Computational Archival Science (CAS) : from research to practice

Esteva, Maria
Texas Advanced Computing Center, University of Texas : Research Associate/Data Archivist

Marciano, Richard
Digital Curation Innovation Center, University of Maryland : Professor/Director

出版情報：2018-01-16
バージョン：
権利関係：

Computational Archival Science (CAS): *from research to practice*

Dr. Maria Esteva
Data Curator – Research Associate
Texas Advanced Computing Center (TACC)
University of Texas at Austin

Dr. Richard Marciano
Professor – Director of the Digital Curation Innovation Center (DCIC)
Maryland's iSchool
University of Maryland

January 16, 2018
Kyushu University
Fukuoka, Japan

TACC

dcic digital curation innovation center

Motivation

- How can we address exploration, understanding and processing of big, digital archives?
- Digital tools for digital records
- Archival analysis = data analysis
- Curiosity and the desire to move the profession forward

CAS, an evolving definition:

- *An interdisciplinary field concerned with the application of computational methods and resources to large-scale records/archives processing, analysis, storage, long-term preservation, and access, with aim of improving efficiency, productivity and precision in support of appraisal, arrangement and description, preservation and access decisions, and engaging and undertaking research with archival material.*

CAS: Anatomy of “big archives” visualization

Dr. Maria Esteva
TACC at University of Texas at Austin

Visualization for big archives

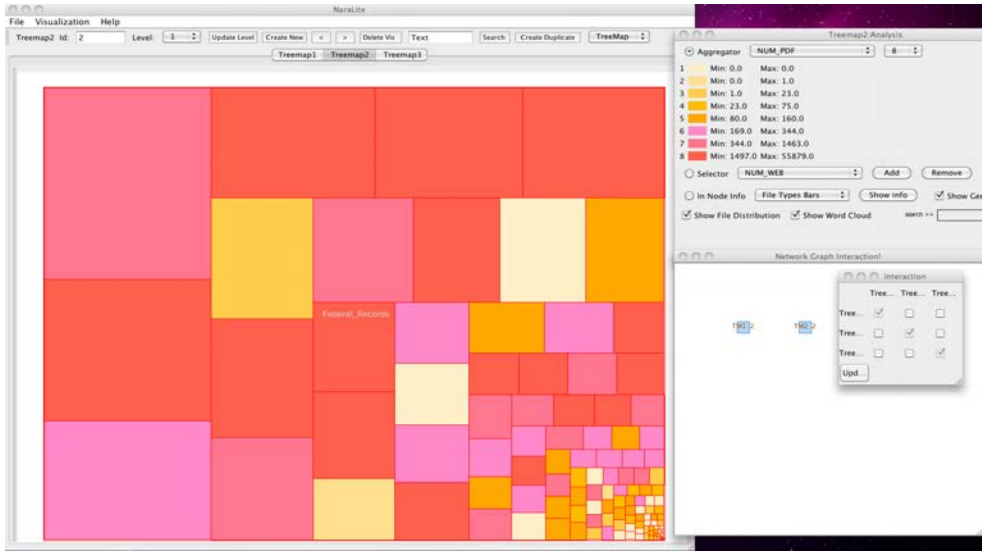
- A problem of “archival” representation
- Two part process: data analysis and interactive visualization design
- Describe the activities involved in archival tasks
- Articulate information requirements according to archival theory and best practices
- Model those requirements to metadata and computational analysis methods
- Decide how to render the results for visual interactive analysis

Visual analytics

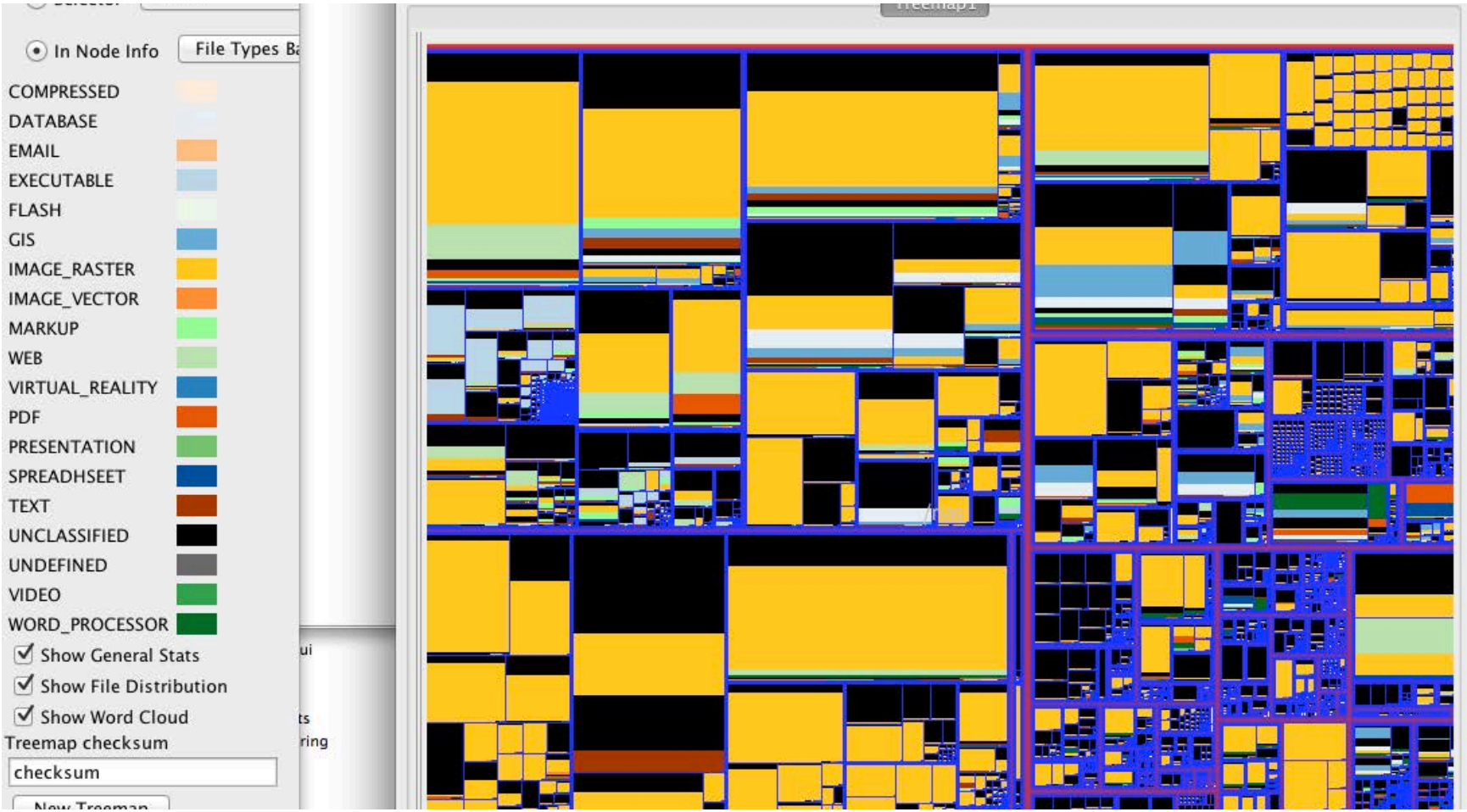
- Data, information, knowledge
- Information assisted visualization
- Knowledge assisted visualization

Anatomy

- Extracted metadata
- Data analysis methods = more metadata
- Visual metaphor
- Visualization manager (Database)
- Design of interactive functions
  - Filter
  - Selector
  - Aggregator
  - Distributions
  - Search
  - Track
  - Timeline

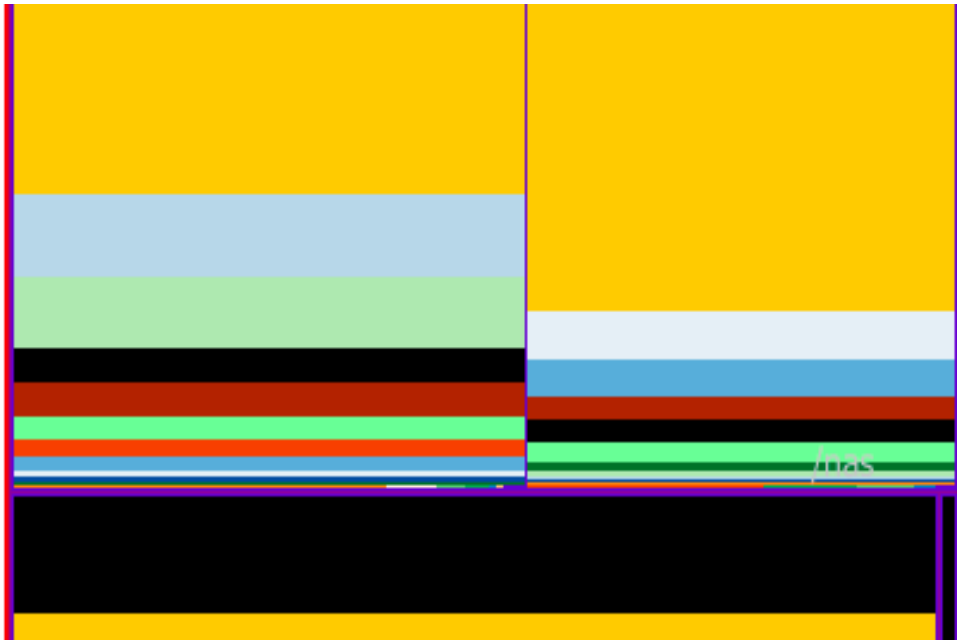


What are 5tb of archaeology data?

