Computer Assisted Appraisal of Contemporary PDF Documents

Presented by: Peter Bajcsy, PhD

- Research Scientist, NCSA

- Adjunct Assistant Professor ECE & CS at UIUC

- Associate Director Center for Humanities, Social Sciences and Arts (CHASS), UIUC

Contributions by: Peter Bajcsy, Sang-Chul Lee and William McFadden

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

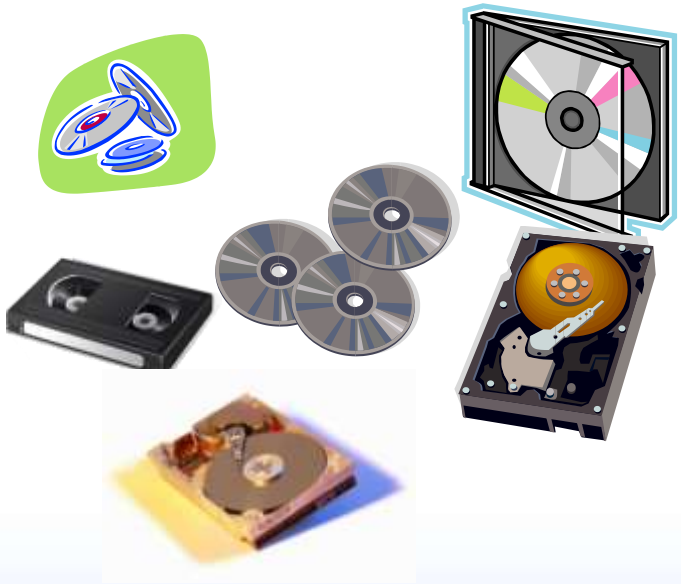
Outline

- Introduction
  - The Strategic Plan of The National Archives and Records Administration 2006–2016
- Motivation
  - Past & current research
- Computer-Assisted Appraisal of Documents
  - Approach
  - PDF documents
  - Methodology
- Experimental Results
  - Grouping, Ranking and Integrity Verification
- Conclusions

Introduction: To Be Preserved!

Information transfer ?

Digital representation of information & knowledge

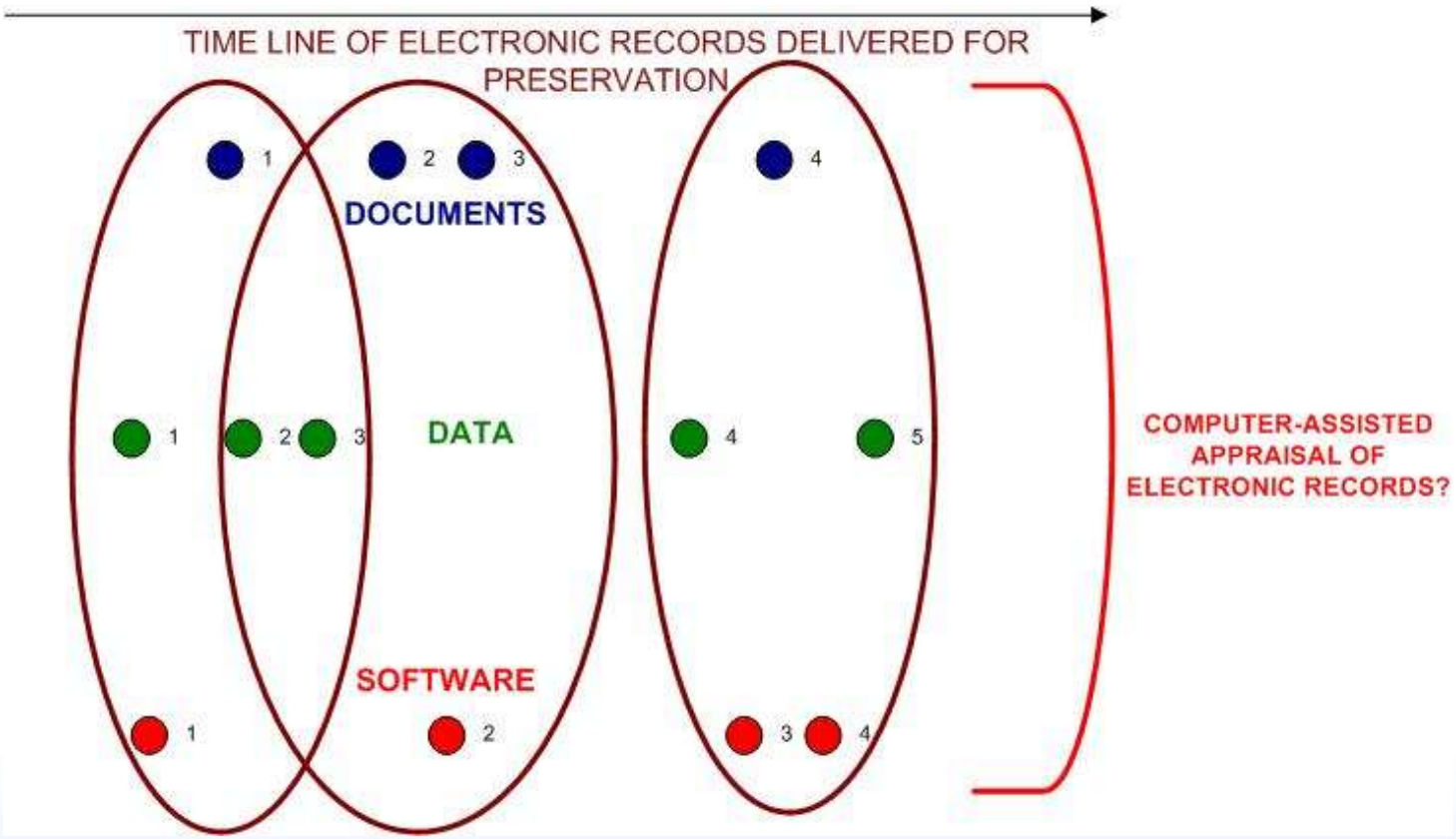


Preservation

AGENCY

ARCHIVES

Introduction: What Should Be Done?



- Can People Do It Manually?
- Human versus Computer or Human with Computer?

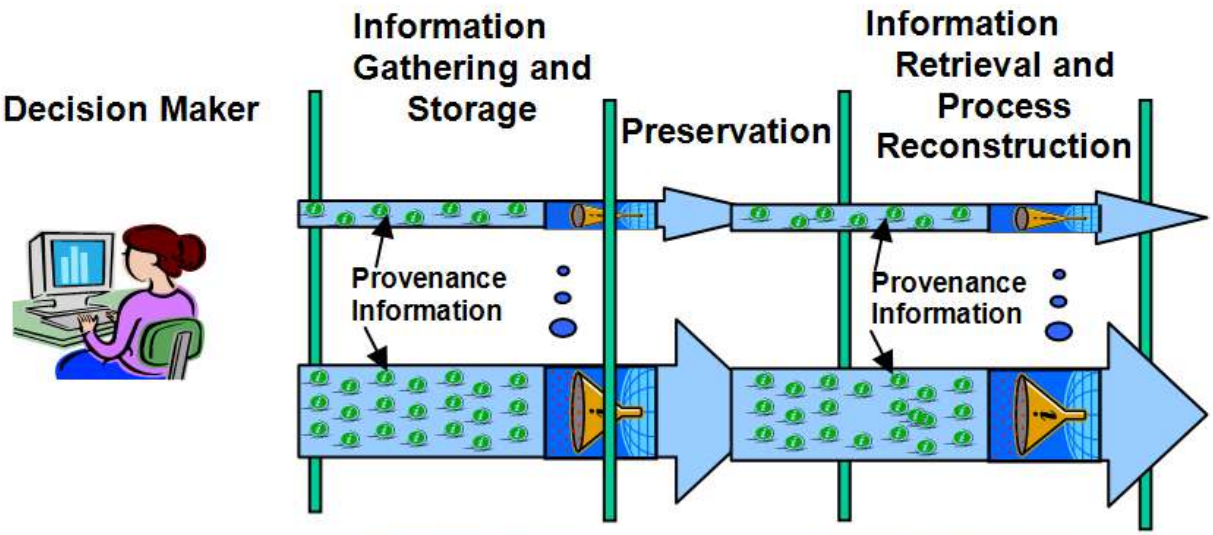
Introduction: Strategic Plan

- According to *The Strategic Plan of The National Archives and Records Administration 2006–2016. "*Preserving the Past to Protect the Future"
  - **"Strategic Goal 2:** 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."*
- The management and appraisal of electronic documents have been identified among the top ten challenges in the 34th Semi-annual Report to Congress by National Archives and Records Administration (NARA) Office of Inspector General (OIG) in 2005.
- Official appraisal policy of NARA adopted in May 17, 2006, and issued as NARA Directive 1441

Motivation (past research)

