

Building Open-Source Digital Curation Services & Repositories at Scale



Dr. Richard Marciano

Professor & Director of the Digital Curation Innovation Center (DCIC)

Greg Jansen / Will Thomas / Sohan Shah / Michael Kurtz

DCIC @ Maryland's iSchool

University of Maryland

Presented on Feb. 20, 2018 at IDCC18

Session on Repository Services

Digital Curation in the DCIC @ U. Maryland



1. Dvt. of Distributed Scalable NoSQL Catalogs and Repositories (DRAS-TIC)
2. Dvt. of Cloud-Based Digital Curation Services (Brown Dog)
3. Creation of a Testbed of Justice, Human Rights, and Cultural Heritage Collections
4. Devt. of a New Trans-Disciplinary Field: Computational Archival Science (CAS)
5. Integration of Digital Curation Education & Research

1. Development of Distributed Scalable NoSQL Catalogs and Repositories (DRAS-TIC)

NoSQL distributed DB technology to support repositories that can scale out horizontally to 1000s of commodity servers. See:

<http://dcicblog.umd.edu/dras-tic-fedora/>

Applies the new Fedora 5 API to the DCIC's open-source software stack called DRAS-TIC. See: <https://github.com/UMD-DRASTIC/drastic>

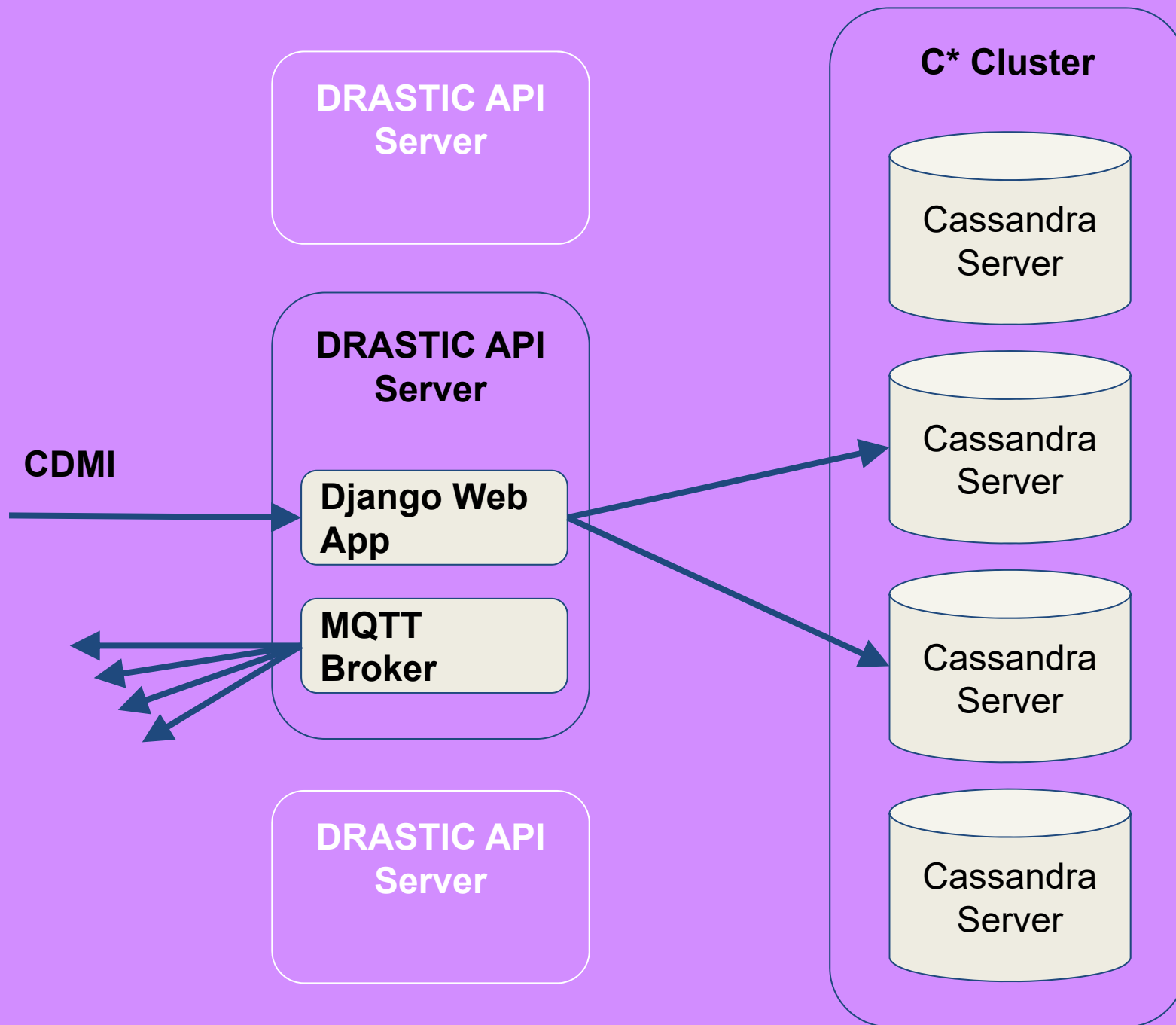
Partners include:

- (1) Fedora Leadership Group and Steering Committee
- (2) Smithsonian Institution (Office of Research Info. Services and National Museum of American History)
- (3) University of Illinois Urbana-Champaign National Center for Supercomputing Applications (NCSA)
- (4) University of Maryland Libraries
- (5) Georgetown University Library



Digital **R**epository **A**t **S**cale - **T**hat **I**nvites **C**omputation
- **T**o **I**mprove **C**ollections

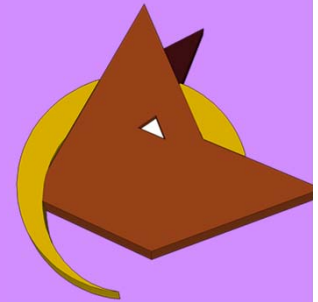
- Product of **2-year startup** by partners, Archival Analytics Ltd.
- **Horizontal scaling to billions of files** and beyond
- Web UI and command-line client
- **Industry standard REST storage** API (CDMI)
- **Key-value** metadata
- Eventing over MQTT message system
- Python source on GitHub (Open AGPL license)
- Based on the **NoSQL Apache Cassandra** (*1,800 companies: CERN, eBay, GitHub, Hulu, Instagram, Netflix, Twitter, and scales to petabytes of storage and billions of objects*)



2. Development of Cloud-Based Digital Curation Services (Brown Dog)

NSF “Brown Dog” Project, “The Super Mutt”