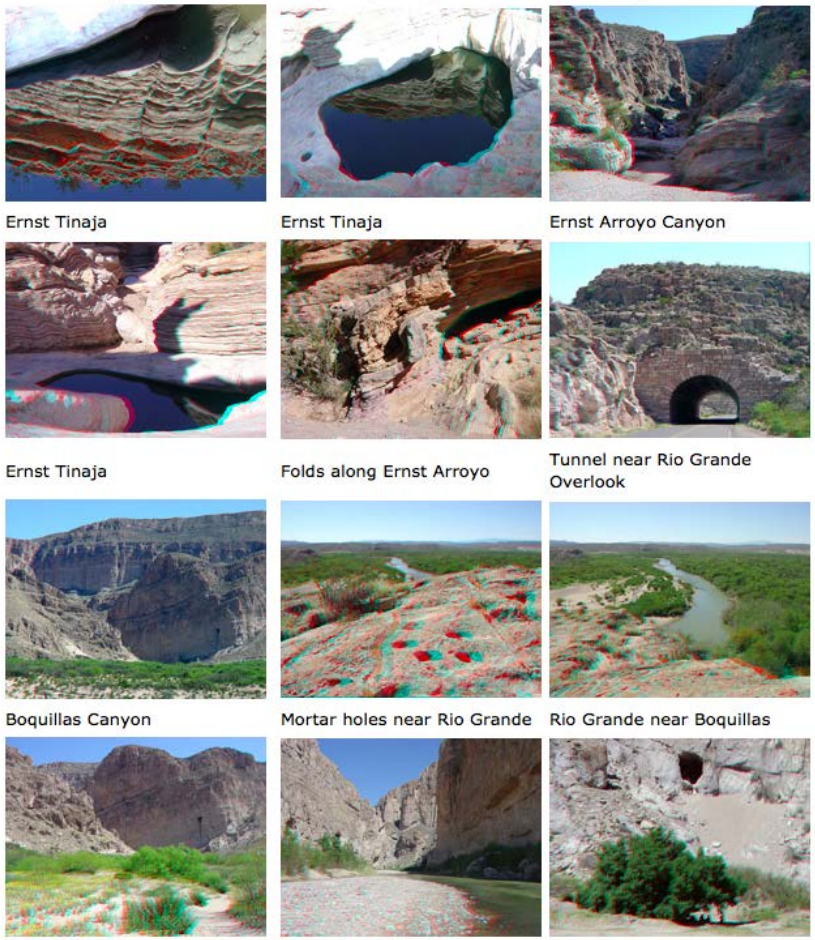


Knowledge-based visualization

	Raw/Field/Primary	Process/Study	Publication/Final/Pre
compressed			
database			
email			
executable			
flash			
gis			
image raster			
image vector			
xml			
web			
virtual reality			
pdf			
presentation			
spreadsheet			
text			
unclassified		black	
undefined		grey	
video			
word processing			
audio		red	

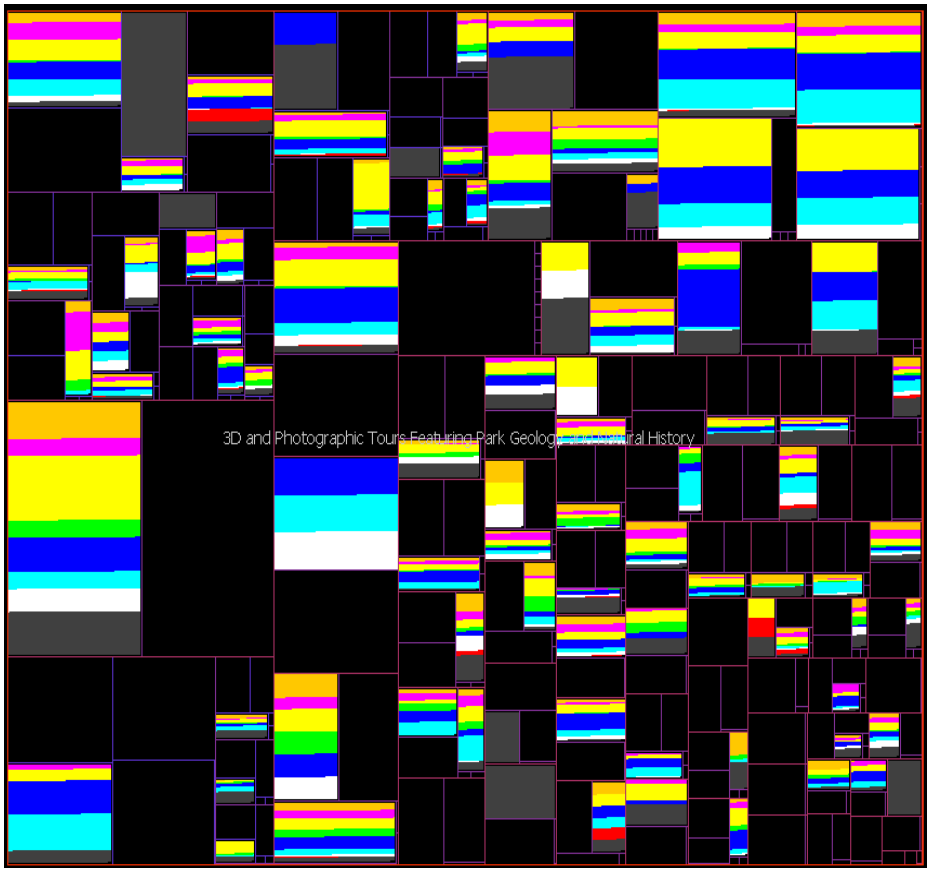


Structure and organization

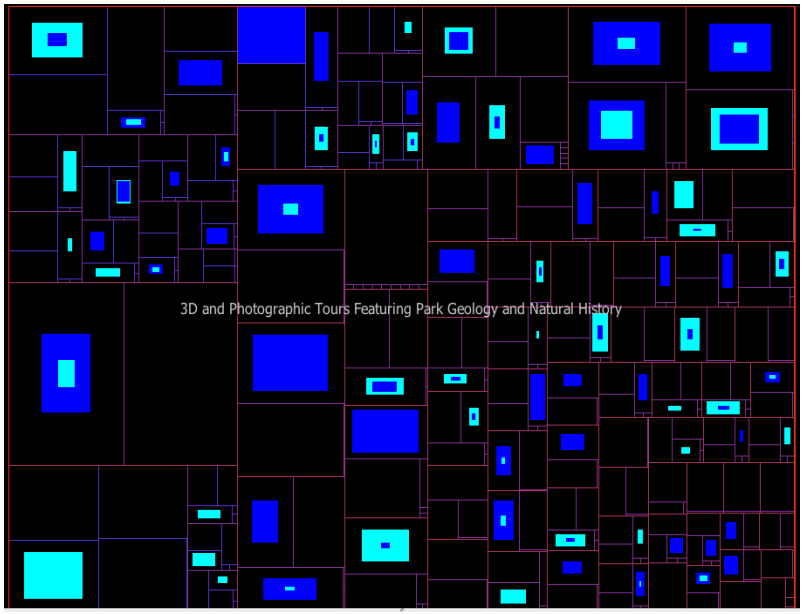


- Images and their captions in the context of park functions and provenance
  - education,
  - Preservation
  - outreach
  - research

