Trusted Digital Repository Design

or

The art of "shaping" our preservation tools

ACA@UBC
Fifth International Seminar and Symposium
We Shape our Tools and our Tools Shape Us
7 February 2013

Richard Marciano

UNC Chapel Hill
richard_marciano@unc.edu

http://salt.unc.edu

Summary

- *This talk addresses policy-driven frameworks to support institution-specific preservation environments.*
- The approach is based on the ISO/DIS 16363 standard on "Audit and Certification of Trustworthy Digital Repositories" and rule-based data management systems.
- The seminar discusses preservation workflows and interface design issues.

Trust


UNITED STATES OF AMER
IN GOD WE TRUST
RECORDS
THE WHITE HOUSE
TWENTY DOLLARS

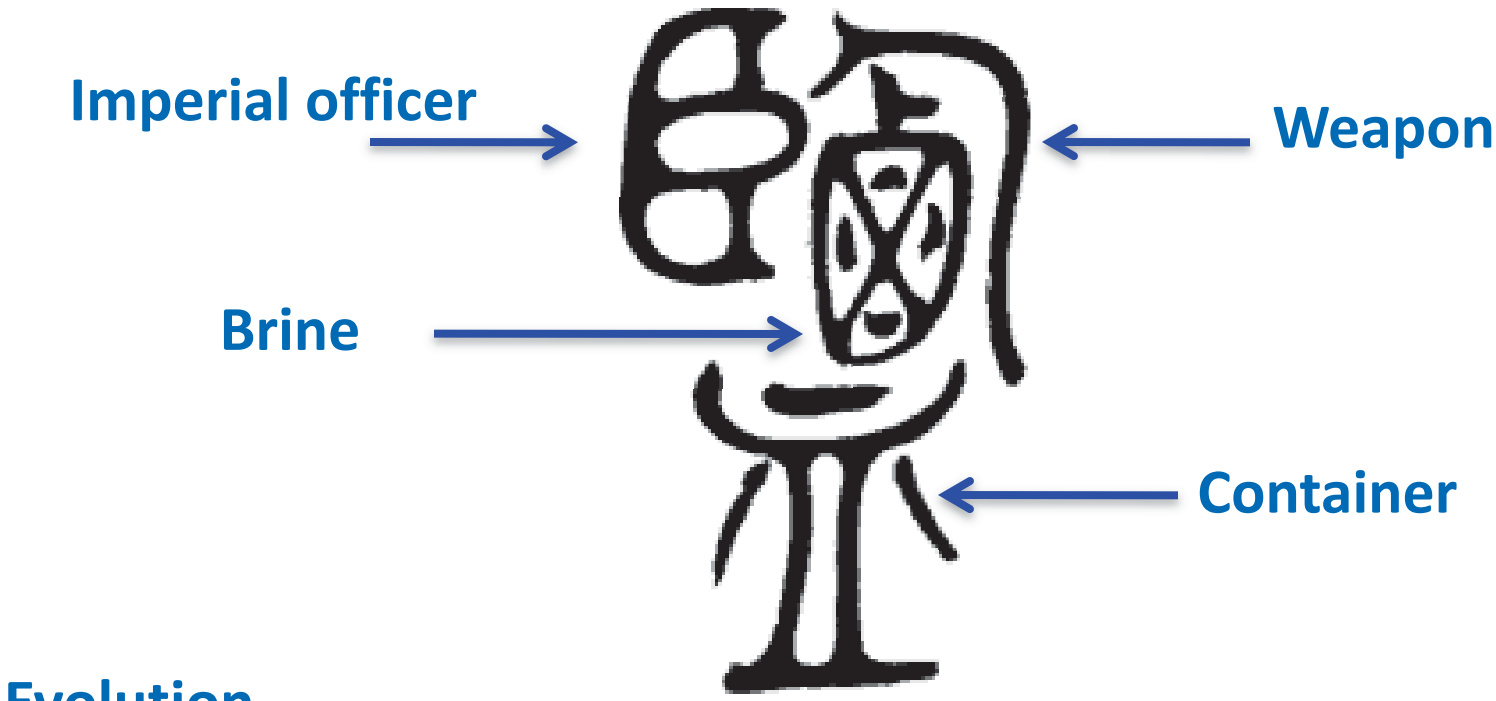
Case-Studies in Trust

1. Software Correctness
2. Incentivizing Volunteers for Crowdsourced Projects
3. Multi-level Information Modeling and Preservation
4. Preservation of the GIS Records of VanMap
5. Data Grids & Federation
6. *Community-driven Policy-based Preservation (DCAPE)*