Public API for: (1) Format migration, and (2) Feature Extraction

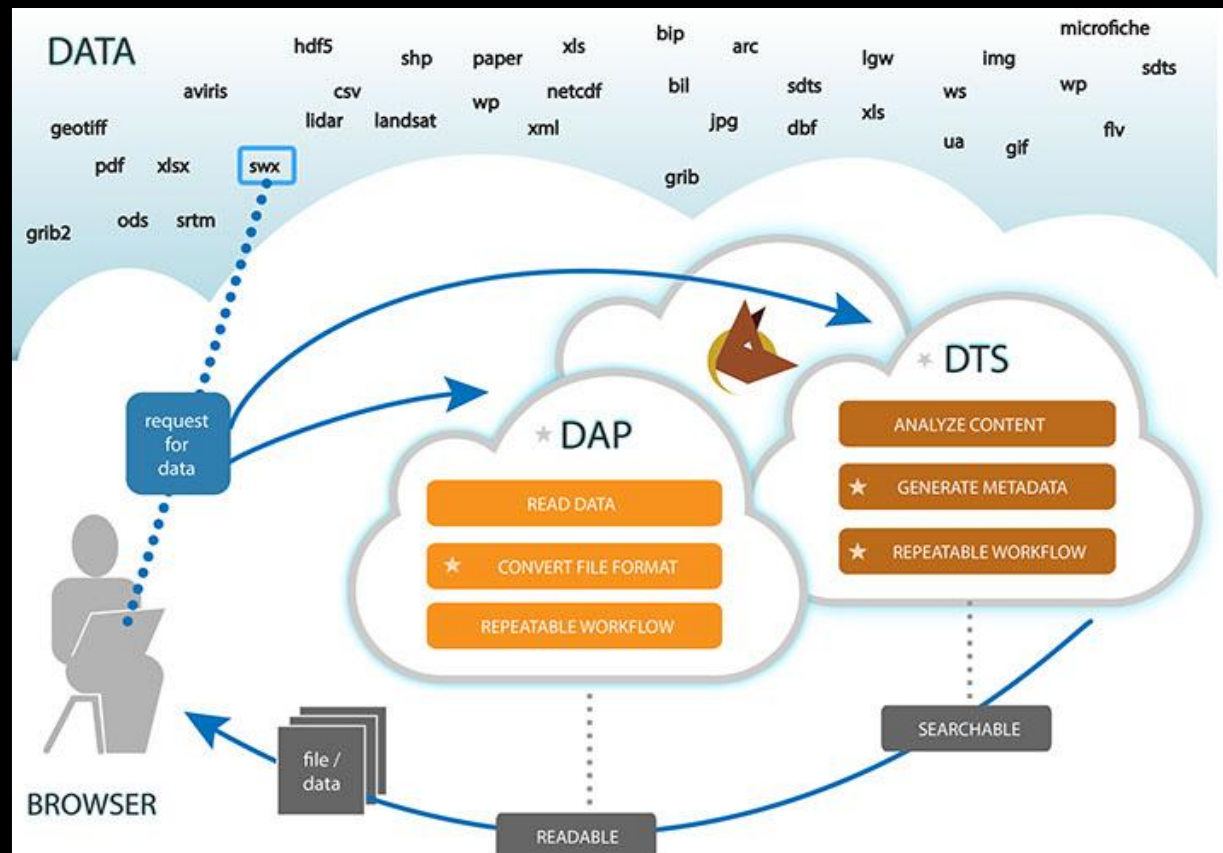


Brown Dog: <http://browndog.ncsa.illinois.edu/>

Web-scale server virtualization
Part of an NSF DiBBs-funded
project (\$10.5M grant).

This service provides web and
API access to 100s of tools,
and is deployed over
DRAS-TIC.

**CNI Fall 2016 (slides and audio) –
Computational Finding Aids:**
<https://www.cni.org/topics/digital-curation/drastic-measures-digital-repository-at-scale-that-invites-computation-to-improve-collections>



3. Creation of a Testbed of Justice, Human Rights, and Cultural Heritage Collections

→ Accelerate the development of Digital Curation processes and services through the creation of a data observatory

Justice, Human Rights, & Cultural Heritage:

THEME
Community Displacement
Racial Zoning:
Refugee Narratives:
Citizen Internment:
Movement of People:
Revealing Untold Stories:

PROJECTS
The Human Face of Big Data
Mapping Inequality
St. Louis Voyage
Japanese American WWII Camps
Overseas Pension Project
Legacy of Slavery

Data Observatory:

- 100TB of data
- 100M files

NSF/NARA:

- 100TB
- 100M files
- 6,000 file types
- 150 fed agencies

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

FUNDING
NSF
IMLS

PROJECTS
Brown Dog
DRAS-TIC Fedora

4. Devt. of a New Trans-Disciplinary Field: Computational Archival Science (CAS)

→ 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 ?

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%]



“Computational Archival Science (CAS)” Portal

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

Goal: Explore computational treatments of archival and cultural content



Pursuing big ideas
and new discoveries.

CAS Founding Partners:

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

Foundational Book Chapter (June 2018)

“Archival Records and Training in the Age of Big Data”

Book: “Advances in Librarianship – Re-Envisioning the MLIS: Perspectives on the Future of Library and Information Science Education”.

Google Group: computational-archival-science@googlegroups.com

- (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 2017: http://dcicblog.umd.edu/cas/ieee_big_data_2017_cas-workshop/

- → **#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, Worcester 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 1200 o'clock noon Tuesday, April 7, 1942.

No Japanese person will be permitted to enter or leave the above described area after 800 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.

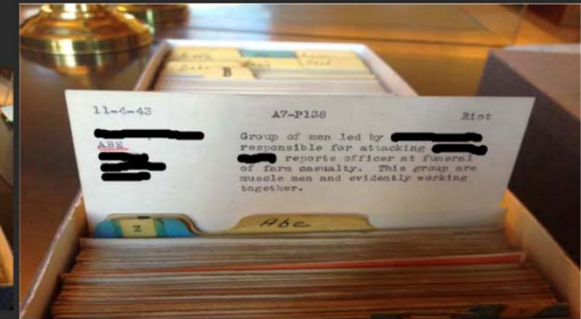
4. Each family, and individual living alone, will be furnished transportation to the Reception Center. Private means of transportation will not be utilized. All instructions pertaining to the movement will be obtained at the Civil Control Station.