- To address the *Strategic Plan of The National Archives and Records Administration – specifically*
  - (1) Understand the tradeoffs between information value and computational/ storage costs by providing simulation frameworks
    - Information granularity, organization, compression, encryption, document format, ...
    - Versus
    - Cost of CPU for gathering information, for processing and for input/output operations; cost of storage media, upgrades, storage room, …
- **Prototype simulation framework**: Image Provenance To Learn available for downloading from isda.ncsa.uiuc.edu

Simulation Framework: Architecture



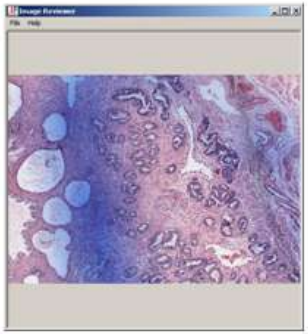
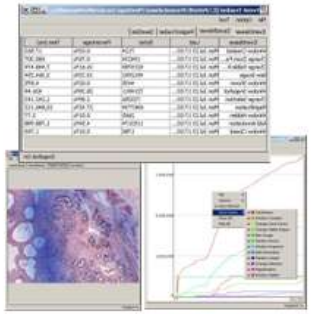
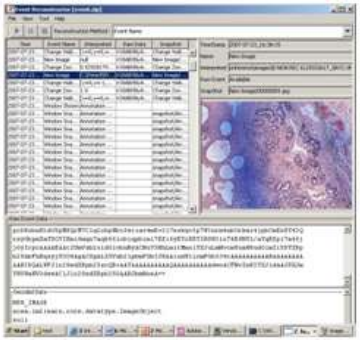


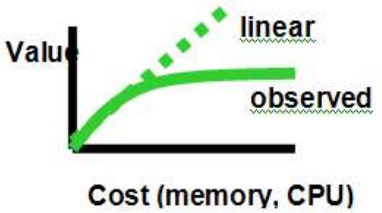
Image Viewer



Event Tracker

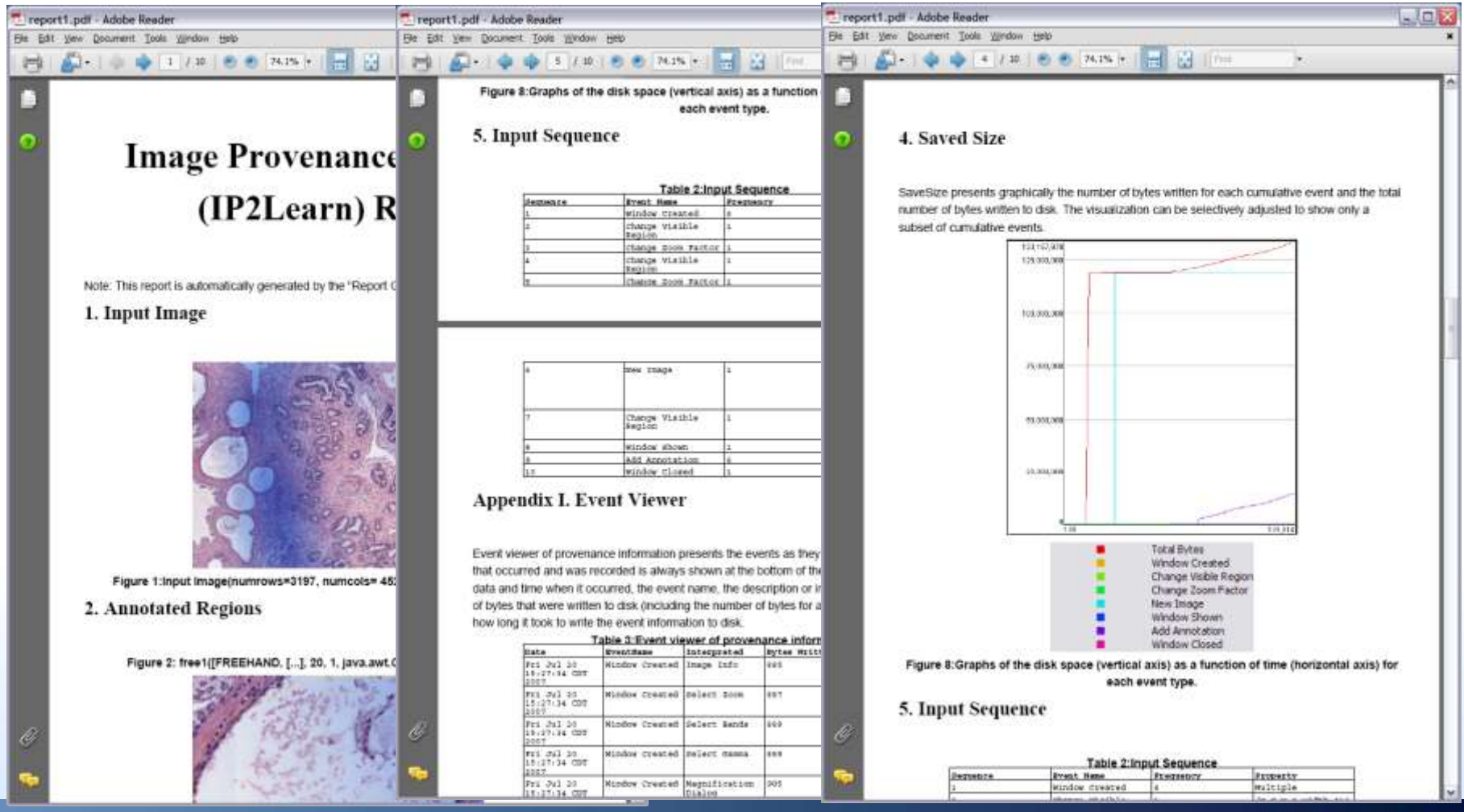


Event Reviewer



Cost / Information Granularity Analysis

Direction: Self-Describing Software with Analytical Capabilities -> Auto Reporting



Motivation (current research)

- To address the *Strategic Plan of The National Archives and Records Administration – specifically*
  - (2) Assist in *improving the efficiency with which archivists manage all holdings from the time they are scheduled through accessioning, processing, storage, preservation, and public use."*
    - Are the records related to other permanent records?
    - What is the timeframe covered by the information?
    - What is the volume of records?
    - Is sampling an appropriate appraisal tool?
- **Prototype computer assisted appraisal framework**: Doc To Learn – work in progress

Objectives

Design a methodology, algorithms and a framework for document appraisal by

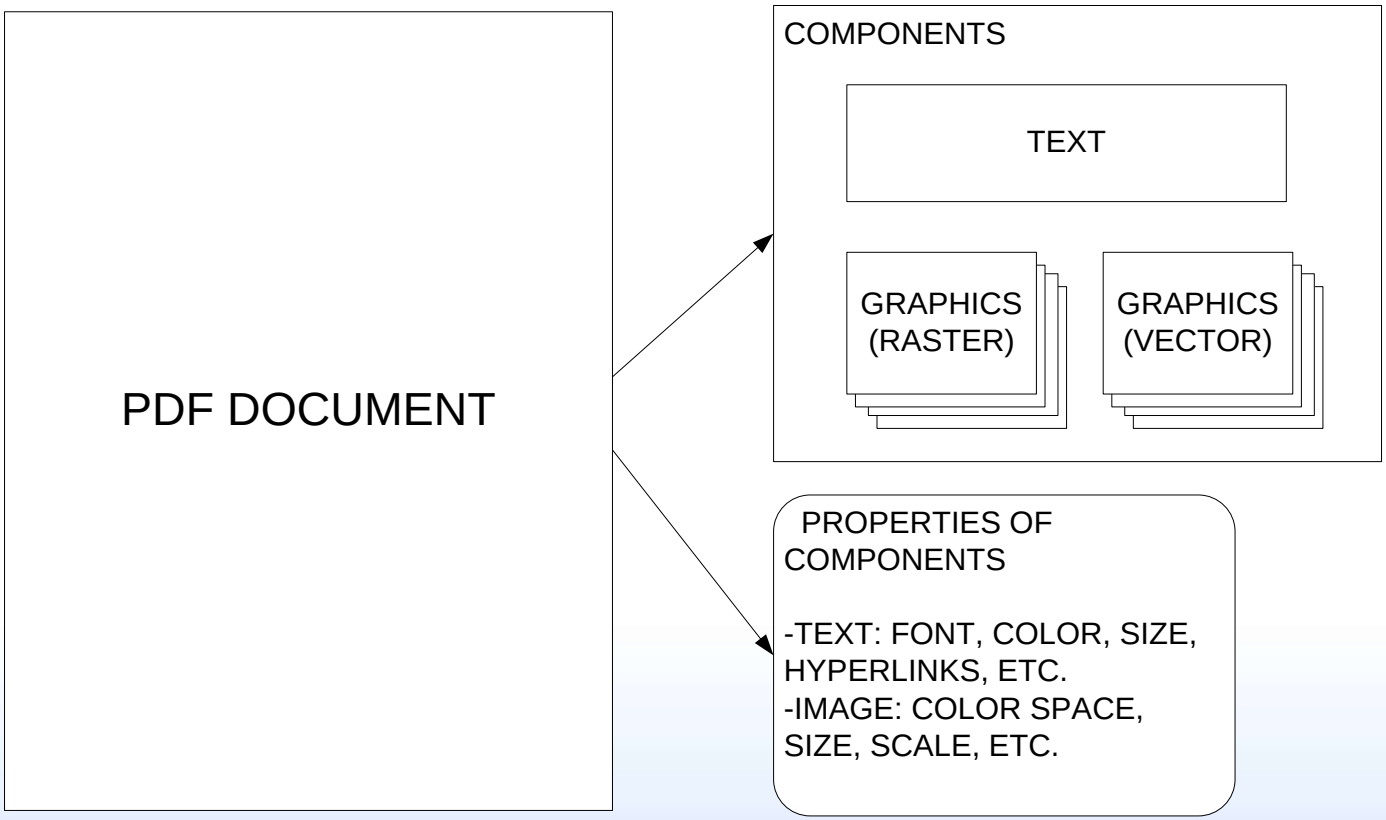
- (a) enabling exploratory document analyses and integrity/authenticity verification,
- (b) supporting automation of some analyses and
- (c) evaluating computational and storage requirements of computer-assisted appraisal processes