SALT…

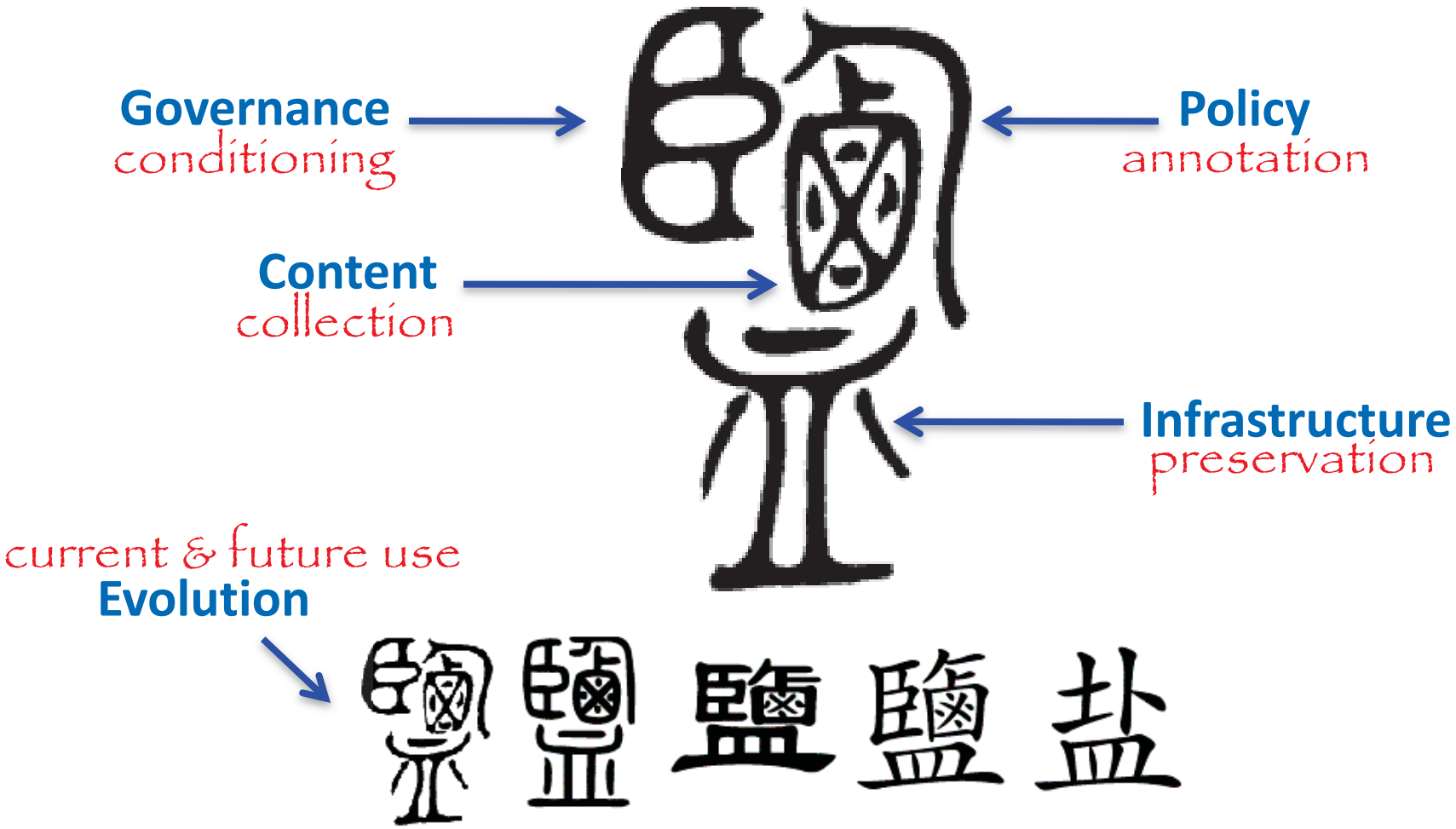
SustainAbiLiTy

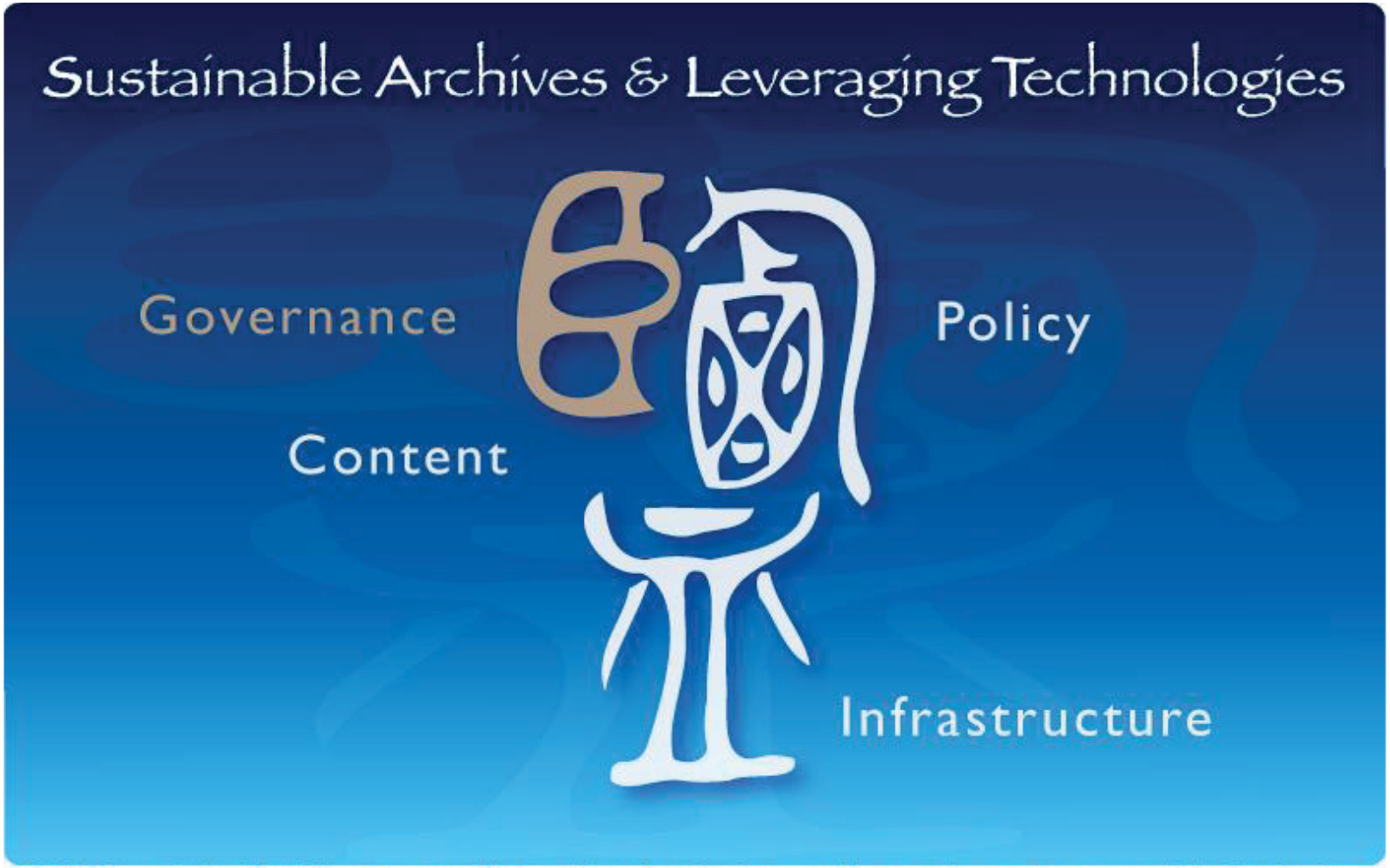
SALT/"yan": a metaphor for long-term preservation



SALT

"We define the 'discipline of data curation' as the practice of collection, annotation, conditioning, and preservation of data for both current and future use"

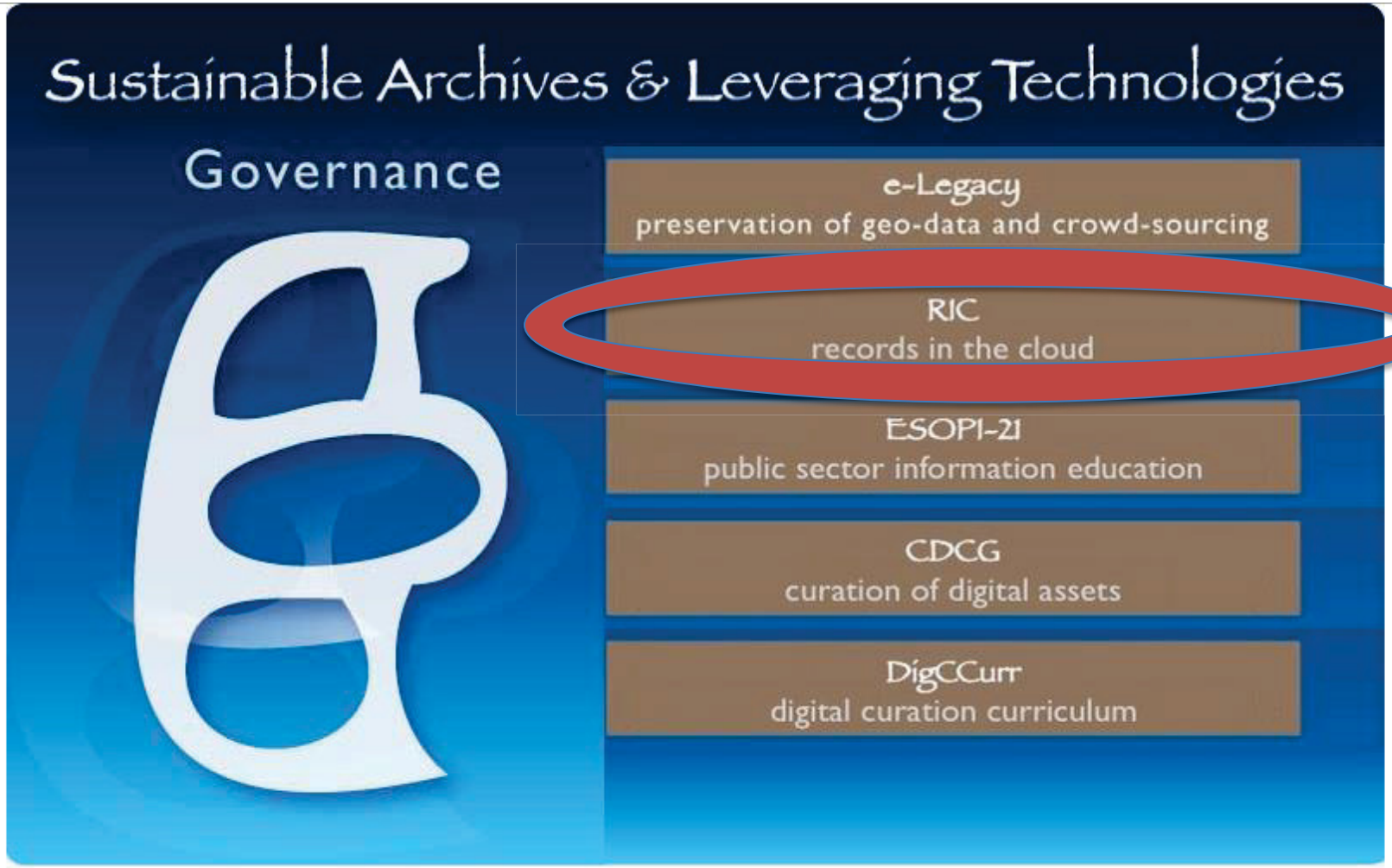




SALT is an interdisciplinary group focused on developing and leveraging resources and technologies to enable collaborations.