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. I. DeWITT
Lieutenant General, U. S. Army
Commanding

SEE CIVILIAN EXCLUSION ORDER NO. 5

Incident Index Cards



List for a Sample Index Card

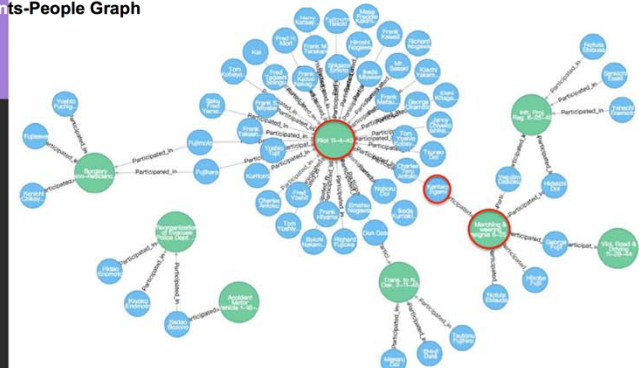
Annotation Sets Annotations List Annotations Stack Co-reference Editor Text					
11-4-43	A-999 P9	Riot			
Amane, Ohashi					
9999-D					
One of possible leaders or trouble makers.					
Type	Set	Start	End	Id Features	
Date		0	7	101	(kind=date, rule=DateNumDash, ruleFinal=DateOnlyFinal)
CaseReportId		17	22	106	(rule=CaseReportId_Page)
CaseReportPage		23	25	107	(rule=CaseReportId_Page)
Offense		40	44	108	(rule=Offenses)
Person		47	60	109	(rule=TupleReverseNamePersonNoSplit)
Residence		62	68	110	(rule=ResidenceAddress)

- ☒ CaseReportId
 - ☒ CaseReportPage
 - ☒ Date
 - ☐ Identifier
 - ☐ Lookup
 - ☐ Offense
 - ☒ Person
 - ☒ Residence
 - ☐ Sentence
 - ☐ SpaceToken
 - ☐ Split
 - ☐ Token
 - ☐ Unknown
- Original markings

GOAL:

Automating
the review
and release
of records
at scale

Events-People Graph



5. Integration of Digital Curation Education & Research



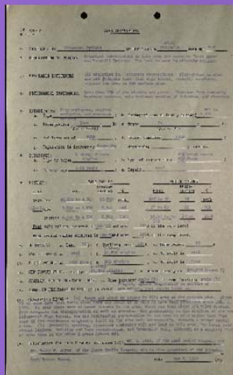
Key components of our initiative:

- Creating a new academic **Specialization, Archives & Digital Curation**, in the MLIS
- Organizing seminars for graduate students to define the theoretical and operational elements of **Computational Archival Science**
- Establishing a **Digital Curation for Information Professionals (DCIP) Certificate** program
- Offering students participation on **interdisciplinary digital curation projects**, at the intersection of archives, digital curation, Big Data, and analytics.
 - E.g. Maryland State Archives' **Legacy of Slavery** project

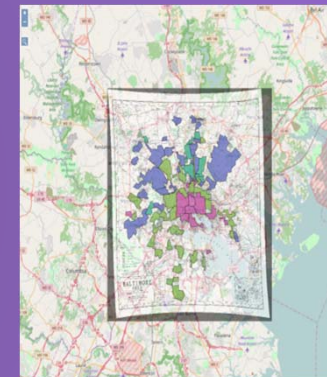
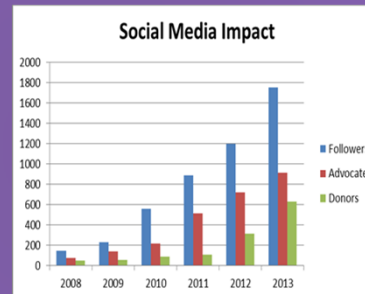
DCIC Digital Curation Projects...

“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]



