The Scope of Computational Archival Science (CAS) : Methods, Resources, and Interdisciplinary Approaches

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

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

The Scope of Computational Archival Science (CAS):
Methods, Resources, and Interdisciplinary Approaches

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

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

January 15, 2018
Kyushu University
Fukuoka, Japan

TACC

dcic digital curation innovation center

The Emergence of Computational XXX's

- XXX=Social Science
  - "Investigating social and behavioral relationships and interactions through: social simulation, modeling, network analysis, and media analysis", Wikipedia
- XXX=Biology
  - "The science of using biological data to develop algorithms or models to better understand biological systems", Wikipedia
- XXX=Journalism
  - "Finding and telling news stories, WITH, BY, or ABOUT algorithms", Nick Diakopoulos
- XXX=Archival Science ?
  - The Focus of this seminar…

What is CAS?

A trans-disciplinary field concerned with the application of:

- computational methods and resources to large-scale records / archives:
  - processing, analysis, storage, long-term preservation, and access,
  - with the 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 materials.

An example *(for Prof. Kohei Hatano !)*

How Artificial Intelligence Could Revolutionize Archival Museum Research *(Nov. 3, 2017)*

https://www.smithsonianmag.com/smithsonian-institution/how-artificial-intelligence-could-revolutionize-museum-research-180967065/

- Deep learning software to help botanists
- Botanical specimen categorization at museums (5 million specimens)
- Two big data analytics questions:
  1. With what accuracy can a trained neural network sort mercury-stained plant specimens from clean ones? [90-94%]
  2. With what accuracy can machine learning algorithms recognize members of two similar plant families? [96-99%]

Related Projects in 2016/2017:

U. California Santa Barbara

1. Always Already Computational – Collections as Data:
   - **Goal:** Hold meetings to develop strategies around library collections that support computationally-driven research and teaching
   - Website: https://collectionsasdata.github.io/ [Look at Resources Tab]
   - Funded by an IMLS National Forum Grant:
     - Award Abstract: https://www.imls.gov/grants/awarded/lg-73-16-0096-16
     - Full Proposal: https://www.imls.gov/sites/default/files/grants/lg-73-16-0096-16/proposals/lg-73-16-0096-16_proposal_documents.pdf

Related Projects in 2016/2017:

Library of Congress (National Digital Initiative)

2. Collections as Data:
   - **Goal:** The rise of accessible digital collections coupled with the development of tools for processing and analyzing data has enabled researchers to create new models of scholarship and inquiry.
   - 2016 Meeting: http://digitalpreservation.gov/meetings/dcs16.html
   - 2017 Meeting: http://digitalpreservation.gov/meetings/asdata/impact.html
     - YouTube videos: https://www.youtube.com/watch?v=OJWMHzgCu3c

Related Projects in 2017:

UK AHRC Heritage Priority Area, Heritage Futures, Alan Turing Institute, British Library workshop & report

3. Heritage and Data: Challenges and Opportunities for the Heritage Sector
   - **Goal:** Discuss challenges and opportunities arising as data becomes ever more significant in the heritage sector.
   - June 2017 British Library Meeting: http://hr.thedistrict.co.uk/app/uploads/2017/11/Heritage-Data-Challenges-Opportunities-Report.pdf
   - The cultural and natural heritage sector holds rapidly increasing volumes of data largely on human society and culture, past and present, which has become a new frontier of digital operations for many institutions.

University of Maryland iSchool (DCIC Center): "Computational Archival Science (CAS)" Portal

http://dcicblog.umd.edu/cas/

- Goal: Explore computational treatments of archival and cultural content
- Google Group: computational-archival-science@googlegroups.com
- Foundational Book Chapter: Jan. 2018
  - Book: "Advances in Librarianship – Re-Envisioning the MLIS: Perspectives on the Future of Library and Information Science Education".
  - Book Chapter: "Archival Records and Training in the Age of Big Data"

CAS Founding Partners:

- United States:
  - TACC: Maria Esteva
  - NARA: Mark Conrad
  - UMD: Richard Marciano, Bill Underwood, Greg Jansen, Michael Kurtz
- Canada:
  - UBC: Vicki Lemieux
- United Kingdom:
  - KCL: Mark Hedges

Foundational Book Chapter

Archival Records and Training in the Age of Big Data

(1) Evolutionary prototyping and computational linguistics (Bill Underwood)

(2) Graph analytics, DH and archival representation (Richard Marciano)

(3) Computational finding aids (Greg Jansen)

(4) Digital curation (Michael Kurtz)

(5) Public engagement with (archival) content (Mark Hedges)

(6) Authenticity (Victoria Lemieux)

(7) Confluences between archival theory and computational methods (Maria Esteva)

(8) Spatial and temporal analytics (Mark Conrad)

CAS Portal: http://dcicblog.umd.edu/cas/

SO FAR:

1. Workshop on CAS at Harvard Library (Dec. 2017)
2. Workshop at IEEE Big Data 2017 in Boston – CAS #2 (Dec. 2017)
3. Workshop at IEEE Big Data 2016 in Washington D.C. – CAS #1 (Dec. 2016)
4. Inaugural workshop at U. Maryland iSchool on CAS (Apr. 2016)

FUTURE:

1. Annual Association of Canadian Archivists (ACA) in Canada (June. 2018)
2. Annual Society of American Archivists (SAA) in D.C. (Aug. 2018)
3. iConference 2019 at the U. Maryland (Mar. 2019)

CAS#1 (2016):
http://dcicblog.umd.edu/cas/ieee_big_data_2016_cas-workshop/

Session 1:

1: Exploring Archives with Probabilistic Models: Topic Modelling for the Valorisation of Digitised Archives of the European Commission [Belgium & Google Germany]

- **Computational Methods:** Topic Modeling for concept extraction from large EC archival holdings
- **Archival Concepts:** Support accessibility to large historical European Commission archival holdings

#2: Traces Through Time: A Probabilistic Approach to Connected Archival Data [The UK National Archives]

- **Computational Methods:** Graph and Probabilistic Databases
- **Archival Concepts:** Going from paper catalog entries to digital catalogs, Matching records in distributed databases.

#3: Opening Up Dark Digital Archives Through The Use of Analytics to Identify Sensitive Content [USA]

- **Computational Methods:** Analytics, predictive coding to address PII
- **Archival Concepts:** technology assisted review accessibility of presidential and federal e-mail accessioned into National Archives

Session 2:

#4: Computational Provenance in DataONE: Implications for Cultural Heritage Institutions [USA]

- **Computatonal Methods:** DataONE extensions to PROV (Provenance data model)
- **Archival Concepts:** Provenance of scientific data records (datasets).Trust in authenticity of the data, transparency and reuse

#5: Content-based Comparison for Collections Identification [USA]

- **Computational Methods:** scalable, robust automated computational service for data content comparison.
- **Archival Concepts:** need for a service to assign globally unique persistent identifiers to data sets in order to support accessibility, reference and reuse.

Session 2 (cont.):

#6: Breaking Down the Invisible Wall to Enrich Archival Science and Practice [USA]

- **Computational Methods:** Linguistic Models and Graph Theory
- **Archival Concepts:** Enriched Archival Science concepts

#7: Mind the explanatory gap: Quality from Quantity [UK]

- **Computational Methods:** Abstraction and ontology construction
- **Archival Concepts:** provenance in terms of why, who and how

#8: Understanding Computational Web Archives Research Methods Using Research Objects [Canada]

- **Computational Methods:** Research Objects Framework used to analyze the computational methods used in web archives research, — Research Objects in Computational Science
- **Archival Concepts:** Web Archives Research Objects — Disciplinary perspective, legal agreements, Motivations, Interpretation, Designs, …

Session 3:

#9: Appraising Digital Archives with Archivematica [USA]

- **Computational Methods:** Analysis Tab — File Format Characterization, File Format policies, Bulk extractor (Identifies PII), Content Preview, Tagging
- **Archival Concepts:** Appraisal

#10: Mining and Analysing One Billion Requests to Linguistic Services [Germany]

- **Computational Methods:** Text mining
- **Archival Concepts:** Corpus — One Billion Requests for Linguistic Services

CAS#2 (2017):
http://dcicblog.umd.edu/cas/ieee_big_data_2017_cas-workshop/

A. Exploring Archival Data:

#1: Building new knowledge from distributed scientific corpus; HERBADROP & EUROPEANA: two concrete case studies for exploring big archival data [France & Netherlands]
- **Computational Methods:** EUDAT automated scalable e-infrastructure, integrated computational services, document scanning, OCR
- **Archival Concepts:** Trusted digital repositories (TDR), digitization, cultural heritage platforms

#2: An Infrastructure and Application of Computational Archival Science to Enrich and Integrate Big Digital Archival Data: Using Taiwan Indigenous Peoples Open Research Data (TIPD) as Example [Taiwan]
- **Computational Methods:** Record linking, GIS
- **Archival Concepts:** Big archival data

→ #3: Computational Curation of a Digitized Record Series of WWII Japanese-American Internment [*USA*]
- **Computational Methods:** NLP, NER, GIS, Graph database, linked data
- **Archival Concepts:** Digital curation, automated metadata extraction

#4: The Cybernetics Thought Collective Project: Using Computational Methods to Reveal Intellectual Context in Archival Material [*USA*]
- **Computational Methods:** NLP, NER, machine learning
- **Archival Concepts:** Geographically dispersed archives

B. Curation and Appraisal:

#5: Towards Automated Quality Curation of Video Collections from a Realistic Perspective [*USA*]
- **Computational Methods:** Feature computing from video records, automated quality prediction, scalable HPC
- **Archival Concepts:** Collection assessment, quality-aware metadata for video collections to inform appraisal, preservation, and access decisions, quality detection in videos

→ #6: Line Detection in Binary Document Scans: A Case Study with the International Tracing Service Archives [*USA*]
- **Computational Methods:** Line detection, image segmentation, OCR
- **Archival Concepts:** Digitization, Classification of archival images

#7: Auto-Categorization & Future Access to Digital Archives [Canada & *USA*]
- **Computational Methods:** Auto-categorization, document-classification, machine learning
- **Archival Concepts:** Recordkeeping, Record disposition

B. Curation and Appraisal (cont.):

#8: Heuristics for Assessing Computational Archival Science (CAS) Research: The Case of the Human Face of Big Data Project [*USA*]

- **Computational Methods:** Heuristics for CAS research, Digital Curation (scanning and adjusting, geo-referencing, geo-tracking), System Design (interface design, database design), Implementation (prototype), Value-sensitive design
- **Archival Concepts:** Records of an urban renewal project (property documents, map)

C. CAS Methodologies:

#9: What Can a Knowledge Complexity Approach Reveal About Big Data and Archival Practice? [Netherlands]

- **Computational Methods:** big data analytics, data science
- **Archival Concepts:** Knowledge complexity in archives, requirements for working with complex collections, context, resistance to computational approaches

#10: Protecting Privacy in the Archives: Preliminary Explorations of Topic Modeling for Born-Digital Collections [*Canada*]

- **Computational Methods:** NLP, NER, sentiment analysis, topic modeling
- **Archival Concepts:** Personally Identifiable Information

#11: Identifying Epochs in Text Archives [United Kingdom]

- **Computational Methods:** Cultural analytics, topic modeling
- **Archival Concepts:** Classification of time-coded collections of textual collections into epochs and periods

#12: GraphQL for Archival Metadata: An Overview of the EHRI GraphQL API [United Kingdom]

- **Computational Methods:** APIs for cultural heritage materials, graph databases, structured data query
- **Archival Concepts:** Query interfaces to archival materials

D. Creation & Mgt. of Current Records

#13: The Blockchain Litmus Test [*USA*]

- **Computational Methods:** Blockchain, secure computing, trustworthiness, risk analysis
- **Archival Concepts:** Decentralized recordkeeping

#14: A Typology of Blockchain Recordkeeping Solutions and Some Reflections on their Implications for the Future of Archival Preservation [Canada]

- **Computational Methods:** Blockchain, computational validation, distributed ledger, computational trust
- **Archival Concepts:** Recordkeeping, digital preservation, archival trust

CAS 2017

- → #6: **Line Detection in Binary Document Scans: A Case Study with the International Tracing Service Archives**

  - Benjamin Lee — United States Holocaust Memorial Museum, USA

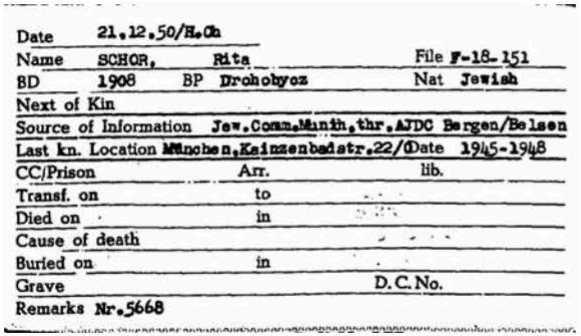
Computational Methods: Line detection, image segmentation

Archival Concepts: Classification of archival images

Context (from Wikipedia):

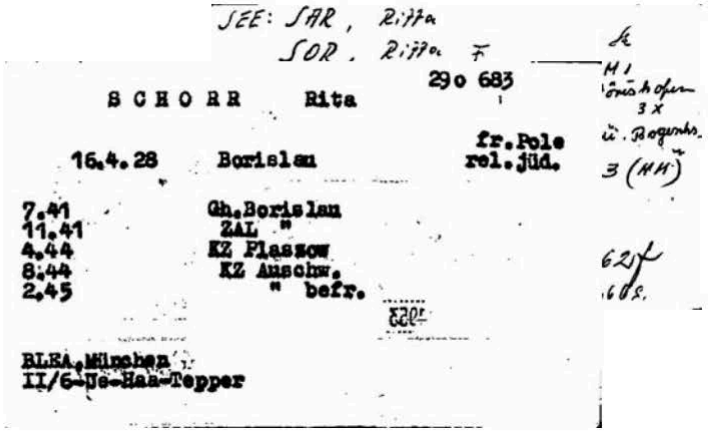
"The International Tracing Service (ITS), in Bad Arolsen, Germany, is an internationally governed center for documentation, information and research on Nazi persecution, forced labor and the Holocaust in Nazi Germany and its occupied regions."

It includes almost 50 million CNI (Central Name Index) Cards.