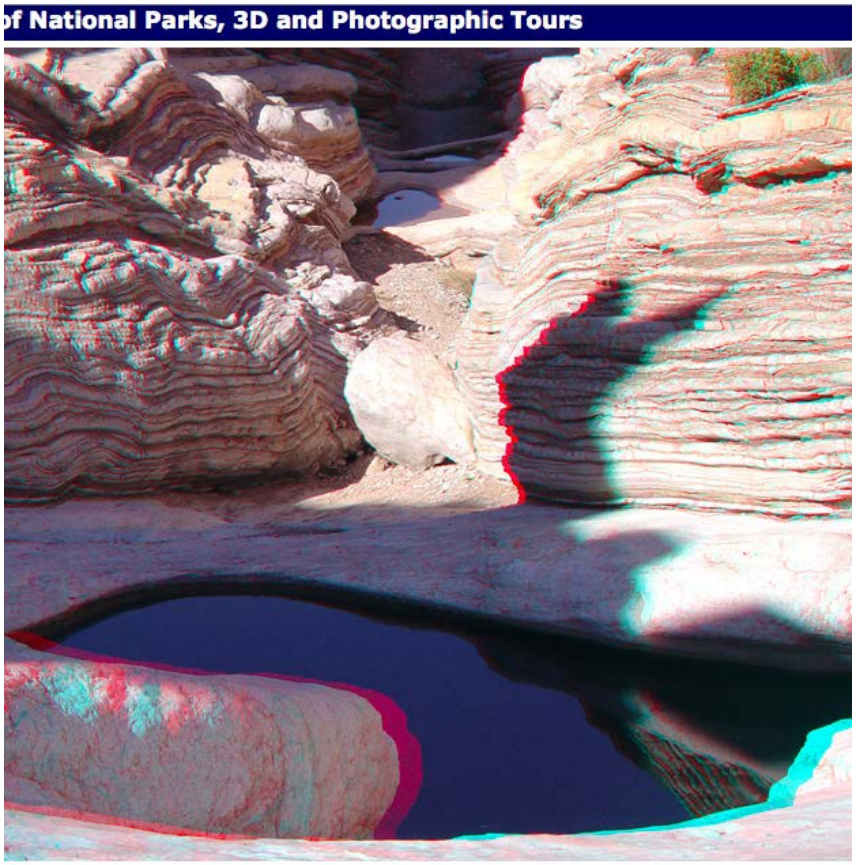
Descriptions in the aggregate



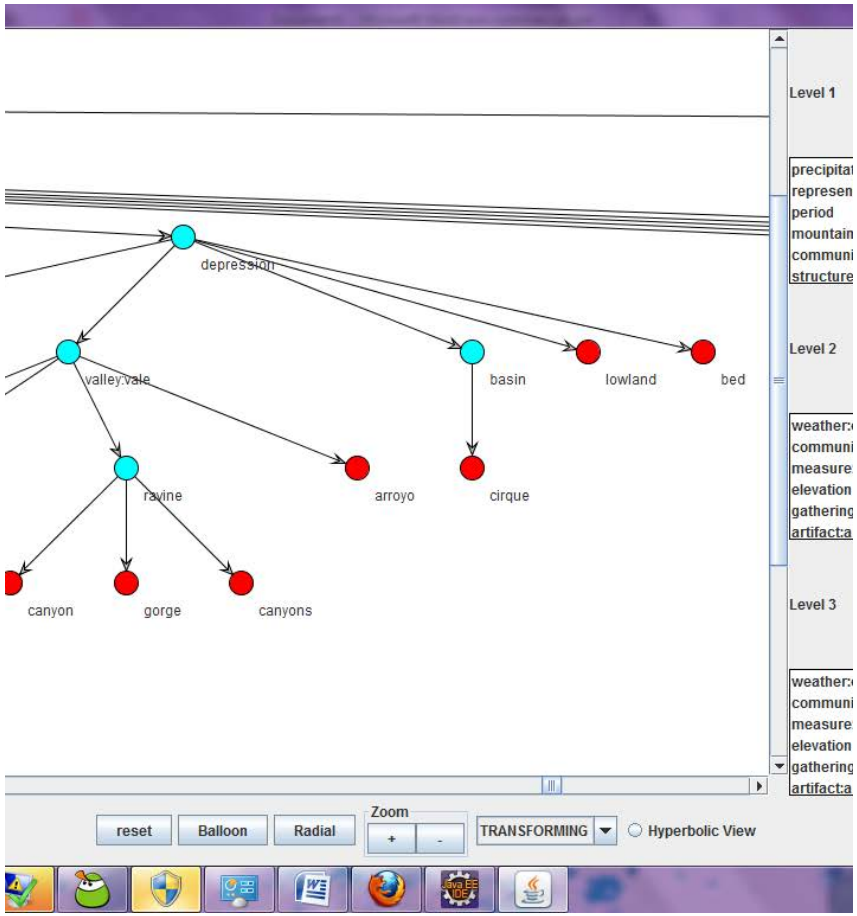
Information-based visualization



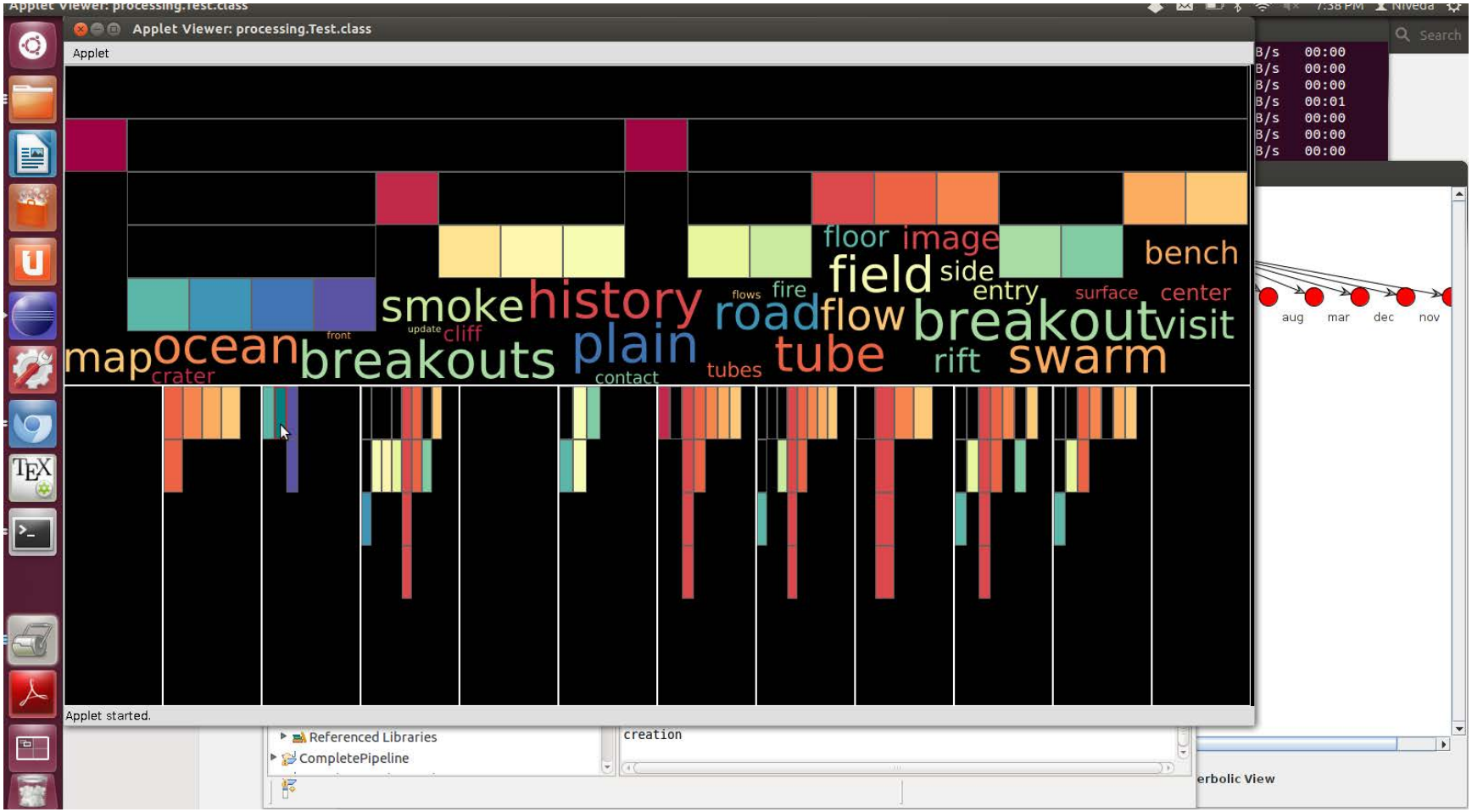
Structure and synthesis



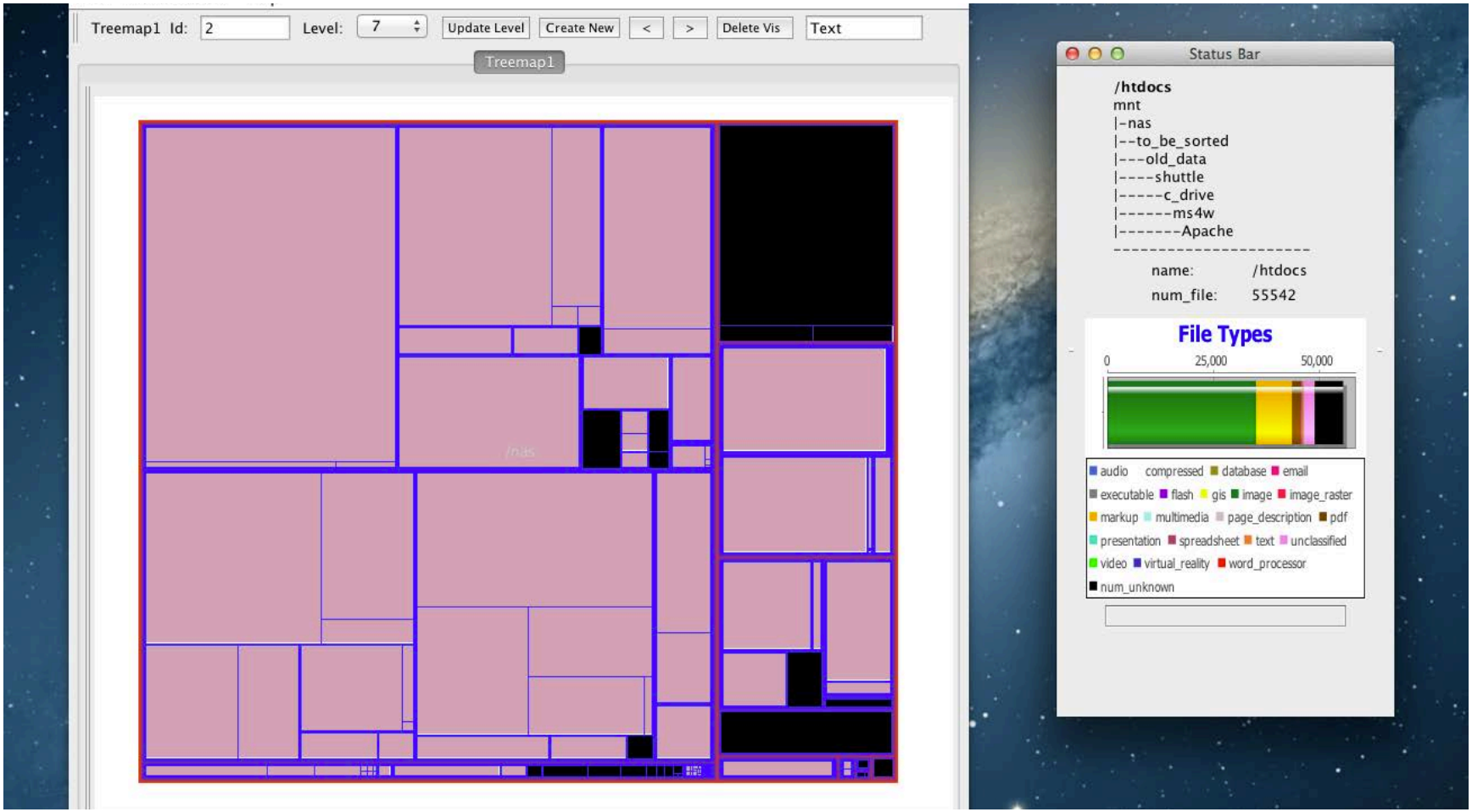
downstream into Ernst Tinaja. The Ernst Member of the Boquillas Formation consists of thin-be
noceramids (an extinct group of clams). These sediments were deposited in the Western Interic
etaceous Period (Scheubel and Mruk, 1994; Cooper and others (2007).