DATA DESCRIPTION
Security Age of
Los Angeles County
1. POPULATION
a. Increasing
b. Decreasing
c. Stable
d. Unknown
e. Other
f. Other
g. Other
h. Other
i. Other
j. Other
k. Other
l. Other
m. Other
n. Other
o. Other
p. Other
q. Other
r. Other
s. Other
t. Other
u. Other
v. Other
w. Other
x. Other
y. Other
z. Other
aa. Other
ab. Other
ac. Other
ad. Other
ae. Other
af. Other
ag. Other
ah. Other
ai. Other
aj. Other
ak. Other
al. Other
am. Other
an. Other
ao. Other
ap. Other
aq. Other
ar. Other
as. Other
at. Other
au. Other
av. Other
aw. Other
ax. Other
ay. Other
az. Other
ba. Other
bb. Other
bc. Other
bd. Other
be. Other
bf. Other
bg. Other
bh. Other
bi. Other
bj. Other
bk. Other
bl. Other
bm. Other
bn. Other
bo. Other
bp. Other
bq. Other
br. Other
bs. Other
bt. Other
bu. Other
bv. Other
bw. Other
bx. Other
by. Other
bz. Other
ca. Other
cb. Other
cc. Other
cd. Other
ce. Other
cf. Other
cg. Other
ch. Other
ci. Other
cj. Other
ck. Other
cl. Other
cm. Other
cn. Other
co. Other
cp. Other
cq. Other
cr. Other
cs. Other
ct. Other
cu. Other
cv. Other
cw. Other
cx. Other
cy. Other
cz. Other
da. Other
db. Other
dc. Other
dd. Other
de. Other
df. Other
dg. Other
dh. Other
di. Other
dj. Other
dk. Other
dl. Other
dm. Other
dn. Other
do. Other
dp. Other
dq. Other
dr. Other
ds. Other
dt. Other
du. Other
dv. Other
dw. Other
dx. Other
dy. Other
dz. Other
ea. Other
eb. Other
ec. Other
ed. Other
ee. Other
ef. Other
eg. Other
eh. Other
ei. Other
ej. Other
ek. Other
el. Other
em. Other
en. Other
eo. Other
ep. Other
eq. Other
er. Other
es. Other
et. Other
eu. Other
ev. Other
ew. Other
ex. Other
ey. Other
ez. Other
fa. Other
fb. Other
fc. Other
fd. Other
fe. Other
ff. Other
fg. Other
fh. Other
fi. Other
fj. Other
fk. Other
fl. Other
fm. Other
fn. Other
fo. Other
fp. Other
fq. Other
fr. Other
fs. Other
ft. Other
fu. Other
fv. Other
fw. Other
fx. Other
fy. Other
fz. Other
ga. Other
gb. Other
gc. Other
gd. Other
ge. Other
gf. Other
gg. Other
gh. Other
gi. Other
gj. Other
gk. Other
gl. Other
gm. Other
gn. Other
go. Other
gp. Other
gq. Other
gr. Other
gs. Other
gt. Other
gu. Other
gv. Other
gw. Other
gx. Other
gy. Other
gz. Other
ha. Other
hb. Other
hc. Other
hd. Other
he. Other
hf. Other
hg. Other
hh. Other
hi. Other
hj. Other
hk. Other
hl. Other
hm. Other
hn. Other
ho. Other
hp. Other
hq. Other
hr. Other
hs. Other
ht. Other
hu. Other
hv. Other
hw. Other
hx. Other
hy. Other
hz. Other
ia. Other
ib. Other
ic. Other
id. Other
ie. Other
if. Other
ig. Other
ih. Other
ii. Other
ij. Other
ik. Other
il. Other
im. Other
in. Other
io. Other
ip. Other
iq. Other
ir. Other
is. Other
it. Other
iu. Other
iv. Other
iw. Other
ix. Other
iy. Other
iz. Other
ja. Other
jb. Other
jc. Other
jd. Other
je. Other
jf. Other
jg. Other
jh. Other
ji. Other
jj. Other
jk. Other
jl. Other
jm. Other
jn. Other
jo. Other
jp. Other
jq. Other
jr. Other
js. Other
jt. Other
ju. Other
jv. Other
jw. Other
jx. Other
jy. Other
jz. Other
ka. Other
kb. Other
kc. Other
kd. Other
ke. Other
kf. Other
kg. Other
kh. Other
ki. Other
kj. Other
kk. Other
kl. Other
km. Other
kn. Other
ko. Other
kp. Other
kq. Other
kr. Other
ks. Other
kt. Other
ku. Other
kv. Other
kw. Other
kx. Other
ky. Other
kz. Other
la. Other
lb. Other
lc. Other
ld. Other
le. Other
lf. Other
lg. Other
lh. Other
li. Other
lj. Other
lk. Other
ll. Other
lm. Other
ln. Other
lo. Other
lp. Other
lq. Other
lr. Other
ls. Other
lt. Other
lu. Other
lv. Other
lw. Other
lx. Other
ly. Other
lz. Other
ma. Other
mb. Other
mc. Other
md. Other
me. Other
mf. Other
mg. Other
mh. Other
mi. Other
mj. Other
