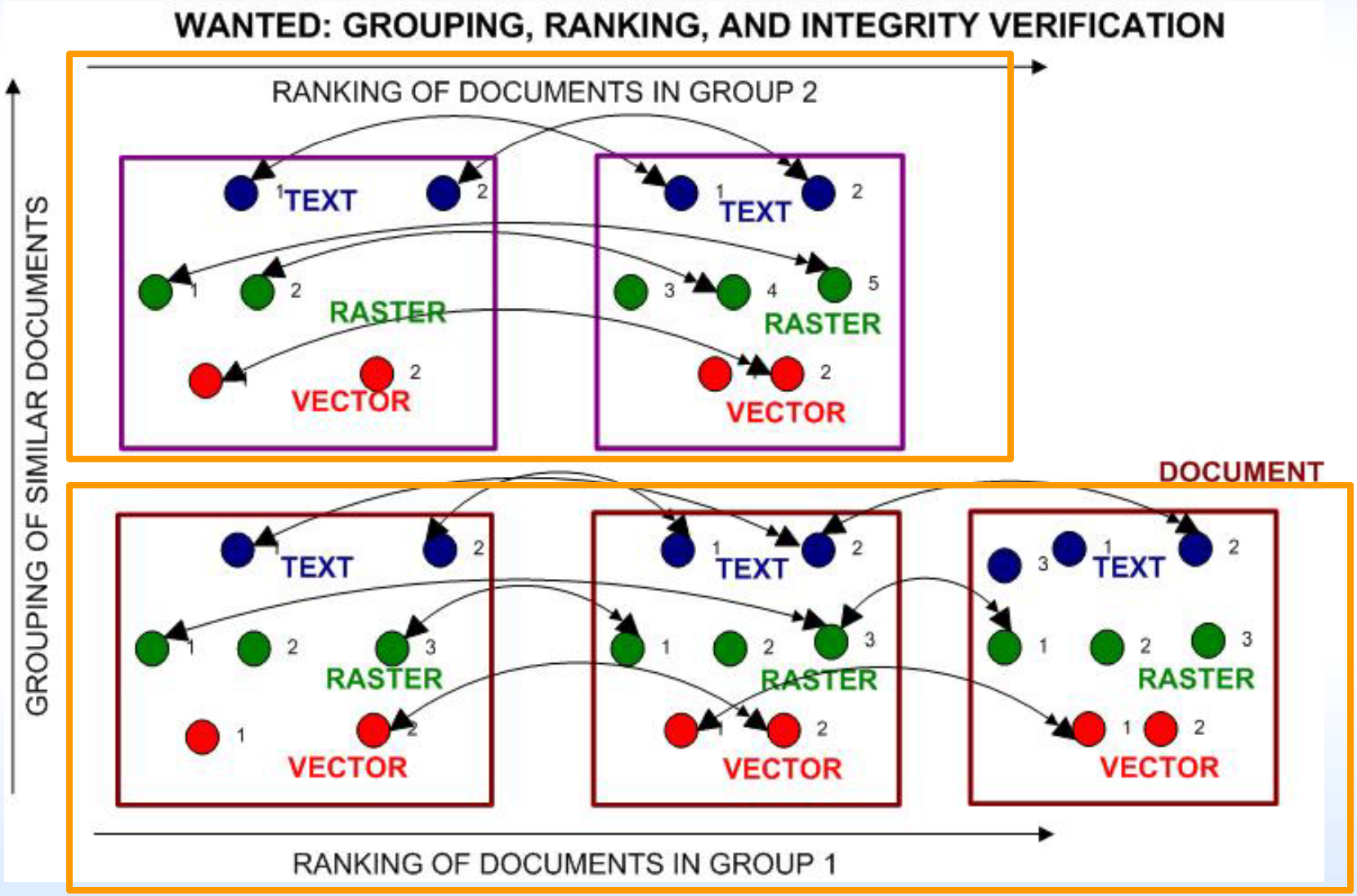


Example: Adobe Portable Document Format (PDF)

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

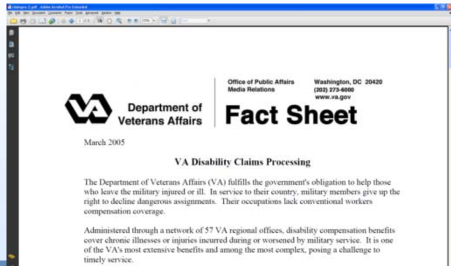


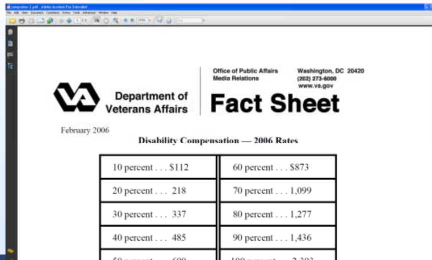
Comparisons



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**
  - **comprates-2.pdf, comprates.pdf**