Date 21.12.50/H.Ch
Name SCHOR, Rita File F-18-151
BD 1908 BP Drohobycz Nat Jewish
Next of Kin
Source of Information Jew.Comm.Minth.thr.AJDC Bergen/Belsen
Last kn. Location München,Kaiserbadstr.22/Date 1945-1948
CC/Prison Arr. lib.
Transf. on to
Died on in
Cause of death
Buried on in
Grave D.C.No.
Remarks Nr.5668

Question:



SEE: SAR, Rita
SOR, Rita F
290 683

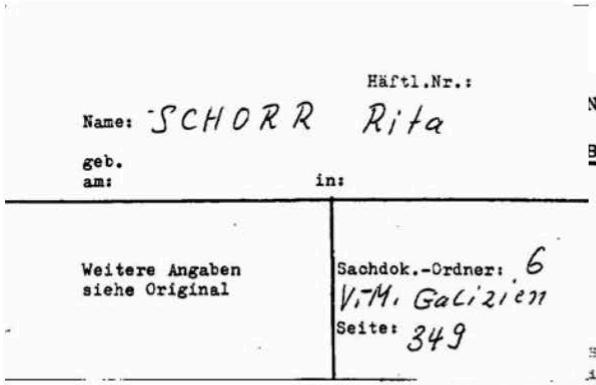
SCHORR Rita

16.4.28 Borislau fr.Pole rel.jüd.

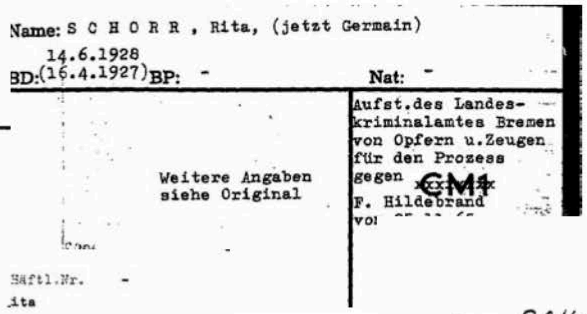
7.41 Gh.Borislau
11.41 ZAL "
4.44 KZ Plaszow
8.44 KZ Auschw.
2.45 " befr.

RLKA,München
II/6-Js-Has-Tepper

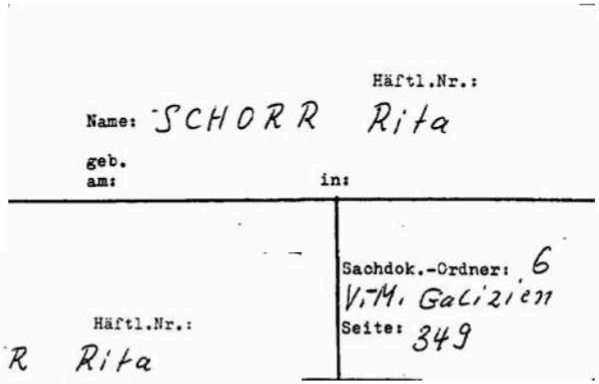
Can we sort the 47+ million CNI cards based on their line structure and form type in order to isolate card types of historical interest?



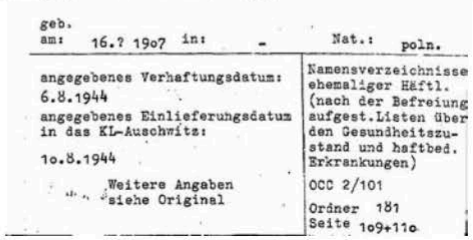
Häftl.Nr.:
Name: SCHORR Rita
geb.
am: in:
Weitere Angaben siehe Original
Sachdok.-Ordner: 6
V.M. Galizien
Seite: 349



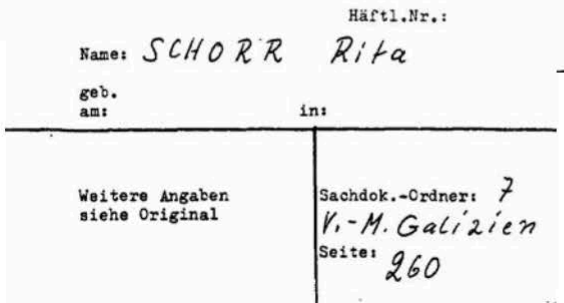
Name: S C H O R R , Rita, (jetzt Germain)
14.6.1928
BD:(16.4.1927) BP: - Nat: -
Weitere Angaben siehe Original
Aufst.des Landes-kriminalamtes Bremen von Opfern u.Zeugen für den Prozess gegen F. Hildebrand
Häftl.Nr. -
ita



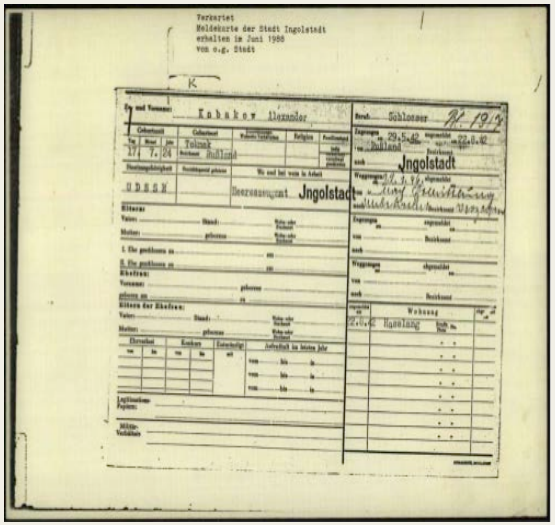
Häftl.Nr.:
Name: SCHORR Rita
geb.
am: in:
Sachdok.-Ordner: 6
V.M. Galizien
Seite: 349



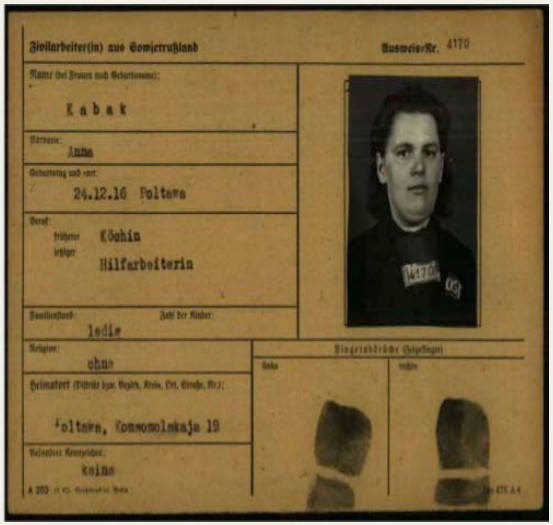
geb.
am: 16.? 190? in: - Nat.: poln.
angegebenes Verhaftungsdatum:
6.8.1944
angegebenes Einlieferungsdatum in das KL-Auschwitz:
10.8.1944
Weitere Angaben siehe Original
Namensverzeichnisse ehemaliger Häftl. (nach der Befreiung aufgest.Listen über den Gesundheitszustand und haftbed. Erkrankungen)
OCC 2/101
Ordner 181
Seite 109+110



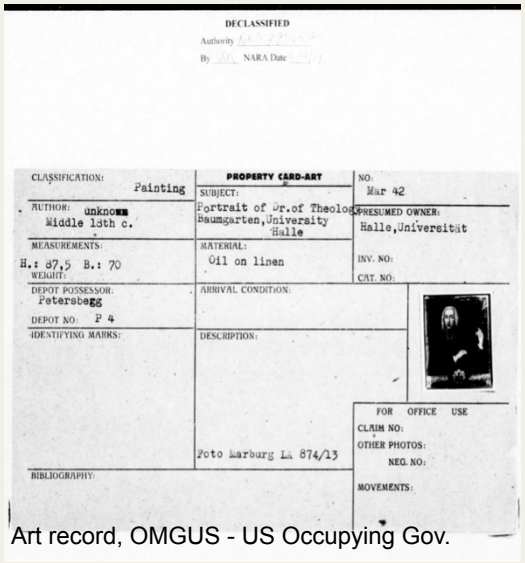
Häftl.Nr.:
Name: SCHORR Rita
geb.
am: in:
Weitere Angaben siehe Original
Sachdok.-Ordner: 7
V.-M. Galizien
Seite: 260



Displacement record, TRACING Service project



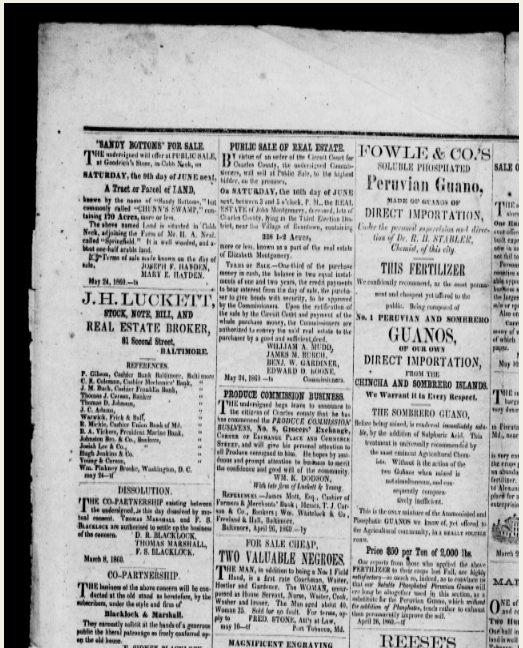
Displacement record, TRACING Service project



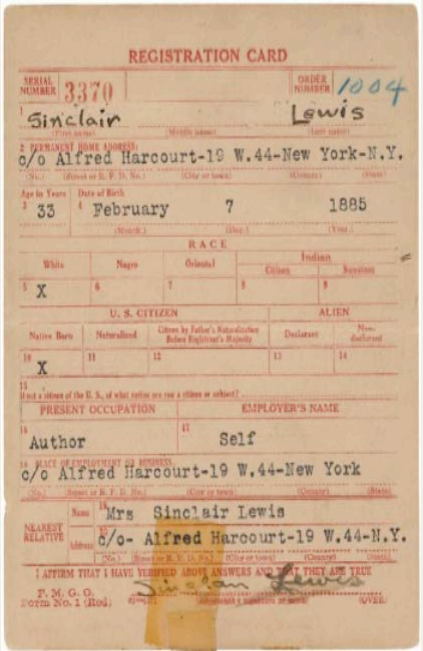
Art record, OMGUS - US Occupying Gov.

Making data from historical forms at the DCIC

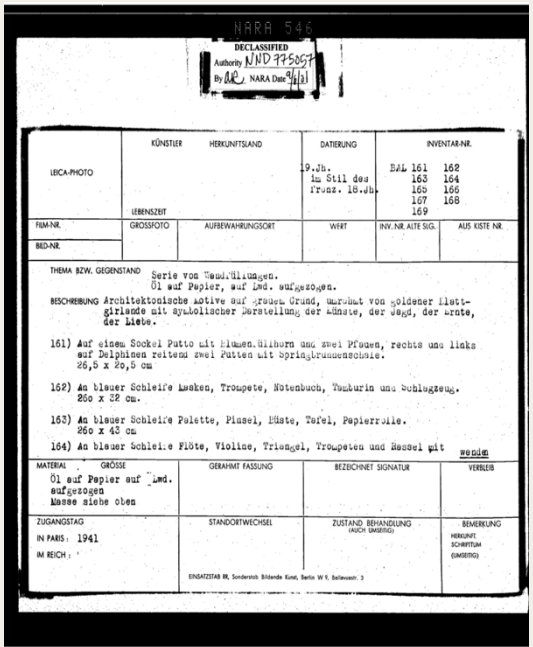
Maryland historical newspaper



United States draft card



Art record, ERR - Nazi looting organization



CAS 2017

- → #3: **Computational Curation of a Digitized Record Series of WWII Japanese-American Internment**
  - William Underwood, Richard Marciano, … — USA

Computational Methods: NLP, NER, GIS, Graph database

Archival Concepts: Digital curation, automated metadata extraction

WESTERN DEFENSE COMMAND AND FOURTH ARMY
WARTIME CIVIL CONTROL ADMINISTRATION

Presidio of San Francisco, California
April 1, 1942

INSTRUCTIONS TO ALL PERSONS OF JAPANESE ANCESTRY

Living in the Following Area:

All that portion of the City and County of San Francisco, State of California, lying generally west of the north-south line established by Junipero Serra Boulevard, Worchester Avenue, and Nineteenth Avenue, and lying generally north of the east-west line established by California Street, to the intersection of Market Street, and thence on Market Street to San Francisco Bay.

All Japanese persons, both alien and non-alien, will be evacuated from the above designated area by 12:00 o'clock noon Tuesday, April 7, 1942.

No Japanese person will be permitted to enter or leave the above described area after 8:00 a. m., Thursday, April 2, 1942, without obtaining special permission from the Provost Marshal at the Civil Control Station located at:

1701 Van Ness Avenue
San Francisco, California

The Civil Control Station is equipped to assist the Japanese population affected by this evacuation in the following ways:

1. Give advice and instructions on the evacuation.
2. Provide services with respect to the management, leasing, sale, storage or other disposition of most kinds of property including: real estate, business and professional equipment, buildings, household goods, boats, automobiles, livestock, etc.
3. Provide temporary residence elsewhere for all Japanese in family groups.
4. Transport persons and a limited amount of clothing and equipment to their new residence, as specified below.

The Following Instructions Must Be Observed:

1. A responsible member of each family, preferably the head of the family, or the person in whose name most of the property is held, and each individual living alone, will report to the Civil Control Station to receive further instructions. This must be done between 8:00 a. m. and 5:00 p. m., Thursday, April 2, 1942, or between 8:00 a. m. and 5:00 p. m., Friday, April 3, 1942.
2. Evacuees must carry with them on departure for the Reception Center, the following property:
   (a) Bedding and linens (no mattress) for each member of the family;
   (b) Toilet articles for each member of the family;
   (c) Extra clothing for each member of the family;
   (d) Sufficient knives, forks, spoons, plates, bowls and cups for each member of the family;
   (e) Essential personal effects for each member of the family.

All items carried will be securely packaged, tied and plainly marked with the name of the owner and numbered in accordance with instructions received at the Civil Control Station.

The size and number of packages is limited to that which can be carried by the individual or family group.

No contraband items as described in paragraph 6, Public Proclamation No. 3, Headquarters Western Defense Command and Fourth Army, dated March 24, 1942, will be carried.

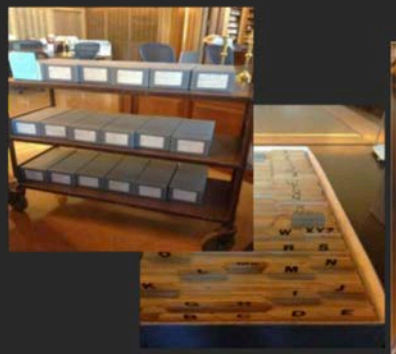
3. The United States Government through its agencies will provide for the storage at the sole risk of the owner of the more substantial household items, such as iceboxes, washing machines, pianos and other heavy furniture. Cooking utensils and other small items will be accepted if crated, packed and plainly marked with the name and address of the owner. Only one name and address will be used by a given family.
4. Each family, and individual living alone, will be furnished transportation to the Reception Center. Private means of transportation will not be utilized. All instructions pertaining to the movement will be obtained at the Civil Control Station.

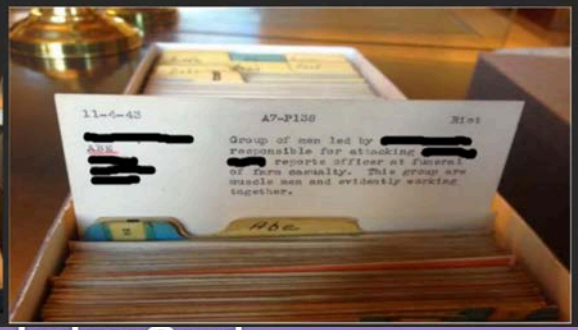
Go to the Civil Control Station at 1701 Van Ness Avenue, San Francisco, California, between 8:00 a. m. and 5:00 p. m., Thursday, April 2, 1942, or between 8:00 a. m. and 5:00 p. m., Friday, April 3, 1942, to receive further instructions.