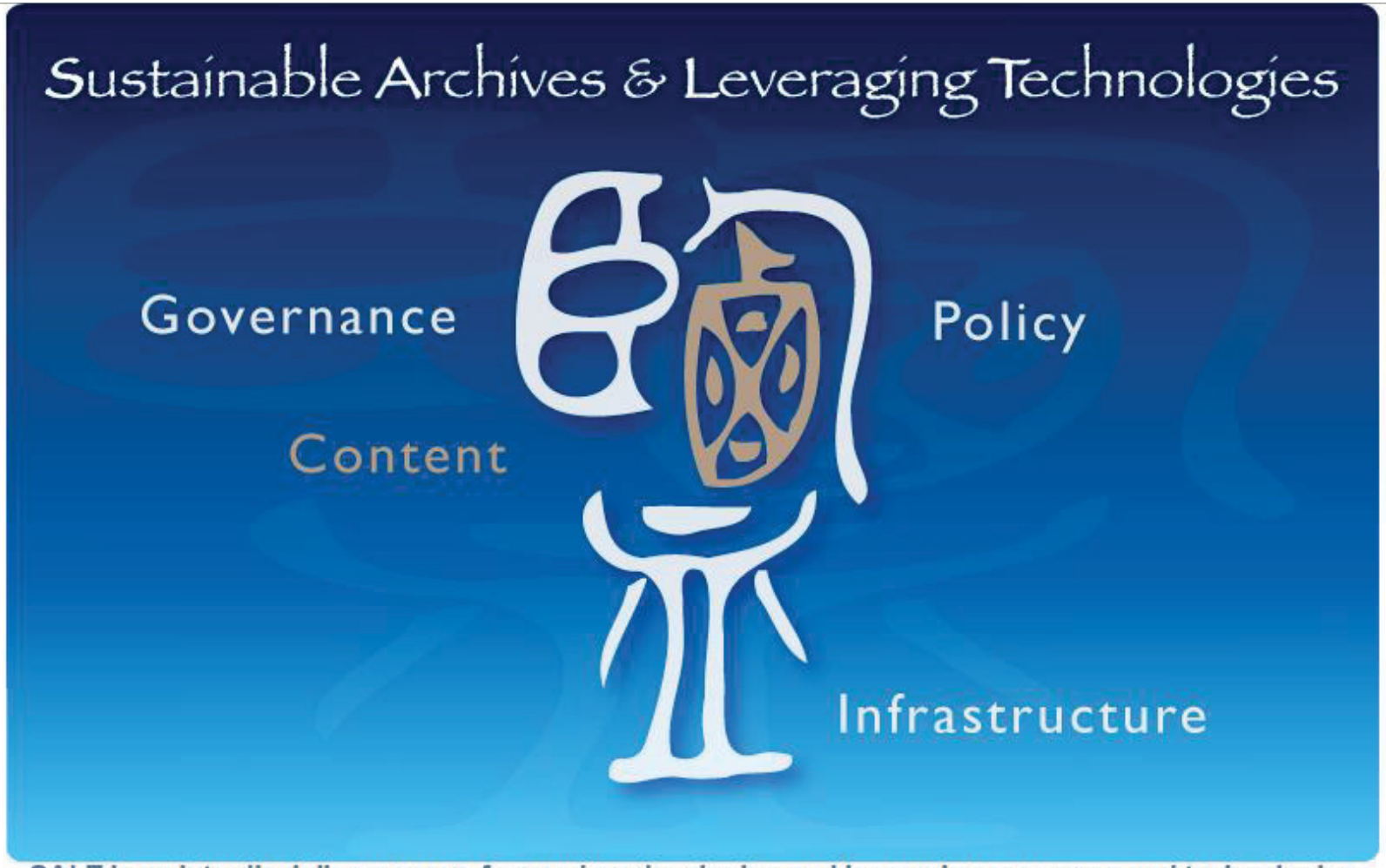
Focusing on the interplay of content, policy, governance, and cyberinfrastructure.

Partners 鹽 鹽 鹽 鹽 盐 About Us



preservation / social networking

HOME

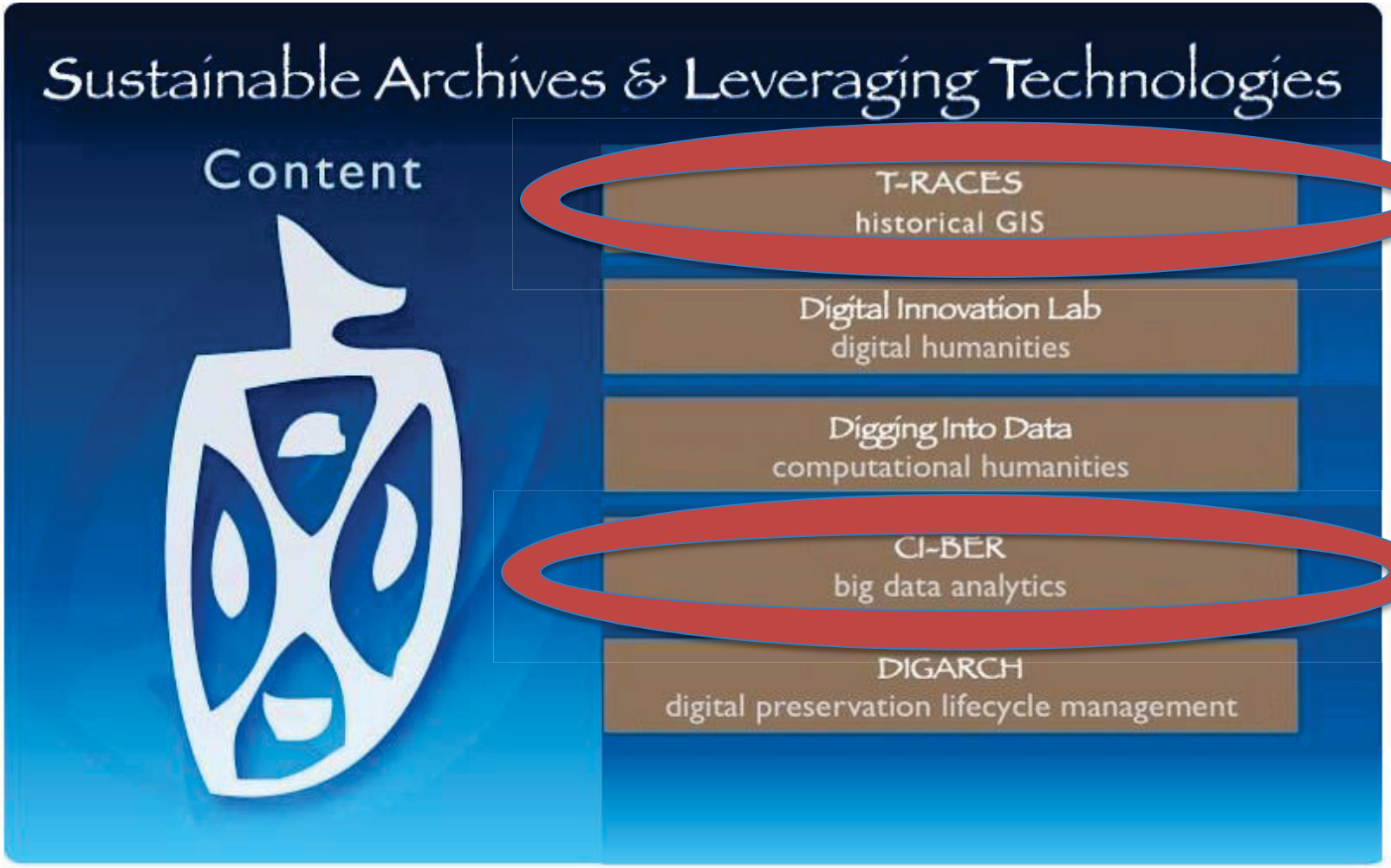


SALT is an interdisciplinary group focused on developing and leveraging resources and technologies to enable collaborations.

Focusing on the interplay of content, policy, governance, and cyberinfrastructure.