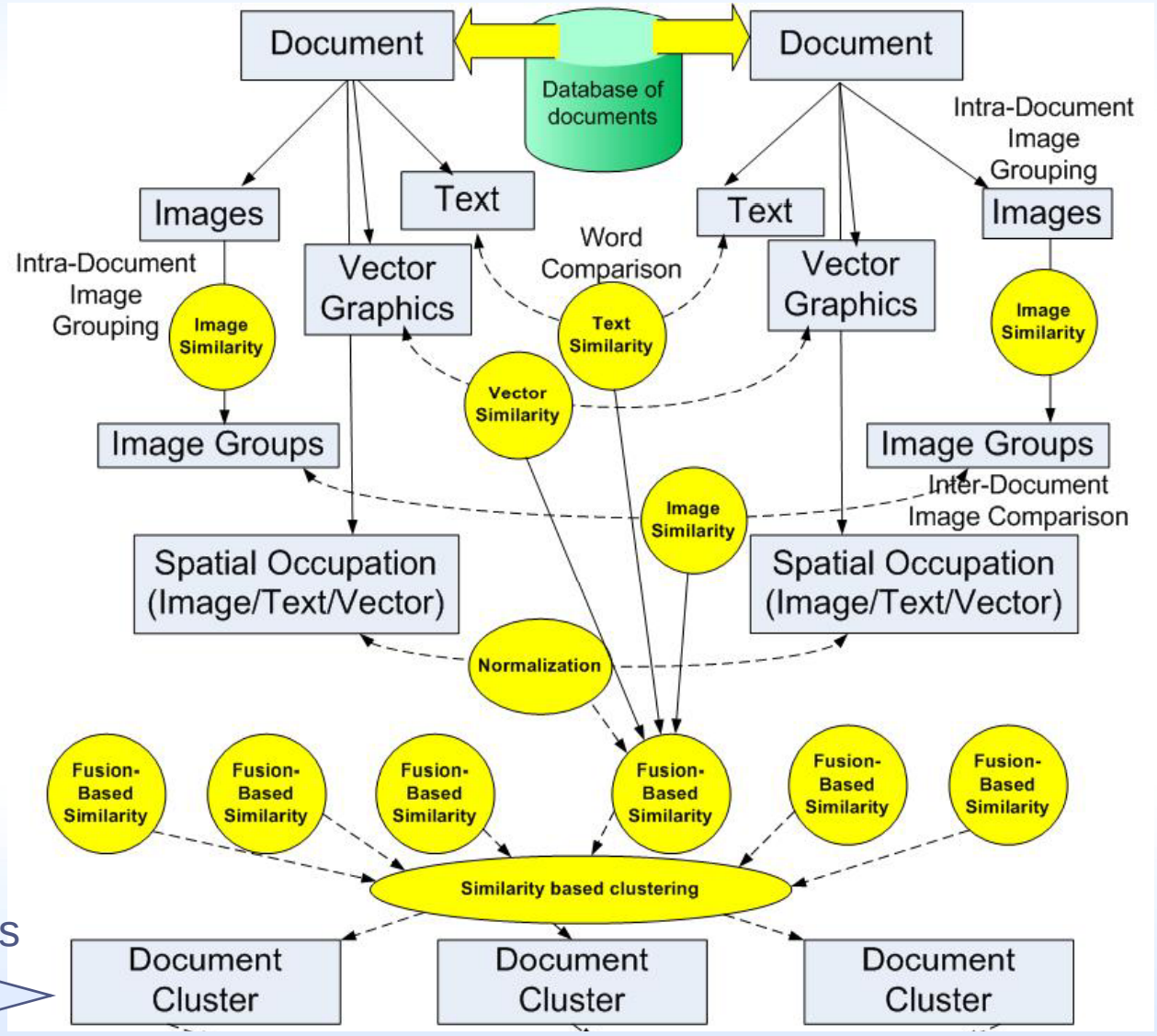


Methodology

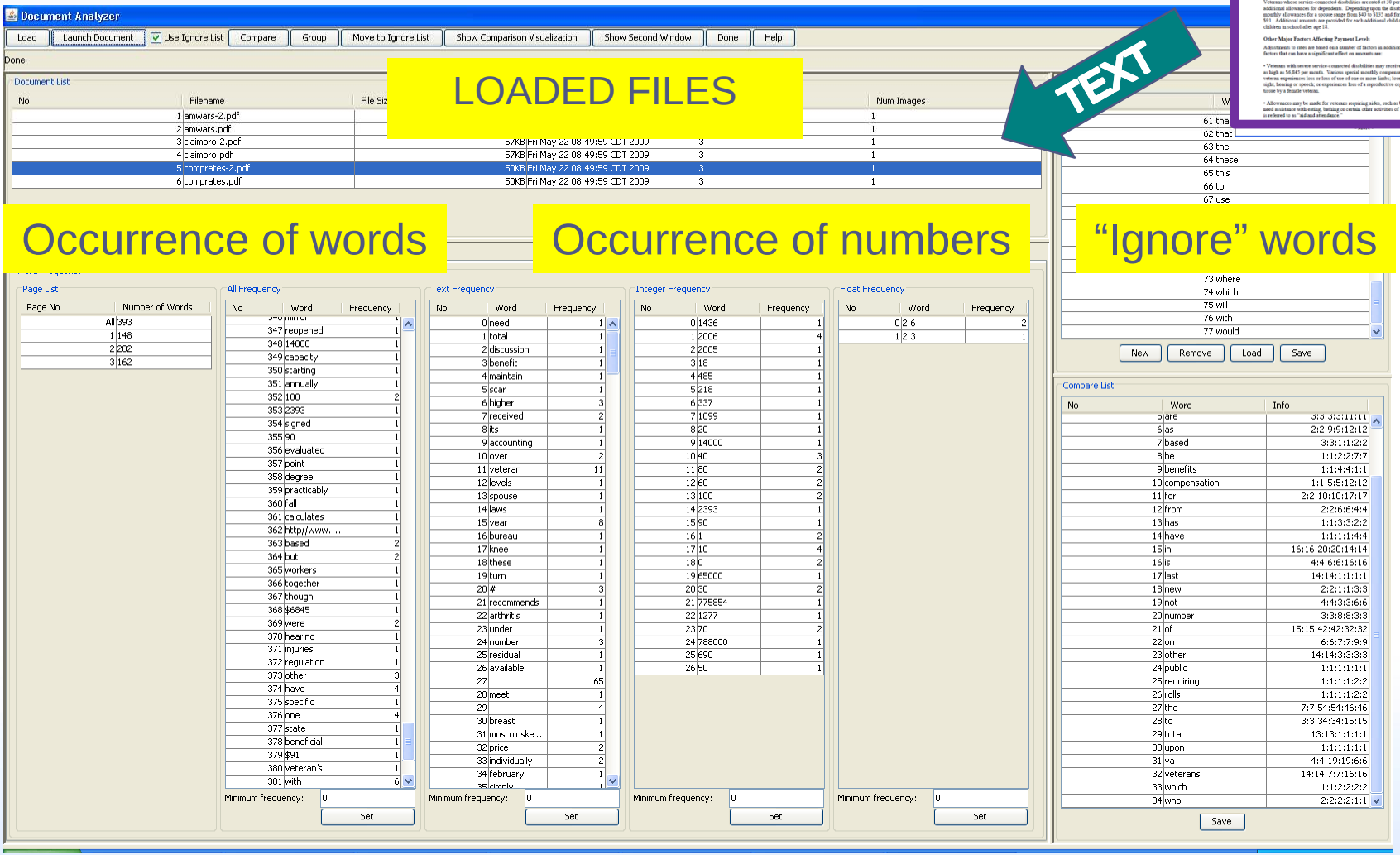
Pair-wise comparison of the same digital objects