J. L. DeWITT
Lieutenant General, U. S. Army
Commanding

SEE CIVILIAN EXCLUSION ORDER NO. 5

Incident Index Cards





List for a Sample Index Card

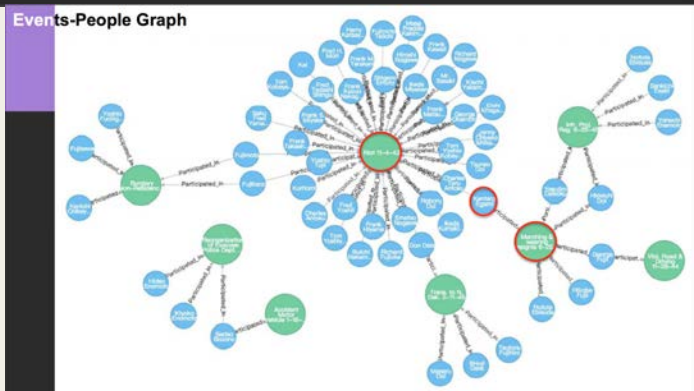
Annotation Sets | Annotations List | Annotations Stack | Co-reference Editor | Text

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}

- CaseReportId
- CaseReportPage
- Date
- Identifier
- Lookup
- Offense
- Person
- Residence
- Sentence
- SpaceToken
- Split
- Token
- Unknown
- Original markups

GOAL:
Automating the review and release of records at scale



THEMES:

Big Records & Archival Analytics

Computational Archival Science (CAS)

dcic

digital curation innovation center

http://dcic.umd.edu

Mission:

Be a leader in the digital curation research and educational fields, and foster interdisciplinary collaborations using Big Records and Archival Analytics with public / industry / government partnerships.

Goals:

Sponsor interdisciplinary projects that explore the integration of archival research data, user-contributed data, and technology to generate new forms of analysis and historical research engagement, particularly in the arenas of social justice, human rights, and cultural heritage.

Motto:

"Integrating Education and Research"

COLLEGE OF INFORMATION STUDIES

People

Research Staff:

Richard **Marciano**	*Director & SALT Lab Director*
Michael **Kurtz**	*Associate Director*
Ricardo **Punzalan**	*Research Associate & ARC Lab Director*
Ken **Heger**	*Research Associate & DigitizationLab Director*
Greg **Jansen**	*Research Software Architect*
Adam **Kriesberg**	*Researcher & Lecturer*
Maria **Esteva**	*Affiliate Professor*
Victoria **Lemieux**	*Affiliate Professor*
William **Underwood**	*Affiliate Professor*

Student Assistants:

Myeong **Lee**, Sohan **Shah**, & Martin **Moreno** *Lab Assistants*

Research Affiliates:

U. Maryland

Tammy **Clegg**, Nick **Diakopoulos**, Jesse **Johnston**, Katy **Lawley**, Trevor **Owens**, Jenny **Preece**, & Katie **Shilton**

External

Bruce **Ambacher**, Natalie **Baur**, John **Burns**, Andrew **Lau**, & Scott **Madry**

Postdoctoral Fellows:

Morgan **Daniels** & Kathryn **Gucer** *(Advisor: Punzalan)*

Doctoral Students:

Andrew **Casertano** & Will **Thomas**	*(Advisor: Marciano)*
Diane **Travis**	*(Advisors: Butler/Marciano)*
Caitlin **Christian-Lamb**, Edward **Summers**, & Amy **Wickner**	*(Advisor: Punzalan)*

curatelab

Hornbake South 4110

Digital lab for group learning, collaborative design, and hands-on digital curation project development (23 seats, 3 interactive screens, 12 workstations with 12TB of storage).

digitizationlab

Hornbake South 4110D

Document scanning, image manipulation, and archival ingestion facility for group projects.

serverfarm

Atlantic Building

On-campus virtual machine farm for research data processing, storage, and hosting (15TB storage, 2 Dell servers, VMWare-powered).

cloudlab

Amazon Cloud

Dashboard-enabled virtual computing lab in the cloud for creating Windows/Ubuntu instances using Amazon Web Services (AWS).

dataCave

UMD Cyberinfrastructure Center at the Rivertech Bldg

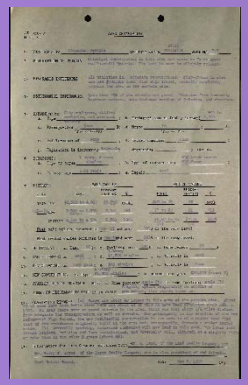
DRAS-TIC

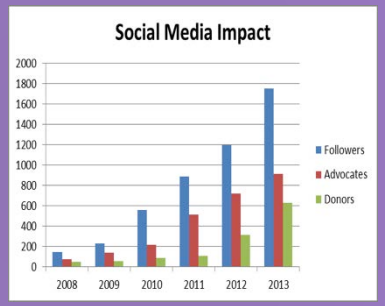
Digital Repository At Scale That Invites Computation (To Improve Collections): a petascale archival storage and preservation repository (based on the **DRAS-TIC** open-source software [NoSQL Cassandra database] and computational infrastructure (4 Dell nodes).

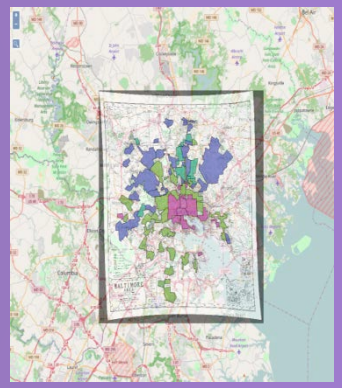
DCIC Digital Curation Projects…

"The active and ongoing management and enhancement of digital assets for current and future use." Digital curation entails more than secure storage and preservation of digital information because curation may add value to digital information and increase its utility.

[Preparing the Workforce for Digital Curation (2015) - NRC / BRDI Report]







Archival Documents **Digitization** **Datafication Data Modeling** **Data Visualization Digital Storytelling**

Archival Analytics

Practical Digital Curation Skills for Archivists in the 21st Century

MARAC, Nov 3, 2016

http://drum.lib.umd.edu/handle/1903/18865

Justice, Human Rights, & Cultural Heritage:

THEME	PROJECTS
Community Displacement	The Human Face of Big Data
Racial Zoning:	Mapping Inequality
Refugee Narratives:	St. Louis Voyage
Citizen Internment:	Japanese American WWII Camps
Movement of People:	Overseas Pension Project
Revealing Untold Stories:	Legacy of Slavery

Cyberinfrastructure for the Cur. & Mgt. of Dig. Assets at Scale:

FUNDING	PROJECTS
NSF	Brown Dog
IMLS	DRAS-TIC Fedora

Projects

Justice, Human Rights, & Cultural Heritage:

1. Overseas Pension Project
A student- and professional society-driven project to collect information documenting payment of pensions to American veterans living overseas. The project documents migration patterns, the flow of money, health conditions, and family connections prior to World War I.

2. IRP² Graph Database Study
This study is a follow-up project to the federated search capability developed for the International Research Portal Project (IRP²) for Records Related to Nazi-Era Cultural Property. In this phase a sample of property cards created by the U.S. Army at the end of World War II to inventory and control property seized by the Nazis will be used to explore the feasibility of using graph database/visualization technology to link property card data captured through OCR.

3. Legacy of Slavery
In partnership with the Maryland State Archives, two projects using digitized records involving Manumissions and Certificates of Freedom and Underground Railway documentation will mine data and create finding aids to enhance access for users.

4. Japanese-American WWII Camps
In partnership with the US National Archives, the project explores the auto-redaction of WWII Japanese-American Camp data. NLP and NER text processing techniques are explored through the development of extraction workflows.

5. Mapping Inequality
In partnership with Johns Hopkins, Virginia Tech, and U. of Richmond, a national collection of New Deal redlining records is being crowdsourced (these unique records capture racial, ethnic, and economic conditions).

6. The Human Face of Big Data
A student-led project that will create access and collaborative opportunities around historically and socially-significant heterogeneous datasets rooted in urban renewal housing records for a number of cities.

Cyberinfrastructure for the curation & management of digital assets at scale:

7. DRAS-TIC Fedora
See project description under Funding.

8. Hands-On Computation in the Archives
This project leverages Brown Dog, DRAS-TIC, Graph Databases, and ElasticSearch to create CAS learning opportunities for students.

2016/2017
Digital Curation
PROJECT OBJECTIVES:

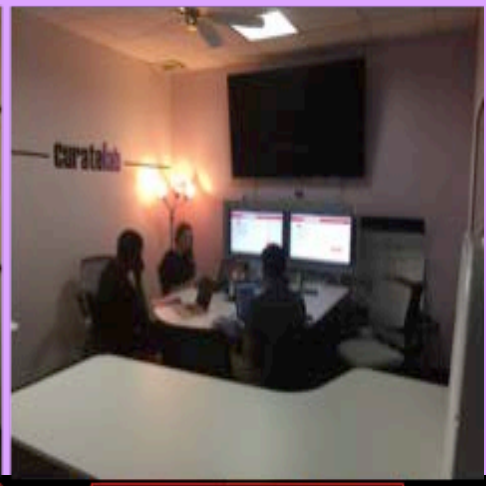
(1) Gain new digital skills

(2) Conduct interdisciplinary research

… at the intersection of archives, big data, & analytics

REFUGEE NARRATIVES (#8)
St. Louis Voyage

COMMUNITY DISPLACEMENT (#7)
Human Face of Big Data

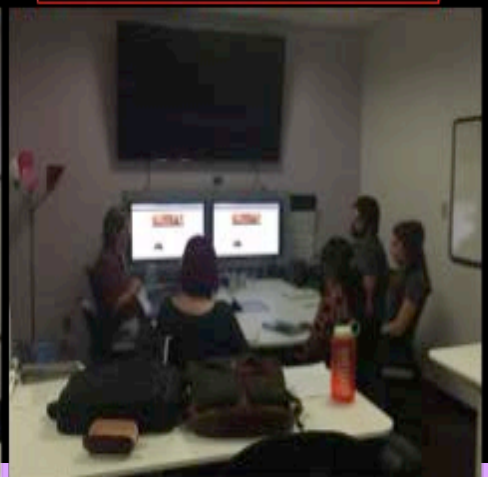


RACIAL ZONING (#11)
Mapping Inequality

DIGITIZATION LAB (#3)
Radio Log Program

CI FOR DIGITAL CURATION (#5)
Brown Dog

MOVEMENT OF PEOPLE (#5)
Overseas Pension



CITIZEN INTERNMENT (#5)
Japanese-American WWII Camps

How Each Project Is Related to Computational Themes:

Project	Computational Linguistics	Data Modeling & Evolutionary Prototyping	Graph Analytics	Crowdsourcing	GIS
1. Human Face of Big Data [Community Displacement]		X		X	X
2. Mapping Inequality [Racial Zoning]	X			X	X
3. St. Louis Voyage [Refugee Narratives]		X	X		X
4. World War II Japanese Camps [Citizen Internment]	X	X	X	X	X

The DCIC is Pursuing a CAS Training / Teaching Agenda

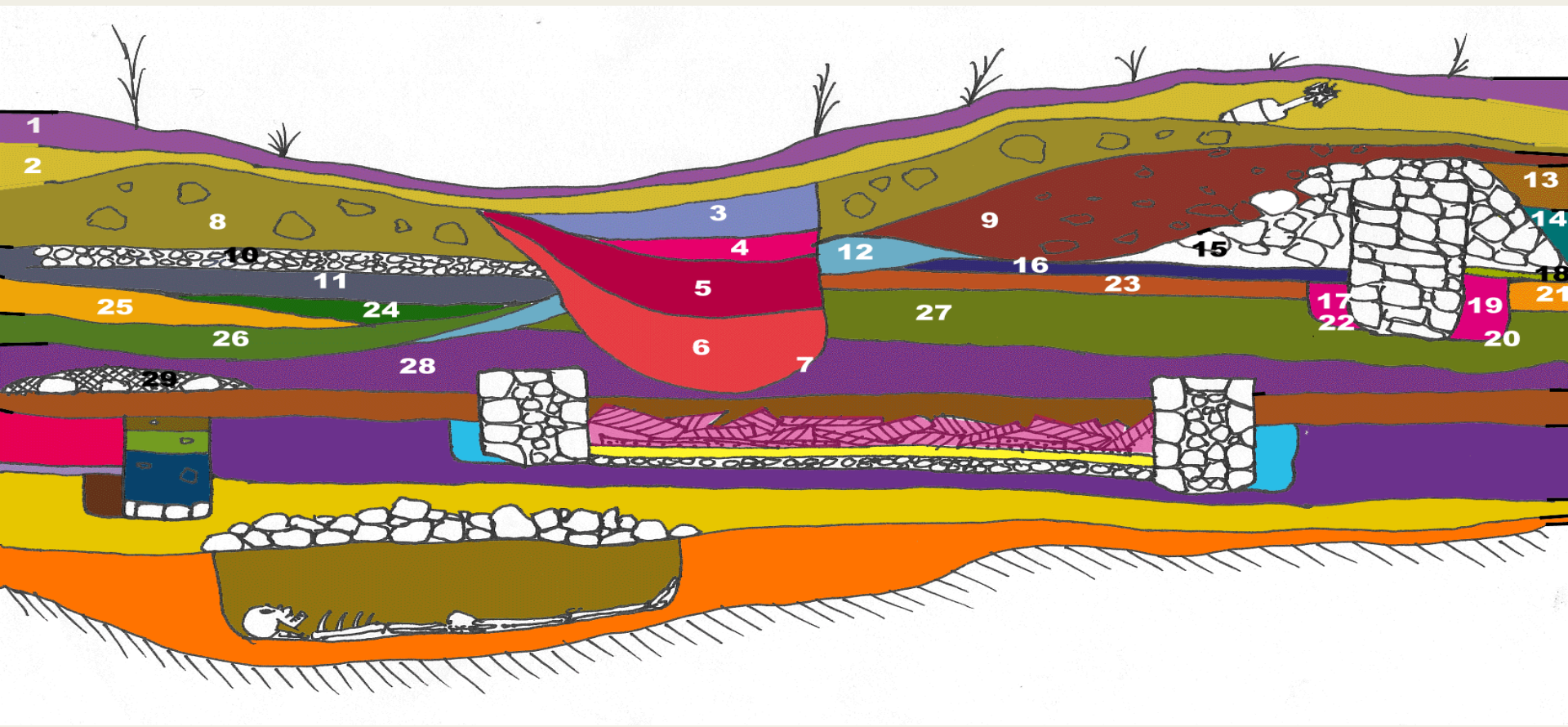
There is a need to :

- 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 (Jeannette Wing, 2006 – Comm. of the ACM)
- 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.
- COLLABORATE! https://datachallenge.ischool.umd.edu/ Feb. 24 – Mar. 3, 2018

CONTACTS

marciano@umd.edu

http://dcicblog.umd.edu/cas/
computational-archival-science@googlegroups.com



Cyberinfrastructure for digital records: a collection architecture for archaeology data

