

# The Perspectives of Digital Curators on Building Distributed Repositories

**Richard Marciano**

*Lead Scientist, Sustainable Archives & Library Technologies lab (SALT) / SDSC*

**Chien-Yi YOU**

*Digital preservation specialist, SDSC*

**Reagan MOORE**

*Director of Data and Knowledge Systems, SDSC*

**Caryn WOJCIK**

*Government Records Archivist, State of Michigan*

**Mark CONRAD**

*Archives Specialist, ERA/NARA*

# Recent Collaborations on Preservation (NARA, NHPRC, LOC, NSF, IMLS)



- NARA:** 1998-2007, NARA - U Md, GTech, SLAC, UC Berkeley  
Transcontinental Persistent Archive Prototype based on data grids.
- IP2:** 2002-2006, NHPRC/SSHRC/NSF - UBC and others.  
InterPARES 2 collaboration with UBC on infrastructure independence
- PERM:** 2002-2004, NHPRC - Michigan, SDSC  
Preservation of records from an RMA. Interoperability across RMAs.
- LoC:** 2003-2004, LoC - SDSC, LOC  
Evaluation of use of SRB for storing American Memory collections
- ICAP:** 2003-2006, NHPRC - UCSD, UCLA, SDSC  
Exploring the ability to compare versions of records, run historical queries
- A&W:** 2000-2003, NHPRC - SDSC  
Methodologies for preservation & access of software- dependent electronic records
- DIGARCH:** 2005-2007, NSF - UCTV, Berkeley, UCSD Libraries, SDSC  
Preservation of video workflows
- eLegislature:** 2005-2007, NSF - Minnesota, SDSC  
Preserving the records of the e-Legislature
- VanMAP:** 2005-2006, UBC - UBC, Vancouver  
Preserving the GIS records of the city of Vancouver
- eLegacy:** 2006-2008, NHPRC - California  
Preserving the geospatial data of the state of California
- T-RACES:** 2006-2008, IMLS - UCHRI, SDSC  
California's redlining archives testbed
- PAT:** 2004-2007, NHPRC - Mi, Mn, Ke, Oh, Slac, SDSC  
Demonstration of a cost-effective system for preserving electronic records.

# Project Summary



- **Participants were digital curators from:**
  - Libraries / archives / historical societies / scientific data environments / museums
  - IT researchers and staff
- **Main Goal:**
  - Design a distributed repository for electronic records management
  - Demonstrate the management of various types of records with a common software infrastructure
- **Approach:** *each site...*
  - chose an archival collection
  - set up access control and update permissions for their preservation environment independently of the other participants
  - implemented a different preferred interface for interacting with their archival collections

# Presentation Goals



- **Comments:**
  - “No repository is an island”, David Giaretta
  - ... PAT fits the archipelago model
- **Examine:**
  - lessons learned and skills needed by digital curators to automate archival functions: appraisal, accessioning, arrangement, description, preservation, and access of records.
  - benefits achieved by using common infrastructure



# Partners



# PAT Project



- Test a community model for electronic records management, with archival and technological functions in a distributed network (using the SRB: Storage Resource Broker – data grid technology)
- Initial Test sites:
  - (1) **Michigan** Department of History, Arts and Libraries,
  - (2) **Ohio** Historical Society,
  - (3) **Kentucky** Department for Libraries and Archives,
  - (4) **Minnesota** Historical Society,
  - (5) **SLAC** Stanford Linear Accelerator Archives and History Office.

## Participants:

- (a) **California** State Archives
- (b) **Kansas** State Historical Society
- (c) University of Illinois Urbana Champaign
- (d) University of California Los Angeles (UCLA):
- (e) Yale Manuscripts and Archives
- (f) Georgia Tech