Comparison of multiple and heterogeneous digital objects

Relationship to Permanent Records

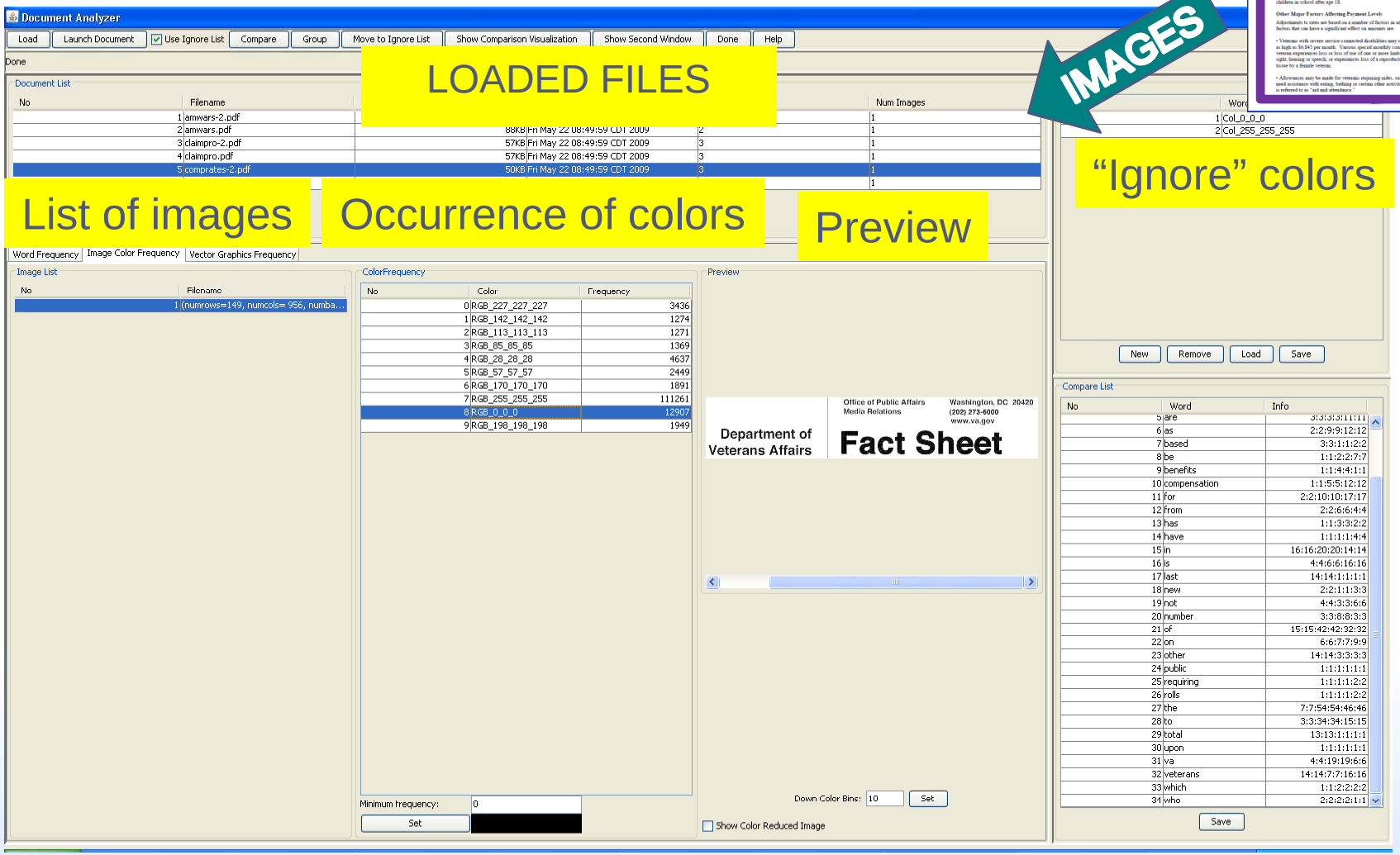


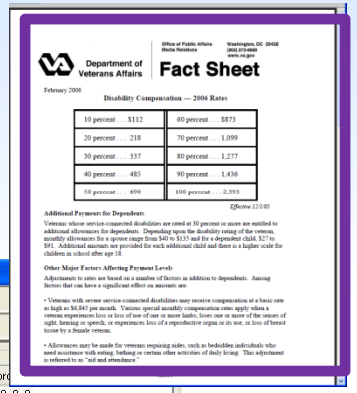
Exploration of Text Components



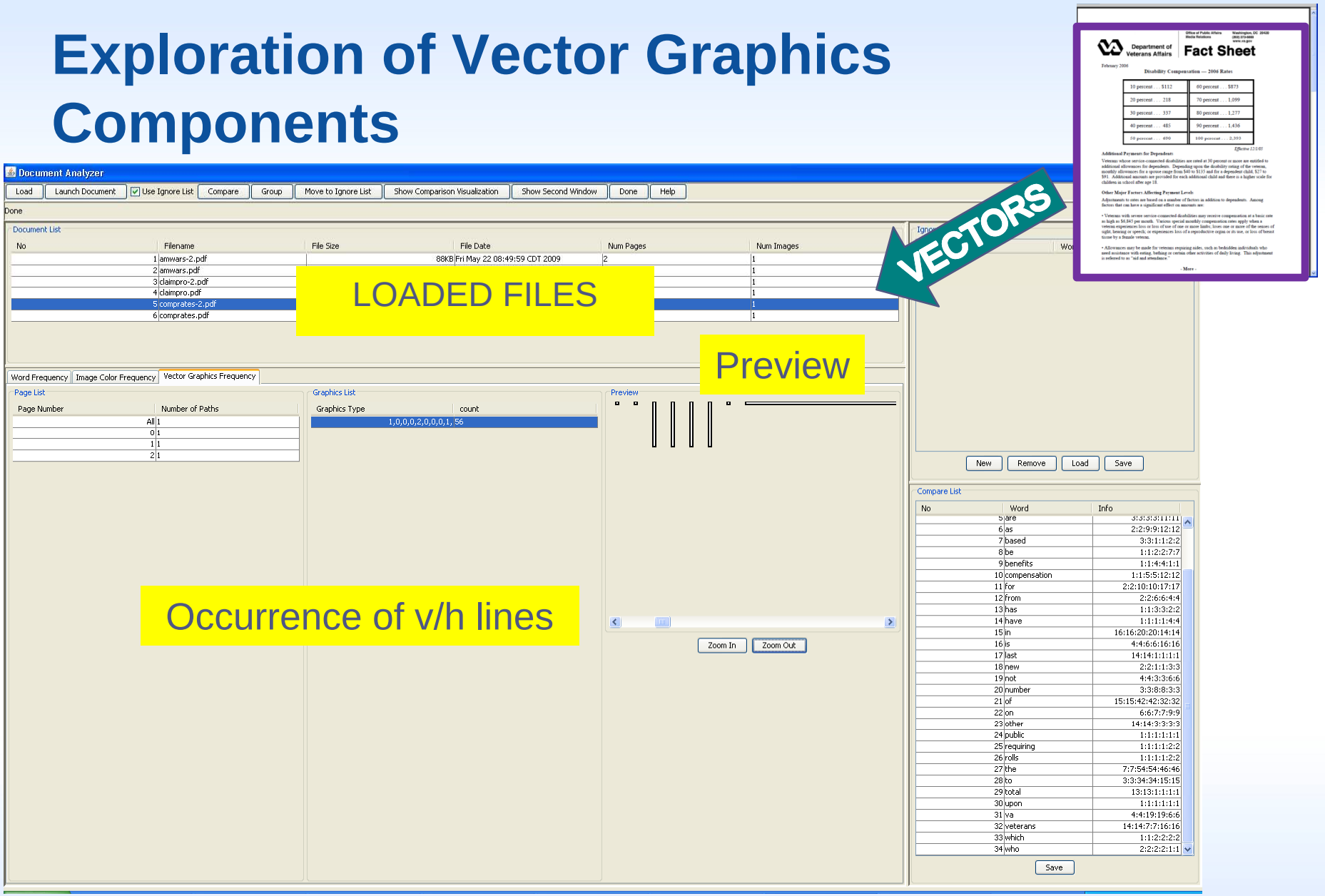
NCSA