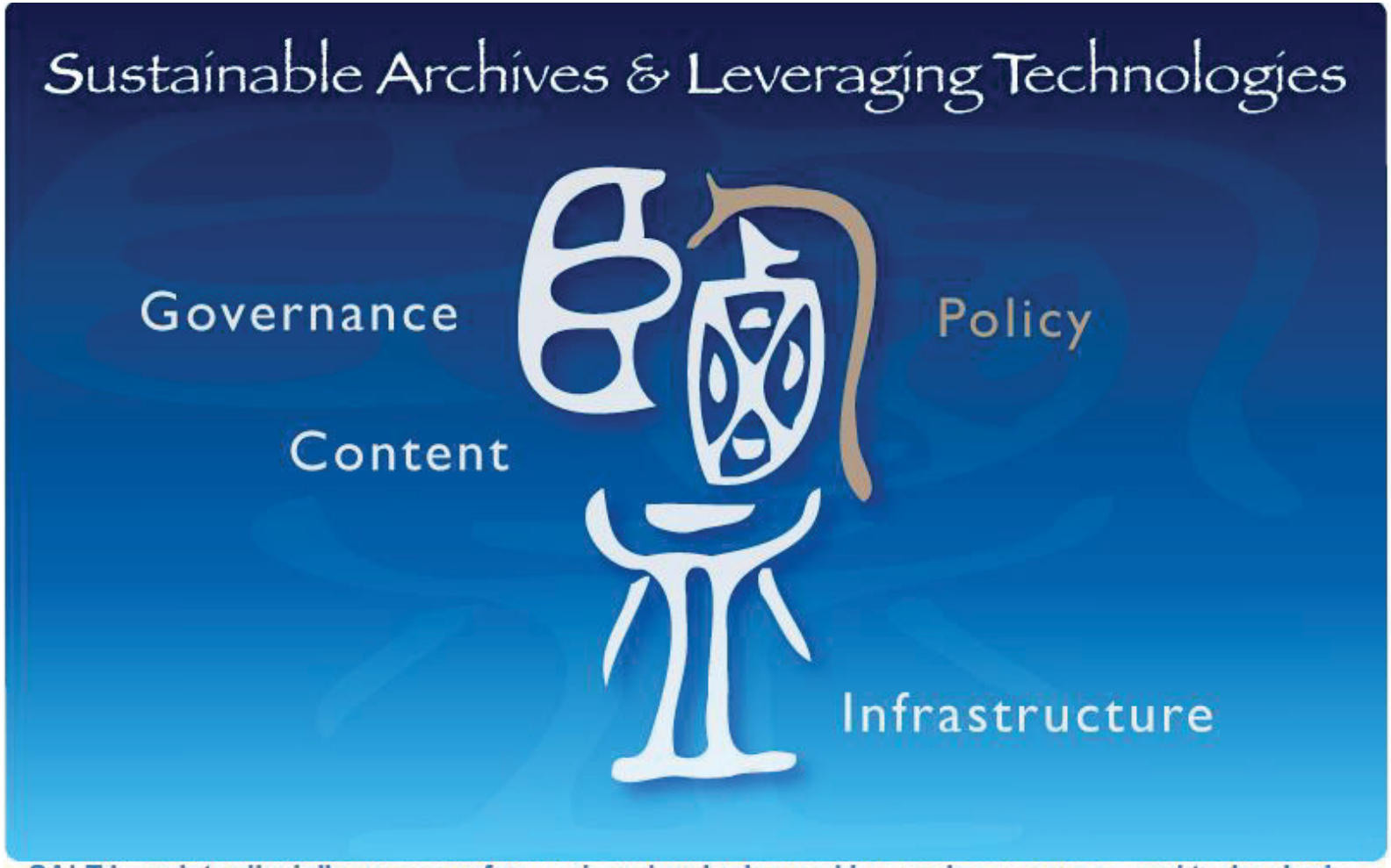
Partners 鹽 鹽 鹽 鹽 盐 About Us

salt.unc.edu/content.html



data grids / digital libraries / digital mapping

HOME

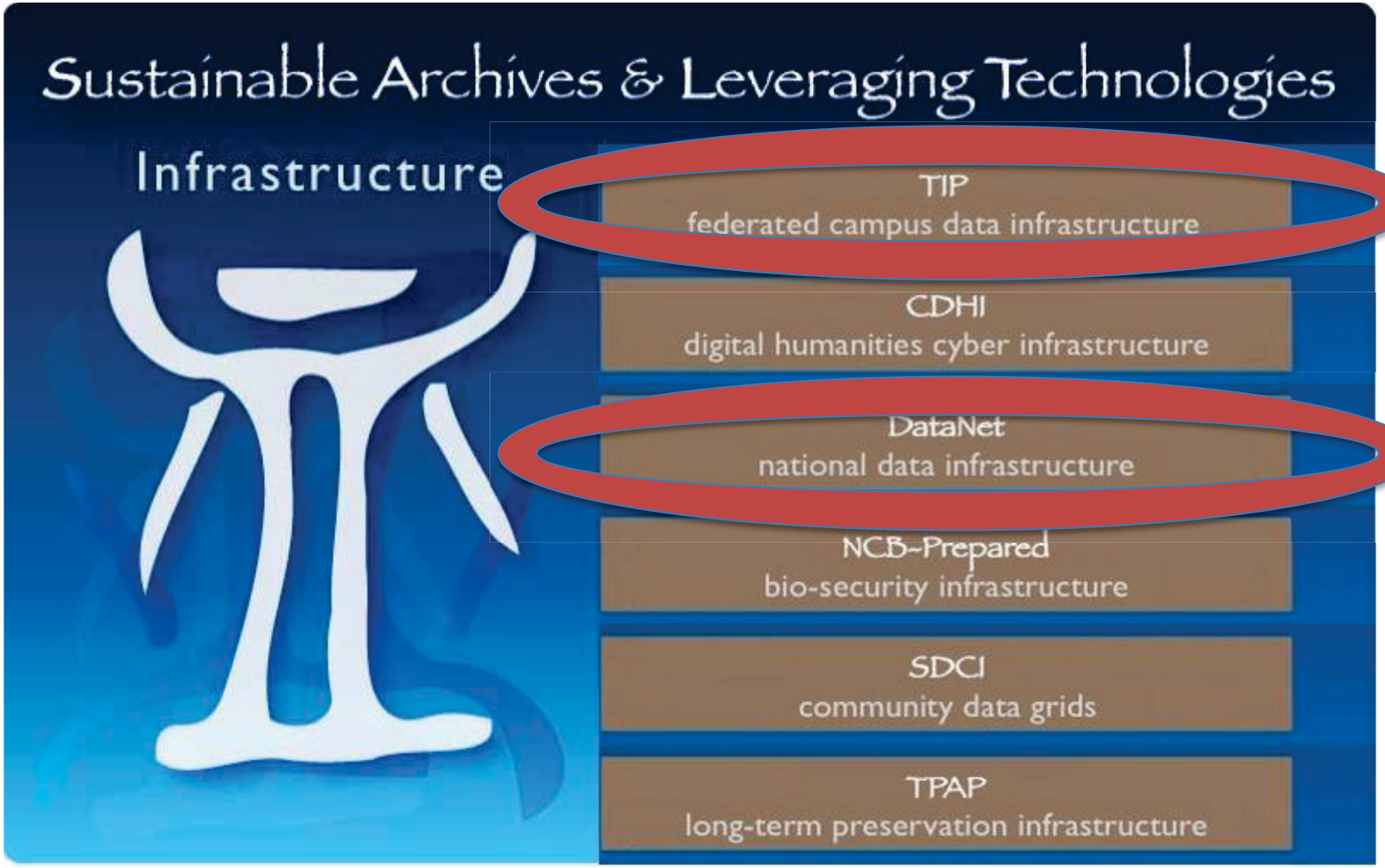


SALT is an interdisciplinary group focused on developing and leveraging resources and technologies to enable collaborations.

Focusing on the interplay of content, policy, governance, and cyberinfrastructure.