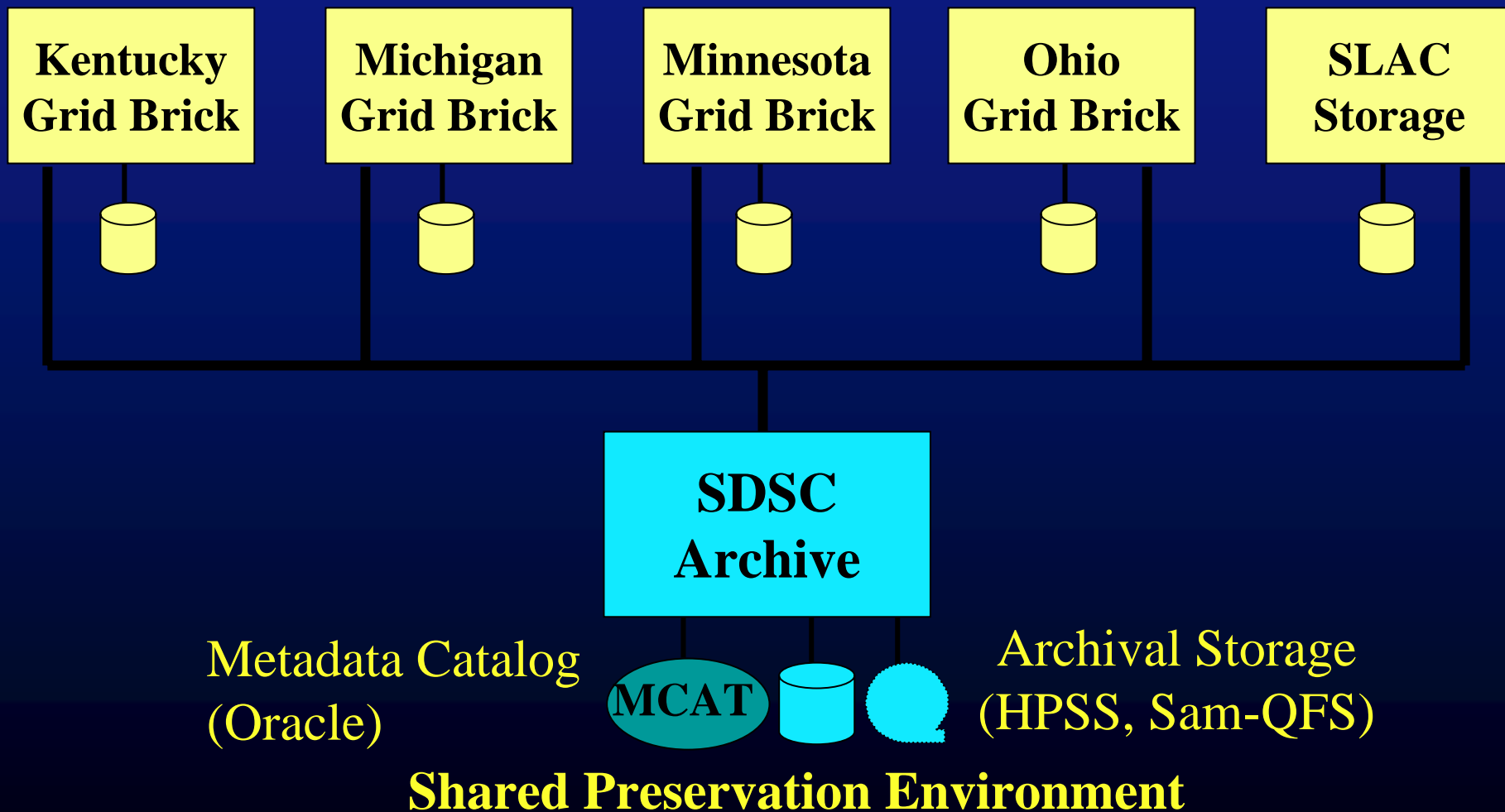
## Observers:

- (a) Getty Research Institute

# PAT Community Grid



## Local Storage Resources



# Automating Archival Processes



	<i>Kentucky Web</i>	<i>Michigan RMA -Precinct Results DB</i>	<i>Minnesota Spatial</i>	<i>Ohio E-mail</i>	<i>SLAC Documents</i>
<i>Appraisal</i>				X	
<i>Accession</i>	X			X	X
<i>Arrangement</i>	X	X		X	X
<i>Description</i>	X	X	X	X	X
<i>Preservation</i>	X	X	X		X
<i>Access</i>	X	X	X		X



# Unique Contributions of the Digital Curators to the Infrastructure



- **Windows-based SRB clients / servers**
- **Development of a Perl for Windows client library**
- **Bulk operations were developed, tested, and refined** (registration, accessioning, metadata extraction from records, metadata loading, validation of data movement into/out of/within the system)
- **End-to-end workflows were developed** (accessioning, replication)
- **SRB bugs revealed: better reliability**
- **MCAT ported to MySQL** (Oracle, DB2, Sybase, Informix)
- **Development of a wiki for documentation**
- **Registration of filenames with unusual characters discovered and fixed**
- **Suggestions on ways to simplify governance issues tied to particular types of data management:**
  - Need to express such policies as rules to be applied to the data mgt. system
  - Development of the next generation of data grid technology: iRODS (integrated Rule-Oriented Data System)
  - Each preservation process is express as a set of micro-services (operations that can be performed using a remote storage system access protocol)

# What Digital Curators Liked...



- Leverage common software and hardware
- Use commodity storage hardware
- Lower the cost of participation
- Reduce the level of expertise required at each site
- Focus on management of the archival collections and outsource the details of the archival repository
- Automate the manipulation of collections to minimize the level of effort

# Conclusions



- **PAT suggests that sustainability is probably beyond the capability of most individual archival repositories (cost of tracking new types of technology, expertise required to manage new technology, costs of the storage systems and databases, expertise necessary to manage multiple types of storage systems)**
- **Outsourcing of the mgt. or records is feasible through use of data grid technology**
- **Preservation environments can be assembled by creating regional community archival partnerships with university data centers**
- **Independence can be maintained:**
- **Service agreements for storage and preservation or archival e-records are needed**

# The Michigan example:



- Preservation of historical election data for the state of Michigan: precinct-level election data
- Process: from tape to archive to web...