mk. Other
ml. Other
mm. Other
mn. Other
mo. Other
mp. Other
mq. Other
mr. Other
ms. Other
mt. Other
mu. Other
mv. Other
mw. Other
mx. Other
my. Other
mz. Other
na. Other
nb. Other
nc. Other
nd. Other
ne. Other
nf. Other
ng. Other
nh. Other
ni. Other
nj. Other
nk. Other
nl. Other
nm. Other
nn. Other
no. Other
np. Other
nq. Other
nr. Other
ns. Other
nt. Other
nu. Other
nv. Other
nw. Other
nx. Other
ny. Other
nz. Other
oa. Other
ob. Other
oc. Other
od. Other
oe. Other
of. Other
og. Other
oh. Other
oi. Other
oj. Other
ok. Other
ol. Other
om. Other
on. Other
oo. Other
op. Other
oq. Other
or. Other
os. Other
ot. Other
ou. Other
ov. Other
ow. Other
ox. Other
oy. Other
oz. Other
pa. Other
pb. Other
pc. Other
pd. Other
pe. Other
pf. Other
pg. Other
ph. Other
pi. Other
pj. Other
pk. Other
pl. Other
pm. Other
pn. Other
po. Other
pp. Other
pq. Other
pr. Other
ps. Other
pt. Other
pu. Other
pv. Other
pw. Other
px. Other
py. Other
pz. Other
qa. Other
qb. Other
qc. Other
qd. Other
qe. Other
qf. Other
qg. Other
qh. Other
qi. Other
qj. Other
qk. Other
ql. Other
qm. Other
qn. Other
qo. Other
qp. Other
qq. Other
qr. Other
qs. Other
qt. Other
qu. Other
qv. Other
qw. Other
qx. Other
qy. Other
qz. Other
ra. Other
rb. Other
rc. Other
rd. Other
re. Other
rf. Other
rg. Other
rh. Other
ri. Other
rj. Other
rk. Other
rl. Other
rm. Other
rn. Other
ro. Other
rp. Other
rq. Other
rr. Other
rs. Other
rt. Other
ru. Other
rv. Other
rw. Other
rx. Other
ry. Other
rz. Other
sa. Other
sb. Other
sc. Other
sd. Other
se. Other
sf. Other
sg. Other
sh. Other
si. Other
sj. Other
sk. Other
sl. Other
sm. Other
sn. Other
so. Other
sp. Other
sq. Other
sr. Other
ss. Other
st. Other
su. Other
sv. Other
sw. Other
sx. Other
sy. Other
sz. Other
ta. Other
tb. Other
tc. Other
td. Other
te. Other
tf. Other
tg. Other
th. Other
ti. Other
tj. Other
tk. Other
tl. Other
tm. Other
tn. Other
to. Other
tp. Other
tq. Other
tr. Other
ts. Other
tt. Other
tu. Other
tv. Other
tw. Other
tx. Other
ty. Other
tz. Other
ua. Other
ub. Other
uc. Other
ud. Other
ue. Other
uf. Other
ug. Other
uh. Other
ui. Other
uj. Other
uk. Other
ul. Other
um. Other
un. Other
uo. Other
up. Other
uq. Other
ur. Other
us. Other
ut. Other
uu. Other
uv. Other
uw. Other
ux. Other
uy. Other
uz. Other
va. Other
vb. Other
vc. Other
vd. Other
ve. Other
vf. Other
vg. Other
vh. Other
vi. Other
vj. Other
vk. Other
vl. Other
vm. Other
vn. Other
vo. Other
vp. Other
vq. Other
vr. Other
vs. Other
vt. Other
vu. Other
vv. Other
vw. Other
vx. Other
vy. Other
vz. Other
wa. Other
wb. Other
wc. Other
wd. Other
we. Other
wf. Other
wg. Other
wh. Other
wi. Other
wj. Other
wk. Other
wl. Other
wm. Other
wn. Other
wo. Other
wp. Other
wq. Other
wr. Other
ws. Other
wt. Other
wu. Other
wv. Other
ww. Other
wx. Other
wy. Other
wz. Other
xa. Other
xb. Other
xc. Other
xd. Other
xe. Other
xf. Other
xg. Other
xh. Other
xi. Other
xj. Other
xk. Other
xl. Other
xm. Other
xn. Other
xo. Other
xp. Other
xq. Other
xr. Other
xs. Other
xt. Other
xu. Other
xv. Other
xw. Other
xx. Other
xy. Other
xz. Other
ya. Other
yb. Other
yc. Other
yd. Other
ye. Other
yf. Other
yg. Other
yh. Other
yi. Other
yj. Other
yk. Other
yl. Other
ym. Other
yn. Other
yo. Other
yp. Other
yq. Other
yr. Other
ys. Other
yt. Other
yu. Other
yv. Other
yw. Other
yx. Other
yy. Other
yz. Other
za. Other
zb. Other
zc. Other
zd. Other
ze. Other
zf. Other
zg. Other
zh. Other
zi. Other
zj. Other
zk. Other
zl. Other
zm. Other
zn. Other
zo. Other
zp. Other
zq. Other
zr. Other
zs. Other
zt. Other
zu. Other
zv. Other
zw. Other
zx. Other
zy. Other
zz. Other