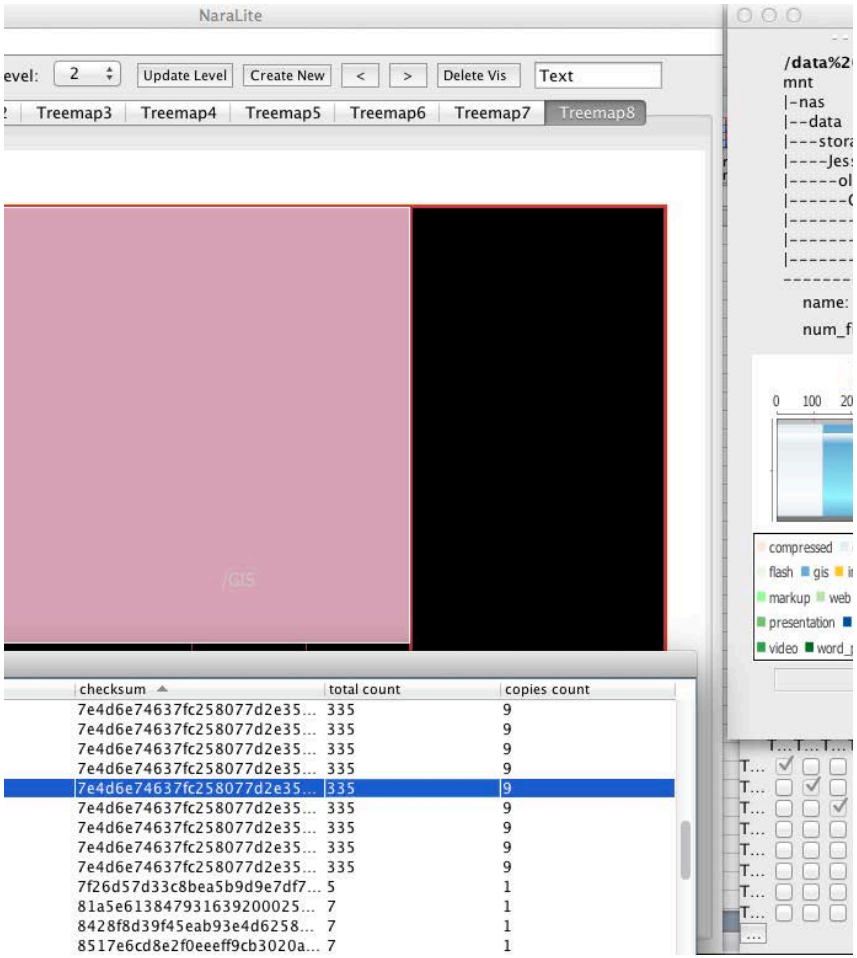
Description and access

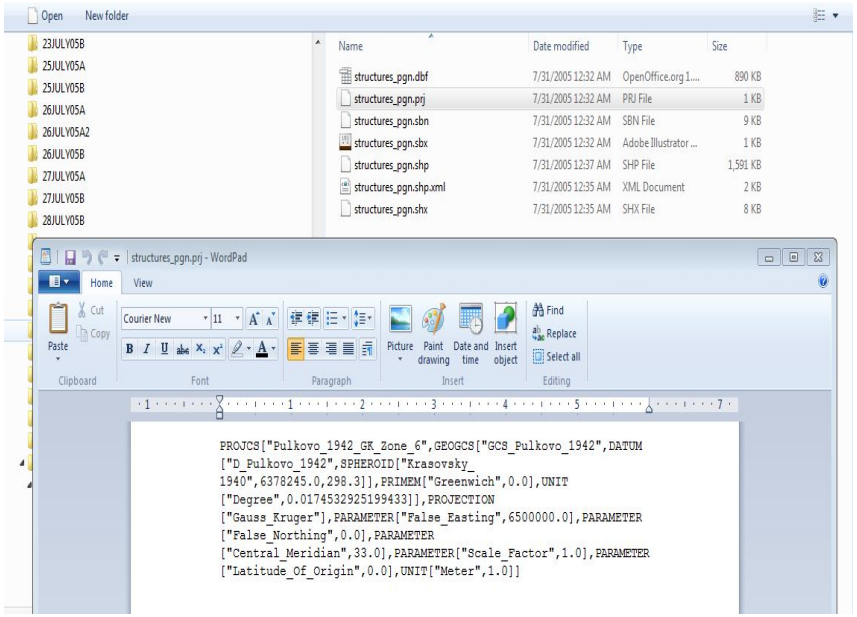


Integrity

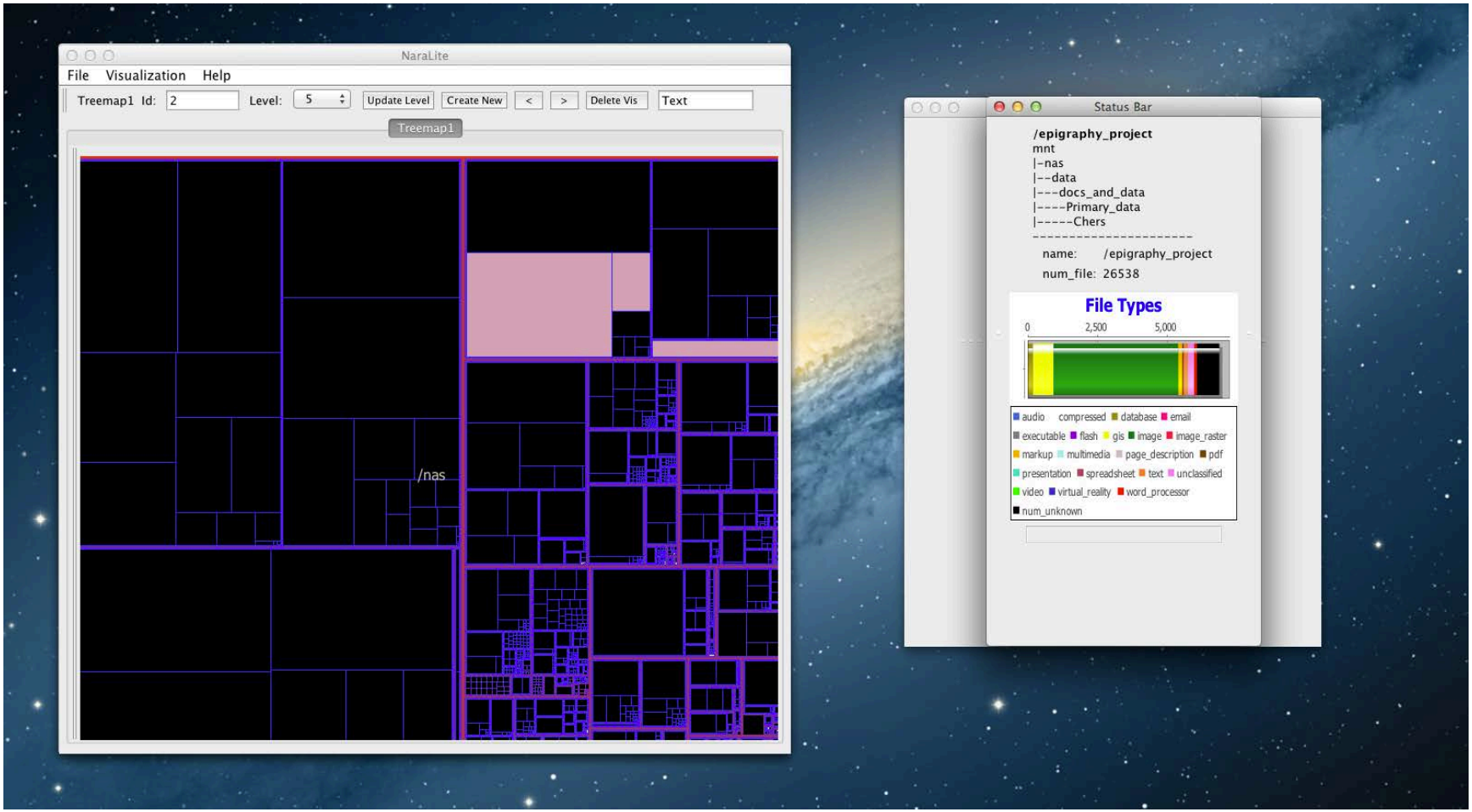

Treemap1 Id: 2
Level: 7
Update Level
Create New
<
>
Delete Vis
Text
Treemap1
Status Bar
/htdocs
mnt
|-nas
|--to_be_sorted
|---old_data
|----shuttle
|-----c_drive
|------ms4w
|-------Apache
name: /htdocs
num_file: 55542
File Types
0
25,000
50,000
audio
compressed
database
email
executable
flash
gis
image
image_raster
markup
multimedia
page_description
pdf
presentation
spreadsheet
text
unclassified
video
virtual_reality
word_processor
num_unknown