# Before

鹽鹽鹽鹽鹽

2004 votes totals.txt - Notepad

File Edit Format Help

2004	GEN	0	00000	0	0	1	2	0	1		435
2004	GEN	0	00000	0	0	1	2	0	901	AVCB	373
2004	GEN	0	00000	0	0	1	4	0	1		512
2004	GEN	0	00000	0	0	1	4	0	901	AVCB	203
2004	GEN	0	00000	0	0	1	6	0	1		533
2004	GEN	0	00000	0	0	1	6	0	901	AVCB	167
2004	GEN	0	00000	0	0	1	8	0	1		820
2004	GEN	0	00000	0	0	1	10	0	1		339
2004	GEN	0	00000	0	0	1	10	0	901	AVCB	61
2004	GEN	0	00000	0	0	1	12	0	1		834
2004	GEN	0	00000	0	0	1	14	0	1		474
2004	GEN	0	00000	0	0	1	14	0	901	AVCB	151
2004	GEN	0	00000	0	0	1	16	0	1		333
2004	GEN	0	00000	0	0	1	16	0	901	AVCB	130
2004	GEN	0	00000	0	0	1	18	0	1		366
2004	GEN	0	00000	0	0	1	18	0	901	AVCB	83
2004	GEN	0	00000	0	0	1	20	0	1		231
2004	GEN	0	00000	0	0	1	22	0	1		267
2004	GEN	0	00000	0	0	1	52	0	901	AVCB	61
2004	GEN	0	00000	0	0	1	52	1	1		70
2004	GEN	0	00000	0	0	1	52	2	1		84
2004	GEN	0	00000	0	0	1	52	3	1		81
2004	GEN	0	00000	0	0	2	2	0	1		651
2004	GEN	0	00000	0	0	2	4	0	1		352
2004	GEN	0	00000	0	0	2	6	0	1		32
2004	GEN	0	00000	0	0	2	8	0	1		225
2004	GEN	0	00000	0	0	2	10	0	1		303
2004	GEN	0	00000	0	0	2	12	0	1	A	868
2004	GEN	0	00000	0	0	2	12	0	2		228
2004	GEN	0	00000	0	0	2	14	0	1		254
2004	GEN	0	00000	0	0	2	16	0	1		200
2004	GEN	0	00000	0	0	2	16	0	2		255
2004	GEN	0	00000	0	0	2	16	0	3		231
2004	GEN	0	00000	0	0	2	52	0	1	A	441
2004	GEN	0	00000	0	0	2	52	0	1	B	770
2004	GEN	0	00000	0	0	3	2	0	1		1462
2004	GEN	0	00000	0	0	3	2	0	2		600
2004	GEN	0	00000	0	0	3	4	0	1		1460
2004	GEN	0	00000	0	0	3	6	0	1		1085
2004	GEN	0	00000	0	0	3	8	0	1		654

Karyn Wojcik

Start

Novell...

Novell ...

Microso...

Microso...

wojcika ...

2004 vot...

N

9:36 AM

Michigan Precinct Results | Use Resource | Container sfs-disk-pat

Attribute	Value					
Name	Size	Owner	Timestamp	Repl	Resource	
2002 county codes.txt	1078	wojcik	2005-04-20-08.44.19	2	patMI-win	
2002 elective office codes.txt	28292	wojcik	2005-01-12-07.53.19	0	hpss-sdsc	
2002 elective office codes.txt	28292	wojcik	2005-03-28-17.00.27	1	sfs-disk-pat	
2002 elective office codes.txt	28292	wojcik	2005-04-20-08.44.30	2	patMI-win	
2002 readme.txt	4899	wojcik	2005-01-12-07.53.23	0	hpss-sdsc	
2002 readme.txt	4899	wojcik	2005-03-28-17.00.29	1	sfs-disk-pat	
2002 readme.txt	4899	wojcik	2005-04-20-08.44.36	2	patMI-win	
2002 vote totals.txt	22127167	wojcik	2005-01-12-07.53.05	0	hpss-sdsc	
2002 vote totals.txt	22127167	wojcik	2005-03-28-17.01.48	1	sfs-disk-pat	
2002 vote totals.txt	22127167	wojcik	2005-04-20-08.39.08	2	patMI-win	
2004 candidate name codes.txt	16633	wojcik	2005-06-16-11.05.53	0	patMI-win	
2004 candidate name codes.txt	16633	wojcik	2005-06-16-11.24.38	1	sfs-disk-pat	
2004 city and township codes.txt	54965	wojcik	2005-06-16-11.05.35	0	patMI-win	
2004 city and township codes.txt	54965	wojcik	2005-06-16-11.24.20	1	sfs-disk-pat	
2004 county codes.txt	1078	wojcik	2005-06-16-11.21.30	0	patMI-win	
2004 county codes.txt	1078	wojcik	2005-06-16-11.24.48	1	sfs-disk-pat	
2004 elective office codes.txt	13045	wojcik	2005-06-16-11.06.05	0	patMI-win	
2004 elective office codes.txt	13045	wojcik	2005-06-16-11.24.56	1	sfs-disk-pat	
2004 readme.txt	4899	wojcik	2005-06-16-11.21.38	0	patMI-win	
2004 readme.txt	4899	wojcik	2005-06-16-11.25.12	1	sfs-disk-pat	
2004 votes totals.txt	15636522	wojcik	2005-06-16-11.21.13	0	patMI-win	
2004 votes totals.txt	15636522	wojcik	2005-06-16-11.30.31	1	sfs-disk-pat	