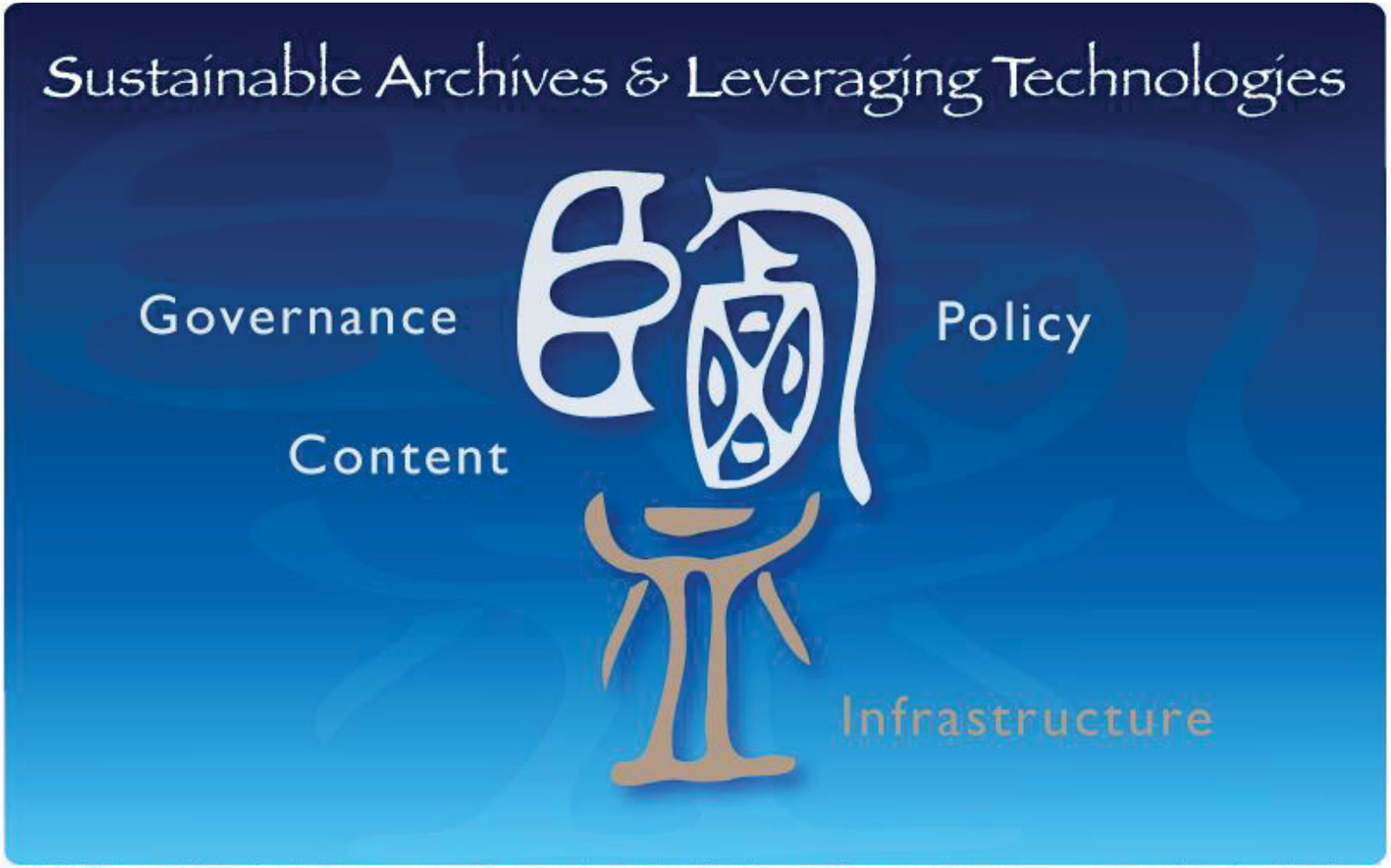
Partners 鹽 鹽 鹽 鹽 盐 About Us

salt.unc.edu/policy.html


Sustainable Archives & Leveraging Technologies
Infrastructure
TIP
federated campus data infrastructure
CDHI
digital humanities cyber infrastructure
DataNet
national data infrastructure
NCB-Prepared
bio-security infrastructure
SDCI
community data grids
TPAP
long-term preservation infrastructure

federation

Sustainable Archives & Leveraging Technologies

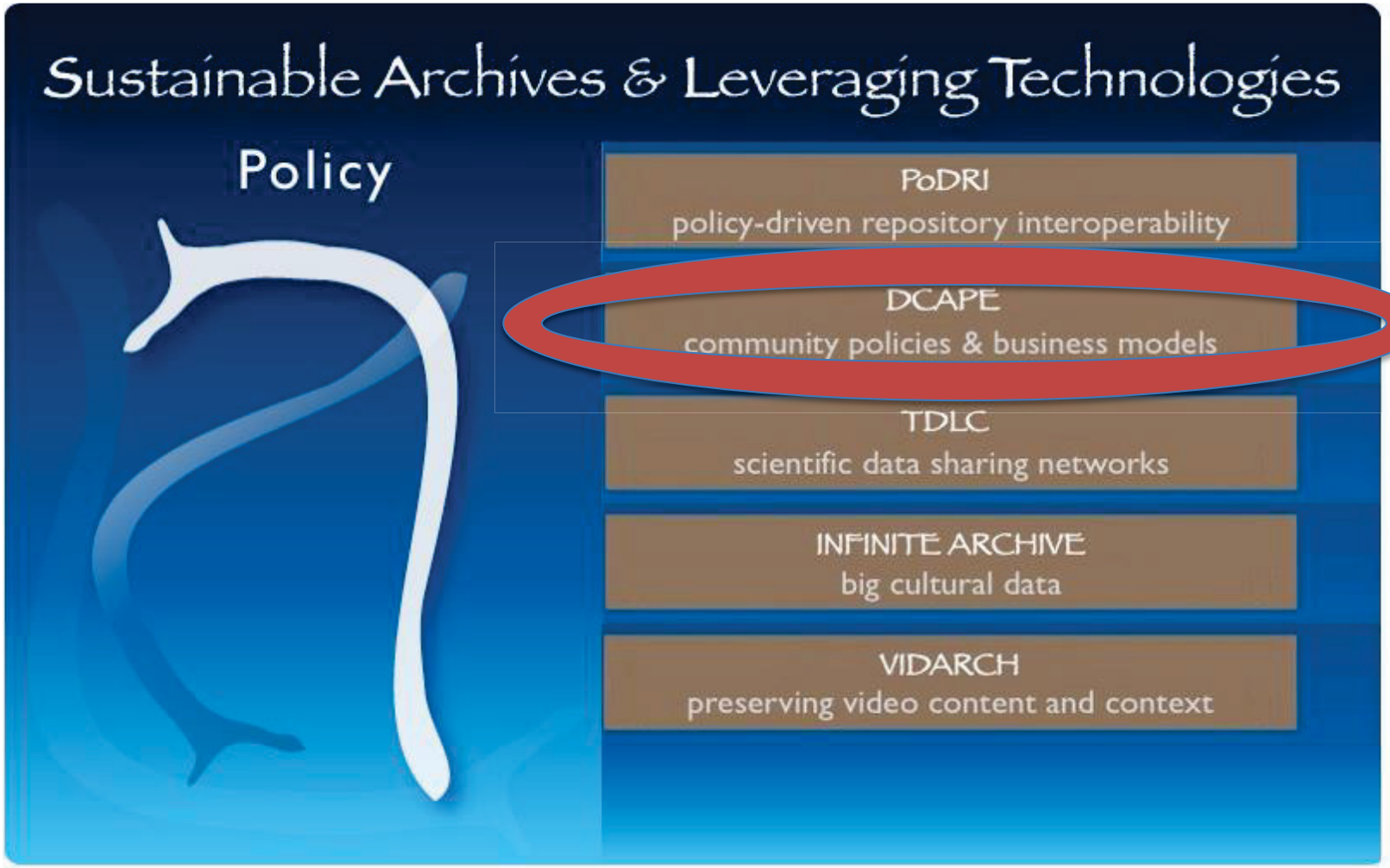


SALT is an interdisciplinary group focused on developing and leveraging resources and technologies to enable collaborations.

Focusing on the interplay of content, policy, governance, and cyberinfrastructure.

Partners 鹽 鹽 鹽 鹽 盐 About Us

salt.unc.edu/infra.html



Business Models / SLAs

1. Software Correctness

- Algebraic Methodology and Software Technology
  - Abstract state machines – algebraic methods – algebraic specifications – correct software design – formal methods – formal verification – process algebra
- Algebraic compiler technology

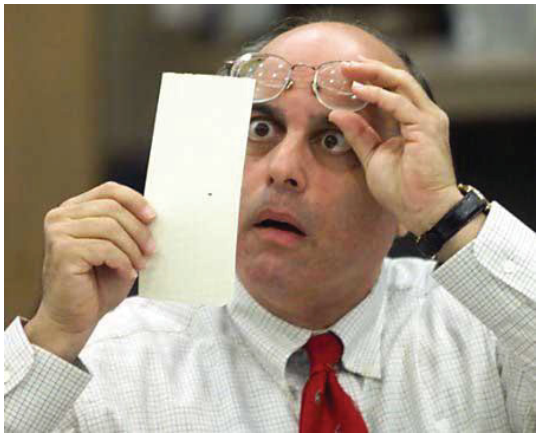
2. Incentivizing Volunteers for Crowdsourced Projects

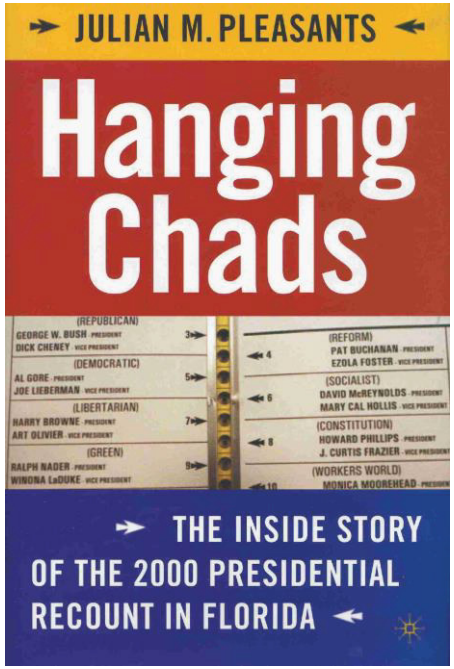
SAA 2013 Lightning Session

- Historical Residential Segregation
  - Richard Marciano (UNC)
- Genealogy Records
  - Emily Schultz (FamilySearch)
- Slave Narratives:
  - Chien-Yi Hou (UNC)
- Civil War Project
  - Colleen Theisen (U. of Iowa)
- Mapping Historical Photos
  - Jon Protas (SepiaTown)
- Engaging the Crowd with Humanities Research
  - Mark Hedges (KCL)
- 1940 Census Indexing
  - Kenton McHenry (NCSA)
- What's on the Menu
  - David Riordan (NYPL)
- Old Weather
  - Mark Mollan (National Archives)