Electronic Records of Interest

- A class of electronic records that
  - (a) correspond to information content found in software manuals or reports (e.g., scientific or government agency reports),
  - (b) have an incremental nature of their content in time, and
  - (c) are represented by office documents used for reporting.
- Selected document file format to work with:
  - Adobe Portable Document Format (PDF) – found open source loader/writer (in comparison with MS Word)

Adobe Portable Document Format (PDF)

- **Contemporary PDF documents**

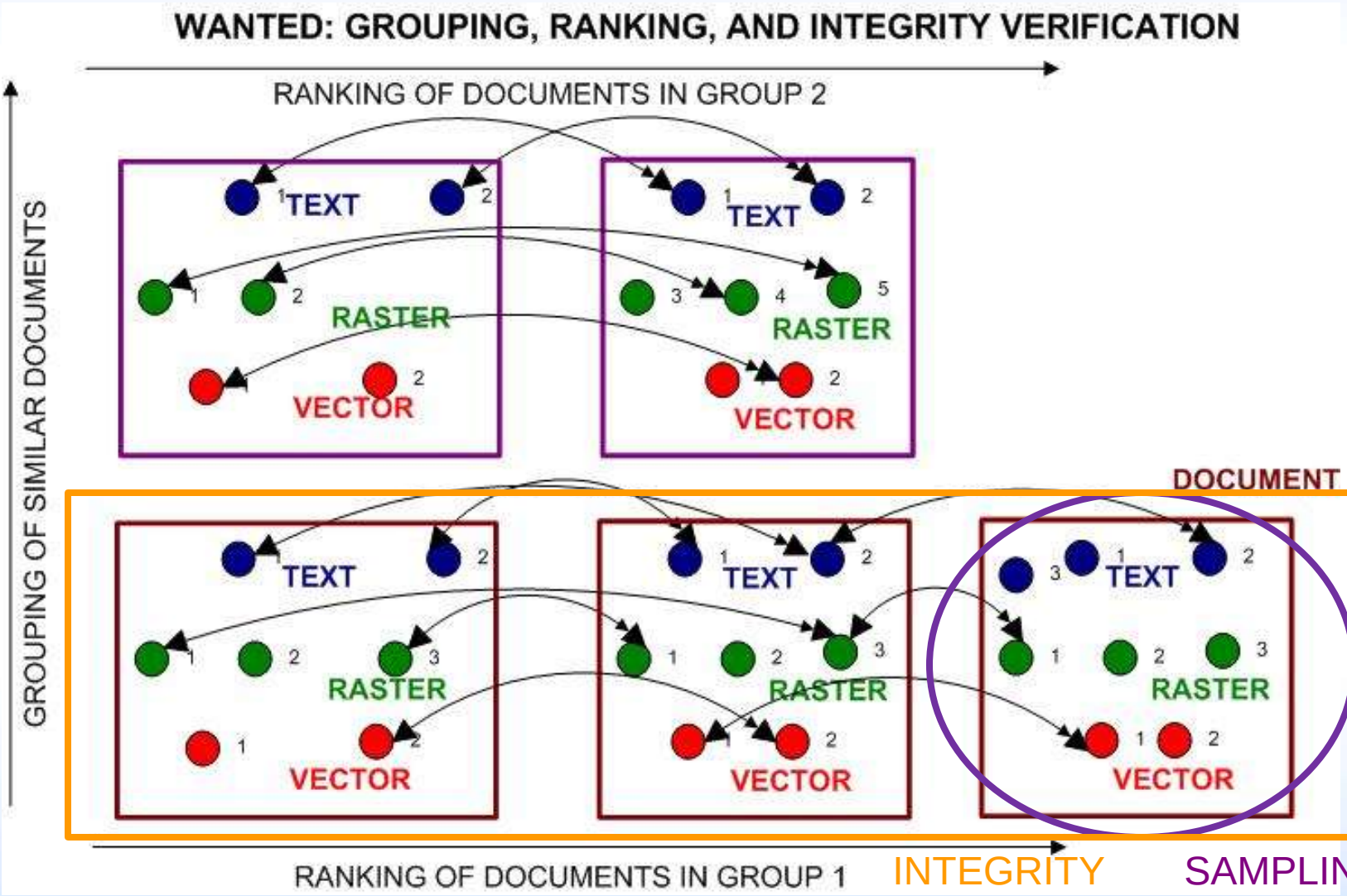


Approach

Decompose the series of appraisal criteria into a set of focused analyses:

- (a) find groups of records with similar content,
- (b) rank records according to their creation/last modification time and digital volume,
- (c) detect inconsistency between ranking and content within a group of records,
- (d) compare sampling strategies for preservation of records.

Overview of the Approach



Related Work

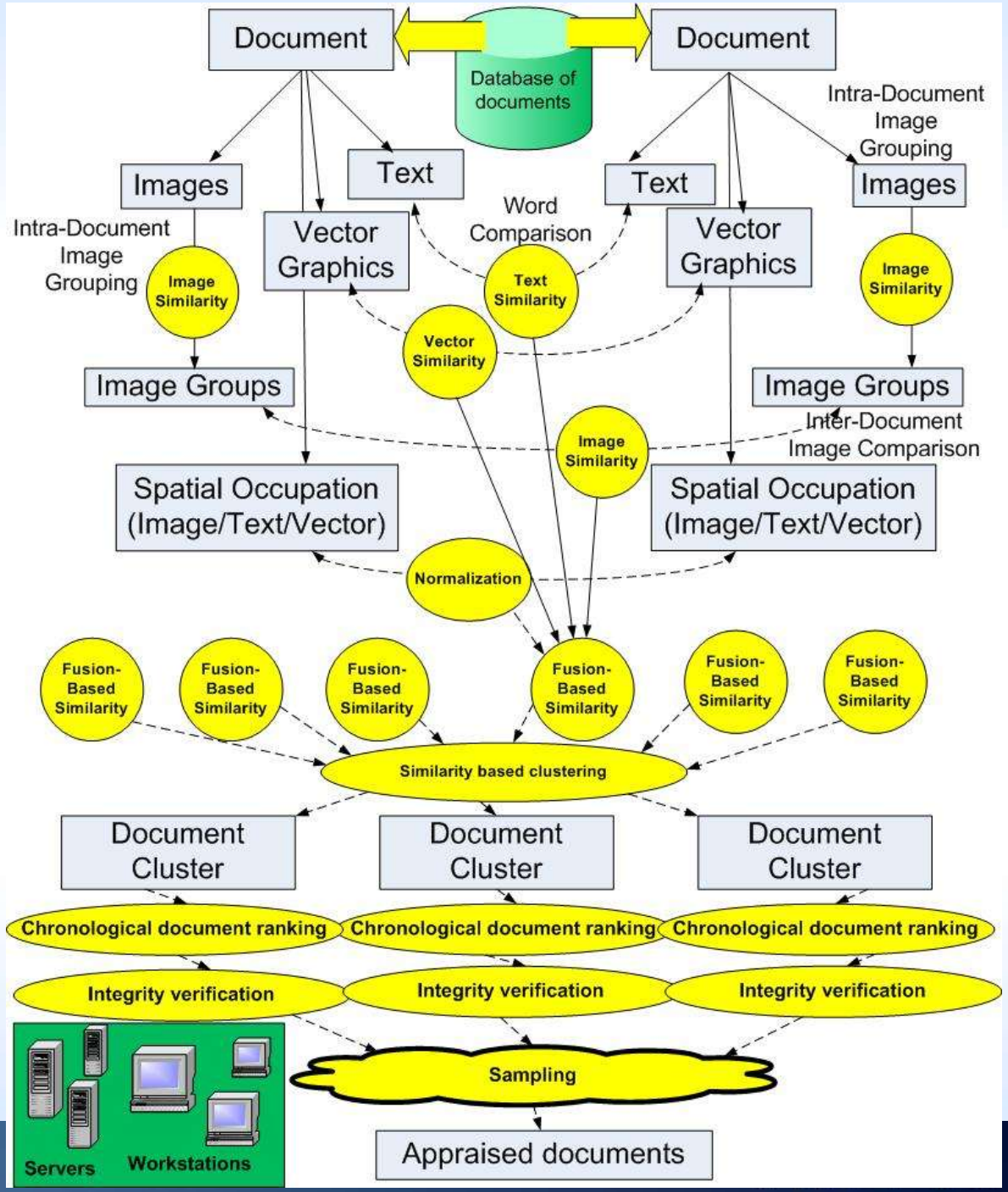
- Past work in the areas of
  - (a) content-based image retrieval,
  - (b) digital libraries, and
  - (c) appraisal studies.
- We adopted some of the image comparison metrics used in (a), text comparison metrics used in (b), and lessons learnt from (c).