wojcik @ michigan: all  
 1972 general election.txt  
 1972 general election.txt  
 1972 general election.txt  
 1972 i SAS setup.sas  
 1972 i SAS setup.sas  
 1972 i SPSS setup.sps  
 1972 i SPSS setup.sps  
 1972 i codebook.pdf  
 1972 i codebook.pdf  
 1972 i data.txt  
 1972 i data.txt  
 1972 i description.pdf  
 1972 i description.pdf  
 1972 i manifest.txt  
 1972 i manifest.txt  
 1972 i osiris dictionary.od  
 1972 i osiris dictionary.od  
 1972 i related literature.tx  
 1972 i related literature.tx  
 1974 general election.txt  
 1974 general election.txt  
 1974 general election.txt  
 1974 i codebook.pdf  
 1974 i codebook.pdf  
 1974 i data.txt

Datasets: 145 - Users: 1

Karyn Wojcik

鹽鹽鹽鹽





## History, Arts and Libraries

### Election Precinct Results Search

Please choose the Office Sought for INGHAM COUNTY  
or go BACK to change other criteria:

POLL BOOK TOTALS (TOTAL VOTERS)

POLL BOOK TOTALS (TOTAL VOTERS)

PRESIDENT OF THE UNITED STATES 4 YEAR TERM (1) POSITION

8TH DISTRICT REPRESENTATIVE IN CONGRESS 2 YEAR TERM (1) POSITION

67TH DISTRICT STATE REPRESENTATIVE 2 YEAR TERM (1) POSITION FILES IN INGHAM COUNTY

68TH DISTRICT STATE REPRESENTATIVE 2 YEAR TERM (1) POSITION FILES IN INGHAM COUNTY

69TH DISTRICT STATE REPRESENTATIVE 2 YEAR TERM (1) POSITION FILES IN INGHAM COUNTY

MEMBER OF THE STATE BOARD OF EDUCATION 8 YEAR TERMS (2) POSITIONS

MEMBER OF THE UNIVERSITY OF MICHIGAN BOARD OF REGENTS 8 YEAR TERMS (2) POSITIONS

MEMBER OF THE MICHIGAN STATE UNIVERSITY BOARD OF TRUSTEES 8 YEAR TERMS (2) POSITIONS

MEMBER OF THE WAYNE STATE UNIVERSITY BOARD OF GOVERNORS 8 YEAR TERMS (2) POSITIONS

JUSTICE OF THE SUPREME COURT 8 YEAR TERMS (2) POSITIONS





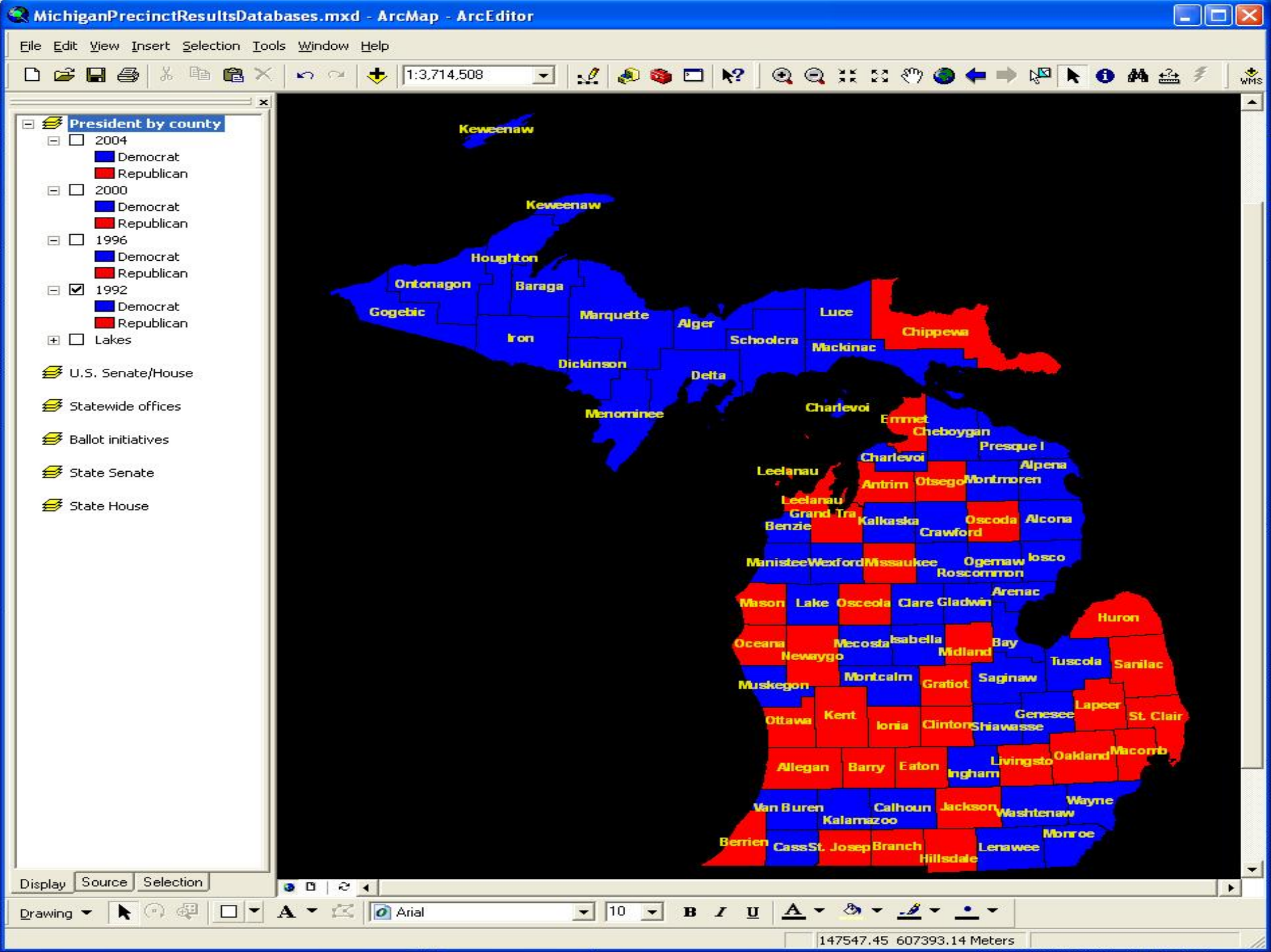
## History, Arts and Libraries

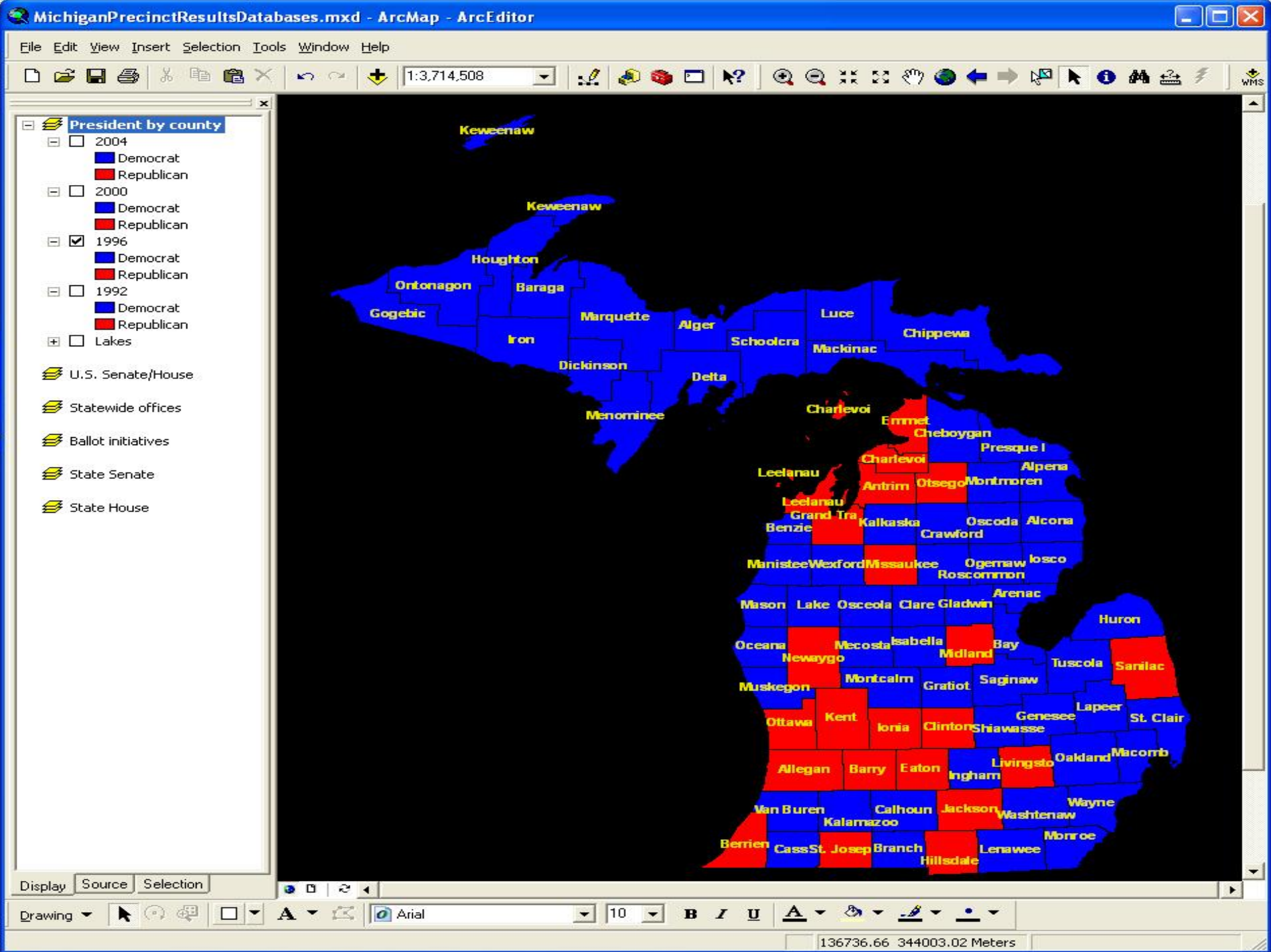
### Results of the Election Precinct Results Search