3. Multi-level Information Modeling and Preservation

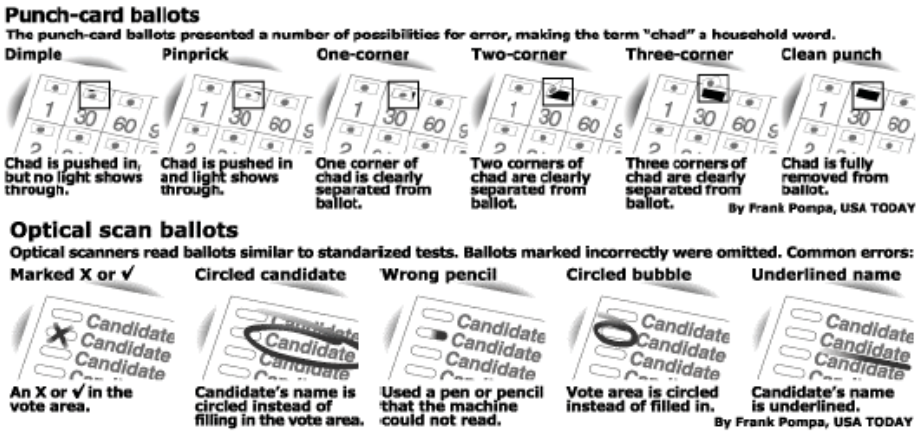
- **"On behalf of the State Elections Canvassing Commission and in accordance with the laws of the State of Florida, I hereby declare Governor George W. Bush the winner of Florida's 25 Electoral Votes,"** said Florida's Secretary of State, Katherine Harris, as she certified George W. Bush the winner over Al Gore, on November 26, 2000.





Official NARA Documents

- The National Archives and Records Administration (NARA), went on to record this 25-Vote result by collecting two documents for permanent retention:
  - Certificate of Ascertainment, containing the proposed Electors:
    - http://www.archives.gov/federal-register/electoral-college/2000_certificates/ascertainment_florida.html
  - Certificate of Vote, capturing the winning Electors:
    - http://www.archives.gov/federal-register/electoral-college/2000_certificates/vote_florida.html



Examples of Undervote ballots for punch-card and optical-scan from the USA Today study.

NORC Codes used to classify each mark on **punch-card** and **optical-scan** ballots.

Punch-card ballots		**Optical-scan ballots**	
Code	**Meaning**	**Code**	**Meaning**
0	blank, no mark seen	00	blank, no mark seen
1	1-corner of chad detached	11	circled party name
2	2-corners of chad detached	12	other mark on or near party name
3	3-corders of chad detached	21	circled candidate name
4	4-corners of chad detached, clean punch	22	other mark on or near candidate name
5	dimpled chad, no sunlight	31	arrow/oval mark other than fill: circle, x, /, check, scribble
6	dimpled chad, with sunlight	32	other mark near oval/arrow
7	dimple within chad area, off chad, with or without sunlight	44	arrow/oval filled
8	dimple on border of chad area, with or without sunlight	88	arrow/oval filled or marked other than fill, then erased or partially erased
9	chad marked with pencil or pen	99	negated mark: scribble-through, cross-out, "NO", and similar

Appendix C: Harmonized Codes

Equivalence Classes	NORC Codes	
	Punch-Card	Optical-Scan
0	0	00 / 99 / 88
1	8	
2	7	
3	5	11 / 12 / 21 / 22 / 31 / 32
4	6	
5	1	
6	2	
7	3	
8	4	44
9	9	

Standards to specify evidence of voter intent.

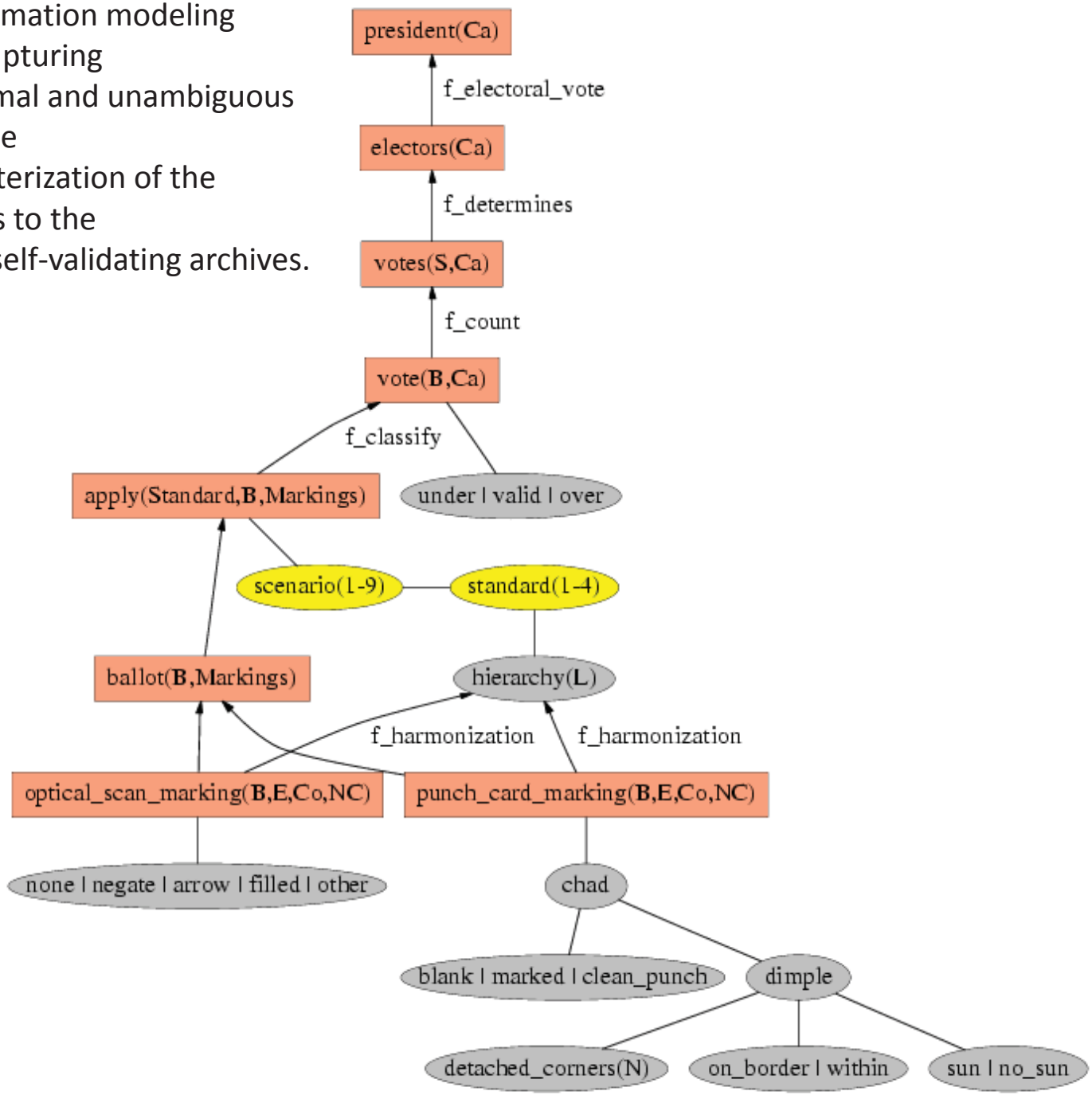
Standards	Equivalence Class Codes	
	Punch-Card	Optical-Scan
1. Dimple or better	>= 3	>= 3
2. One-corner detached	>= 5	>= 3
3. Two-corner detached	>= 6	>= 3
4. Dimple (if rest of ballot is dimpled)	(>= 6) && (3, 4, 5)	>= 3

Appendix E: Scenarios

Scenarios
1. Prevailing statewide standard
2. Supreme Court "simple"
3. Supreme Court "complex"
4. 67-county custom standards
5. Two-corners-detached statewide
6. "Most inclusive" statewide
7. "Most restrictive" statewide
8. The Gore 4-county recount strategy
9. "Dimples when other dimples present"

For example, *Scenario 5.*, *Two-corners-detached statewide*, is based on arguments made by George W. Bush's attorneys during the 36-day period following Election Day, where Standard 3. is applied statewide.
Also, *Scenario 8.*, *The Gore 4-county recount strategy*, is based on early post-election results, where the Gore camp requested hand counts in 4 heavily Democratic counties: Miami-Dade, Broward, Palm Beach and Volusia. Standard 2 is applied to some of Miami-Dade precincts.