Methodology

Yellow indicates computations

Relationship to Permanent Records

Appraisal & Sampling



Mathematical Framework

- Similarity of two documents

$$sim(D_i, D_j) = w_{TEXT} \cdot sim(T_i, T_j) + w_{RASTER} \cdot sim(\{I_{ik}\}_{k=1}^{K}, \{I_{jl}\}_{l=1}^{L}) + w_{VECTOR} \cdot sim(V_i, V_j)$$

- Weighting coefficients

$$W_{IMAGE}(D_i, D_j) = \frac{R_{IMAGE}(D_i) + R_{IMAGE}(D_j)}{2} \qquad R_{IMAGE}(D) = \frac{Area_{IMAGE}(D)}{Area_{IMAGE}(D) + Area_{VECTOR}(D) + Area_{TEXT}(D)}$$

$$W_{IMAGE}(D_i, D_j) + W_{VECTOR}(D_i, D_j) + W_{TEXT}(D_i, D_j) = 1 \qquad R_{IMAGE}(D) + R_{VECTOR}(D) + R_{TEXT}(D) = 1$$

- Intra- and inter-doc image-based similarity

$$sim(I_{ik} \in D_i, I_{il} \in D_j) = \sum_{k1,k2} \omega_{i,k1}\omega_{i,k2} \qquad sim(\{I_{ik}\} \in D_i, \{I_{jl}\} \in D_j) = \sum_{k1,k2} \omega_{i,k1}\omega_{j,k2}$$

Intra-document Inter-document

$$\omega_{ik} = \frac{f_{ik} \log(N / n_k)}{\sqrt{\sum_{l=1}^{L} (f_{il})^2 (\log(N / n_l))^2}}$$

- Text-based and v/h line count similarity

$$sim(T_i, T_j) = \sum_{k1,k2} \omega_{i,k1}\omega_{j,k2}$$

f –frequency of occurrence of a feature (word/color)
L - number of all unique feature primitives
n - number of documents that contain the feature (n=1 or 2)
N – number of documents evaluated

Prototype: Text Comparison

Document Analyzer

Document List

No	Filename	File Size	File Date	Num Pages	Num Images
1	pf01-418.pdf	1644KB	Wed Aug 29 15:30:04 CDT 2007	29	4
2	pubAboutLakeTahoe_fs-100-9...	1744KB	Mon Apr 16 15:15:58 CDT 2007	6	6

LOADED FILES

Occurrence of words

Occurrence of numbers

“Ignore” words

Word Frequency

All Frequency

No	Word	Frequency
0	analyzing	2
1	Service's	1
2	Changes	4
3	materials	2
4	slope	1
5	event	1
6	Surface	1
7	Nev	1
8	generation	1
9	discharges	1
10	hydrology	1
11	timber	2
12	James	1
13	influen	1
14	$	1
15	http://www....	1
16	answer	1
17	+	3
18	regard	2
19	meet	1
20	panchromatic	3
21	arameters	1
22	.	969
23	Team	1

Minimum frequency: 0 Set

Text Frequency

No	Word	Frequency
0	analyzing	2
1	Service's	1
2	Changes	4
3	materials	2
4	slope	1
5	event	1
6	Surface	1
7	Nev	1
8	generation	1
9	discharges	1
10	hydrology	1
11	timber	2
12	James	1
13	http://www....	1
14	$	1
15	influen	1
16	answer	1
17	+	3
18	regard	2
19	Team	1
20	.	969
21	arameters	1
22	panchromatic	3
23	meet	1

Minimum frequency: 0 Set

Integer Frequency

No	Word	Frequency
0	36	1
1	39	1
2	155	1
3	1970	4
4	1971	2
5	103365	1
6	43	1
7	103366	3
8	42	1
9	1975	2
10	1976	5
11	40	6
12	1978	1
13	1979	2
14	12921	1
15	200	1
16	1900	2
17	22	2
18	23	1
19	24	2
20	25	4
21	1990	6
22	26	2
23	27	2

Minimum frequency: 0 Set

Float Frequency

No	Word	Frequency
0	21.5	1
1	9.6	1
2	0.46	1
3	0.011	1
4	0.006	1
5	0.043	1
6	56.5	1
7	0.009	1
8	0.027	1
9	41.2	1

Minimum frequency: 0 Set

Ignore

No	Word
1	
2	&
3	=
4	A
5	Data
6	For
7	In
8	Only
9	The
10	These
11	This
12	We

New Remove Load Save

Compare List

No	Word	Info

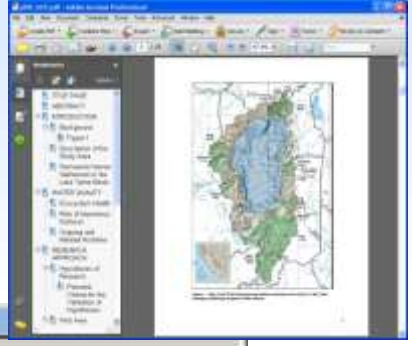
Save

Load | Launch Document | Extract Images | Remove | Use Filter | Compare | Group | Move to Ignore List | Reset | Done

Performing Word Comparison...

NCSA

Prototype: Image Comparison



Document Analyzer

Document List

No	Filename	File Size	File Date	Num Pages	Num Images
1	of01-418.pdf	1644KB	Wed Aug 29 15:30:04 CDT 2007	29	4
2	pubAboutLakeTahoe_fs-100-9...	1744KB	Mon Apr 16 15:15:58 CDT 2007	6	6

LOADED FILES

List of images

Occurrence of colors

Preview

“Ignore” colors

Ignore List

No	Word
1	Col_0_0_0
2	Col_255_255_255

New | Remove | Load | Save

Compare List

No	Word	Info

Save

No	Filename
1	(numrows=630, numcols= 43...
2	(numrows=397, numcols= 29...
3	(numrows=436, numcols= 43...
4	(numrows=574, numcols= 61...

No	Color	Frequency
0	RGB_85_198_170	1
1	RGB_142_198_255	487
2	RGB_113_85_28	3
3	RGB_28_85_85	41
4	RGB_198_227_227	4176
5	RGB_170_170_0	2
6	RGB_227_198_227	82
7	RGB_142_198_113	51
8	RGB_255_255_28	1
9	RGB_142_142_170	336
10	RGB_57_113_28	4
11	RGB_113_113_113	2197
12	RGB_255_198_198	7
13	RGB_57_113_170	75
14	RGB_142_170_198	2556
15	RGB_198_255_255	594
16	RGB_170_113_85	1
17	RGB_227_255_227	1110
18	RGB_142_170_227	76
19	RGB_227_227_85	28
20	RGB_255_227_85	1
21	RGB_85_170_170	46
22	RGB_198_198_255	14
23	RGB_0_85_170	1
24	RGB_170_170_170	4507
25	RGB_198_198_113	8