Granularity

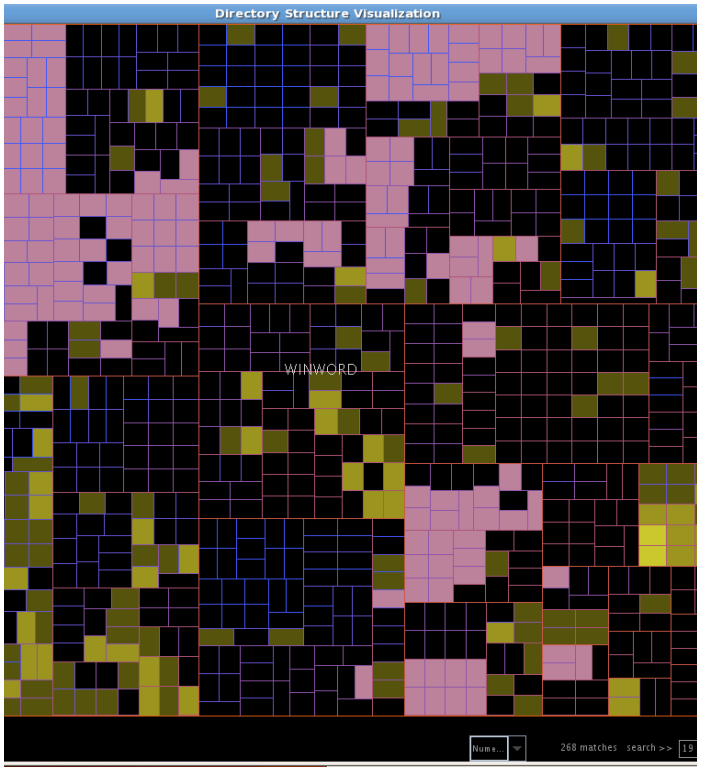


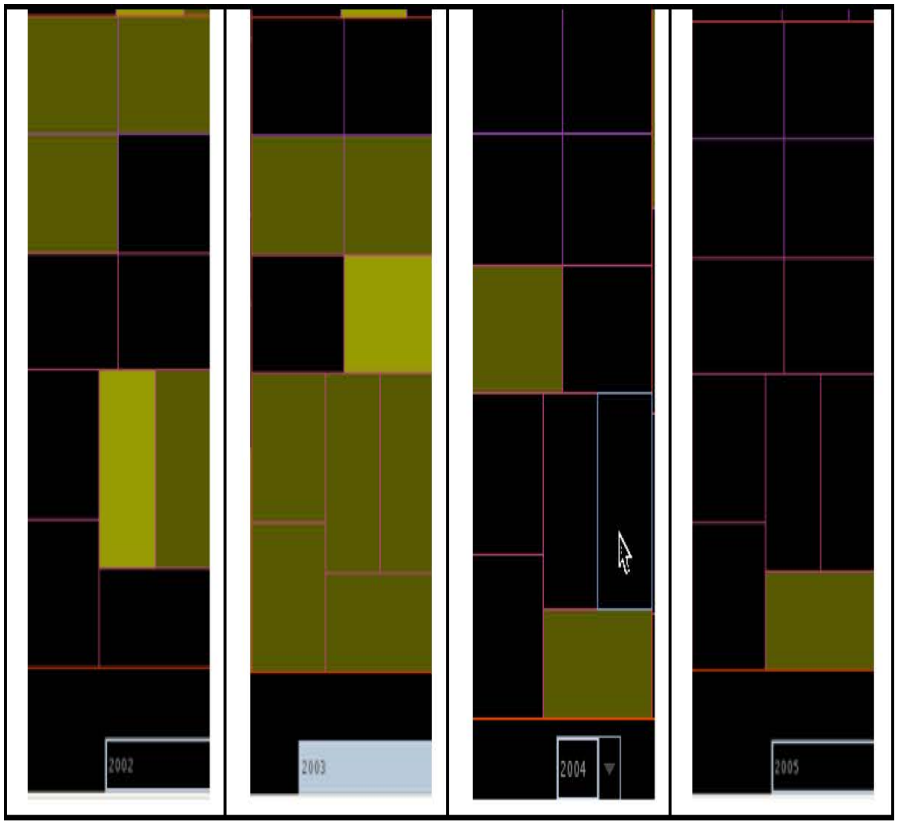


Duplicates in context



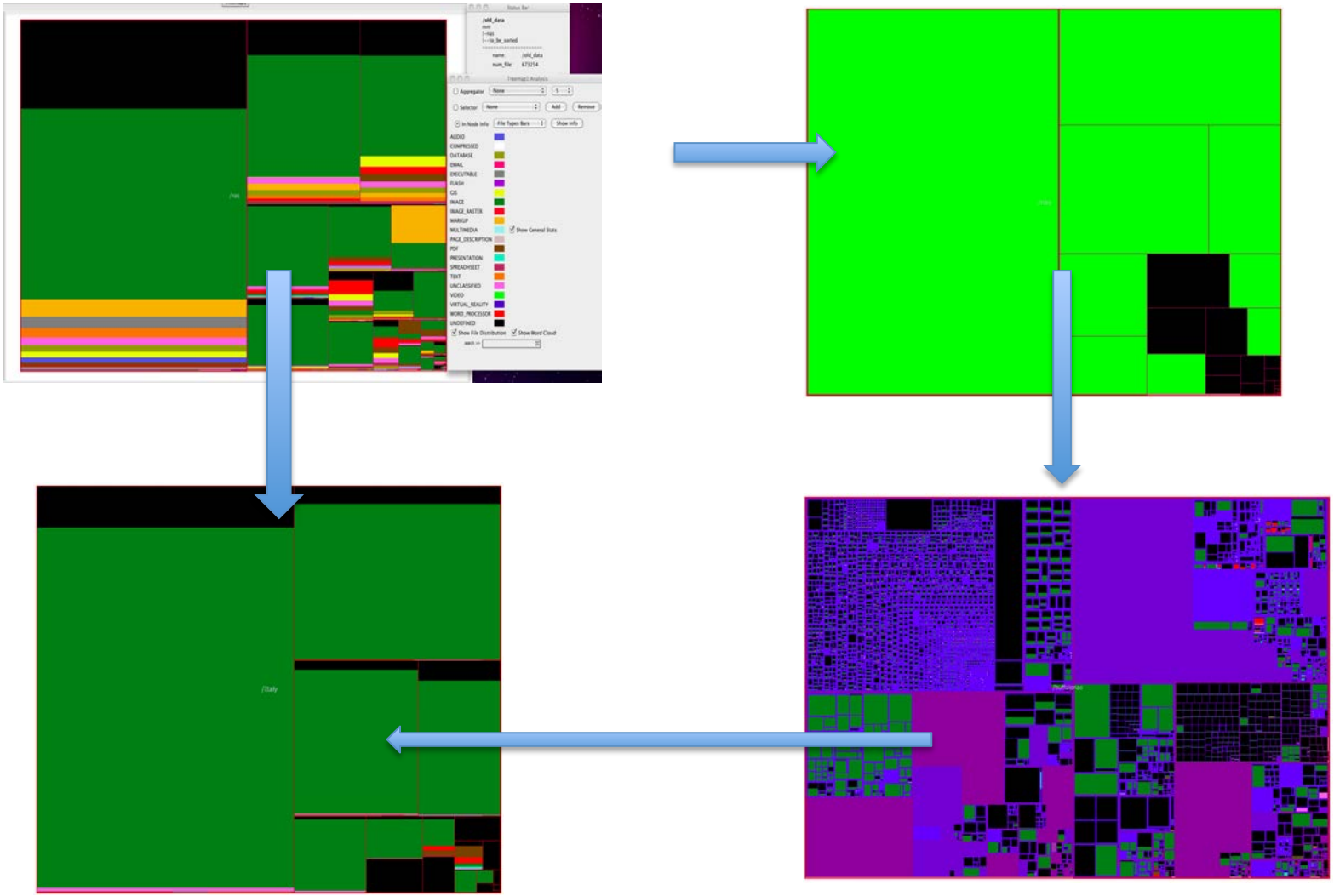
Context: organizational records



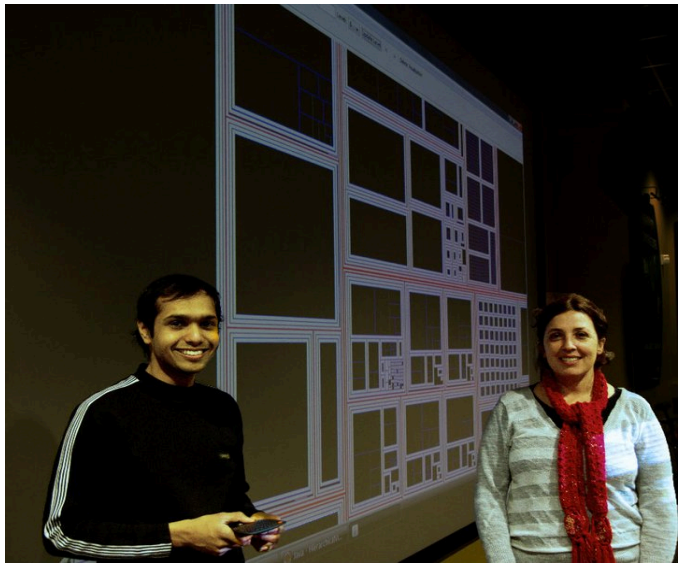


1 2 3 4

Triage workflows



Interaction and cooperation



Continuum

- Where do we want our profession to go
- Visual analytics as aid to archival thinking and imagination
- Make archives relevant today
- Understanding big archives/big data
  - Focus on aggregations and patterns
  - Relationships are highlighted
  - Learn what we do not know
- Context and structure for interpretation
- Research, research, research, research
- We have to drive it, in collaboration

Thanks

- National Archives and Records, Administration: Research Cooperative Agreement
- Institute of Classical Archaeology: Jessica Trelogan
- Dr. Weijia Xu
- Andrew Solis
- Nicholas Lauland

Bibliography

- Weijia Xu, Maria Esteva, Suyog Jain, Varun Jain, (2013). Interactive Visualization for Curatorial Analysis of Large Data Collections. Journal of Information Visualization, April 2013 12(2), doi: 10.1177/1473871612473590
- Maria Esteva, Weijia Xu, Suyog Jain Dott, Jennifer Lee, Wendy K. Martin, (2011). Assessing the Preservation Condition of Large and Heterogeneous Electronic Records Collections with Visualization. International Journal of Digital Curation, Vol 6. No 1. UKLON, University of Bath. Digital Curation Center. Available at: http://www.ijdc.net/index.php/ijdc/article/view/162
- Maria Esteva, Weijia Xu, Suyog Jain, (2010). Visualizing Personal Digital Collections. JCDL/ICADL' 10 Proceedings of the 2010 ACM/IEE JCDL/ICADL Joint Conference on Digital Libraries and Joint Conference on Asia-Pacific Digital Libraries, The University of Queensland, Australia. June 21 – 25 2010. ACM: New York, doi10.1145/1816123.1816147

The WWII Japanese-American Internment Camp Project

Dr. Richard Marciano

Maryland's iSchool & DCIC Center

and

William Underwood & Sandra Laib

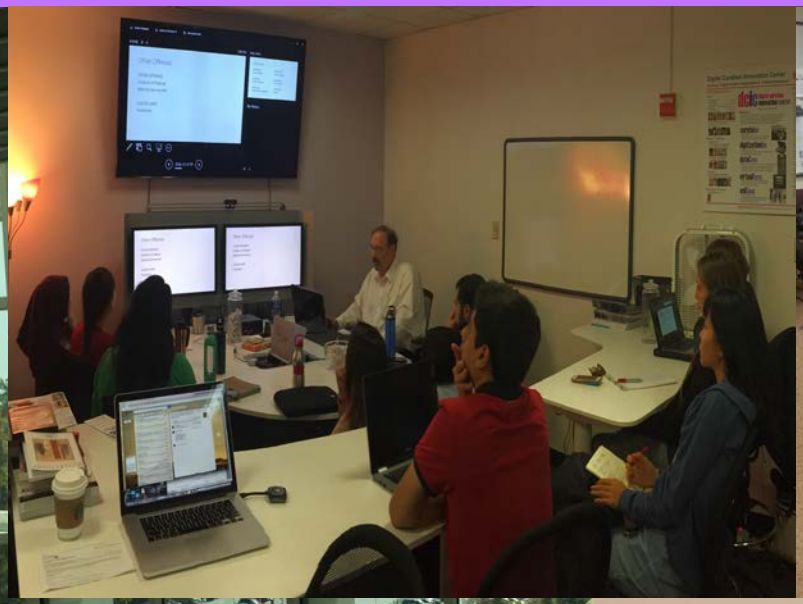
UNIVERSITY OF MARYLAND

COLLEGE OF INFORMATION STUDIES

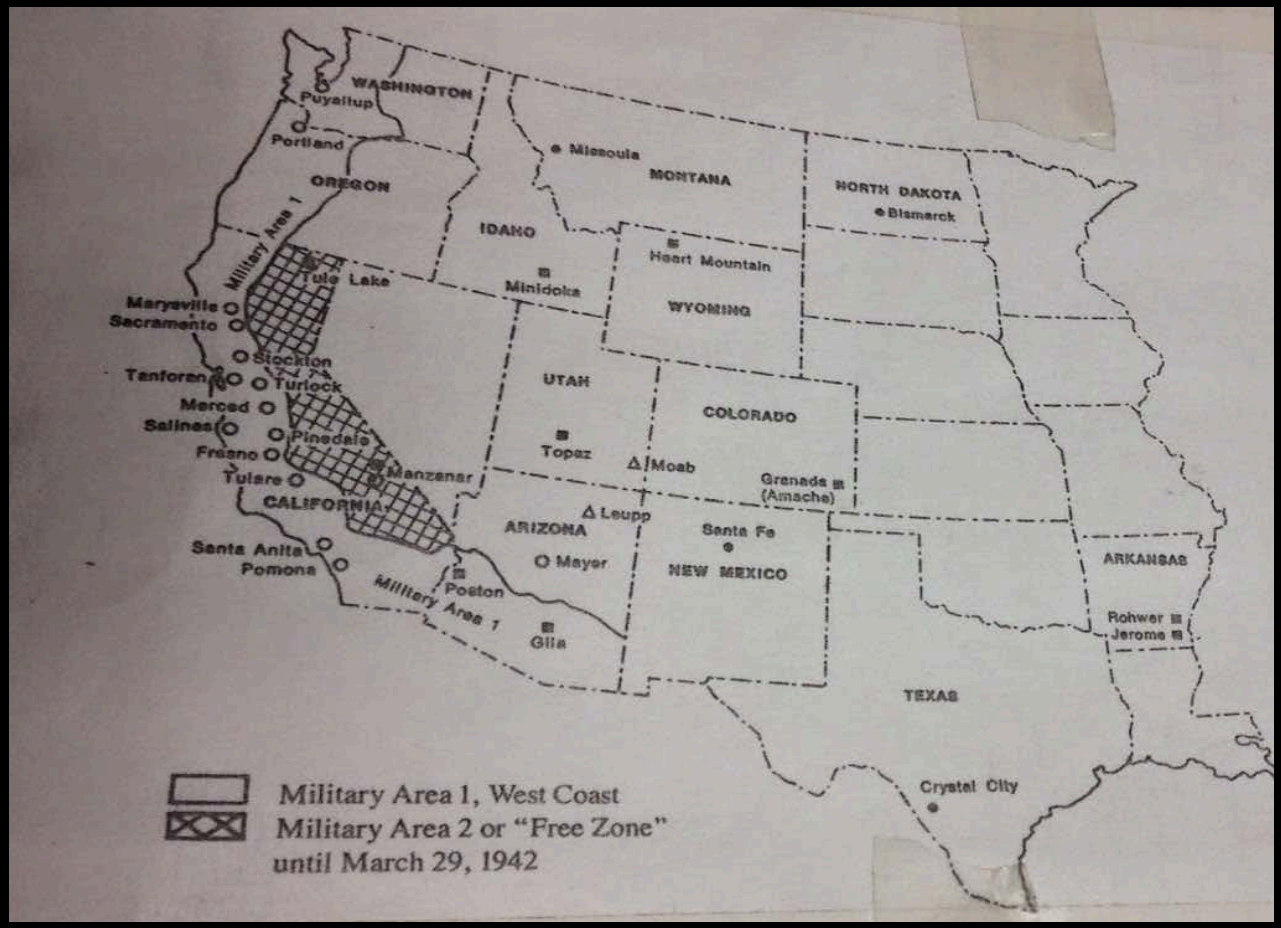
dcic digital curation innovation center

U. Maryland College of Information Studies Student Team

Carl Apgar
Luis Beteta
Waleed Falak
Marisa Gilman
Riss Hardcastle
Keona Holden
Yun Huang
David Baasch
Brittni Ballard
Tricia Glaser
Adam Gray
Leigh Plummer
Zeynep Diker
Mayanka Jha
Aakanksha Singh
Namrata Walanj