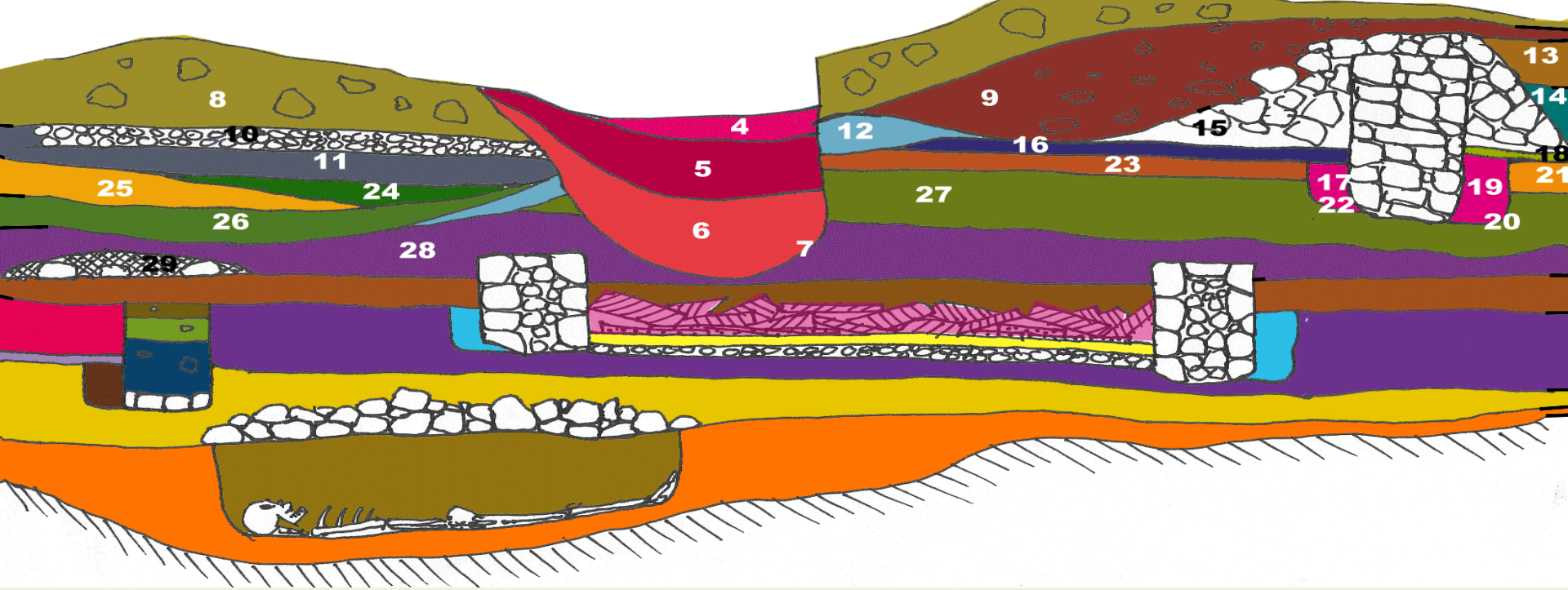
Maria Esteva, Texas Advanced Computing Center, University of Texas at Austin

Fukuoka, Japan, 2018

Drawing courtesy of Adam T. Rabinowitz

In the process of discovery, the archaeological site is destroyed

Images courtesy of the Institute of Classical Archaeology, University of Texas at Austin

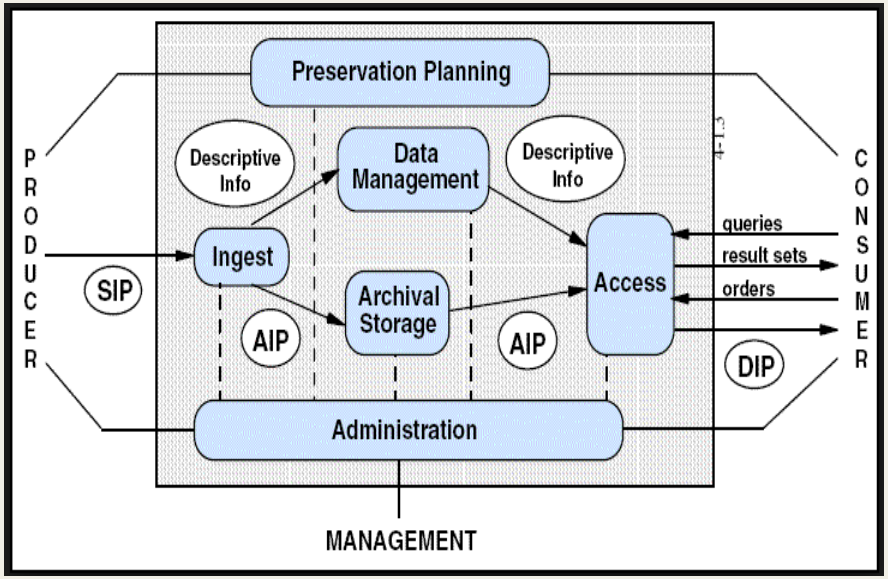


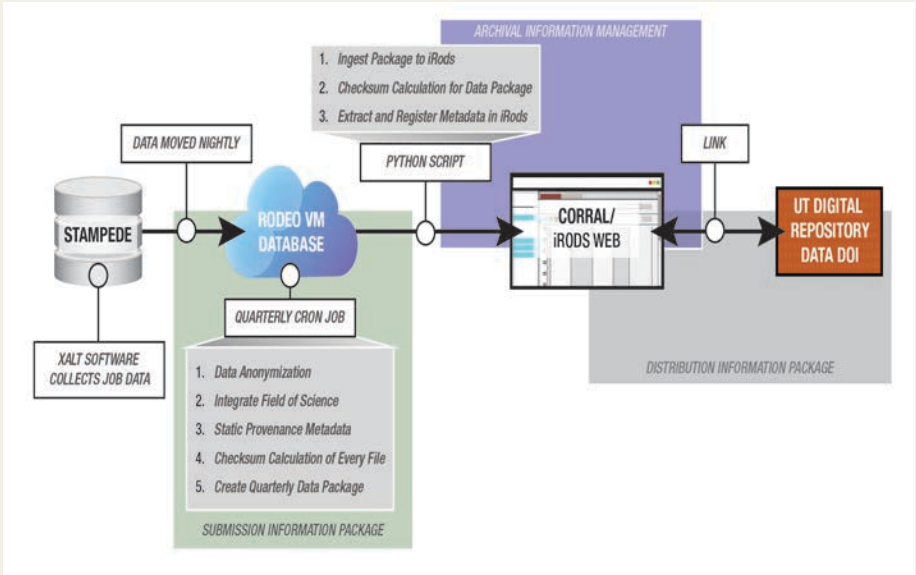
- Archaeological research is increasingly digital
  - Involves multiple systems and data types
- Documentation traces the materials from the site to store-room and museum exhibition
  - Preserve relationships between primary data, process history and the researcher's interpretation
- Sole evidence of destructed excavation
  - Data and system's preservation is fundamental

Drawings courtesy of Adam T. Rabinowitz

Collection's architecture

- Diverse computational resources as building blocks
- Enables active and archival data functions
  - Appraisal, arrangement, description, study, archiving, preservation, publication, and access
  - Research data lifecycle
- Processes can be automated
- Efficiency, security, scalability

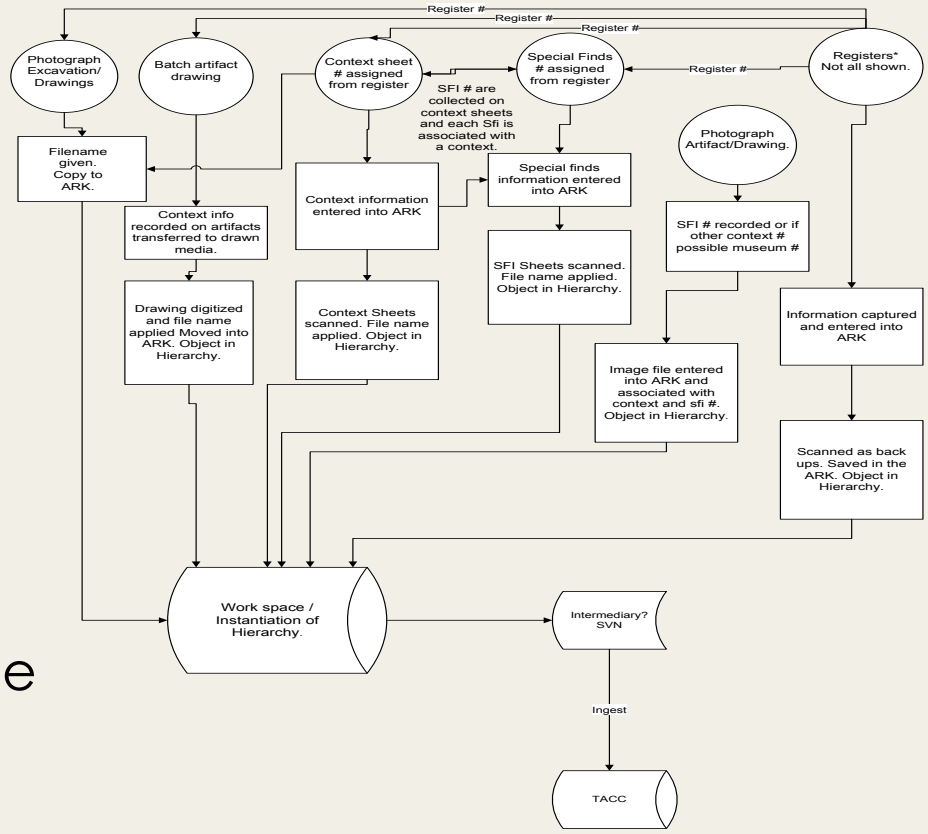




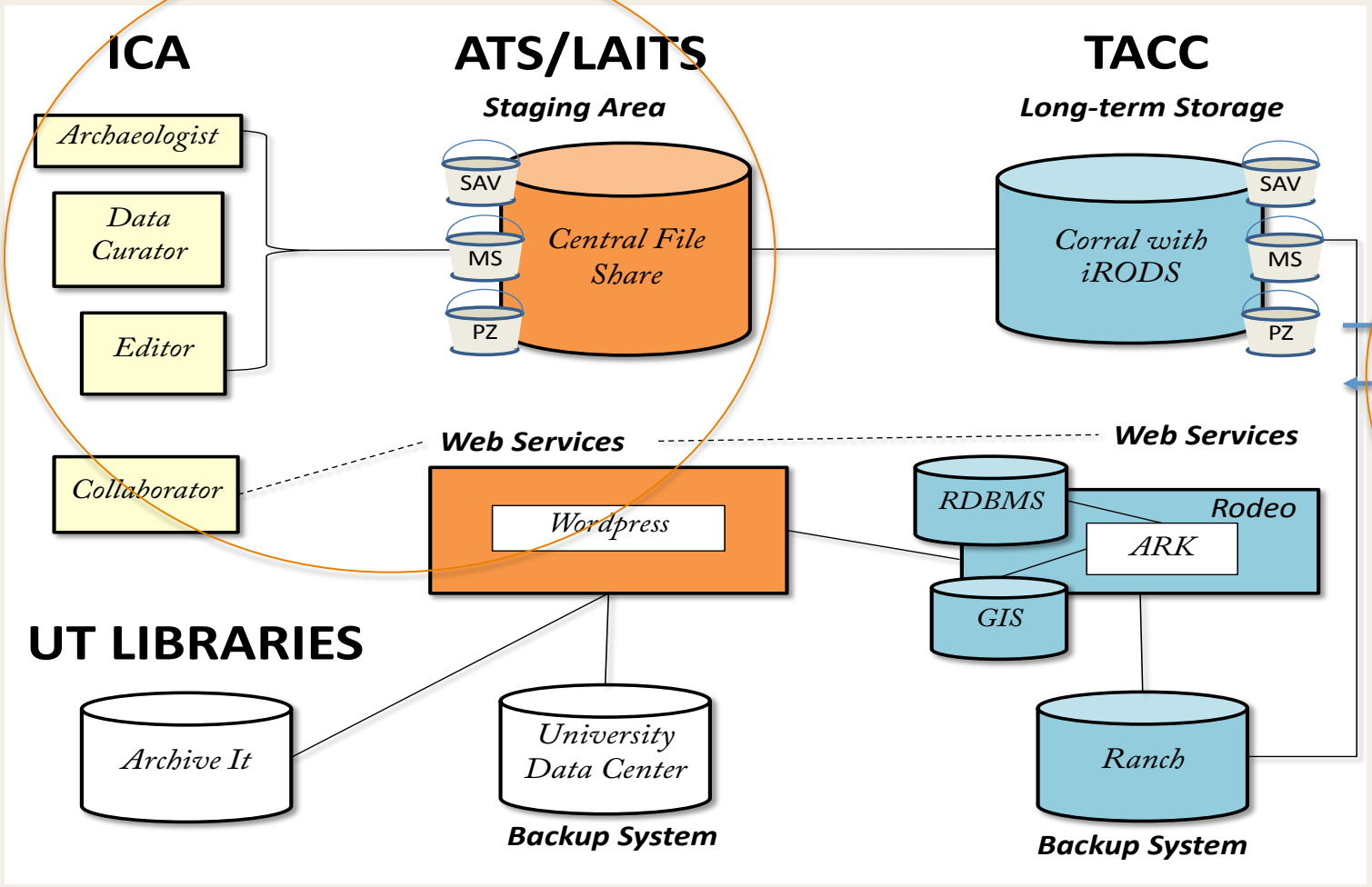
Functions of the archaelogical collection's architecture

- Appraisal and assessment
- Arrangement and description
- Research/study workflows
- Use of powerful computers to improve processes
- Digital preservation
  - while it is more vulnerable (active) and when it is archived