Minimum frequency: 0 Set

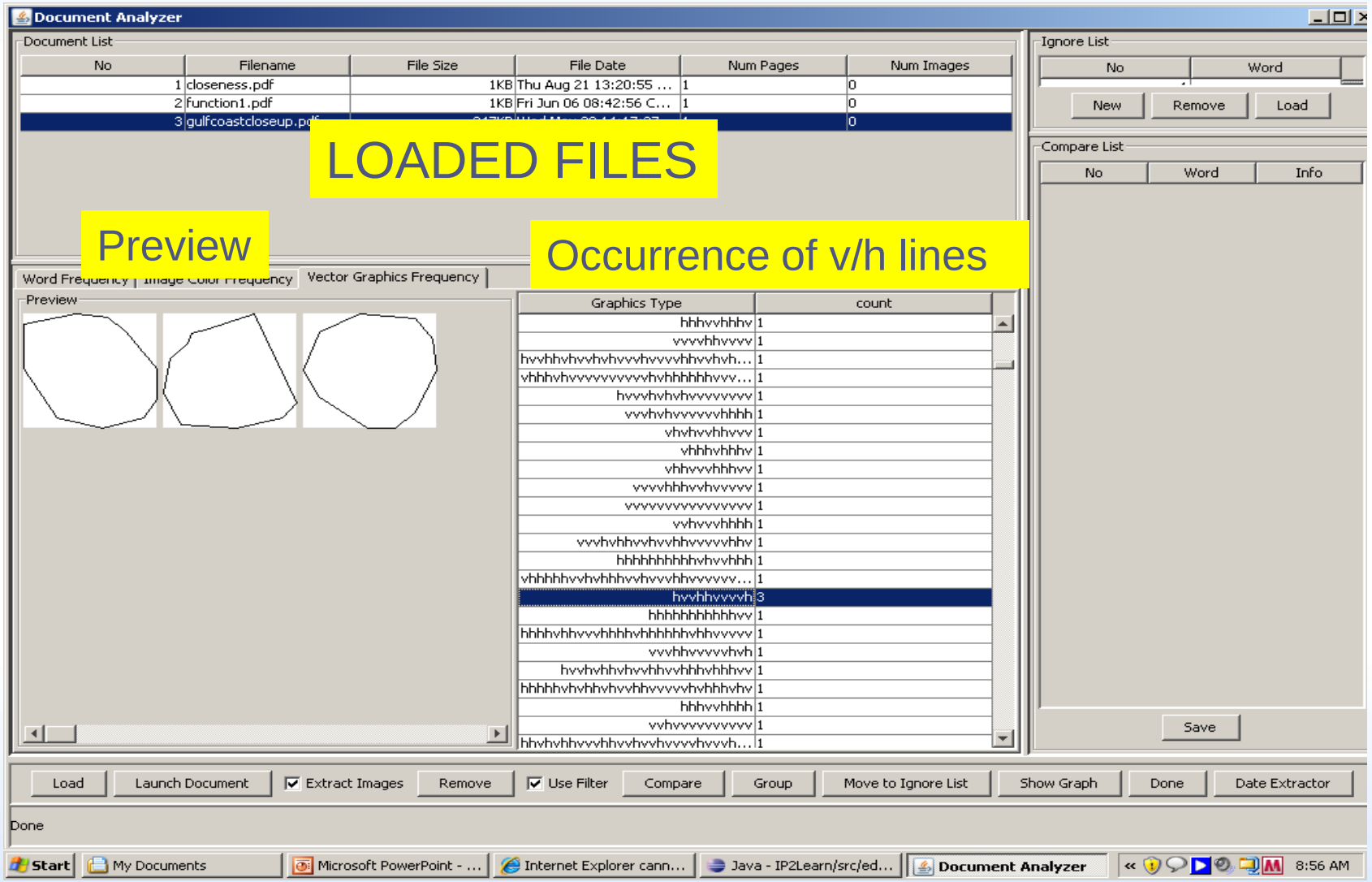
Down Color Bins: 10 Set

Show Color Reduced Image

Load | Launch Document | Extract Images | Remove | Use Filter | Compare | Group | Move to Ignore List | Reset | Done

Performing Word Comparison...

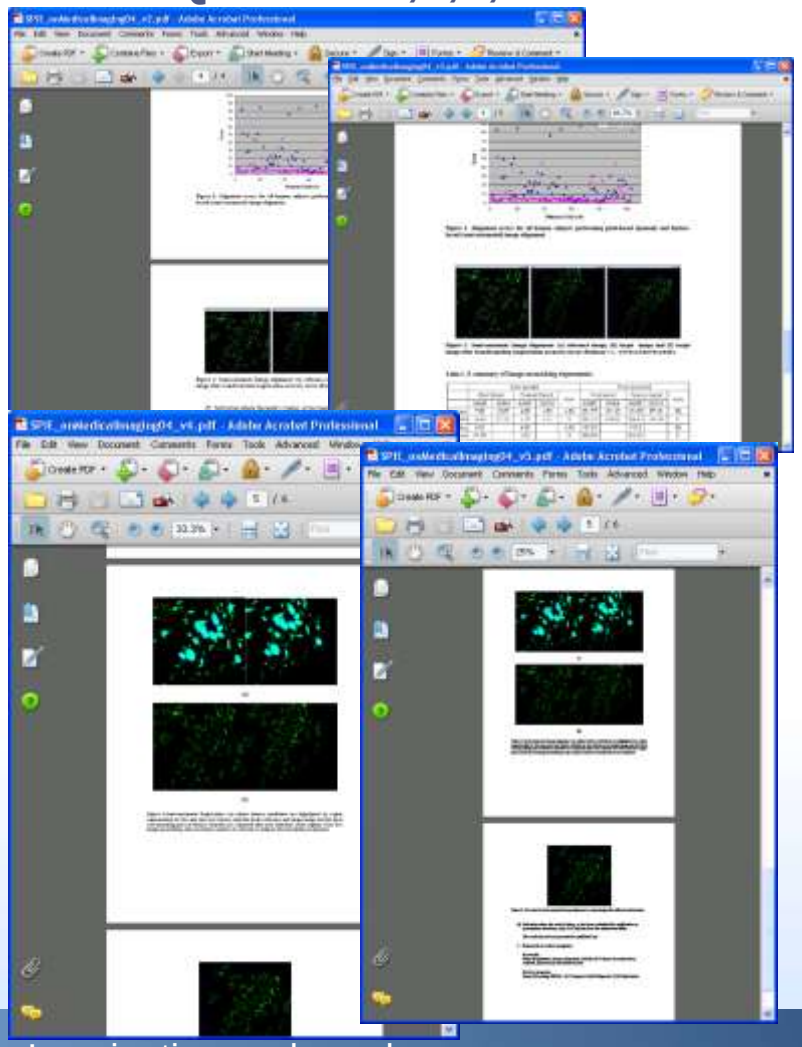
Prototype: Vector Graphics Comparison



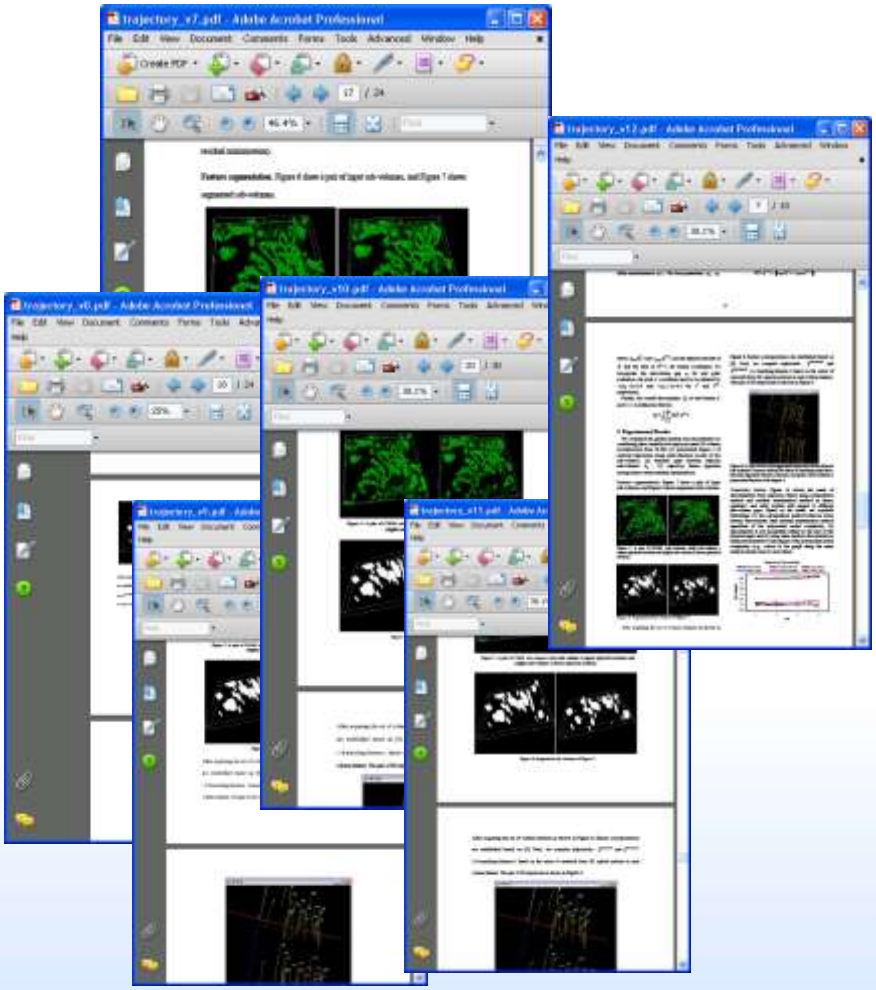
Illustrative Experimental Study

INPUT = 10 PDF docs (4 & 6 Groups)