Archival Documents

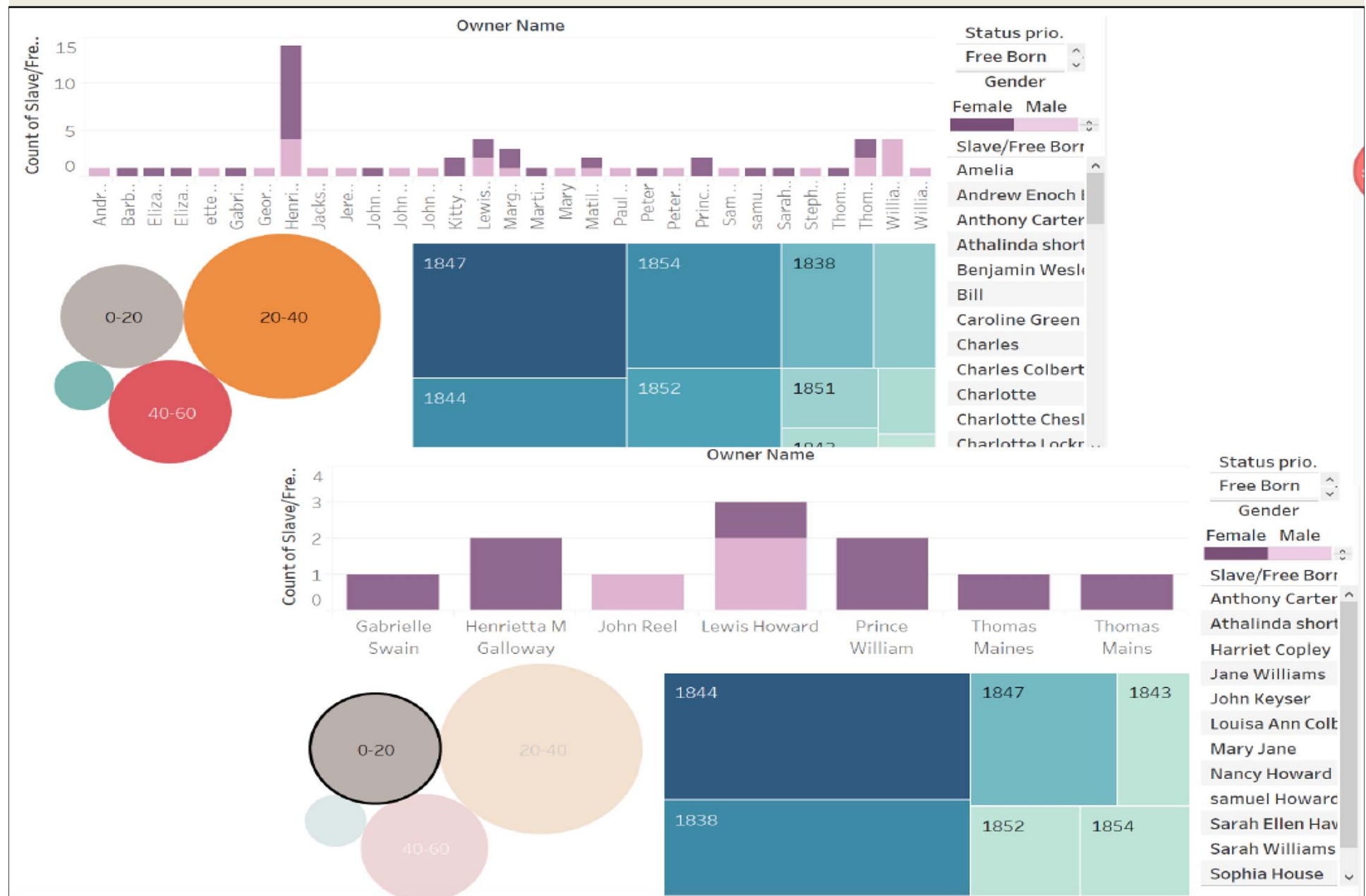
Digitization

Datafication
Data Modeling

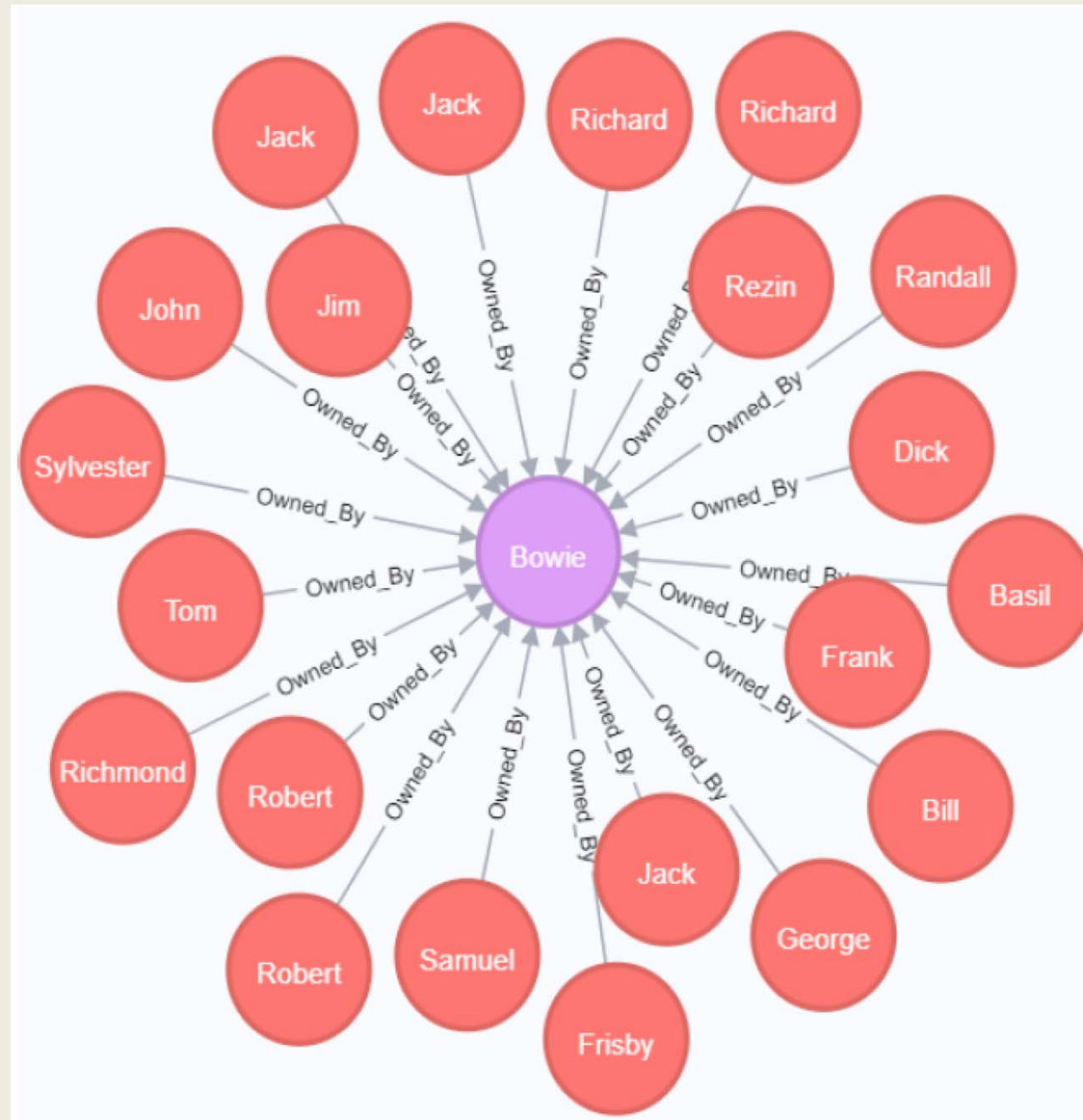
Data Visualization
Digital Storytelling

Archival Analytics

Interactive Tableau dashboard representing slavery statistics



Which slave names pertain to “Robert Bowie”. The purple node is the “Owner”, the red nodes are “Slave” names. The relationship is “Owned_By”.



CONTACTS

marciano@umd.edu

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