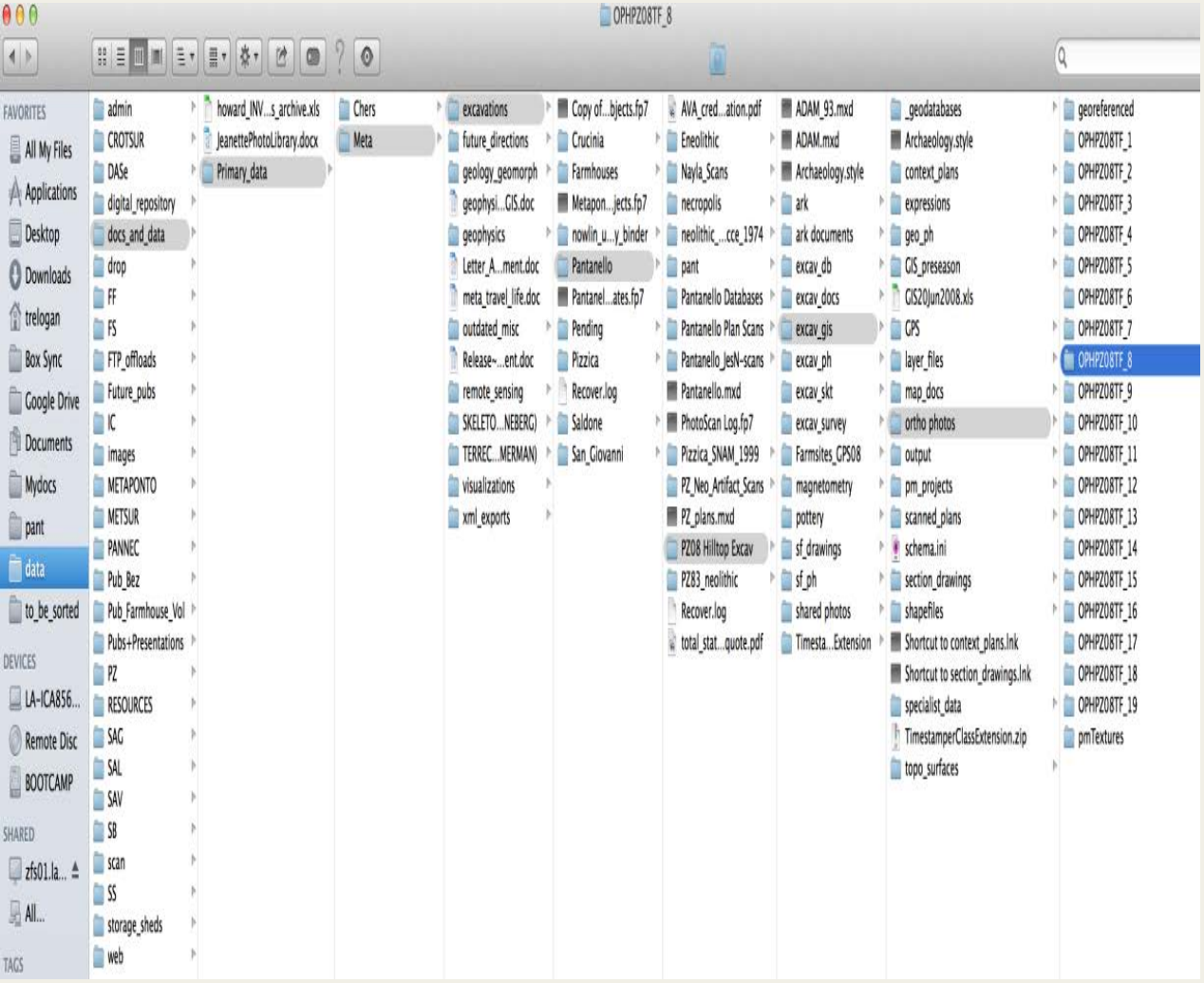
First step is to understand the data processes, record types and research requirements



Collection's (distributed) architecture

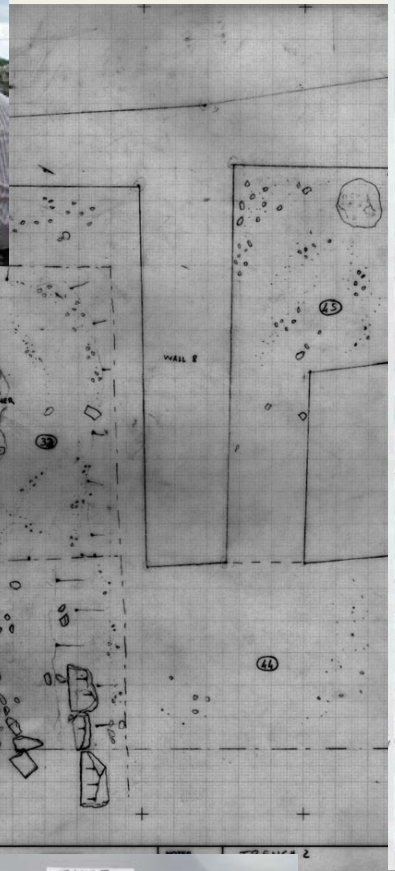


Directory view of the Institute of Classical Archaeology collection



- The collection is 4.3 TB and ~1 million files in deeply nested directories
- 40 + years of excavations including digitized and born-digital data
- Hundreds of researchers across the world
- In a shared drive with no metadata, just based on memory of the staff
- Nobody could find anything

Diverse data and record types



SITE	YEAR	ROOM # or AREA NAME	CONTEXT NUMBER
PZ08TF	2008	Room #1	31

TYPE (positive, negative, wall)	DEFINITION (cut in floor of X, fill of pit Y, north wall of room Z, etc.)
Pos	Possible floor level in Room #1

DESCRIPTION (include consistency, color, composition, inclusions, dimensions; for cuts, also include shape and orientation)
The surface is very hard and made up of white stones with some type of mortar(?)-like, chalky substance that further makes it hard. While excavating it, it left a powdery residue in the ~~south~~ north baulk. It sat on top of 46, large river cobbles that may have been used as a packing surface. It is only partially preserved and there were two sections of it, one join the north baulk and one area slightly to the south.

MUNSELL COLOR	2.5Y 4/4	Olive Brown

FORMATION PROCESSES NATURAL ARTIFICIAL (circle one) CASUAL INTENTIONAL (circle one) TYPE: Floor

STATE OF PRESERVATION	STRATIGRAPHIC RELIABILITY
☑ Very poor ☐ Poor ☐ Medium ☐ Good ☐ Very Good	☐ Very poor ☑ Poor ☐ Medium ☐ Good ☐ Very Go
COVERS 46	COVERED BY 32 (partially)
CUTS —	CUT BY —
FILLS —	FILLED BY —
ABUTS —	ABUTTED BY —
EQUALS —	BONDS WITH —

INTERPRETATION
A floor level placed on a packing surface (46) which was very poorly preserved.

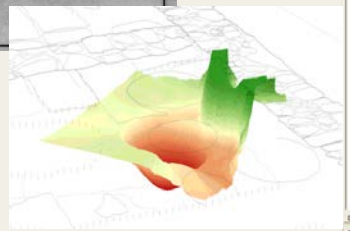
MATERIAL DATABLE IN FIELD? (if yes, specify) | PRELIMINARY DATE

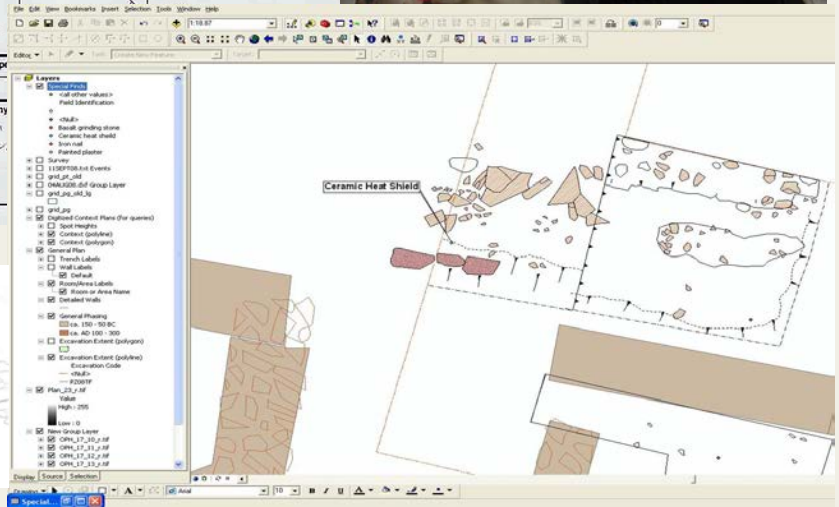
FINDS (tick) ☐ Pot ☐ Bone ☐ Glass ☐ Metal ☐ Wood ☐ Leather ☐ Tile ☐ Other ☐ None | TILE: (quantity buckets?) 1/4 bucket | SPECIAL FINDS (number and description, continue on back if necessa

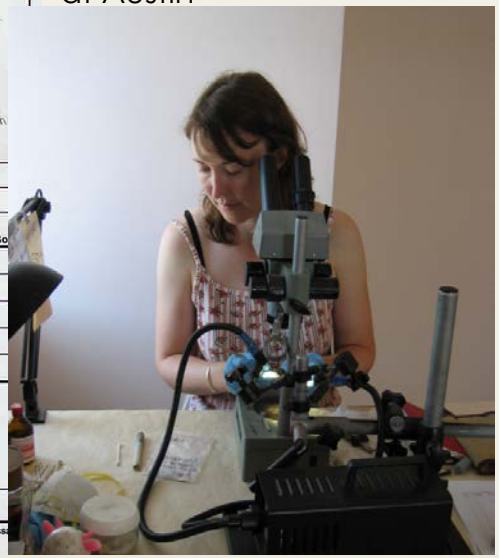
FLOAT? (Y/N) Y | SAMPLE? (if yes, # and typ

PHOTO? (Y/N) Y | PLAN ? (Y/N) Y

STRATIGRAPHIC NOTES/DOUBTS (include any that seems relevant) When we began exposed. This may actu

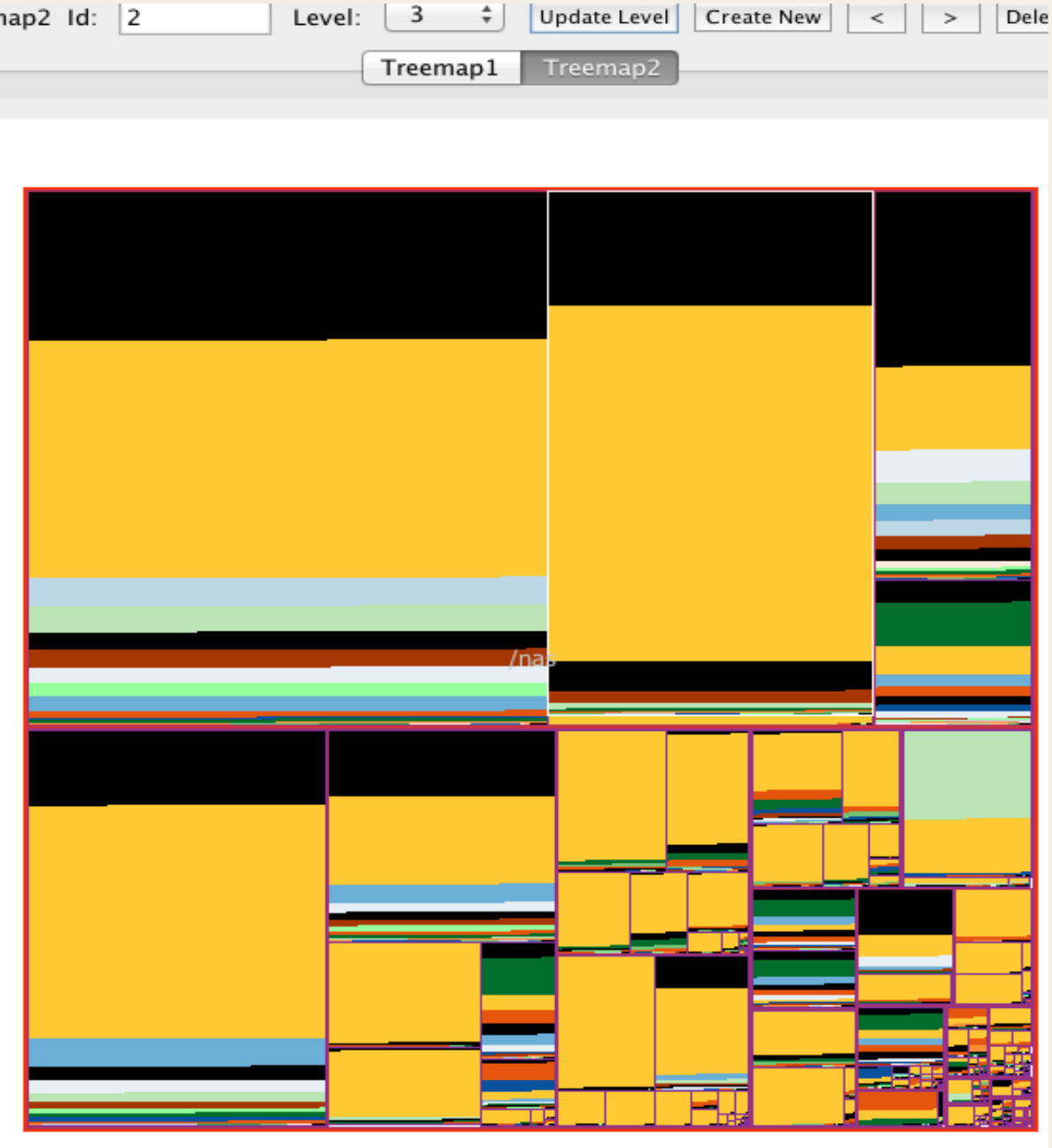






Images courtesy of the Institute of Classical Archaeology, University of Texas at Austin

Appraisal via visual representation

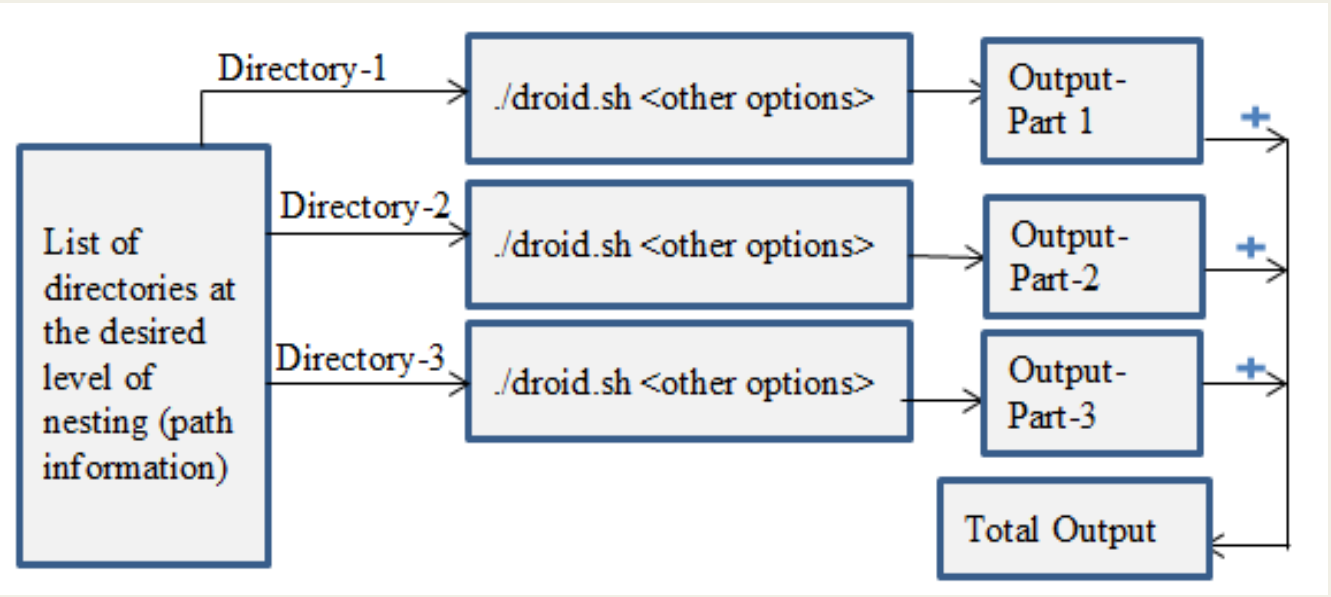


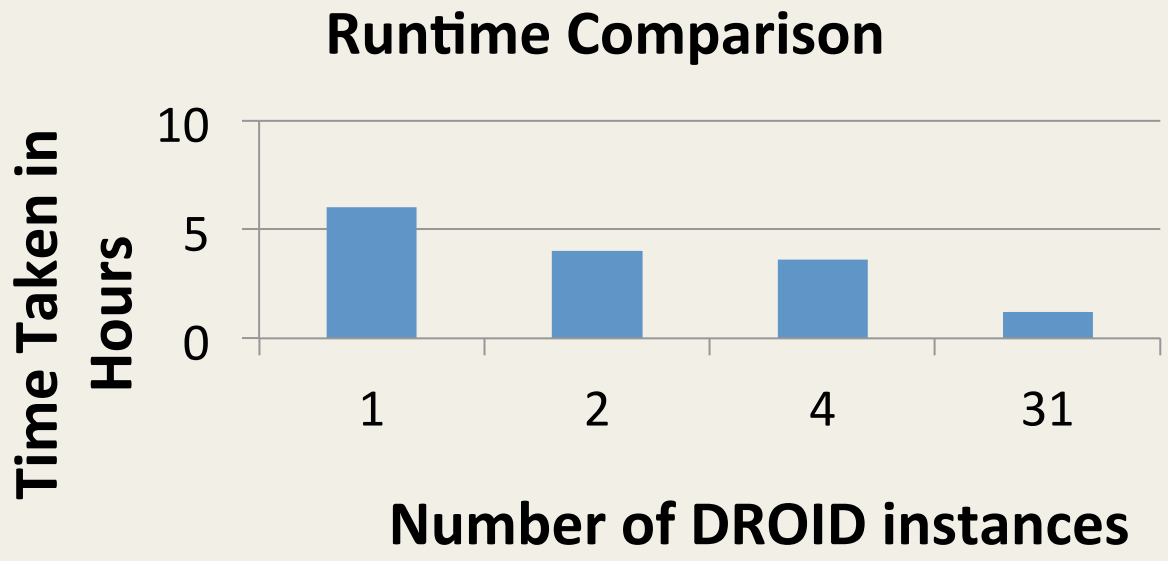
- Developed with metadata extracted using DROID
- Allows understanding file types, size and content of directories,
- Research stages are represented by colors
  - pdfs are publications, gis files are research data, images are raw/ site data, etc.