Background: Executive Order 9066



Issued by President Franklin Roosevelt on February 19, 1942, this order authorized the evacuation of all persons deemed a threat to national security from the West Coast to relocation centers further inland. In the next 6 months, over 100,000 men, women, and children of Japanese ancestry were removed from CA, Oregon and Washington state to assembly centers. They were then evacuated to and confined in isolated, fenced, and guarded relocation centers, known as internment camps.

Background: NARA Record Group 210

The National Archives and Records Administration (NARA) is the repository of the records of this program. Record Group 210, Records of the War Relocation Authority, includes paper records of internal security cases and associated paper index cards for the 10 Relocation Centers. These records have not been released to the public due to access restrictions on some of the records.

Objectives

- Curate the information in these cards by improving their quality (scanning, OCR, text correction, analysis, and extraction) and adding value to the repository of digital information by providing to NARA descriptive metadata supporting access to items in this record series.

- Extract item-level metadata from the Relocation Center index cards in order to supply archivists at the National Archives and Records Administration (NARA) with the information needed to support withdrawal decisions for items in this record series.

- Explore archival analytics approaches through social networking analysis and geospatial processing of the resource database of information extracted from the index cards.

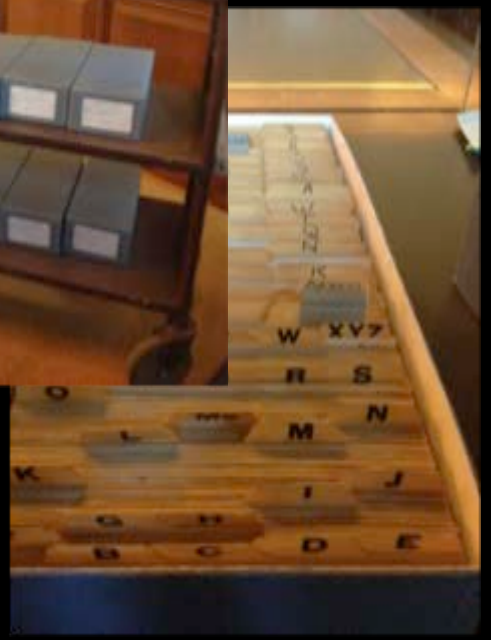
The DCIC is Pursuing a CAS Training / Teaching Agenda

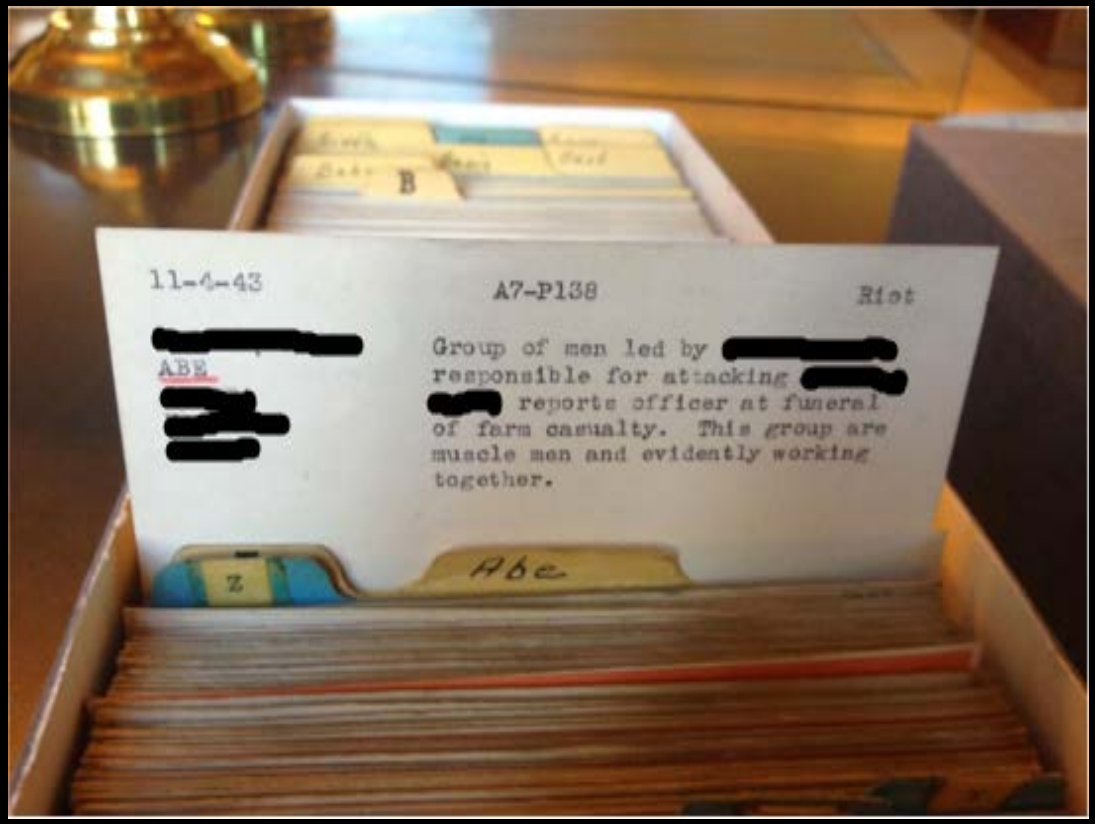
There is a need to :

- create innovative classes that emphasize new modes of collaboration, and interdisciplinary work.
- blend elements of archival thinking and computational thinking:
  - problem solving that uses modeling, decomposition, pattern recognition, abstraction, algorithm design, and scale.
- develop inter-disciplinary iSchools with faculty from Computer Science, Archival Science, and Data Science.
- develop extensive hands-on experience working with cyberinfrastructure to carry out archival functions.

WE WELCOME PARTNERSHIPS -- CONTACT US!

Incident Index Cards





Analysis of the Tule Lake Index Cards

Index card to case reports

11-4-43 A-999 P9 Riot

Amane, Ohashi
9999-D

One of possible leaders or trouble makers.

Meaning of the information on card

- Incident date,
- the case report ID (A-999),
- the relevant page number in the report (P9),
- the subject of the case report (incidents that are offenses such as a Riot),
- the Japanese-American or Japanese internee name,
- the residence ID in the camp (9999-D) and
- a remarks section

Analysis: Other Information on Index Cards

- names of administrative and staff members of the center
- document identifiers (C-816)
- document types (letter, memo)
- names of relocation center facilities (hospital, stockade)
- job titles (chief cook, #35 Block Manager, secretary)
- organizations such as young Japanese men's militant groups (Hokoku Seinen Dan)
- relocation center organizations (Administrative Police, Agricultural Division),
- locations (Honolulu, Hawaii, San Francisco, Calif.)
- time and time periods (2:45 pm, 36 days)
- Remarks: relations such as membership in militant organizations and actions/events (arrested, interviewed by, sentenced to time period, and released).

Styles of Index Cards

- Index card to Case Reports
  - With Other IDs (Family number (F#8504), Individual number (8504-A)
  - With Multiple Names
  - With Other Cases (Accidents, Transfers)
- Index card to Case Report using Keywords (Name, Subject, Location: Remarks)
- Index card to Registration Arrests (Name, Arrested, Sentenced, Released)
- Index card to Documents (Memo, Teletype, Letter)
- Continuation of Remarks – Remarks on back of Index Card

Data Cleaning through OCR

Untitled document [1] - ABBYY FineReader 12 Professional

FILE EDIT VIEW DOCUMENT PAGE AREA TOOLS HELP

Task Open Scan Full Color Edit Image Read Document language: English Save Document layout: Plain text Redo Undo Verify Text Redaction Find Options

IMAGE Read Page Analyze Page Preprocess image for best OCR results.

3-23-45 A-1067 INFRAC. PROJ. REG.

[redacted]
1514-A

The above and [redacted] were put in project jail for military marching, blowing of bugles, display of Japanese emblems. Occurred in the colony.

3-23-45

A-1067

INFR.AC . PROJ• REG

[redacted] 1514-A

The above and [redacted] were put in project jail for military inarching, blowing of bugles, display of Japanese emblems* Occurred in the colony*

Text Annotation using the ANNIE Pipeline

Selected Processing resources

!	Name	Type
●	Document Reset PR	Document Reset PR
●	ANNIE English Tokeniser	ANNIE English Tokeniser
●	ANNIE Gazetteer	ANNIE Gazetteer
●	Tule Gazetteer	ANNIE Gazetteer
●	ANNIE Sentence Splitter	ANNIE Sentence Splitter
●	ANNIE POS Tagger	ANNIE POS Tagger
●	ANNIE NE Transducer	ANNIE NE Transducer
●	Tule NE Transducer	ANNIE NE Transducer
●	ANNIE OrthoMatcher	ANNIE OrthoMatcher

Wordlists Specific to the Index Cards

Messages | Tule Informatio... | Tule Gazetteer

japanese_female_gi | Add

List name	Major	Minor	Language	Annotation type
Japanese_Militant_Organizations.lst	organization	militant		Lookup
Japanese_female_given_names.lst	person_first	female		Lookup
WAR_cities.lst	location	city		Lookup
WAR_location_facilities.lst	facility	building		Lookup
japanese_male_given_names.lst	person_first	male		Lookup
japanese_names_initialcaps.lst	person_full			Lookup
offenses.lst	offense			Lookup
tule_lake_job_titles.lst	jobtitle	tulelake		Lookup
tule_lake_organizations.lst	organization	tulelake		Lookup

Color-Coded Annotations and Annotation List for a Sample Index Card

Annotation Sets | Annotations List | Annotations Stack | Co-reference Editor | Text

11-4-43 A-999 P9 Riot
Amane, Ohashi
9999-D
One of possible leaders or trouble makers.