Multi-level (or “deep”) information modeling provides a mechanism for capturing process information in a formal and unambiguous way as a network of database transformations. The characterization of the modeling process itself leads to the notion of self-instantiating, self-validating archives.

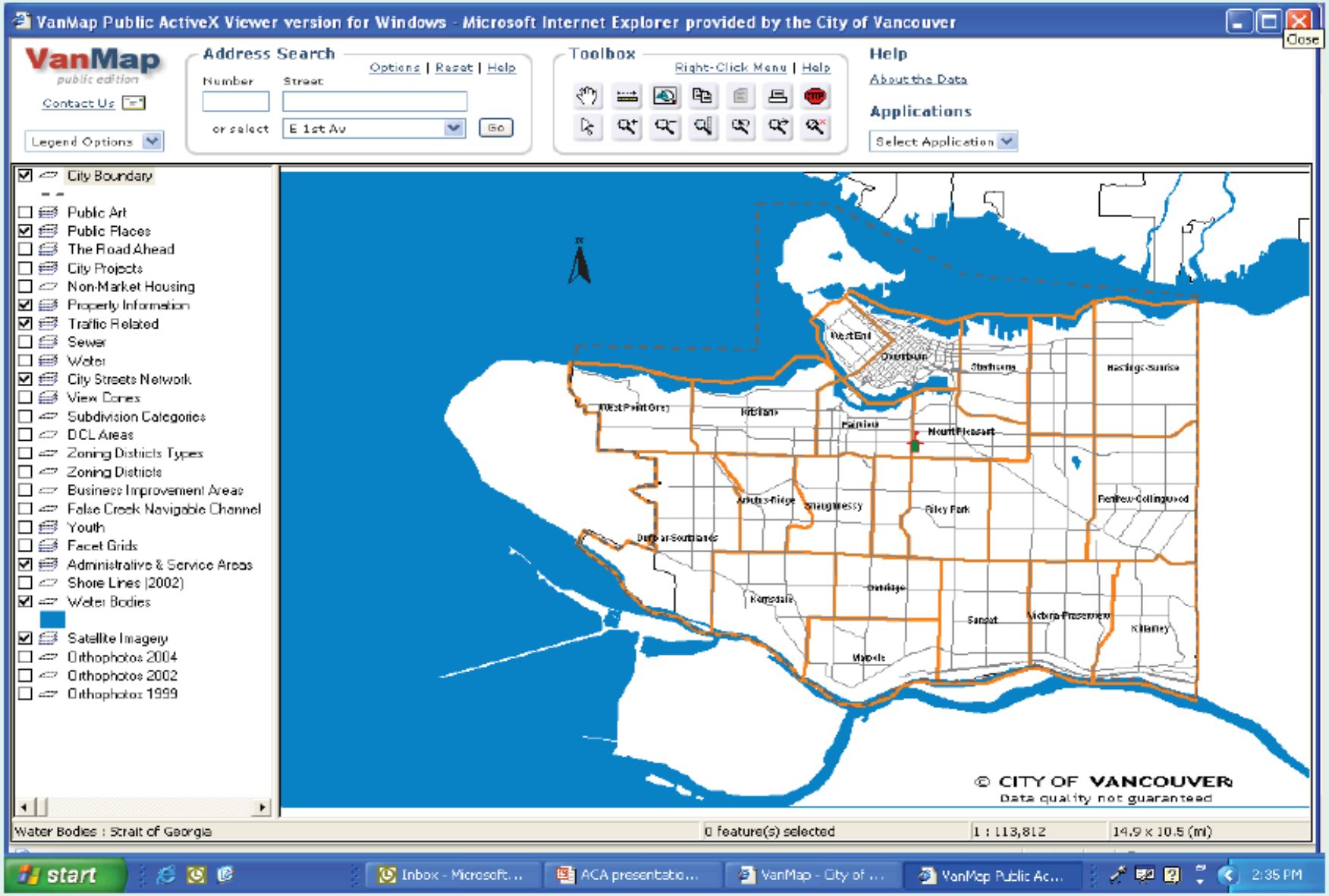


4. Preservation of the GIS Records of VanMap

- *Phase I: 2004-2005*
  - *Insurance Corporation of British Columbia:*
    - **Evelyn McLellan** (case study leader)
  - *UBC:*
    - **Luciana Duranti**, Director InterPARES2
    - **Eleanor Kleiber**, Research Assistant
    - **Catherine Miller**, Research Assistant
  - *City of Vancouver Records & Archives:*
    - **Glenn Dingwall**, Digital Archivist
    - **Liz Wright**, Corporate Records Administrator
    - **Andrew Power**, Corporate Information Analyst
    - **Sue Bigelow**, Conservator
    - **Scott Redgrove,** Digital Archivist
    - **Heather Gordon**
    - **Reuben Ware**, Director
  - *Information Technology, VanMap GIS:*
    - **Jonathan Mark**, Manager
    - **Meng Li,** Chief VanMap Architect
    - **Frank DeWith,** Oracle Database Specialist
  - *Artefactual Systems Inc.:*
    - **Peter Van Garderen**, President / Consultant

- *Phase II: 2006*
  - *SDSC:*
    - **Reagan Moore**
    - **Richard Marciano**

www.vancouver.ca/vanmap



Is VanMap a Record?

- Data is overwritten without being saved *(lack of fixity)*
- Nothing is saved and set aside for future action or reference
  - No fixed documentary form
  - No stable content
  - No archival bond with other records
  - No record context
- Information used to inform a government decision could be kept as a formal record of government activity
- This record could be archived by preserving snapshots of the city databases or by preserving the components of the VanMap system over time
- If the archived databases can be connected to the VanMap system, then a view can be recreated of the City of Vancouver at a prior point in time

VanMap is a "Potential Record"

- VanMap presentations can be turned into records by creating fixed representations: "setting the records aside"

- The impetus to preserve VanMap as a record is driven by public use of government records to support litigation, appraisal and review of prior government decisions, and as a valuable historical resource

What Should be Saved and Set Aside?

- Preserving of all the city databases?
  - Very difficult: housed at different locations, updated at different times, managed by different organizations
- Preserving snapshots of the databases?
  - Very difficult decision
- Solution: "preserve a record of evidence of City actions". Options:
  - Preserve the view represented by the composition of the VanMap layers that were displayed
  - Preserve each layer used to compose the VanMap presentation
  - Preserve the database from which the information for each layer is extracted

- Frequency of preservation:
  - Monthly snapshot, each time the data within the system is updated, each time the system is accessed?
- Solution: look at the amount of time over which government decisions are made

On Using Data Grids

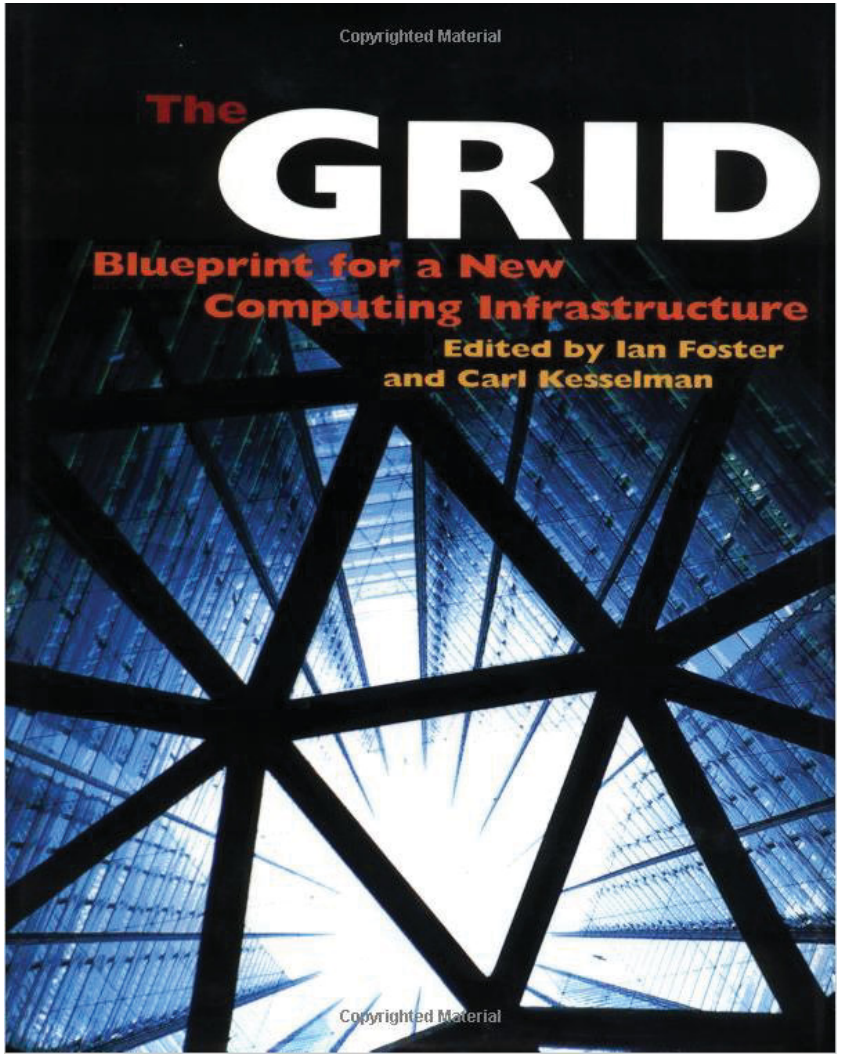
- Possible data grid usage:
  - Each layer within VanMap can be preserved as a record in a data grid
  - By selecting which layers to compose, any desired presentation of the archived data can be assembled
  - The layers can be organized in a logical collection hierarchy by date of preservation
  - Preservation attributes are stored to identify all snapshots of desired layers that are available for a desired time period