Parallel computing

- One job, several processors: a *parallel* solution
- Parallel programming works by breaking problems into pieces, and computing those pieces at the same time

Running DROID in parallel





We ran 31 parallel instances of DROID after coarsely dividing the workload of file-profiling

dump in Buckets on fileshare

copy collection to Stampede

unrecognized

rename, file

useful and recognized

analyze

run DROID, extract metadata

primary data

secondary data

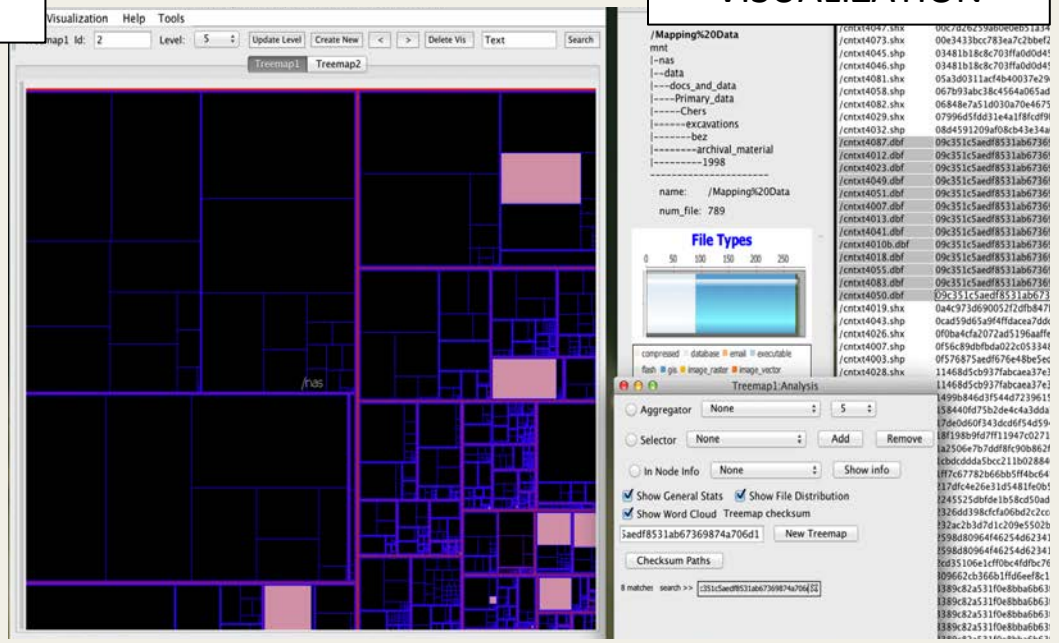
deposit in fileshare/field

deposit in fileshare/ specialist_study

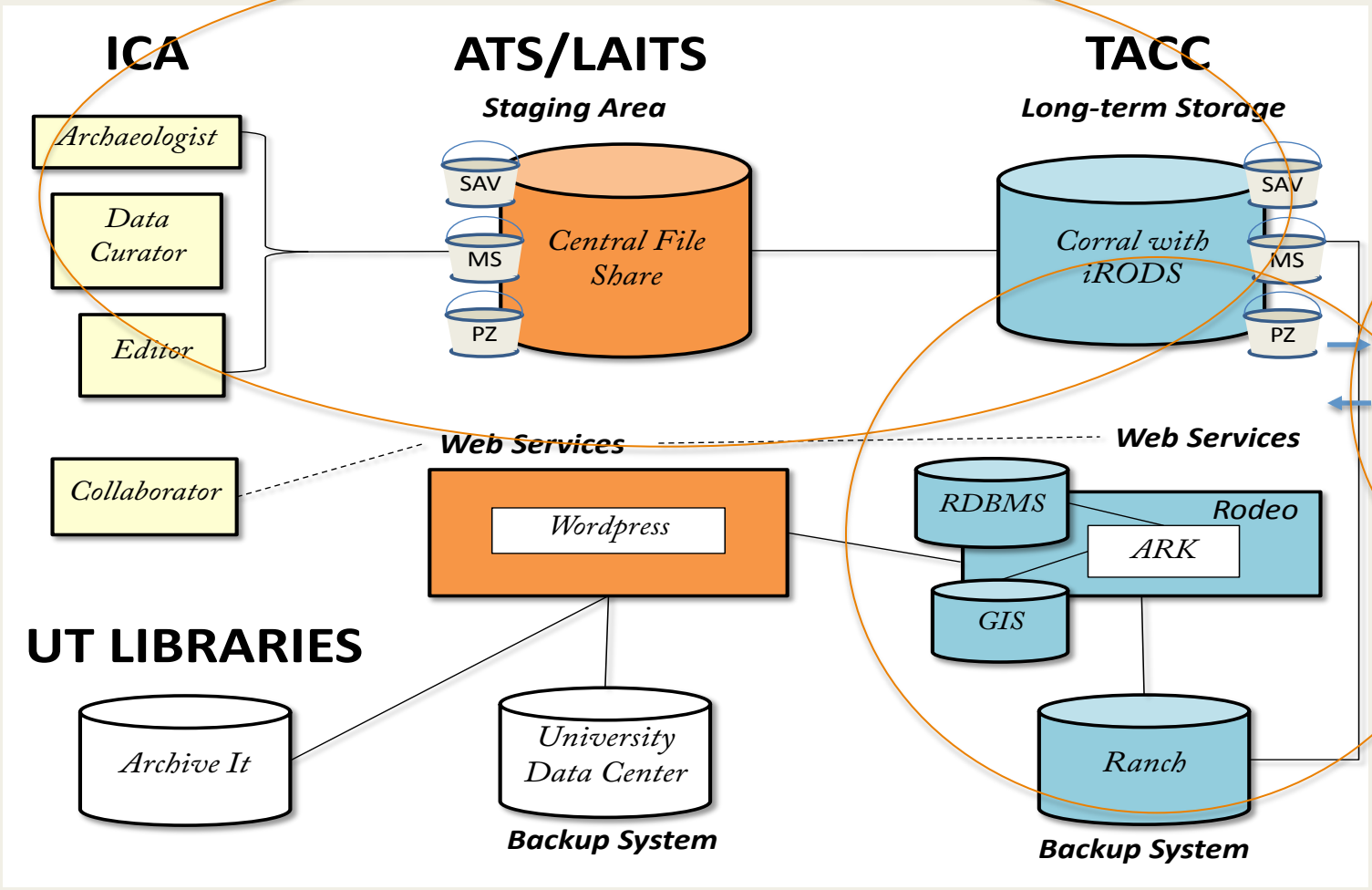
metadata into VISUALIZATION

static/stable

iRODS



Collection's (distributed) architecture



Archaeological collection

Material Studies

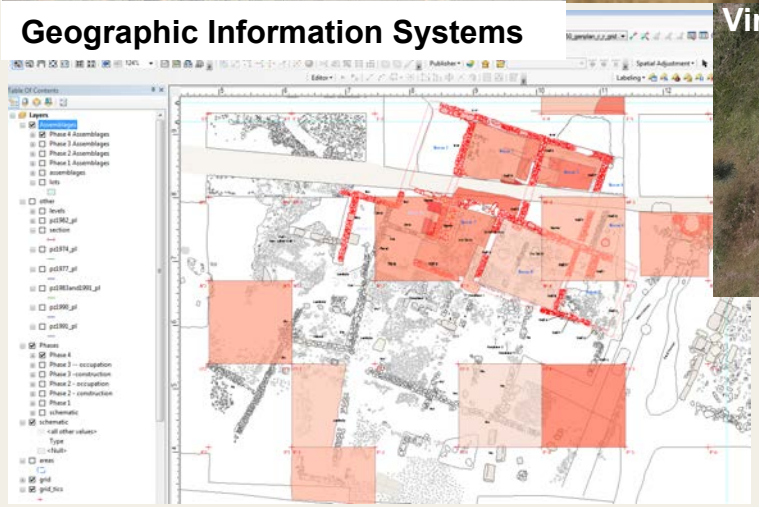
Surface Survey

Excavation

Geomorphology

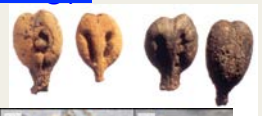
Physical Anthropology

Geographic Information Systems

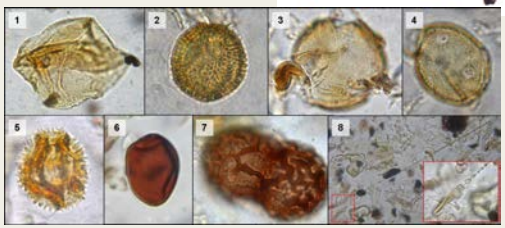


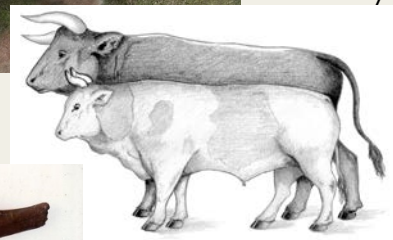
Virtual Reconstruction

Images courtesy of the [Institute of Classical Archaeology,](#) University of Texas at Austin



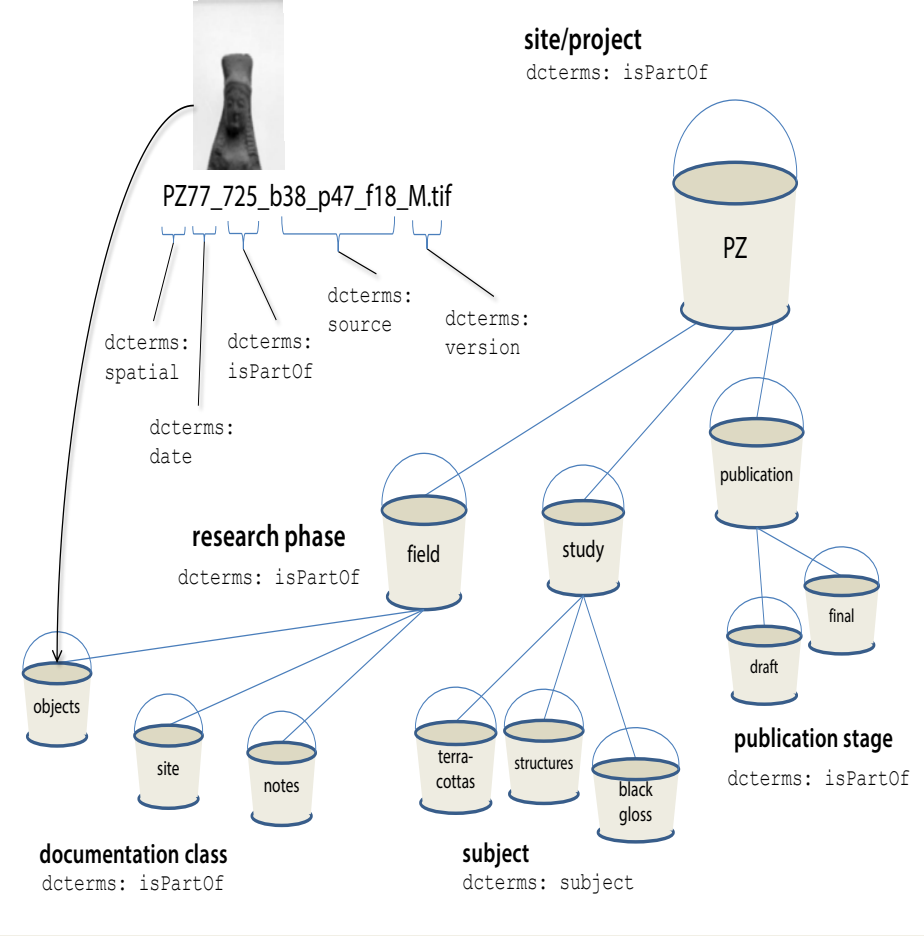
Paleobotany





Archaeozoology

Arrangement



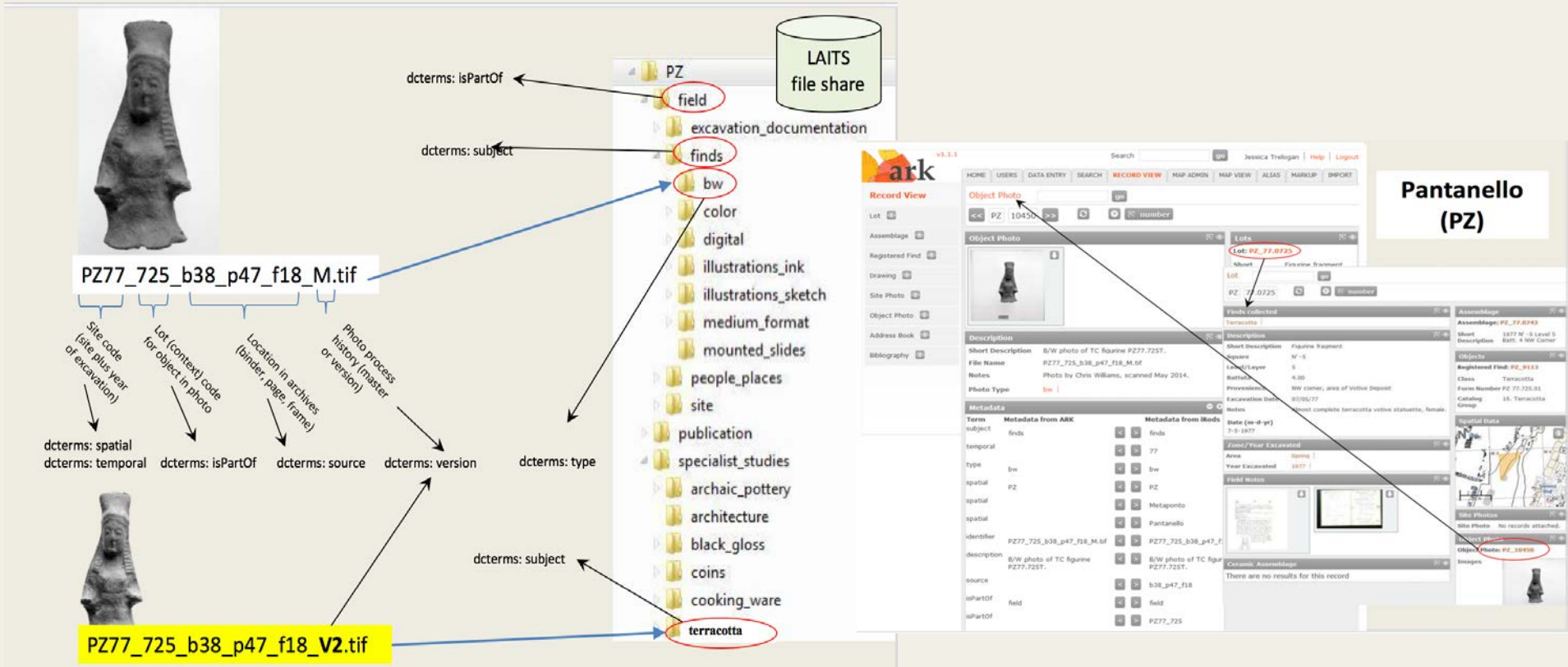
- Objects get arranged in:
  - Big Buckets according to functions
    - records management
- Buckets' labels are mapped to descriptive metadata
- Metadata gets formed automatically at data ingest to iRODS