UNIQUE ID= 1,2,3,4

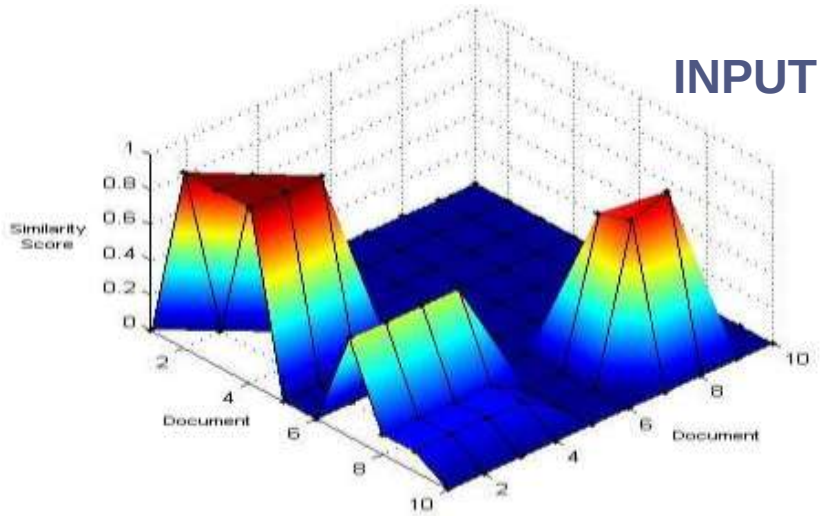


UNIQUE ID= 5,6,7,8,9,10

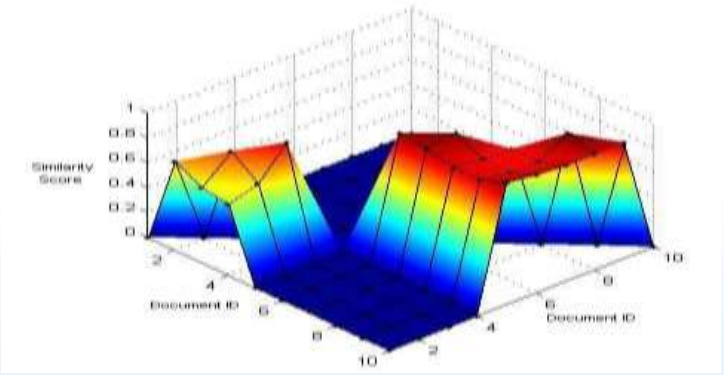


Comparative Experimental Results

INPUT = 10 PDF docs (6 & 4 Groups)



Vector-based similarity



Text-based similarity

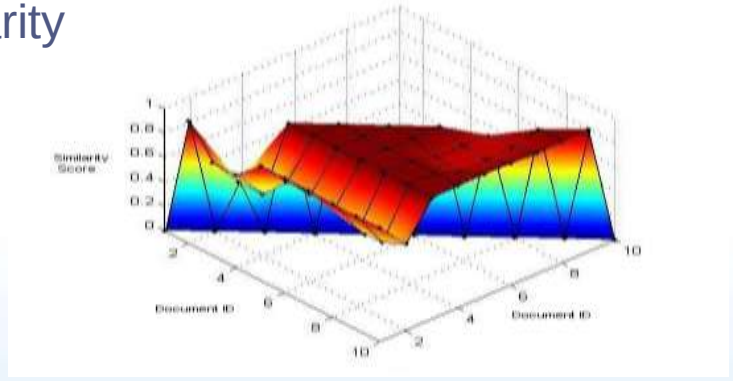
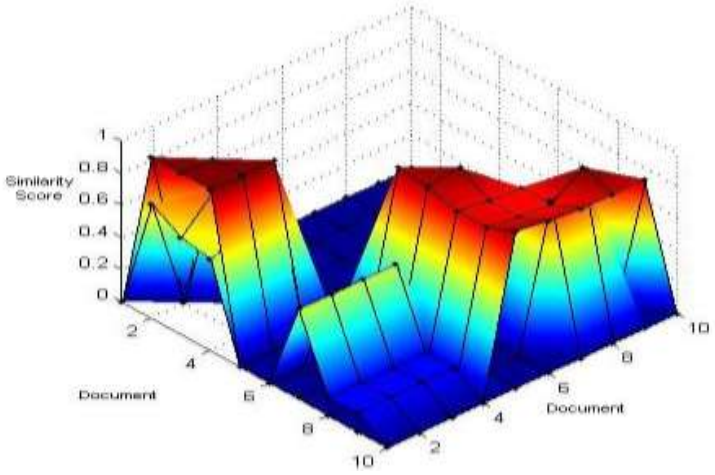
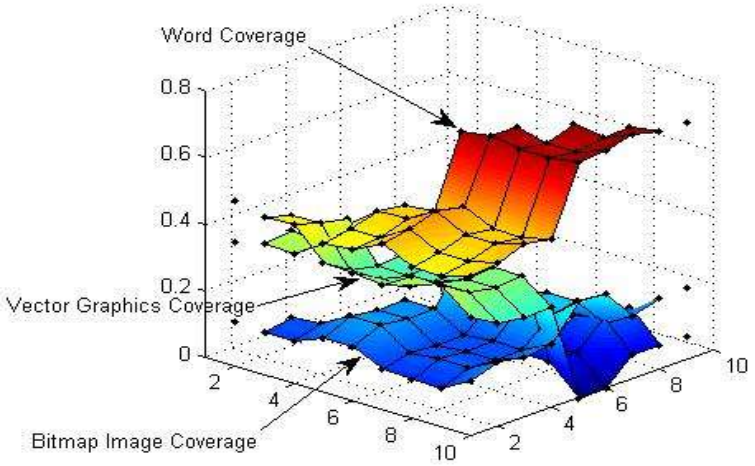