Exploration of Image Components

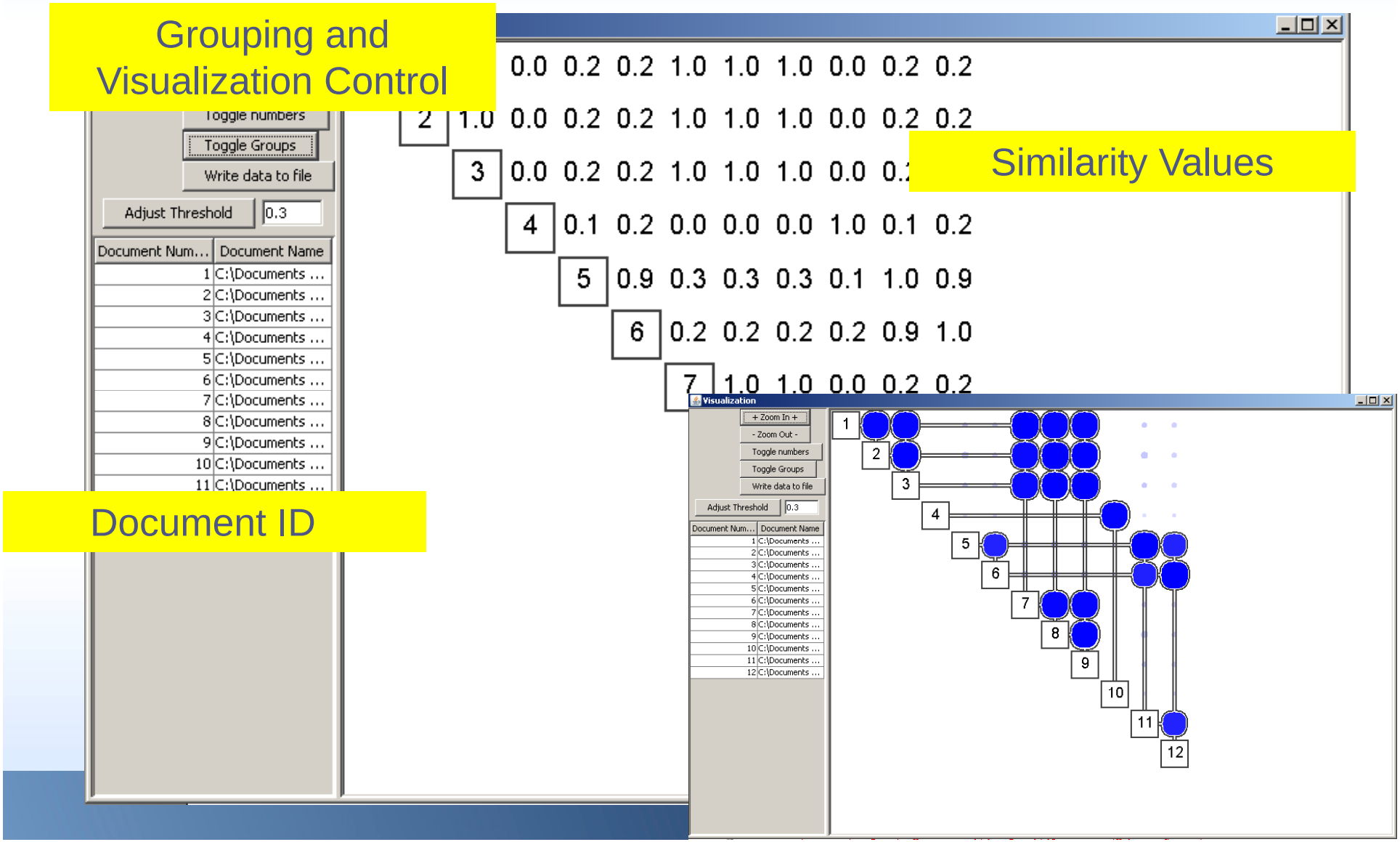




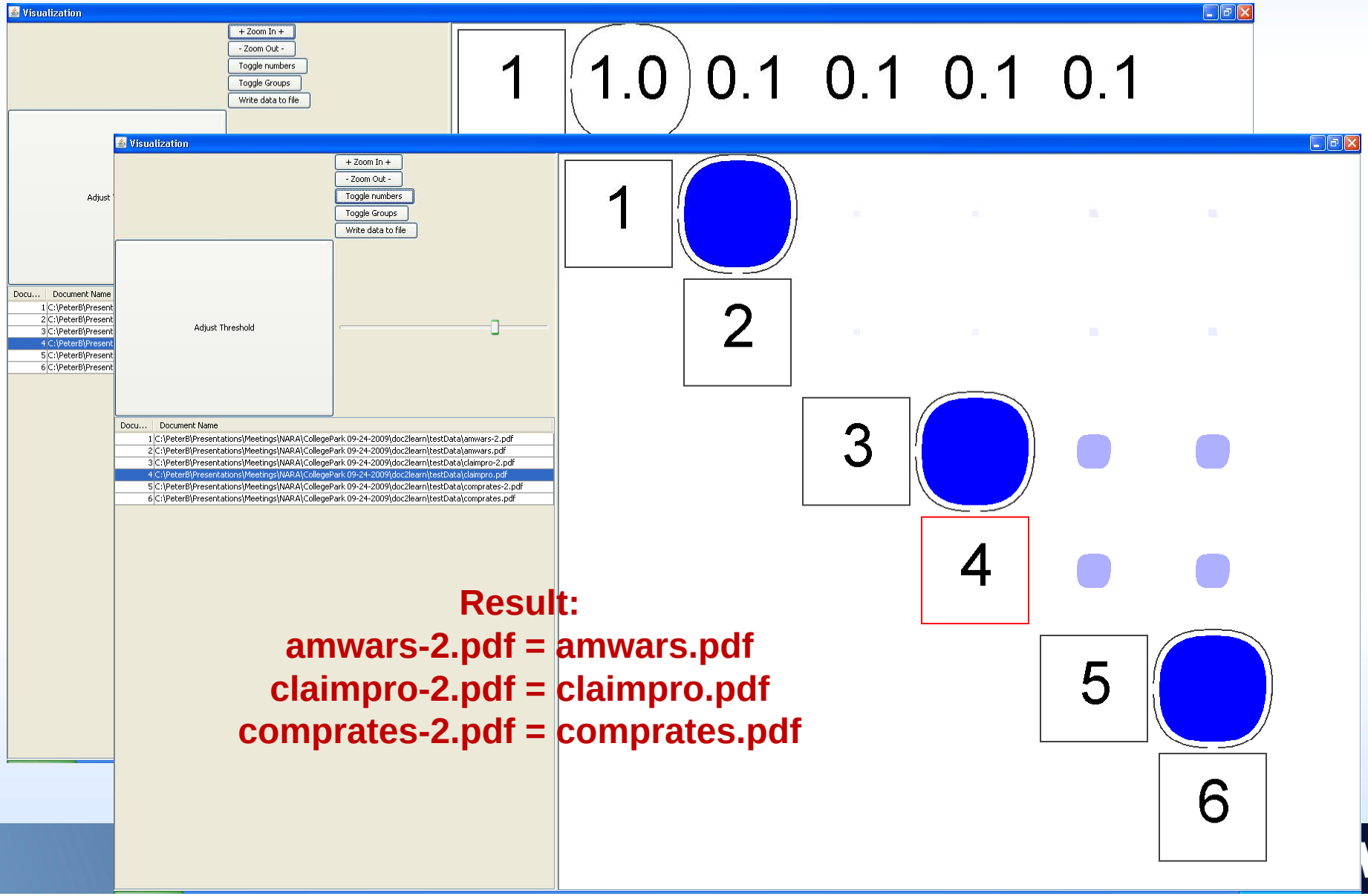
Exploration of Vector Graphics Components



Comprehensive Pair-Wise Comparison of Documents



Visual Comparison for 6 Test Files



Result:

amwars-2.pdf = amwars.pdf

claimpro-2.pdf = claimpro.pdf

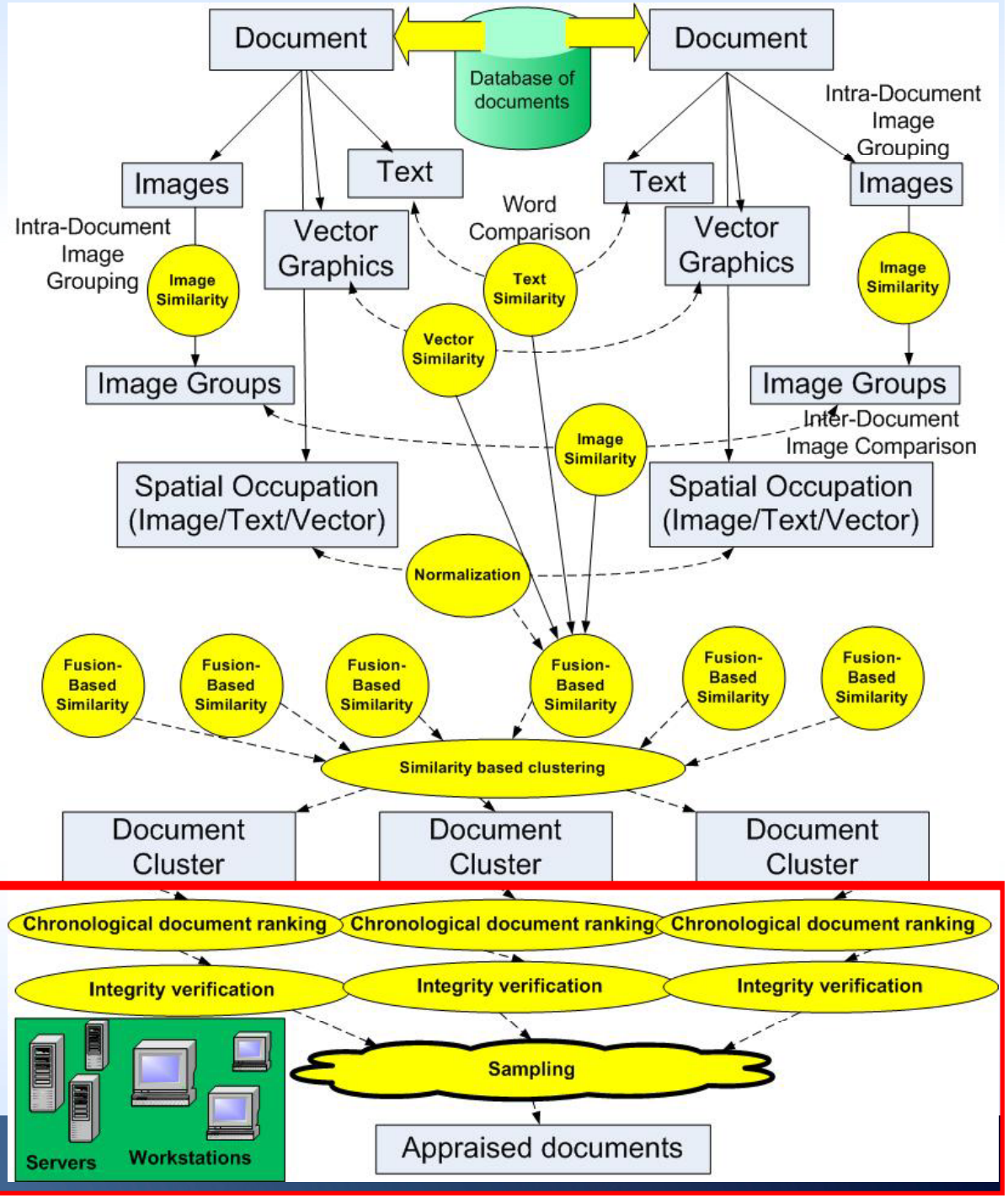
comprates-2.pdf = comprates.pdf

Computational Requirements for Executing the Methodology

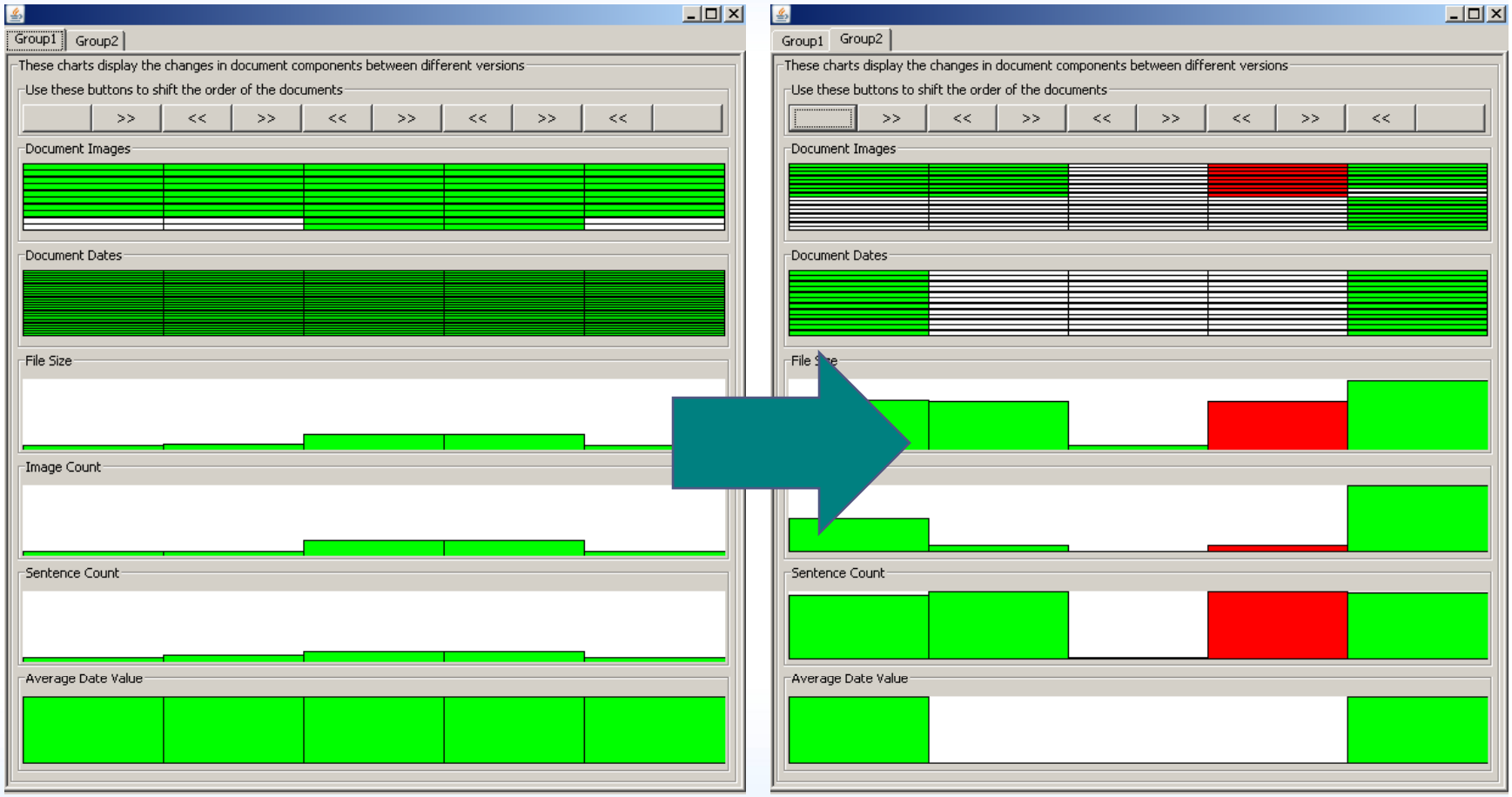
Yellow indicates computations

Relationship to Permanent Records

Appraisal & Sampling



Work in progress: Group and Validate Documents

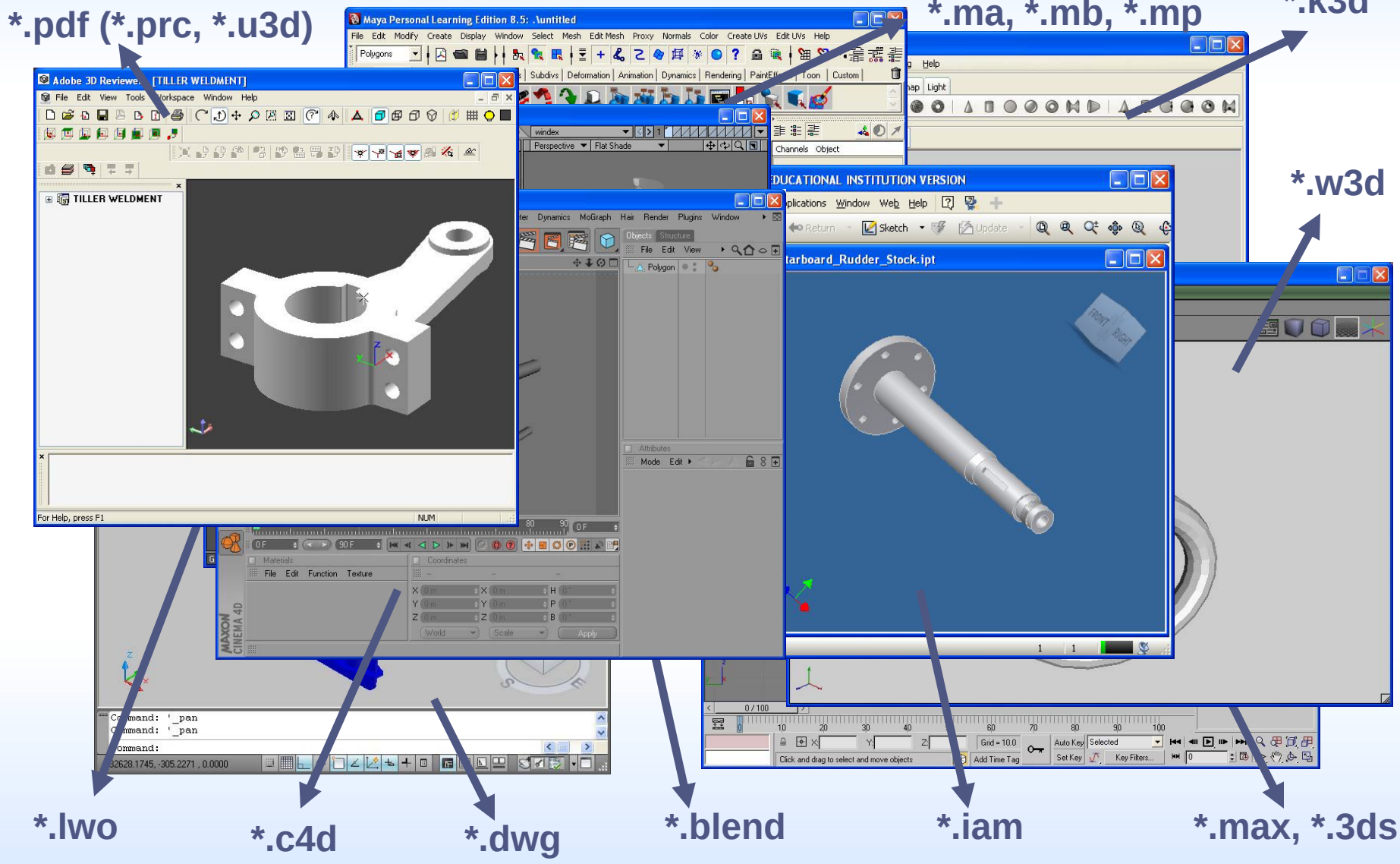


Automated File Format Conversions and Conversion Quality Assessment

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

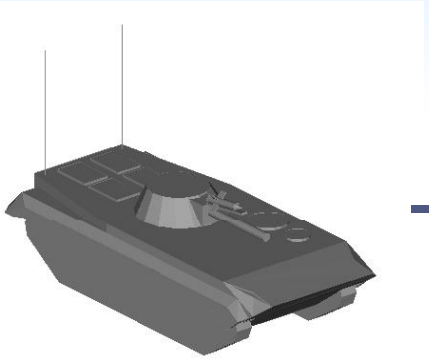
- **Q: How Many 3D File Formats Exist?**
- A: We have found more than 140 3D file formats. Many are proprietary file formats. Many are extremely complex (1,200 and more pages of specifications).
- **Q: How Many Software Packages Support 3D File Format Import, Export and Display?**
- A: We have documented 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

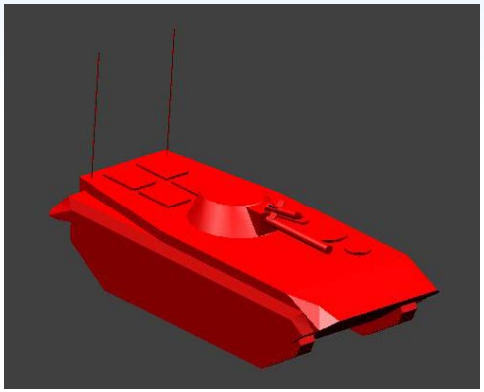
Example: Conversion of X3D to STEP to X3D



X3D

Software:

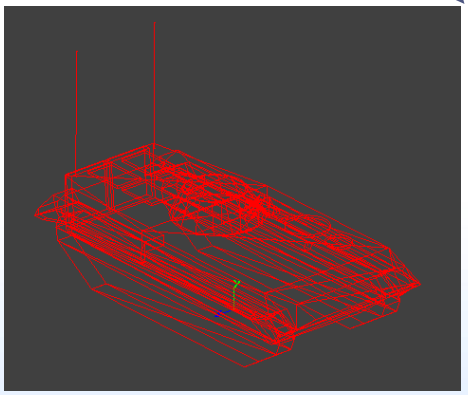
X3dToVrml97



WRL

Software:

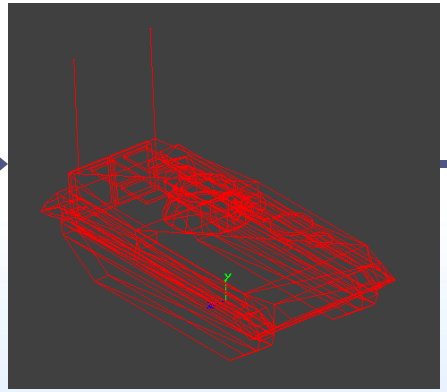
A3D Reviewer



STEP

Software:

A3D Reviewer



WRL

Software:

Vrml97ToX3d

Nothing!