Metadata integration

Record Keeping Example	ARK term	ARK value	DC Term	DC Value
PZ/study/terracotta	Find Type		subject	terracotta
	Short Description	Archaic female figurine	description	Archaic female figurine
PZ77_725T_b38_p47_f18_M.tif	File Name		identifier	PZ77_725T_b38_p47_f18_M.tif
PZ/field/finds/bw	Photo Type		format	BW
1977	Date Excavated		date	1977
	Date Photographed	1978	created	1978
	Photographed by	Chris Williams	creator	Chris Williams
Pantanello (PZ)	Site		spatial	Pantanello (PZ)
	Zone	Sanctuary	spatial	Sanctuary

- Metadata preserves relations between data object, site and location where it was found, changes and versions

Description and interoperability

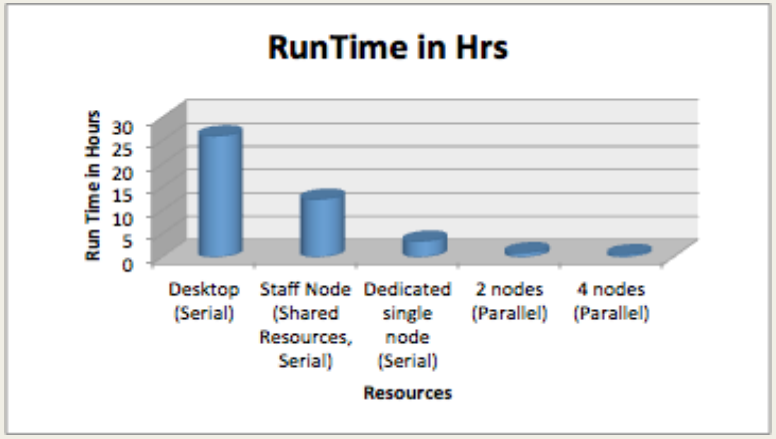


- Metadata is ingested into database Archaeological Recording Toolkit (ARK)
- Allows researchers to access already described data for study purposes

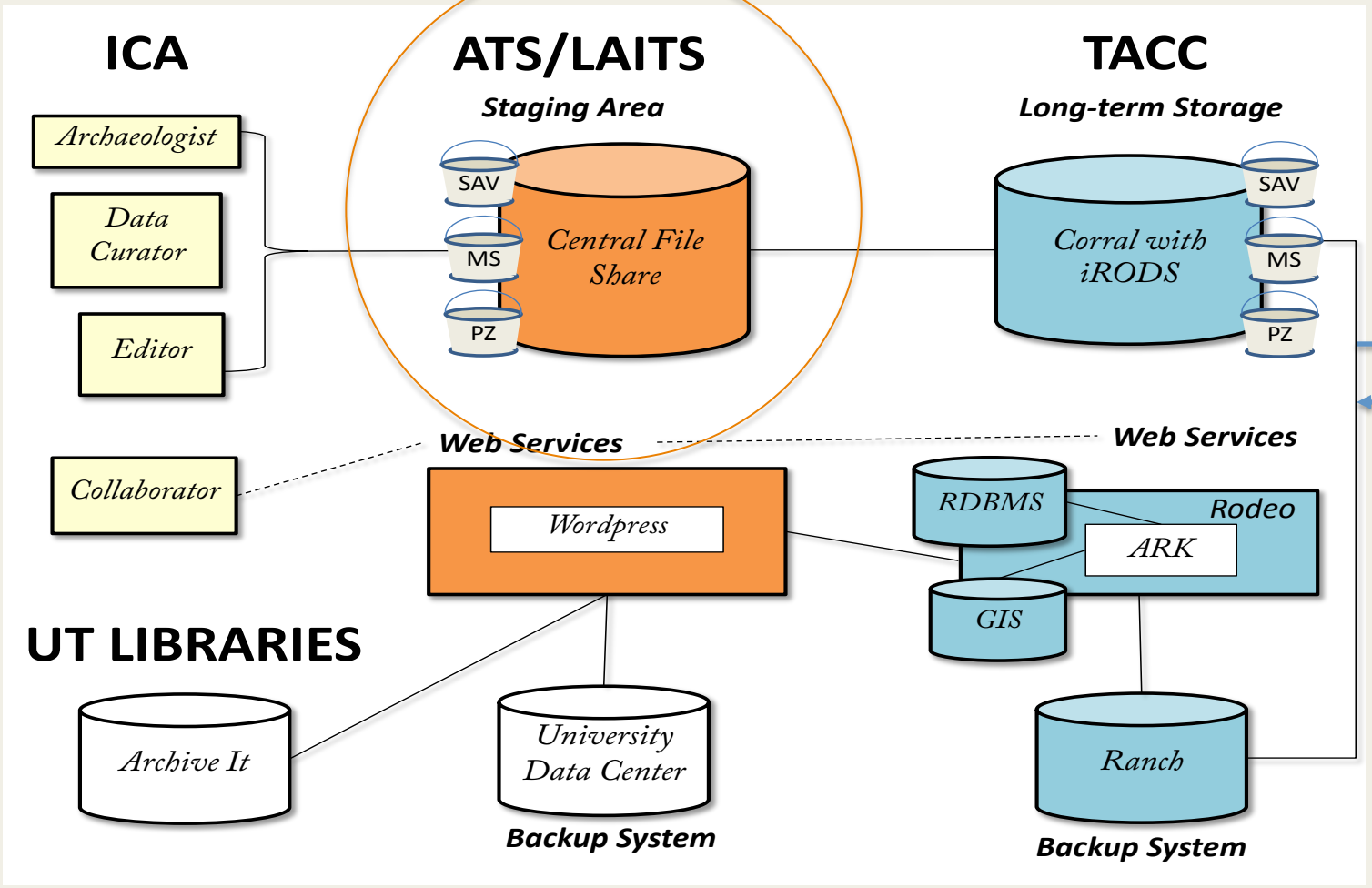
Arrangement in parallel

Site Values:	Parent Directory	Subdirectory
Bezymyannaya	DASe\Ukraine\Chersonesos\BEZ	then according to type
Capo Alfiere	DASe\Italy\Croton\CA\field	then according to type
Capo Cimiti	DASe\Italy\Croton\CROTSUR\field	then according to type
Capo Colonna	DASe\Italy\Croton\CROTSUR\field	then ac
Chersonesos	DASe\Ukraine\Chersonesos\ChersonesosMisc	then ac
Chersonesos Ancient City	DASe\Ukraine\Chersonesos\AncientCity	then ac
Chersonesos Chora	DASe\Ukraine\Chersonesos\Chora	then ac
Chersonesos Museum	DASe\Ukraine\Chersonesos\Museum	then ac
Croton	DASe\Italy\Croton\CROTSUR\field	then according to type
Fattoria Fabrizio	DASe\Italy\Metaponto\FF\field	then according to type
Fattoria Stefan	DASe\Italy\Metaponto\FS\field	then according to type
Fondi	DASe\Ukraine\Chersonesos\Museum	
Fusco Necropolis	skip	
Herakleia	skip	
Herakleian Peninsula	skip	then according to type
Incoronata	DASe\Italy\Metaponto\IC\field	then according to type
Incoronata Greca	DASe\Italy\Metaponto\IC\field	then according to type

TYPE Values:	Parent Directory	Subdirectory
Chart/Data Plot	charts	mounted_slides
Map/Plan/Section	maps	mounted_slides
Object	finds	mounted_slides
People/Places Photo	people_places	mounted_slides
Reconstruction	drawings	mounted_slides
Site Photo	site	mounted_slides

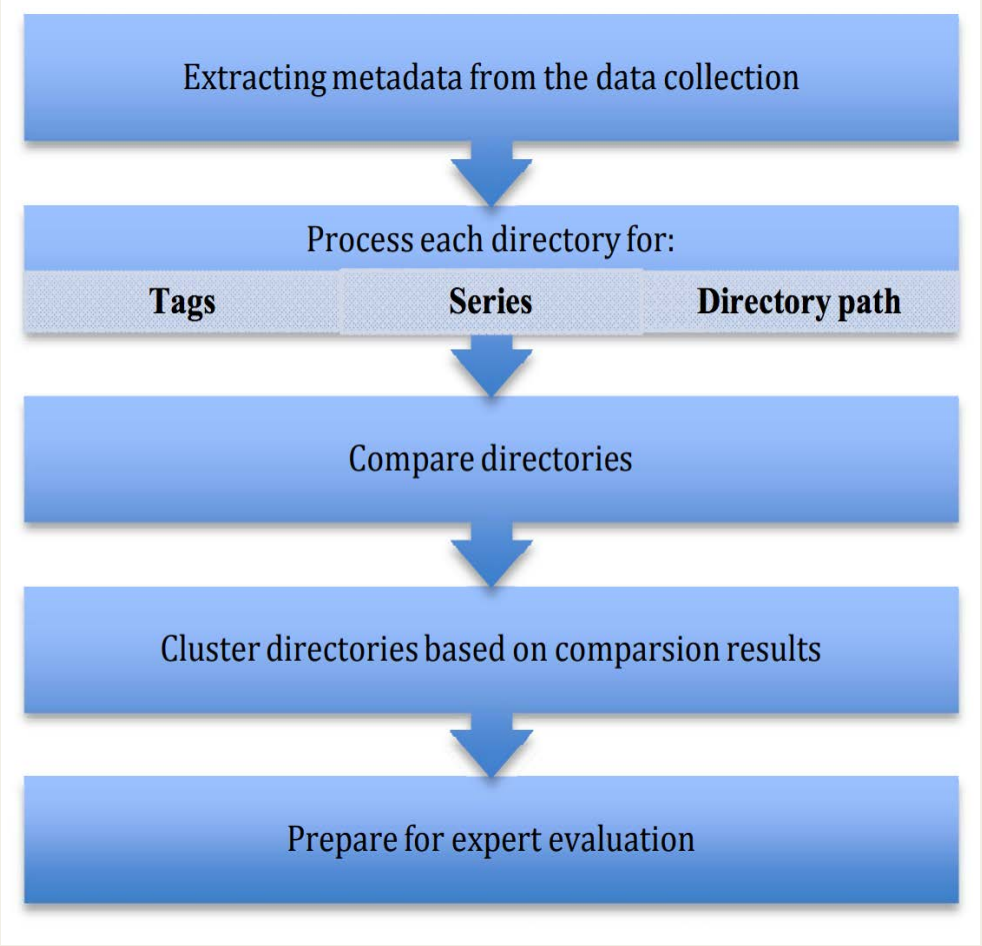


Collection's (distributed) architecture

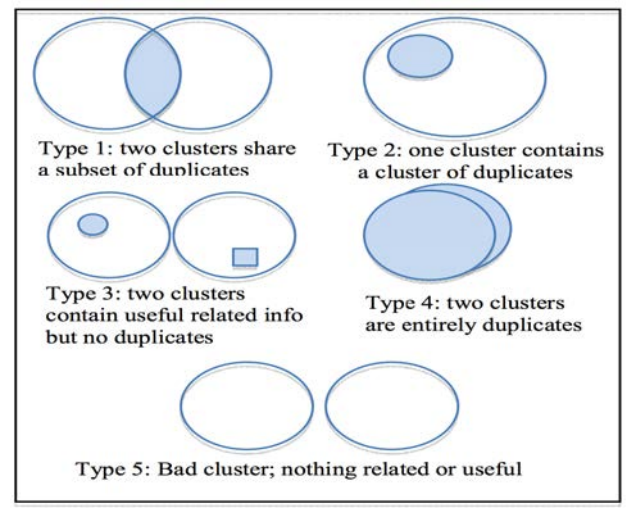


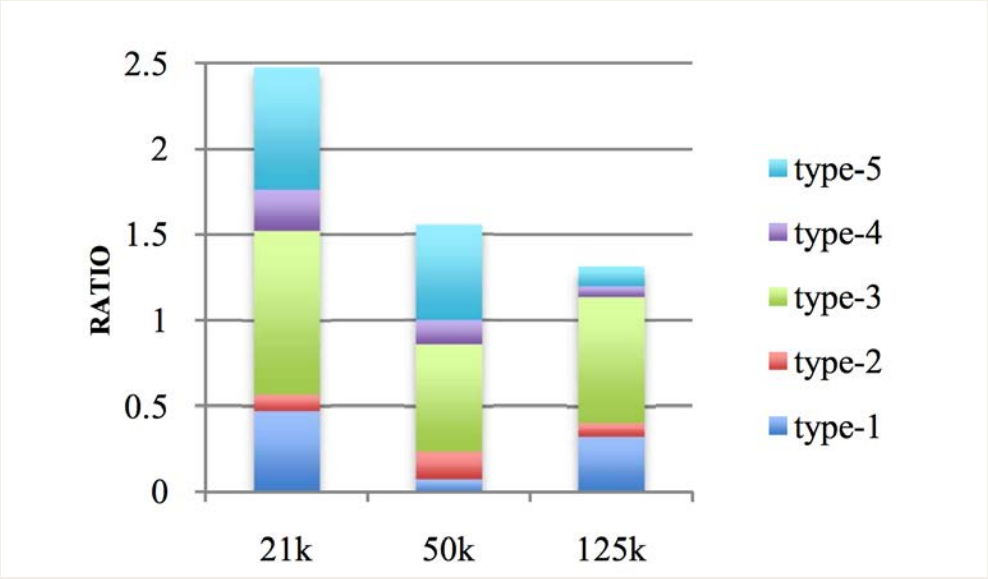
Entity resolution

- Classic Artificial Intelligence problem
- Find duplicate and redundant data
- Available data pointers are processed in a computational workflow
- Requires expert evaluation of results