Image-based similarity

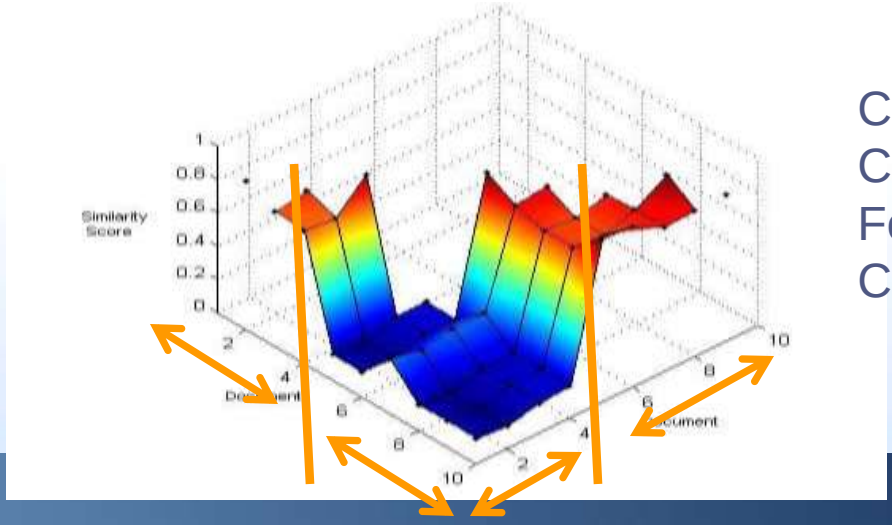
Comparative Experimental Results



Vector Graphics Similarity and Word Similarity Combined

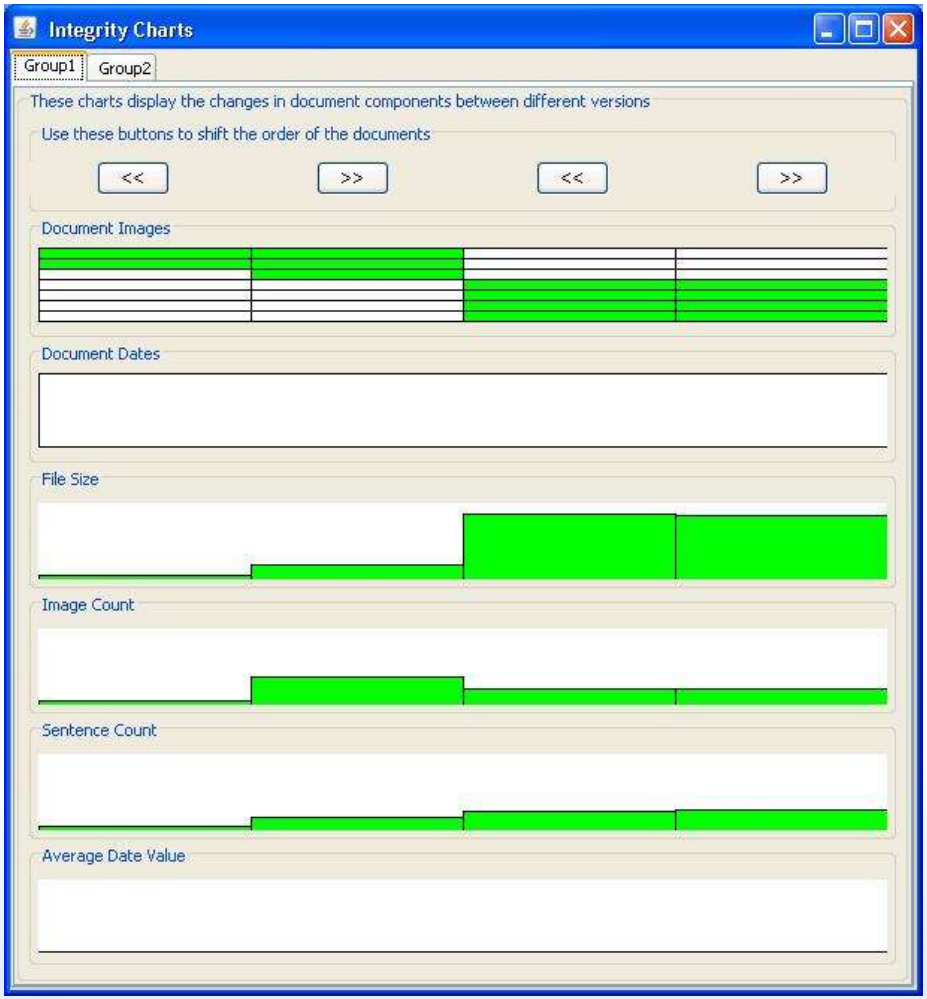


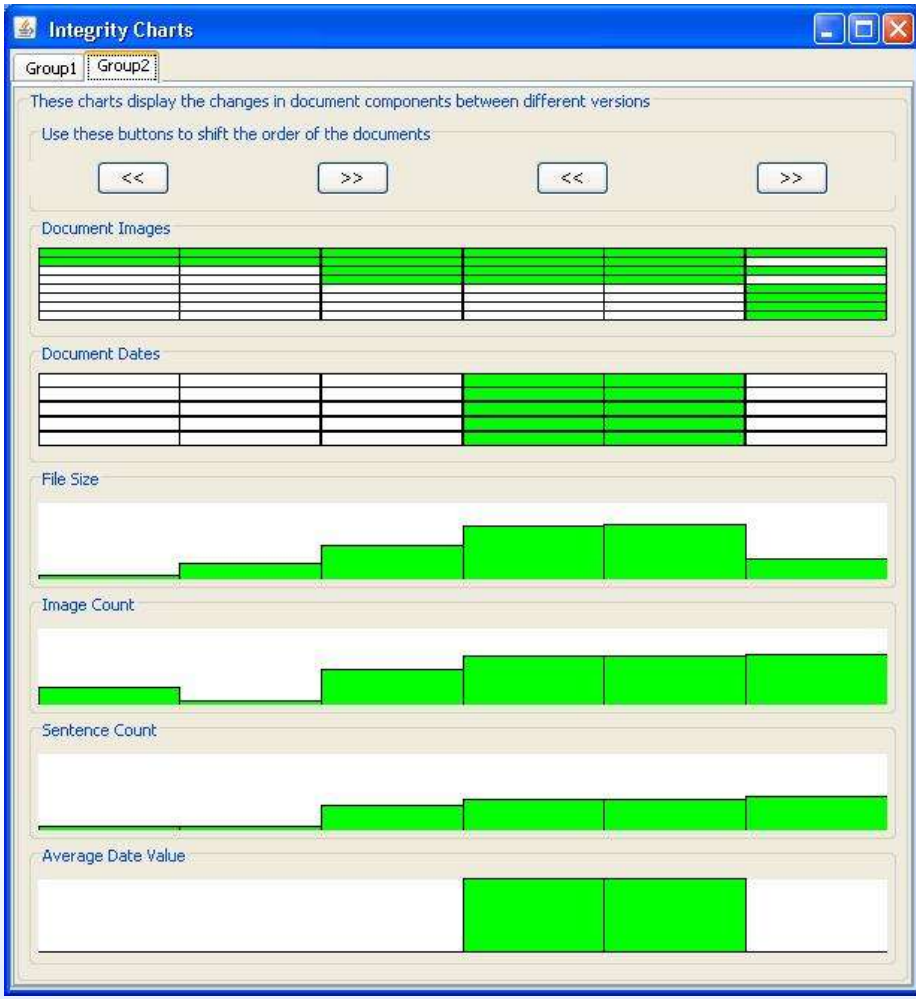
Portion of Document Surface Allotted to Each Document Feature



Comparison Using Combination of Document Features in Proportion to Coverage

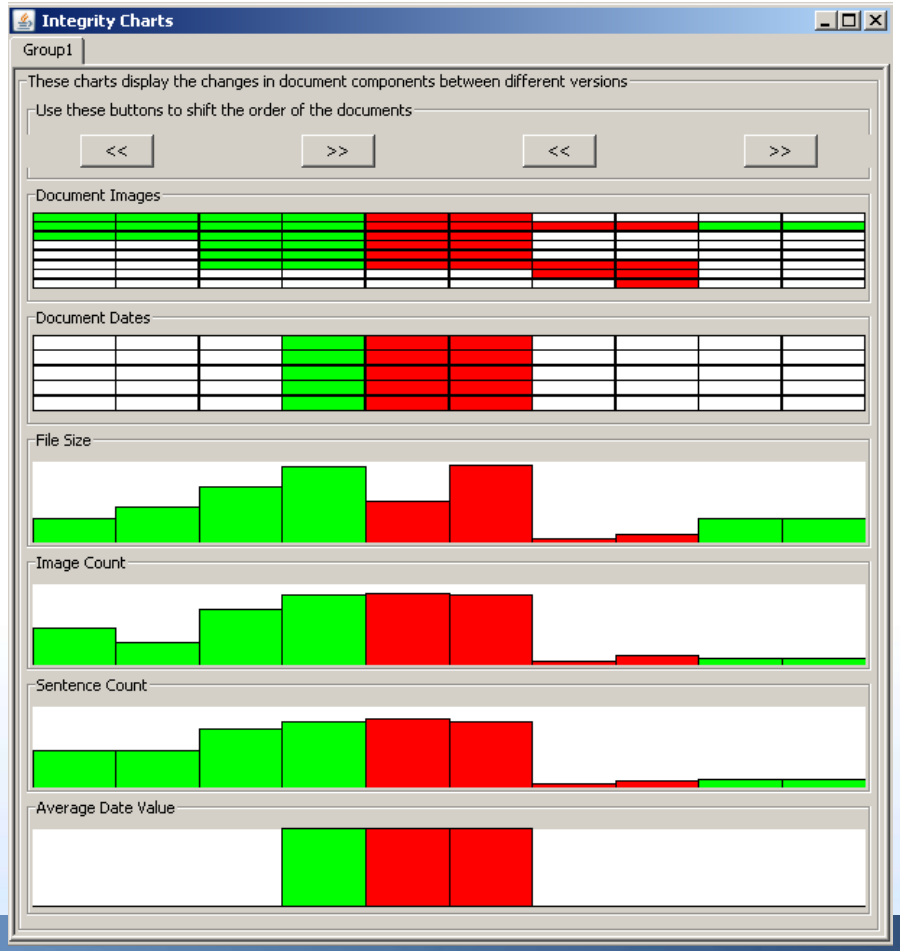
Integrity Verification – Two Groups





Example of Integrity Verification with Detected Inconsistencies

TIME



(1) appearance or disappearance of document images,
(2) appearance and disappearance of dates appearing in documents,
(3) file size,
(4) image count,
(5) number of sentence, and
(6) average value of dates found in document.

Conclusions

- **Accomplishments:** We have designed a framework for computer assisted document appraisal
  - A methodology
  - A prototype for grouping, ranking and integrity verification of PDF documents – support for document explorations
  - Identified computational challenges
- **Key contributions**:
  - Automation
  - Comprehensive comparison of PDF documents (text, images & graphics objects)
  - Initial integrity verification metrics
- **Future work**
  - Sampling is still an open question
  - Scalability of document analyses
    - Each file is large and the number of files is large
    - Exploring the TeraGrid resources