X3D

Towards a Universal Converter

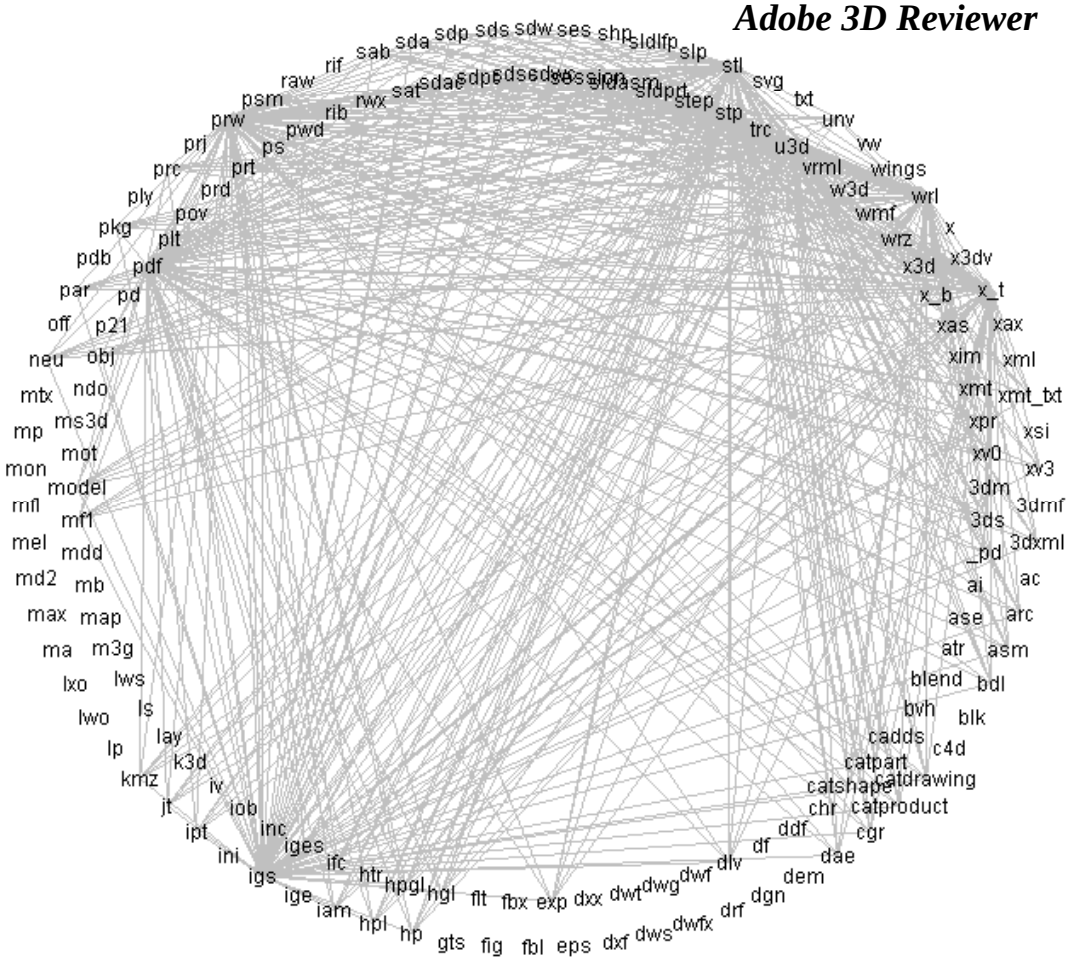
- Use what is available in 3rd party software to perform conversions
  - Document what formats can be opened/imported by each application
  - Document what formats can be saved/exported by each application
  - Automate the use of each application and combine their abilities to perform conversions over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