Clustering analysis

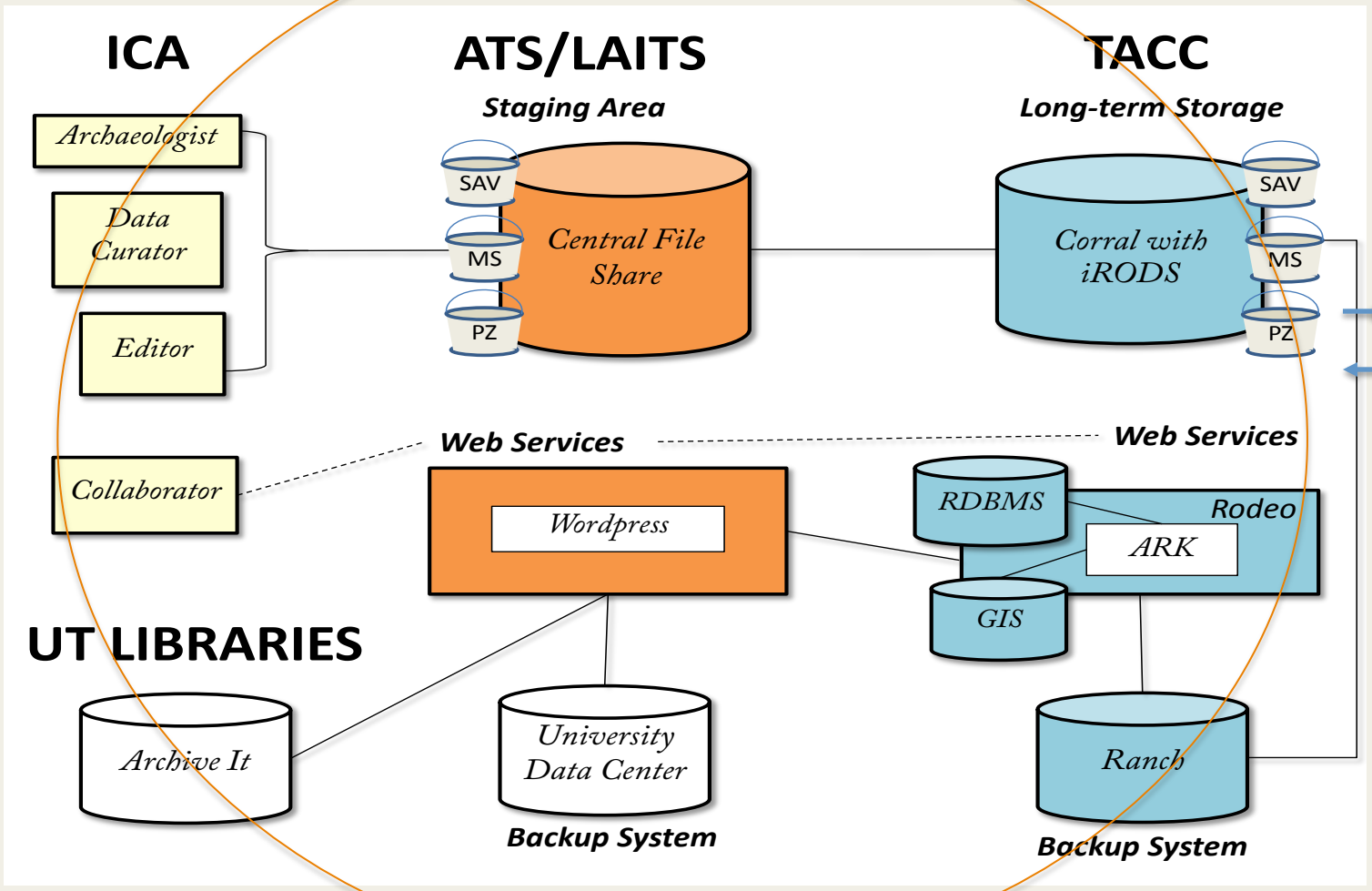




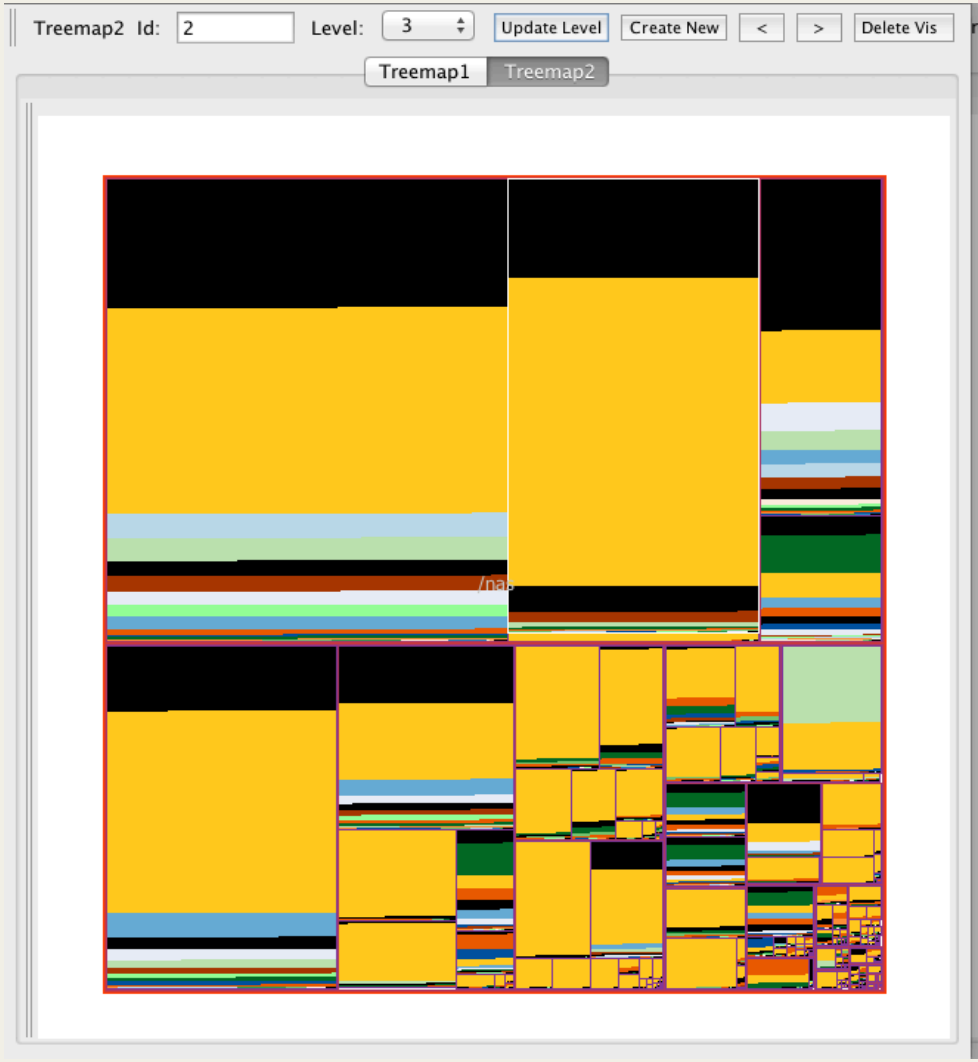
- /Ukraine/shared_photos/2006/CathyD/bones/134.jpg
  - Series
  - Tags
- Quality/expert assessment of clustering results:

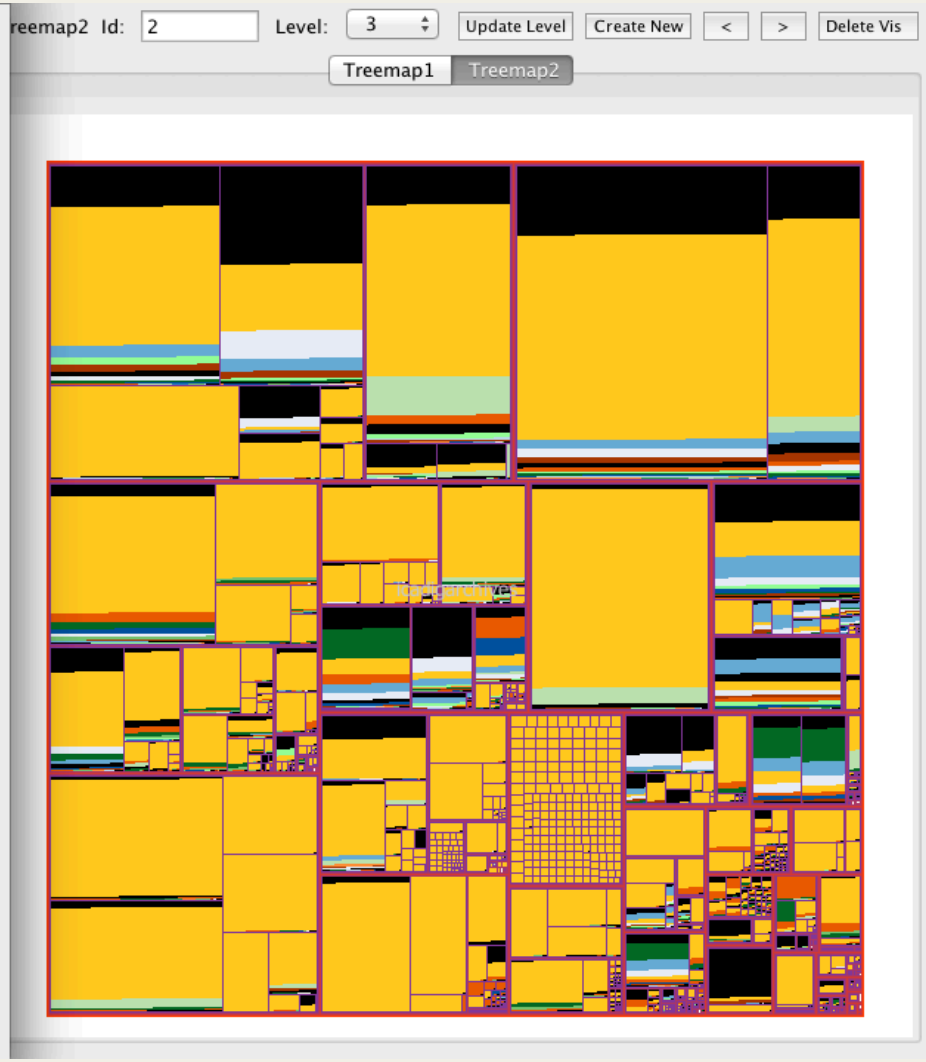
  Do they allow making decisions about what to keep and what to discard?
  - Still a difficult problem to solve

Collection's (distributed) architecture



Collection before and after arrangement





Multiple preservation strategies

- Data on high performance storage resource
  - Managed by iRODS
  - Authenticity/integrity
- Technical metadata extraction
  - Mapped to PREMIS
  - Stored in iRODS with the data object
- Geographical replication
  - On disk in Arlington TX
  - On tape
- ARK Database
  - MySql dumps
  - Entire database snapshots
  - Cascading on tape
- WordPress site
  - ArchiveIT

Conclusions

- Data collections are safe
- Transitions from active to archival
- Workflows get modified
- Archive “on the go”
- Interdisciplinary work
- Architecture and archival functions are seamlessly integrated
- Involves archivists as designers and evaluators of systems and processes

Bibliography

- Sivakumar Kulasekaran, Jessica Trelogan, Maria Esteva, Michael Johnson, (2014). Metadata Integration for an Archaeology Collection Architecture. Proceedings of the DCMI International Conference on Dublin Core and Metadata Applications DC-2014, Austin Texas, USA. Available at: http://dcevents.dublincore.org/IntConf/dc-2014/paper/view/242
- Xu, Weijia, Esteva Maria, Trelogan Jessica, Swinson, Todd. A Case Study on Entity Resolution for Distant Processing of Big Humanities Data. Proceedings of the Big Data, 2013 IEEE International Conference, pp. 113-120 doi: 10.1109/BigData.2013.6691678
- Maria Esteva, Jessica Trelogan, Weijia Xu, Andrew Solis, (2013). Lost in the Data, Aerial Views of an Archaeological Collection. Proceedings of the 2013 Digital Humanities Conference. University of Nebraska, Lincoln. June 16-19. Available at: http://dh2013.unl.edu/abstracts/author.html?q=author:%22esteva%2C%20maria||Esteva%2C%20Maria%22
- Maria Esteva, Jessica Trelogan, Adam Rabinowitz, David Walling, Stephen Pipkin. (2010). From the Site to Long-Term Preservation: A Reflexive System to Manage and Archive Digital Archaeological Data. Proceedings of the IS&T's Archiving 2010 Conference, June 1-4, 2010, Den Haag, The Netherlands. Available at: http://ist.publisher.ingentaconnect.com/contentone/ist/ac/2010/00002010/00000001/art00001
- David Walling, Maria Esteva (2011). Automating the Extraction of Metadata from Archaeological Data Using iRods Rules. International Journal of Digital Curation, Vol 6. No 2. Available at: http://www.ijdc.net/index.php/ijdc/article/view/196
- Adam Rabinowitz, Maria Esteva, Jessica Trelogan, (2012). Ensuring a future for the past: long-term preservation strategies for digital archaeological data. In Proceedings of the UNESCO Conference Memory of the World in the Digital Age. Vancouver, Canada, September 26-28. Available at: http://www.unesco.org/new/en/communication-and-information/events/calendar-of-events/events-websites/the-memory-of-the-world-in-the-digital-age-digitization-and-preservation/
- Arora Ritu, Esteva Maria, Trelogan Jessica, (2014). Leveraging High Performance Computing for Managing Large and Evolving Data Collections. International Journal of Digital Curation. Vol. 9, No 2, p 17 – 27. http://dx.doi.org/10.2218/ijdc.v9i2.331
- Sivakumar Kulasekaran, Maria Esteva, Jessica Trelogan. Si Liu, (2015). A Framework for Multitasking Data-Intensive Management Services in High Performance Computing Environments. Proceedings of the IEE Big Data Service 2015. March 30 – April 2, 2015, San Francisco, USA. doi: 10.1109/BigDataService.2015.42
- Jessica Trelogan, Lauren Jackson, Maria Esteva, (2015). Unlocking the data behind the Chora of Metaponto publication series: "on-the-fly" solutions for sharing and archiving an evolving collection. Proceedings of the 80th Annual Meeting of the Society for American Archaeology 2015 Conference, San Francisco, April 15-19. doi:10.6067/XCV8RB75ZC
- § Jessica Trelogan, Maria Esteva, Lauren Jackson, (2015). Preserving an Evolving Collection: "On-The-Fly" solutions for the Chora of Metaponto Publication Series. Proceedings of iPRES 2015, Chapel Hill, USA, November 2 – 6. Available at http://phaidra.univie.ac.at/o:429572