Criteria used: -- Election: **2004 GEN**;  
 -- Office sought: **PRESIDENT OF THE UNITED STATES 4 YEAR TERM (1) POSITION**;  
 -- County: **INGHAM**;

County	Office Sought	City/Township Name	Precinct Identifier	COBB DAVID	BUSH GEORGE W.	KERRY JOHN F.	BADNARIK MICHAEL	PEROUTKA MICHAEL ANTHONY	NADER RALPH	BROWN WALTER	Precinct Total
INGHAM	PRESIDENT OF THE UNITED STATES 4 YEAR TERM (1) POSITION	ALAIEDON TOWNSHIP	1	0	220	178	1	0	1	0	400
INGHAM	PRESIDENT OF THE UNITED STATES 4 YEAR TERM (1) POSITION	ALAIEDON TOWNSHIP	2	0	264	196	0	0	3	1	464
INGHAM	PRESIDENT OF THE UNITED STATES 4 YEAR TERM (1) POSITION	ALAIEDON TOWNSHIP	3	1	360	267	1	0	1	0	630
INGHAM	PRESIDENT OF THE UNITED STATES 4 YEAR TERM (1) POSITION	ALAIEDON	4	1	244	197	2	1	2	1	446

Karyn Wojcik







File Edit View Insert Selection Tools Window Help



## President by county

- ☐ 2004
  - ☒ Democrat
  - ☐ Republican
- ☒ 2000
  - ☒ Democrat
  - ☐ Republican
- ☐ 1996
  - ☒ Democrat
  - ☐ Republican
- ☐ 1992
  - ☒ Democrat
  - ☐ Republican
- ☐ Lakes