Input/Output Graphs

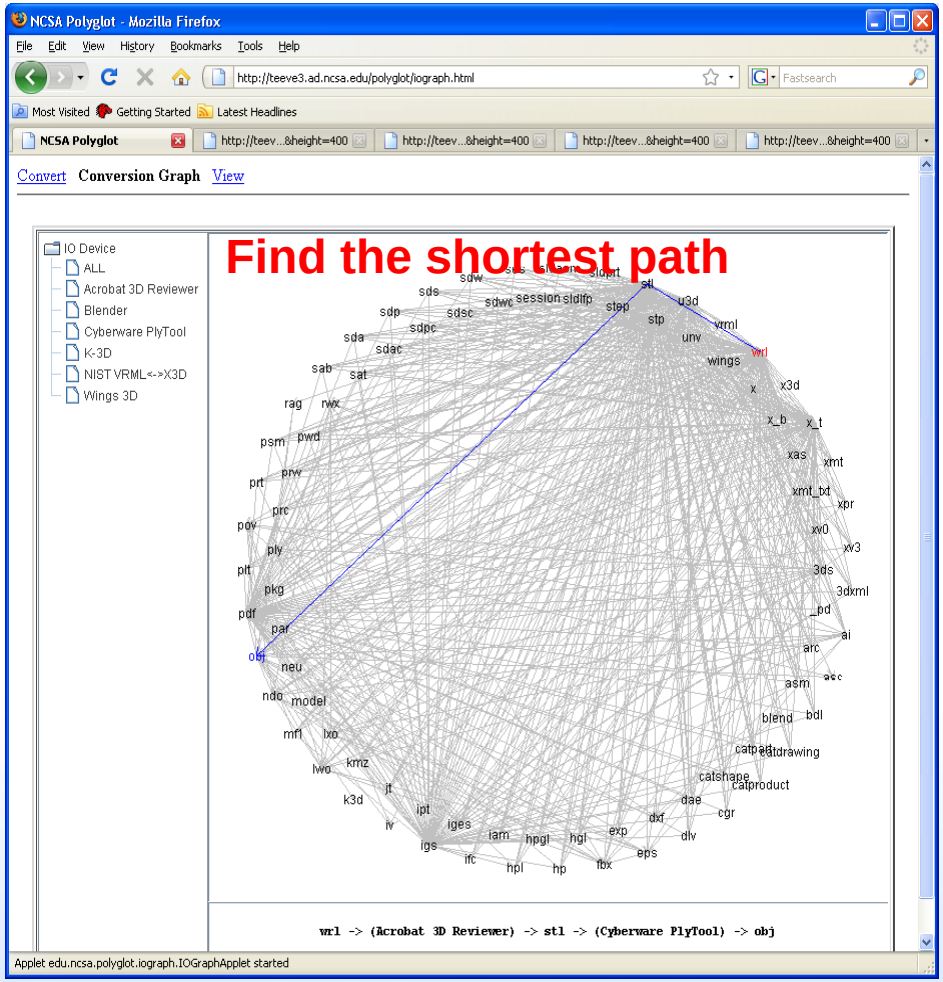
Adobe 3D Reviewer

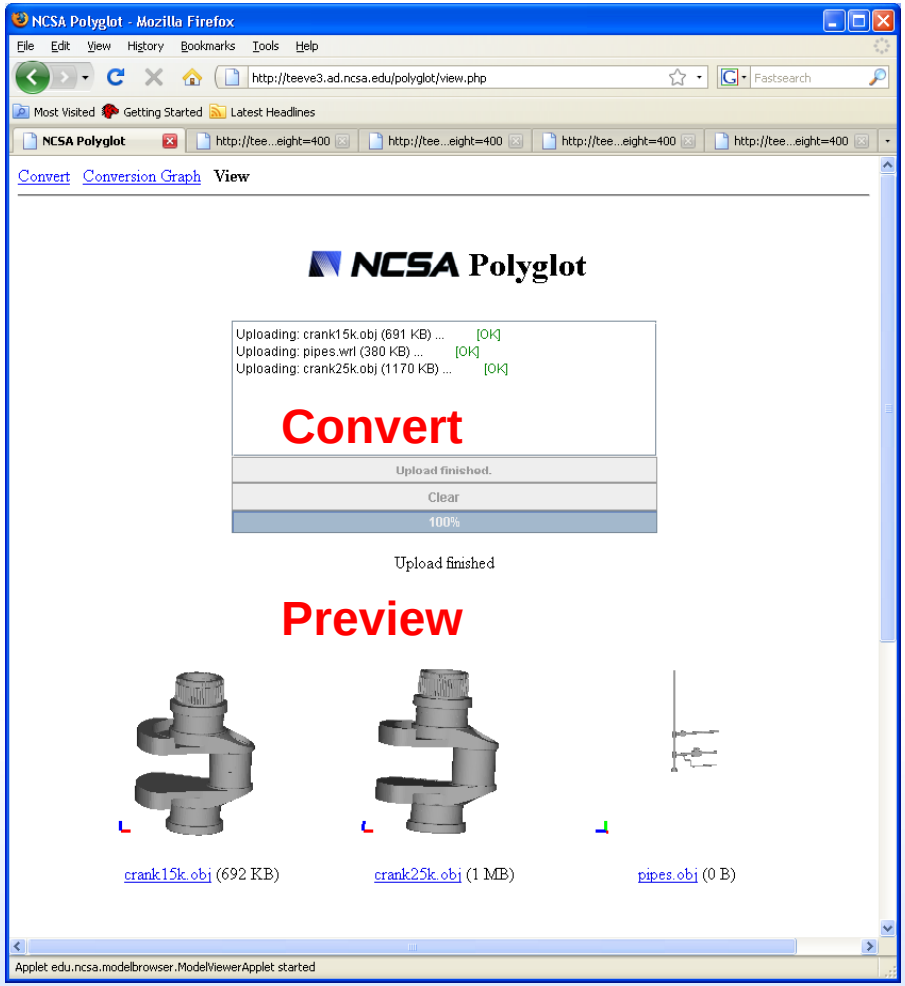
Manufacturer	Adobe
Version	9, Trial
URL	http://www.adobe.com
Cost	
Opens	3ds, 3dxml, arc, asm, bdl, catdrawing dae, dlv, exp, hgl, hp, hpgl, hpl, iam, model, neu, obj, pd, par, pdf, pkg, pl sda, sdac, sdp, sdpc, sds, sdsc, sdw, sldprt, stl, step, stp, u3d, unv, wrl, vr xv0, xv3
Saves	prw
Imports	3ds, 3dxml, arc, asm, bdl, catdrawing dae, dlv, exp, hgl, hp, hpgl, hpl, iam, model, neu, obj, pd, par, pdf, pkg, pl sda, sdac, sdp, sdpc, sds, sdsc, sdw, sldprt, stl, step, stp, u3d, unv, wrl, vr xv0, xv3
Exports	igs, pdf, stl, stp, u3d, wrl, x t