Type	Set	Start	End	Id	Features
Date		0	7	101	{kind=date, rule=DateNumDash, ruleFinal=DateOnlyFinal}
CaseReportId		17	22	106	{rule=CaseReportId_Page}
CaseReportPage		23	25	107	{rule=CaseReportId_Page}
Offense		40	44	108	{rule=Offenses}
Person		47	60	109	{rule=TuleReverseNamePersonNoSplit}
Residence		62	68	110	{rule=ResidenceAddress}

- [x] CaseReportId
- [x] CaseReportPage
- [x] Date
- [ ] Identifier
- [ ] Lookup
- [x] Offense
- [x] Person
- [x] Residence
- [ ] Sentence
- [ ] SpaceToken
- [ ] Split
- [ ] Token
- [ ] Unknown

Original markups

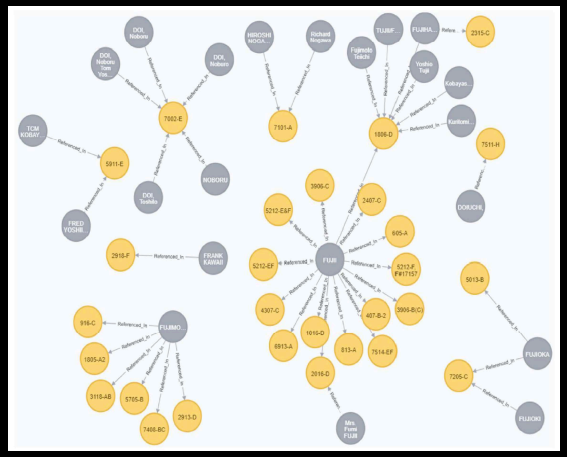
Interpretation of the Annotated Index Cards

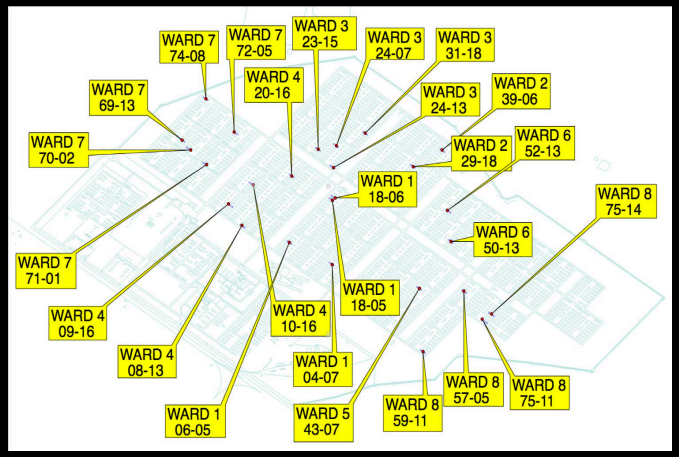
```
(qlf =[
    filename( “Box8-Tule_Lake-999.jpg” ),
    style( “Box8-Tule_Lake-999.jpg”,
           “index_card_to_case_reports” ),
    case_no( “Box8-Tule_Lake-999.jpg”,“A-999” ),
    page_no( “A-999”, “9” ),
    date( “A-999”, “11-4-43” ),
    subject( “A-999”, “Riot” ),
    person_name( ”Amane, Ohashi” ),
    gender( “Amane, Ohashi”, “male” ),
    indexname( “Amane, Ohashi”,
           “Box8-Tule_Lake-999.jpg” )
    residence_no( “Amane, Ohashi”, “999-D” )
])
```

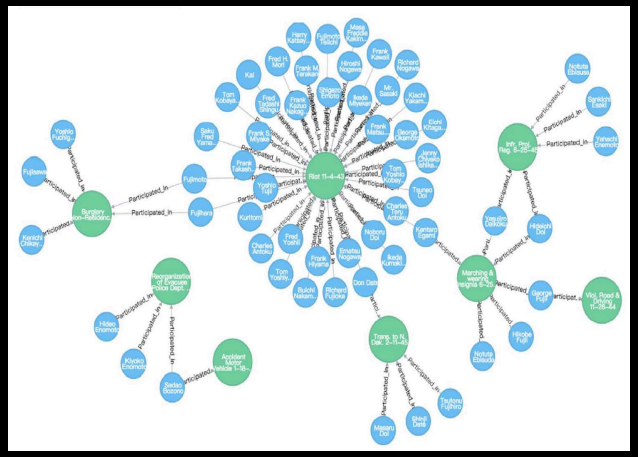
Workflow for Automated Review of the OCR-ed Index Cards for Restrictions on Release to Public

1. Scan, OCR and correct OCR errors of the "index cards".
2. Annotate the person's names, gender, and, if included, the age, family number, and individual number of the persons named on the OCRed incident cards.
3. Interpret the annotated terms using an Ontology of the concepts on the index cards and export the interpretation
4. Check the person names and associated information against the full database of internee information (also part of RG210), which includes person names, birthdates, family numbers and individual numbers.
5. Recommend index cards for withdrawal with person names and information matching information in the internee database of juveniles (under 18) and the release to the public of the other cards.

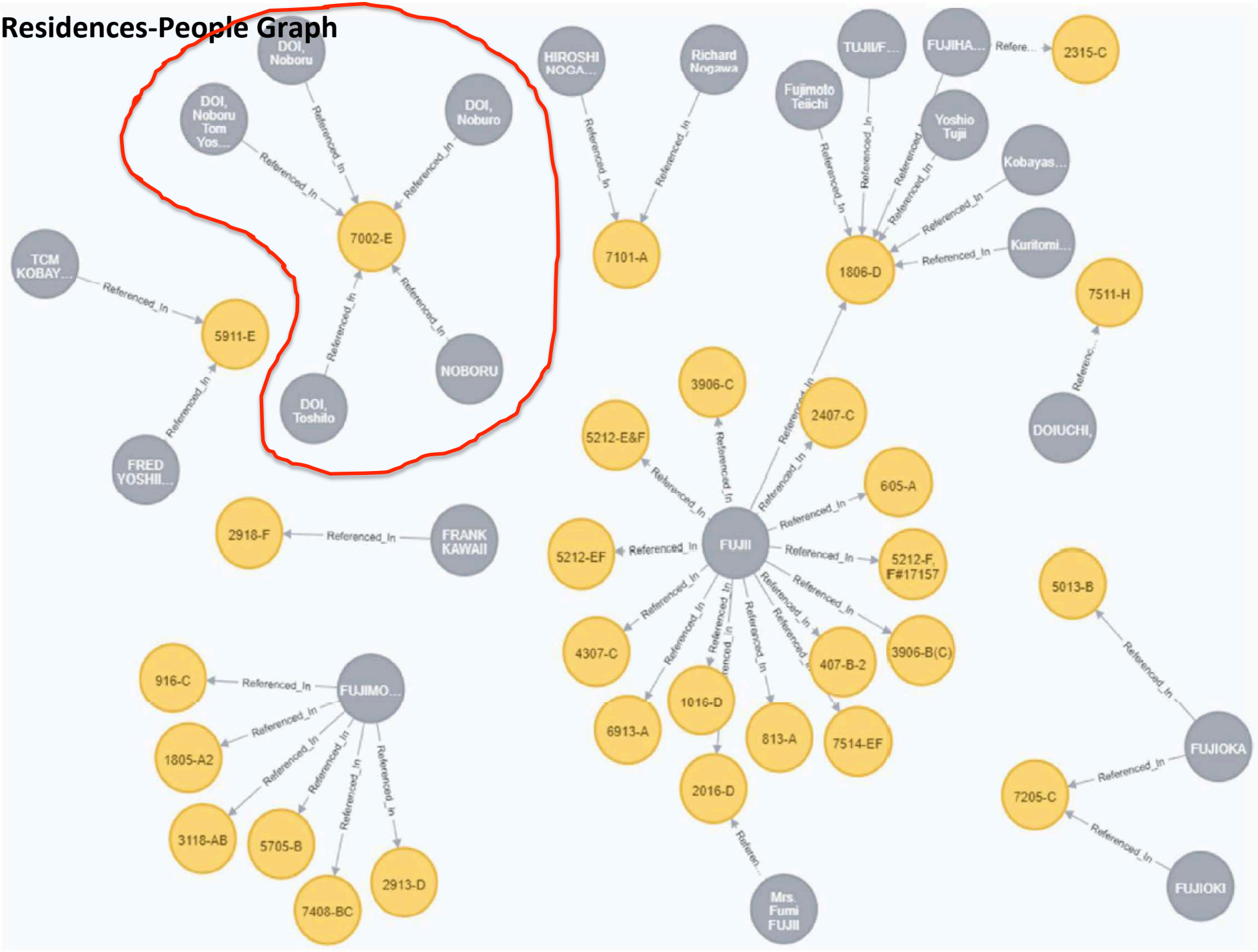
Visual Analytics of the Annotated Index Cards