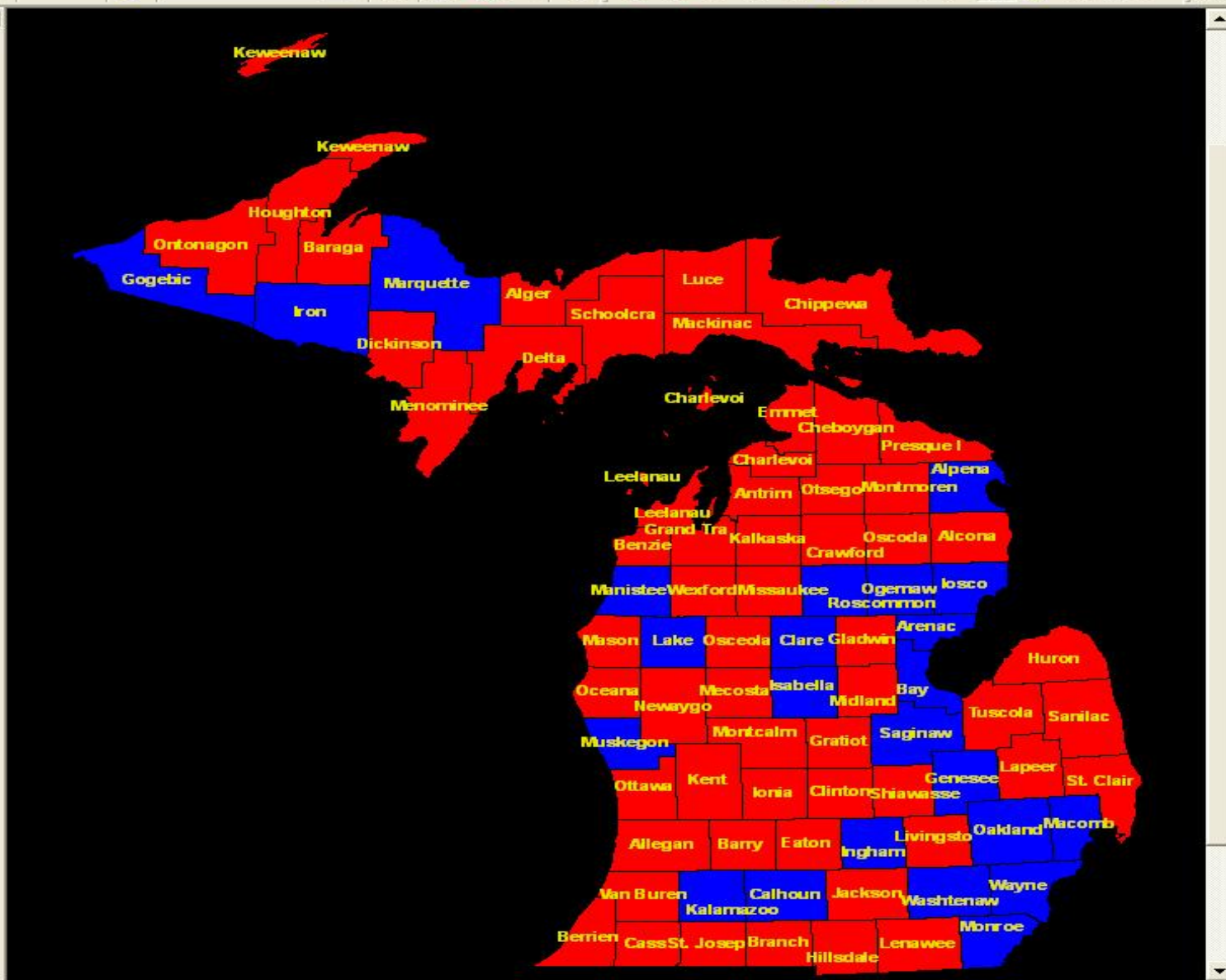
U.S. Senate/House

Statewide offices

Ballot initiatives

State Senate

State House

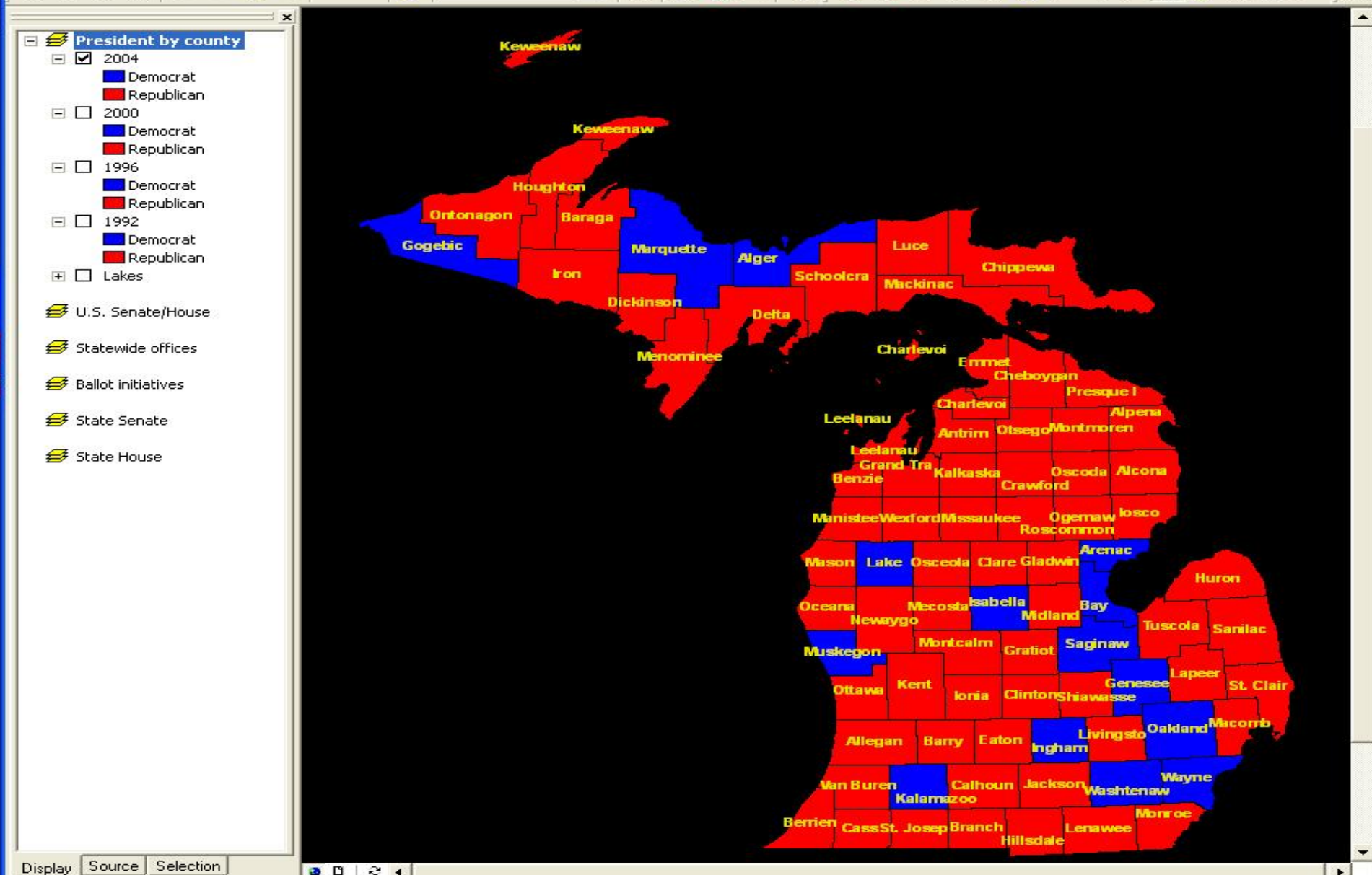