Adobe 3D Reviewer



Automation of 3D File Format Mapping





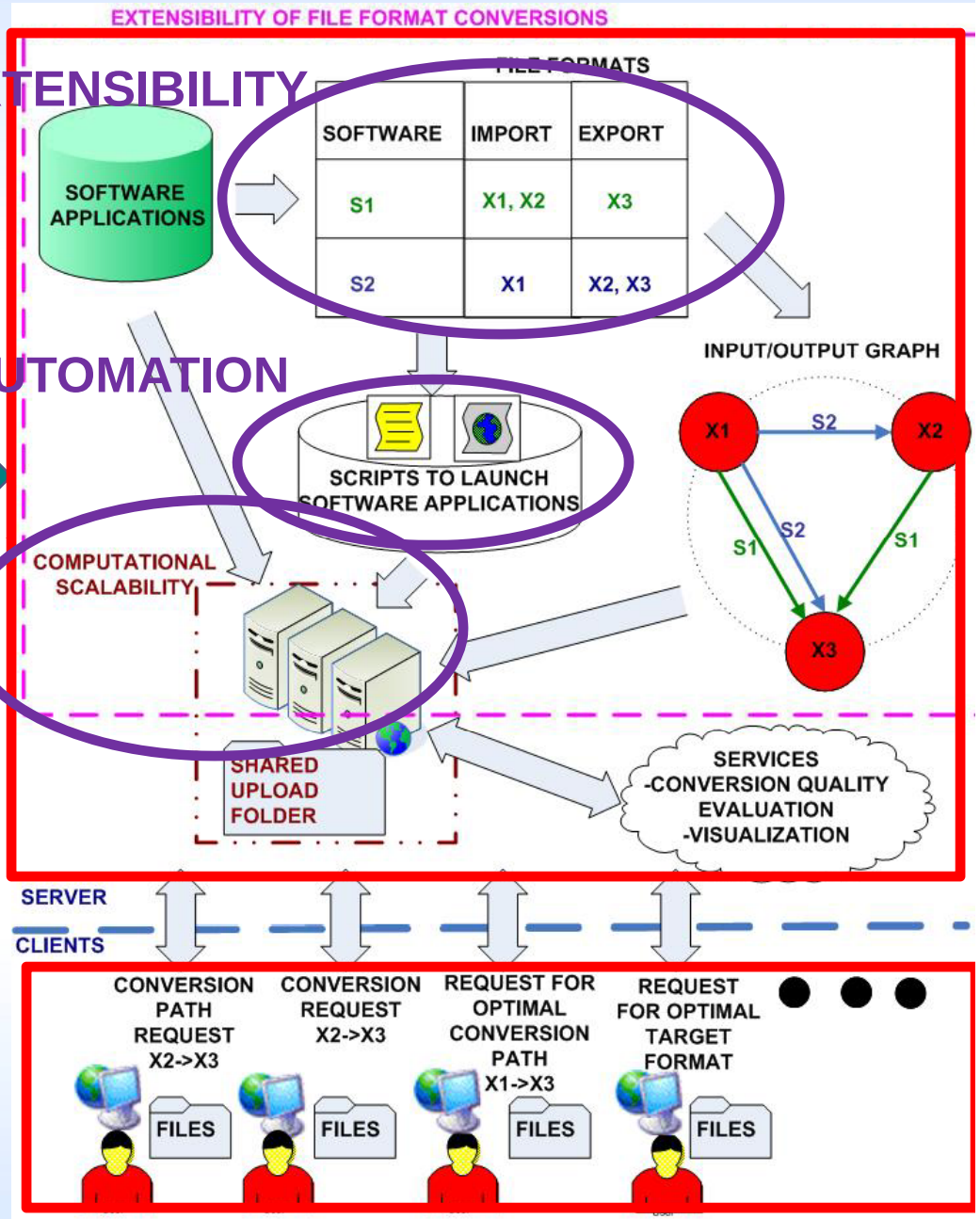
Automation of 3D File Format Conversion

- 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

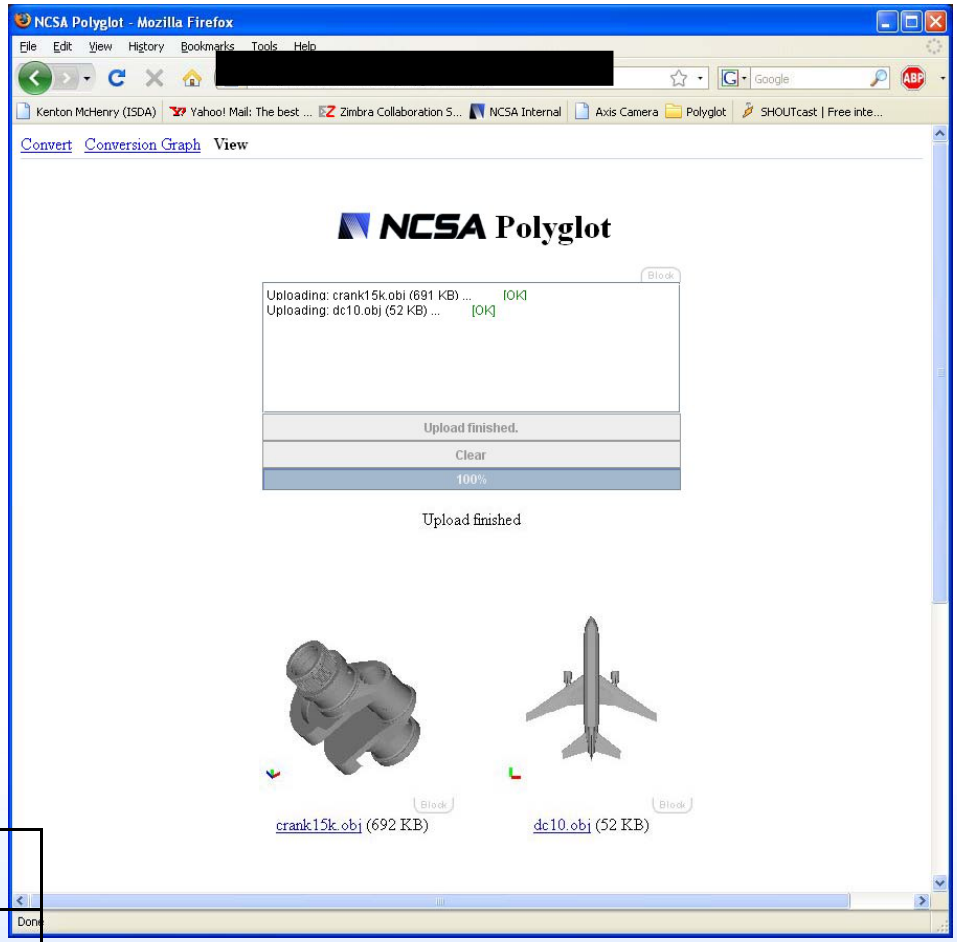
Cloud Computing

Services to Archivists



NCSA Polyglot – Conversion Services

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



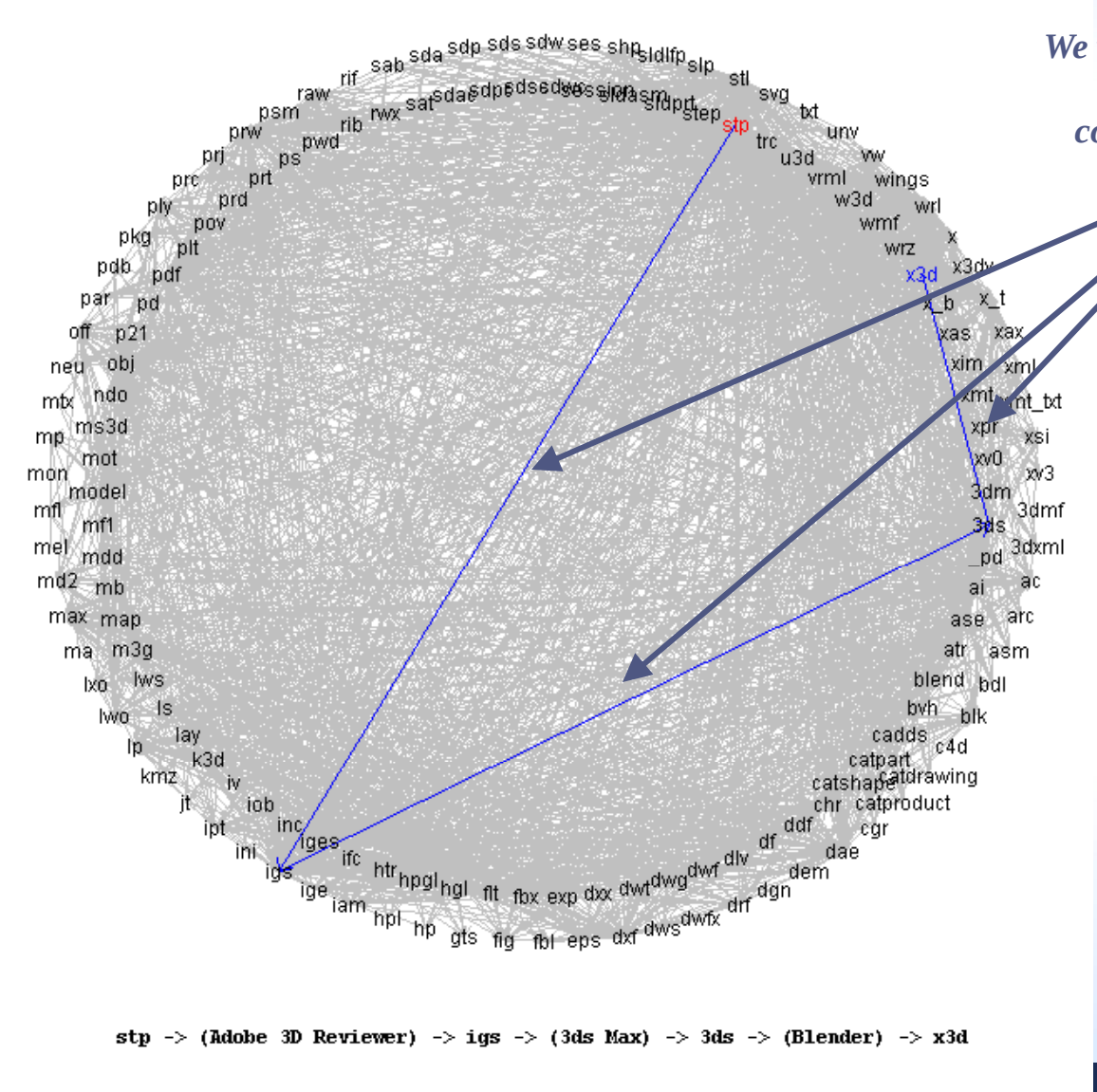
- Java interface:

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

- Scalability Test

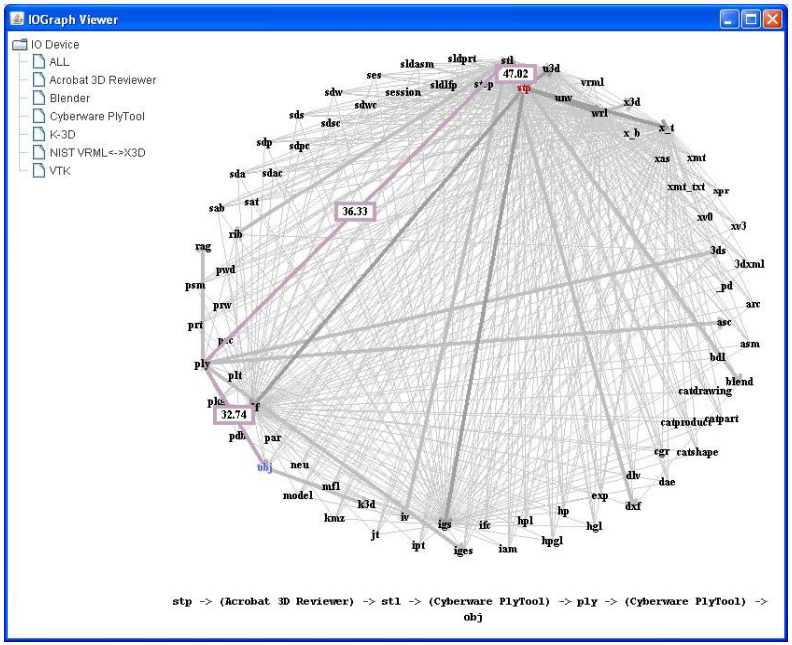
Number of PCs	One PC	Two PCs
Processing Time	33 minutes 6 seconds	16 minutes 40 seconds

NCSA Polyglot – Data Loss Measurement Services



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

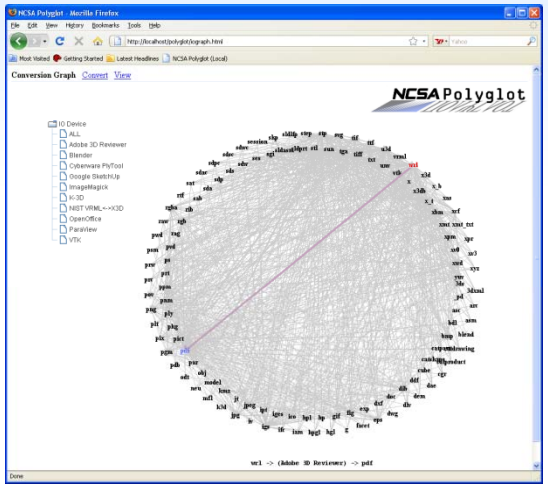
Summary

- 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

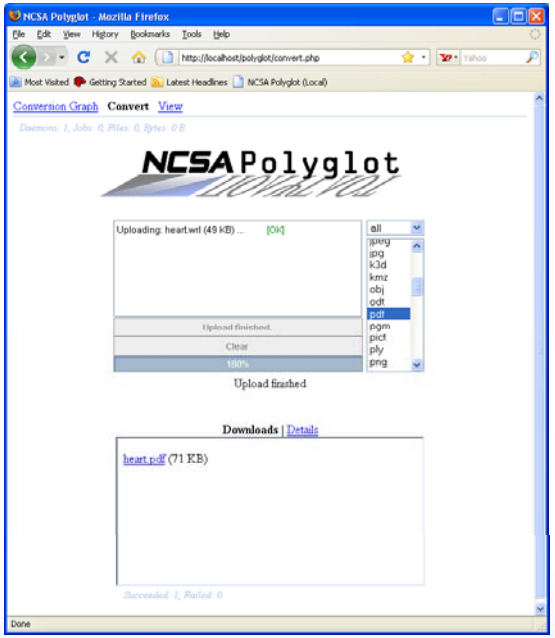
- Questions: Peter Bajcsy – pbajcsy@ncsa.uiuc.edu

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