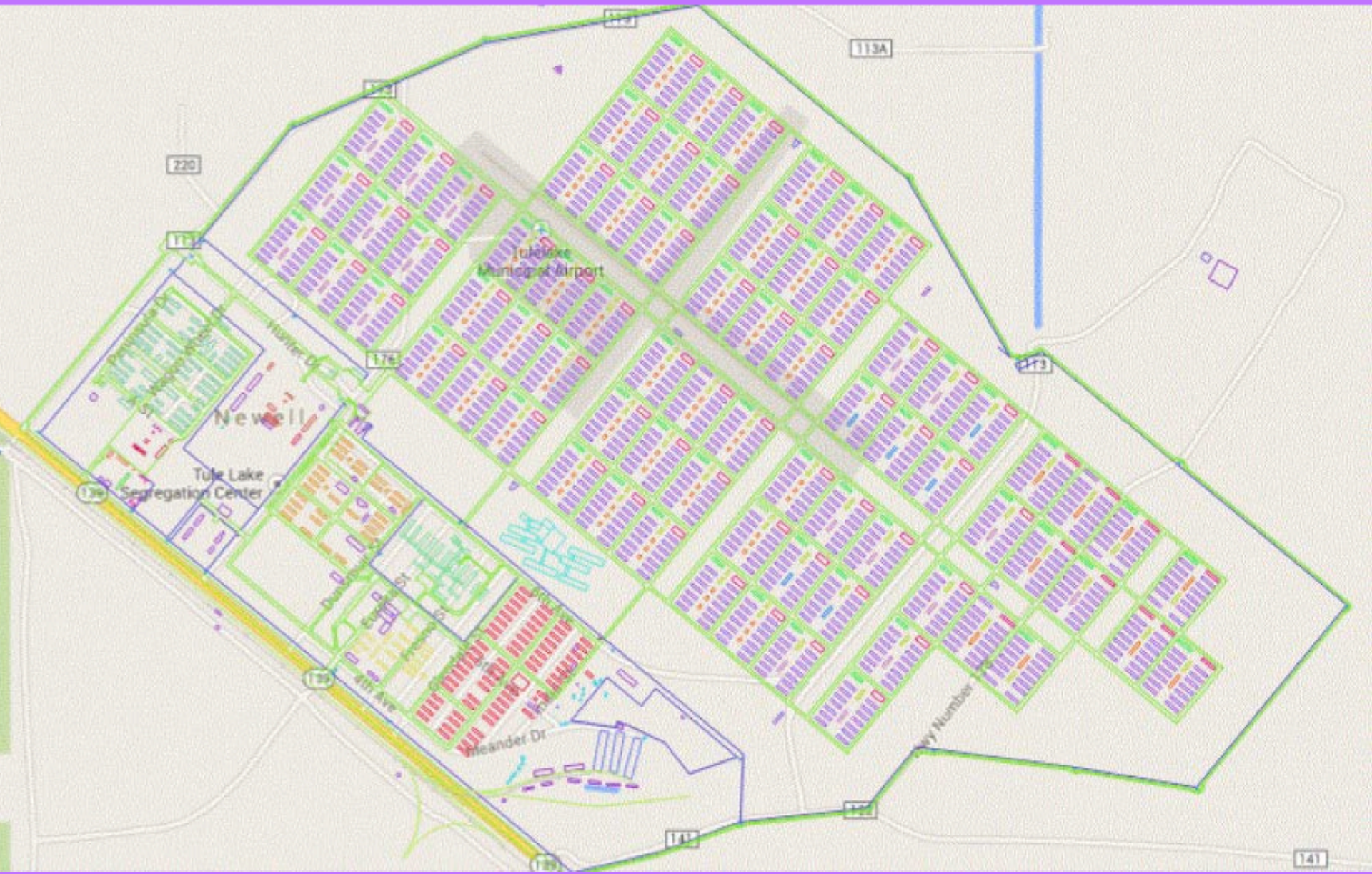
Residences-People Graph

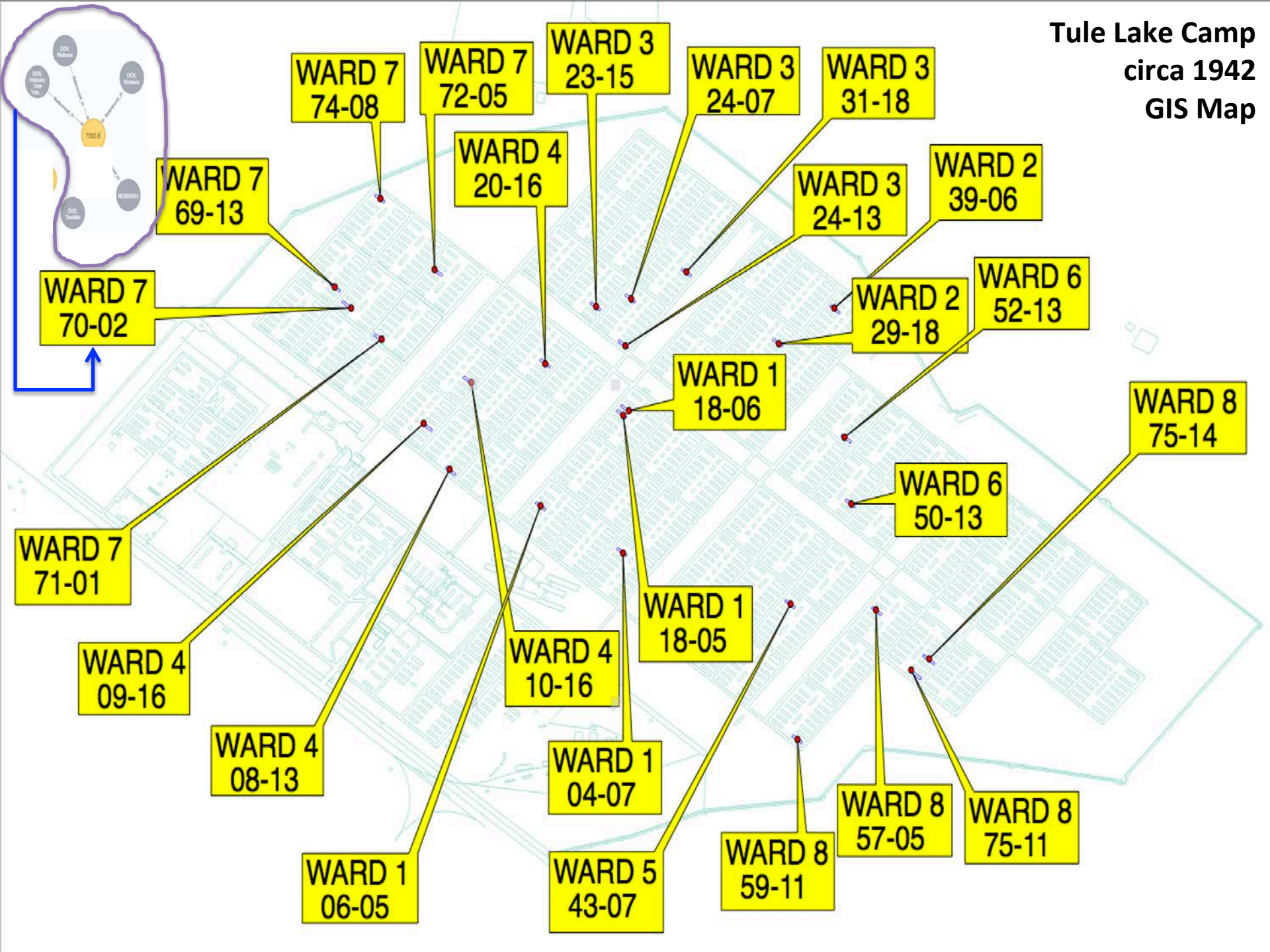


Interactive Map of Tule Lake

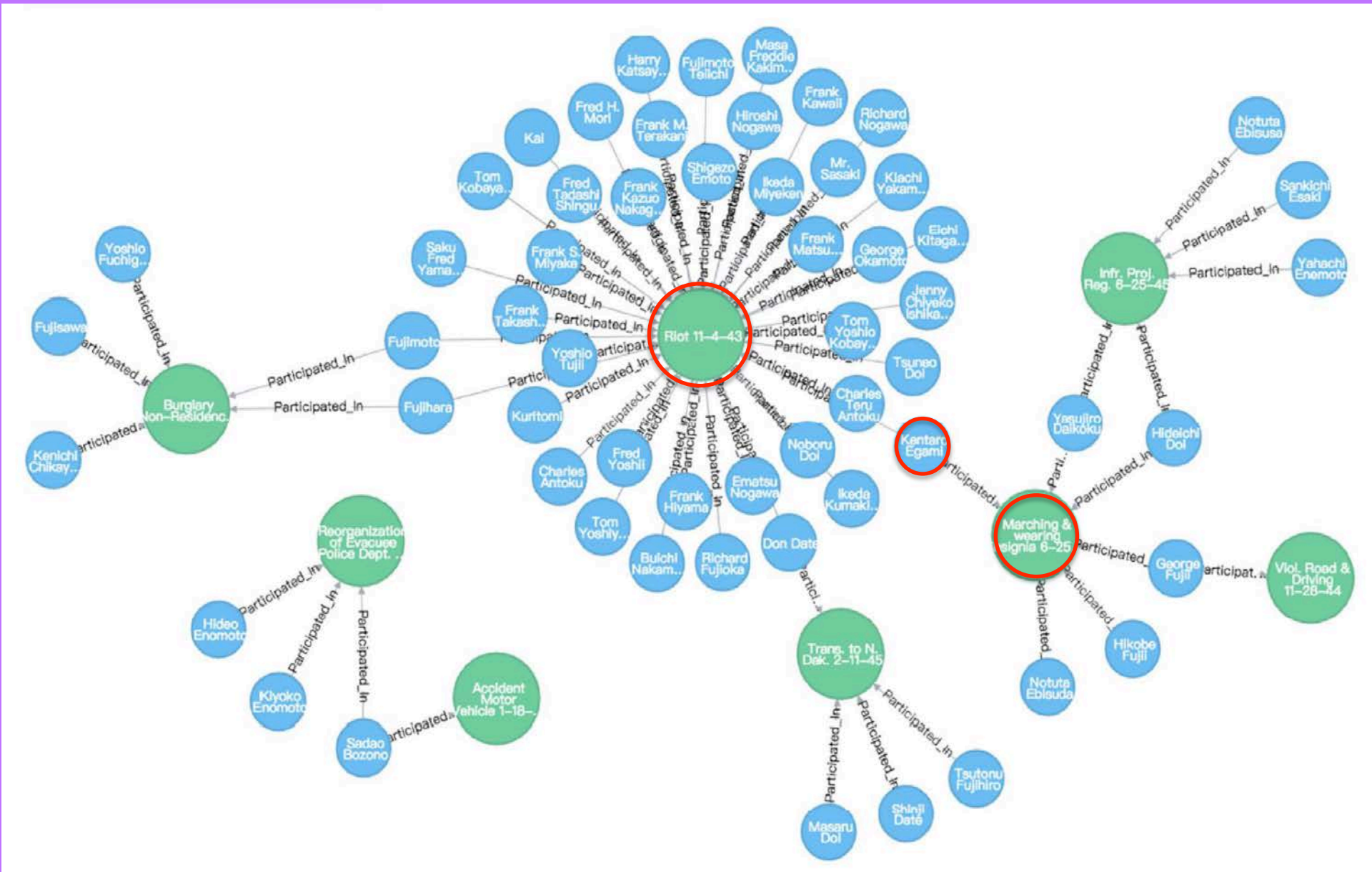
Digital Curation Innovation Center (DCIC)

Drew Barker with Dilip Bharadwaj and Scott Madry -- and initial support from CyArk




Tule Lake Camp
circa 1942
GIS Map
WARD 7
74-08
WARD 7
72-05
WARD 3
23-15
WARD 3
24-07
WARD 3
31-18
WARD 4
20-16
WARD 7
69-13
WARD 3
24-13
WARD 2
39-06
WARD 7
70-02
WARD 2
29-18
WARD 6
52-13
WARD 1
18-06
WARD 8
75-14
WARD 6
50-13
WARD 7
71-01
WARD 1
18-05
WARD 4
09-16
WARD 4
10-16
WARD 4
08-13
WARD 1
04-07
WARD 8
57-05
WARD 8
75-11
WARD 8
59-11
WARD 1
06-05
WARD 5
43-07

Events-People Graph



QUESTIONS