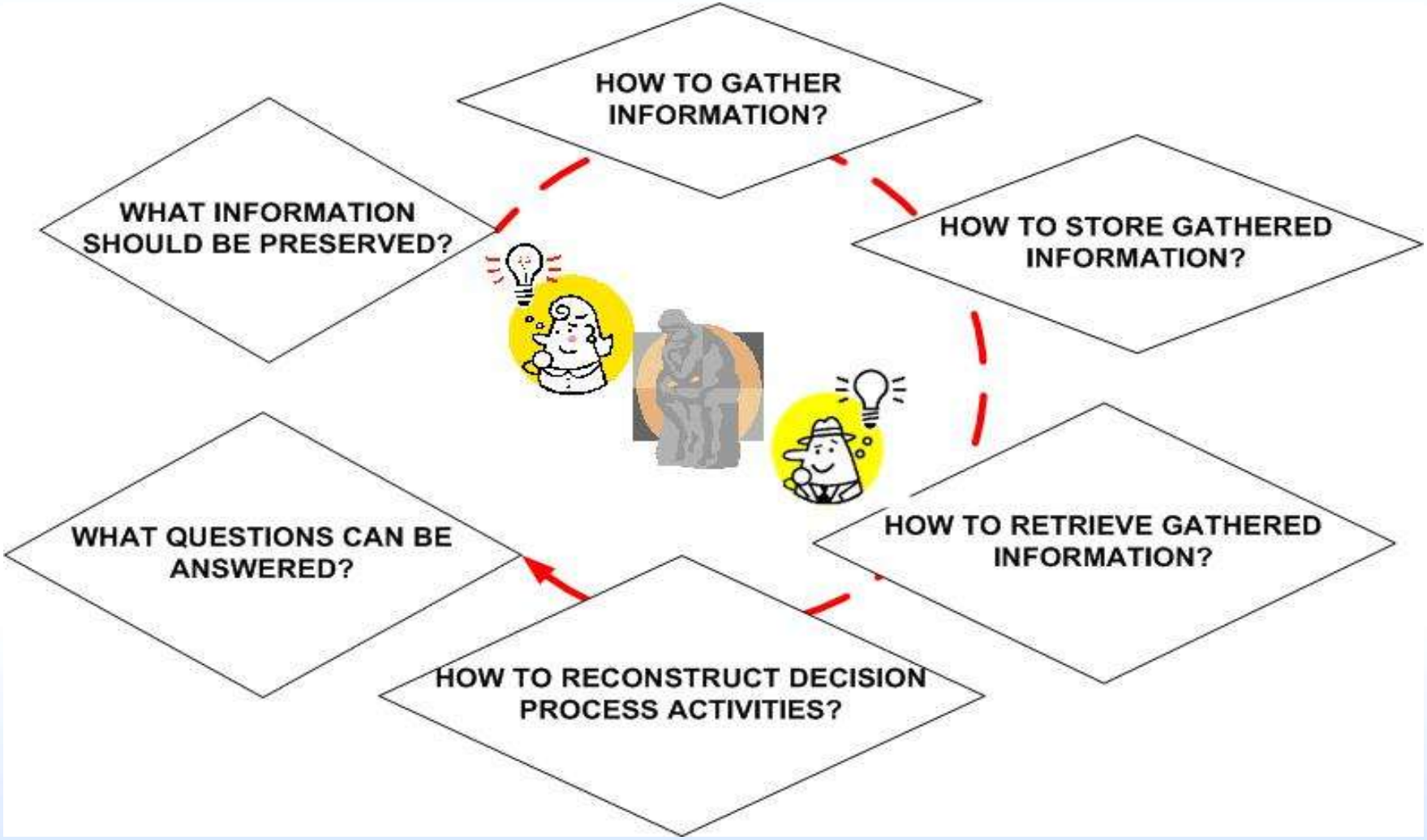
Acknowledgement

- Funding provided by NARA and NCSA Industrial Partners

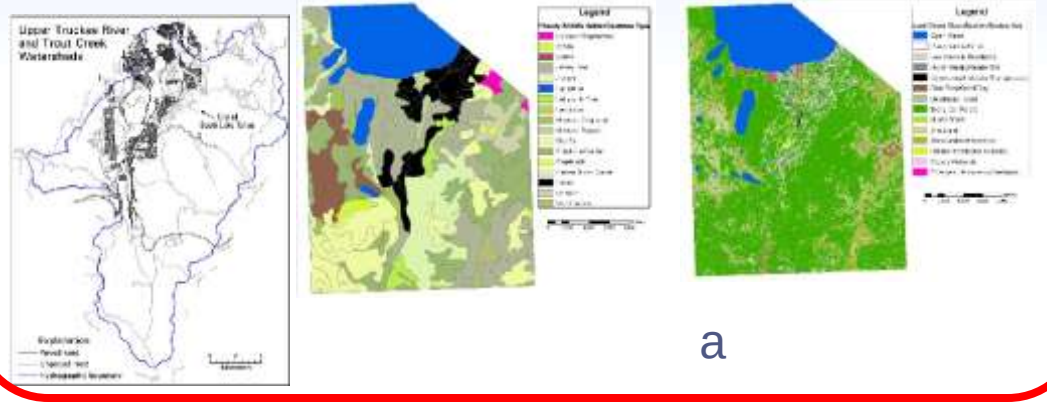
- Questions:
  - Peter Bajcsy; email: pbajcsy@ncsa.uiuc.edu
  - Project URL: http://isda.ncsa.uiuc.edu/

Backup

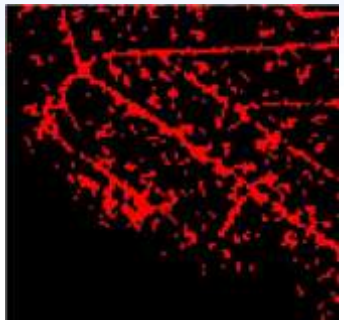
Basic Questions



Example: Image Grouping

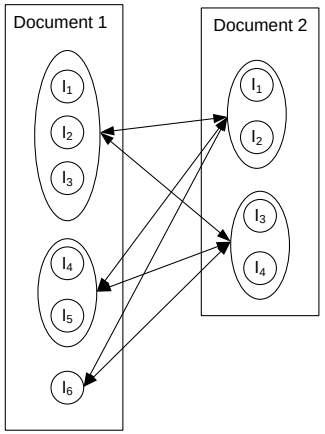


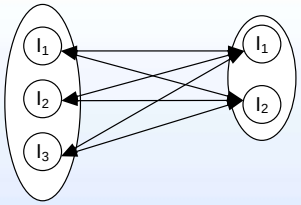
a



c

b

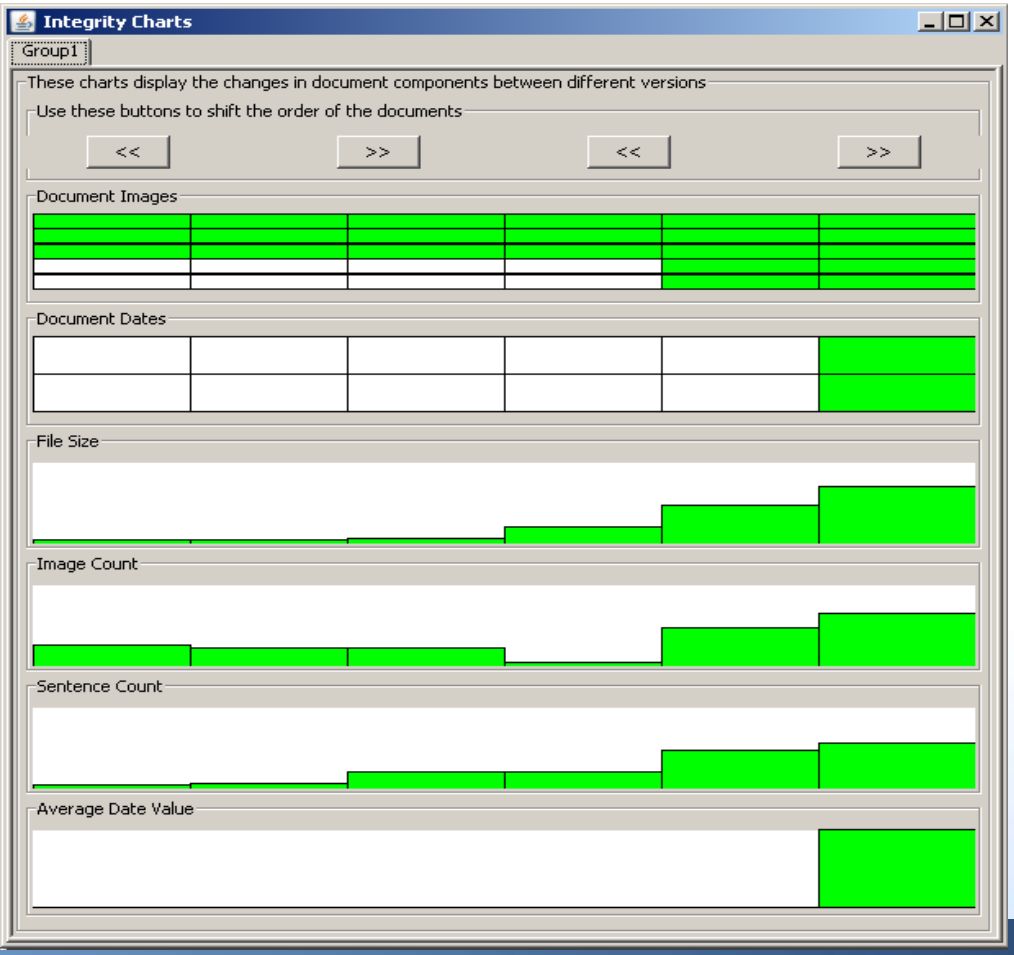




	Average similarity between image pairs	Standard deviation of the similarity
Group (a)	0.9565310641762074	0.045131416130196965
Group (b)	0.873736726083776	0.1746431238539268
Group (c)	1.0	0.0

Integrity Verification - Passed

TIME



(1) appearance or disappearance of document images,
(2) appearance and disappearance of dates appearing in documents,
(3) file size,
(4) image count,
(5) number of sentence, and
(6) average value of dates found in document.