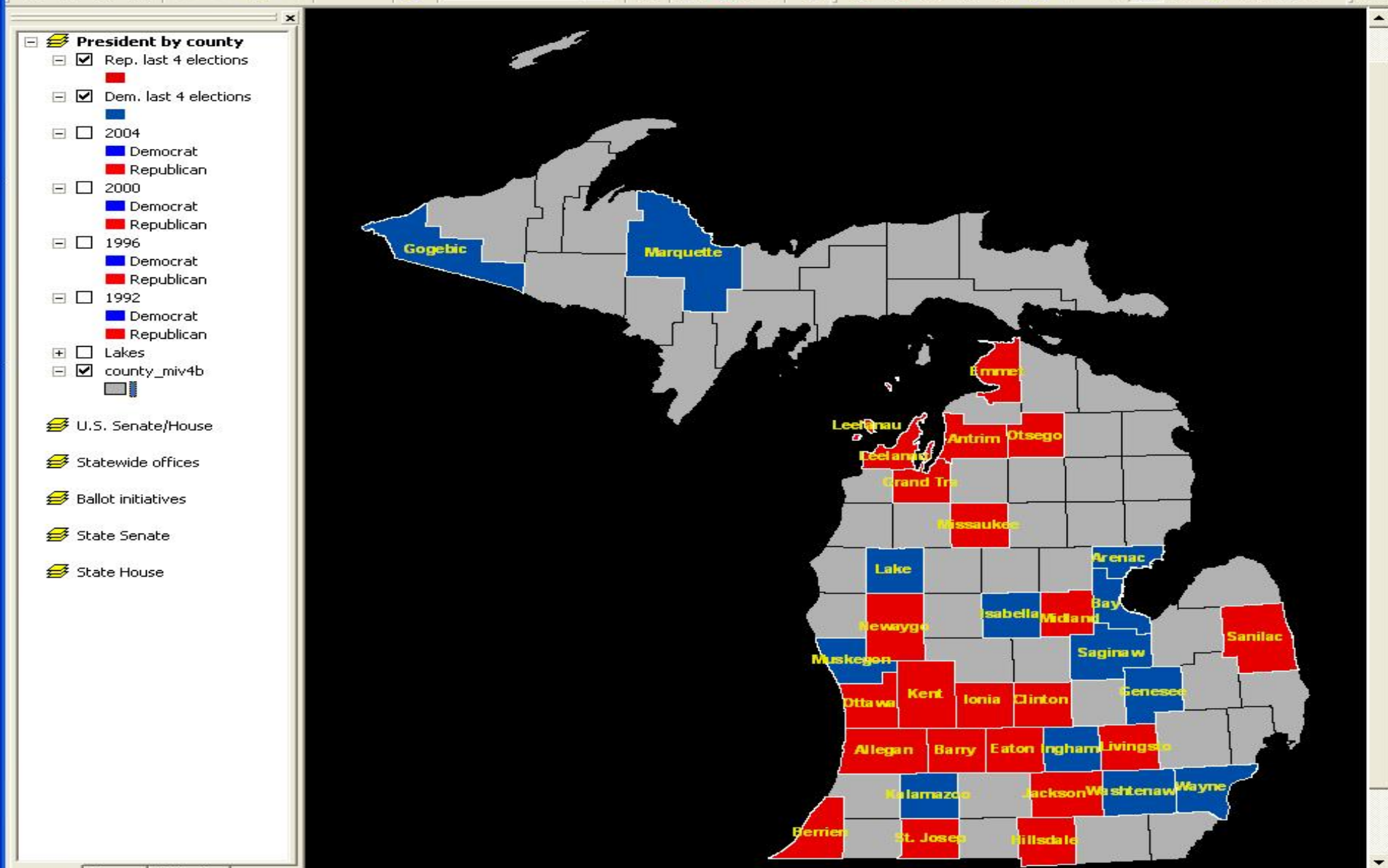
Display Source Selection

Drawing Arial 10 B I U

136736.66 344003.02 Meters







# For More Information

Richard Marciano  
marciano@sdsc.edu