5. Data Grids & Federation

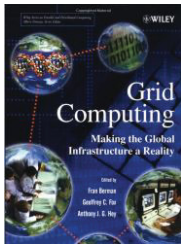
- Data grids provide the ability to name, organize, and manage data on distributed storage resources

- Federation provides a way to name, organize, and manage data on multiple data grids.

1998

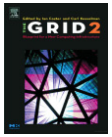


2003



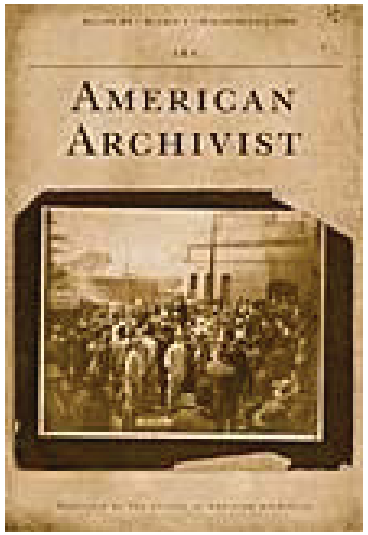
Tony Hey:
"The Data Deluge: An e-Science Perspective"

2004



Collaborative Science

- Grids in Context
  - Larry Smarr
- Computational Grids
  - Ian Foster and Carl Kesselman
- Distributed Supercomputing Applications
  - Paul Messina
- Realtime Widely Distributed Instrumentation
  - William E. Johnston
- **Data-Intensive Computing**
  - Reagan Moore, … Richard Marciano, …
- Teleimmersion
  - Tom DeFanti and Rick Stevens
- Application-Specific Tools
  - Henri Casanova, Jack Dongarra, …
- Compilers, Languages, and Libraries
  - Ken Kennedy
- Object-Based Approaches
  - Dennis Gannon, Andrew Grimshaw
- High-Performance Commodity Computing
  - Geoffrey Fox, Wojtek Furmanski
- The Globus Toolkit
  - Ian Foster, Carl Kesselman
- High-Performance Schedulers
  - Francine Berman
- High-Throughput Resource Management
  - Miron Livny, Rajesh Raman
- Instrumentation and Measurement
  - Jeffrey Hollingsworth, Bart Miller
- Performance Analysis and Visualization
  - Daniel Reed, Randy Ribler
- Security, Accounting, and Assurance
  - Clifford Neuman
- Computing Platforms
  - Andrew Chien
- Network Protocols
  - P.M. Melliar-Smith, Louise Moser
- Network Quality of Service
  - Roch Guerin, Henning Schultzrinne
- Operating Systems and Network Interfaces
  - Peter Druschel, Larry Peterson
- Network Infrastructure
  - *Jon Postel, Joe Touch*
- Testbeds: Bridges from Research to Infrastructure
  - *Charlie Catlett, John Toole*



Approach: Data Grids

American Archivist Journal, Volume 69 – Number 1 *(Spring/Summer 2006)*

"Building Preservation Environments with Data Grid Technology"
Reagan W. Moore

- *Manage technology evolution* for software and hardware systems
  - Data virtualization - manage data collection properties independently of the storage systems
    - Assert fixity properties on the data collection while storing in an evolving storage system
  - Trust virtualization - manage access controls and authentication independently of the storage systems
    - Assert fixity properties on the name spaces while storing across administrative domains

Infrastructure Independence

- Concept that the preserved records can be migrated from the current preservation environment into another choice of technology, while preserving authenticity and integrity.

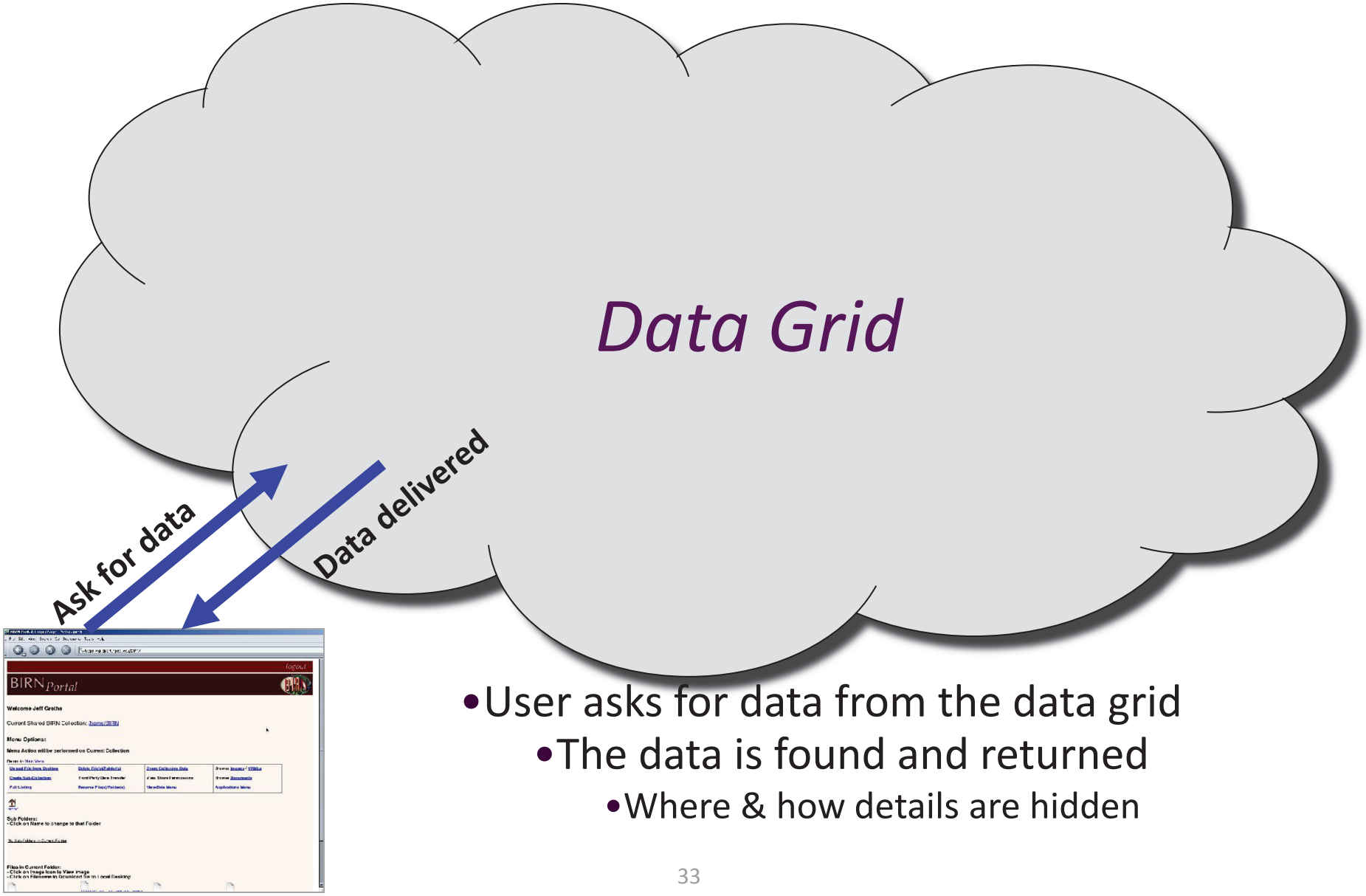
- Challenge is that all components of the preservation environment will evolve:
  - Hardware systems
  - Software systems
  - Encoding formats
  - Access mechanisms
  - Preservation attributes

What are Data Grids?

Data Grids are middleware services

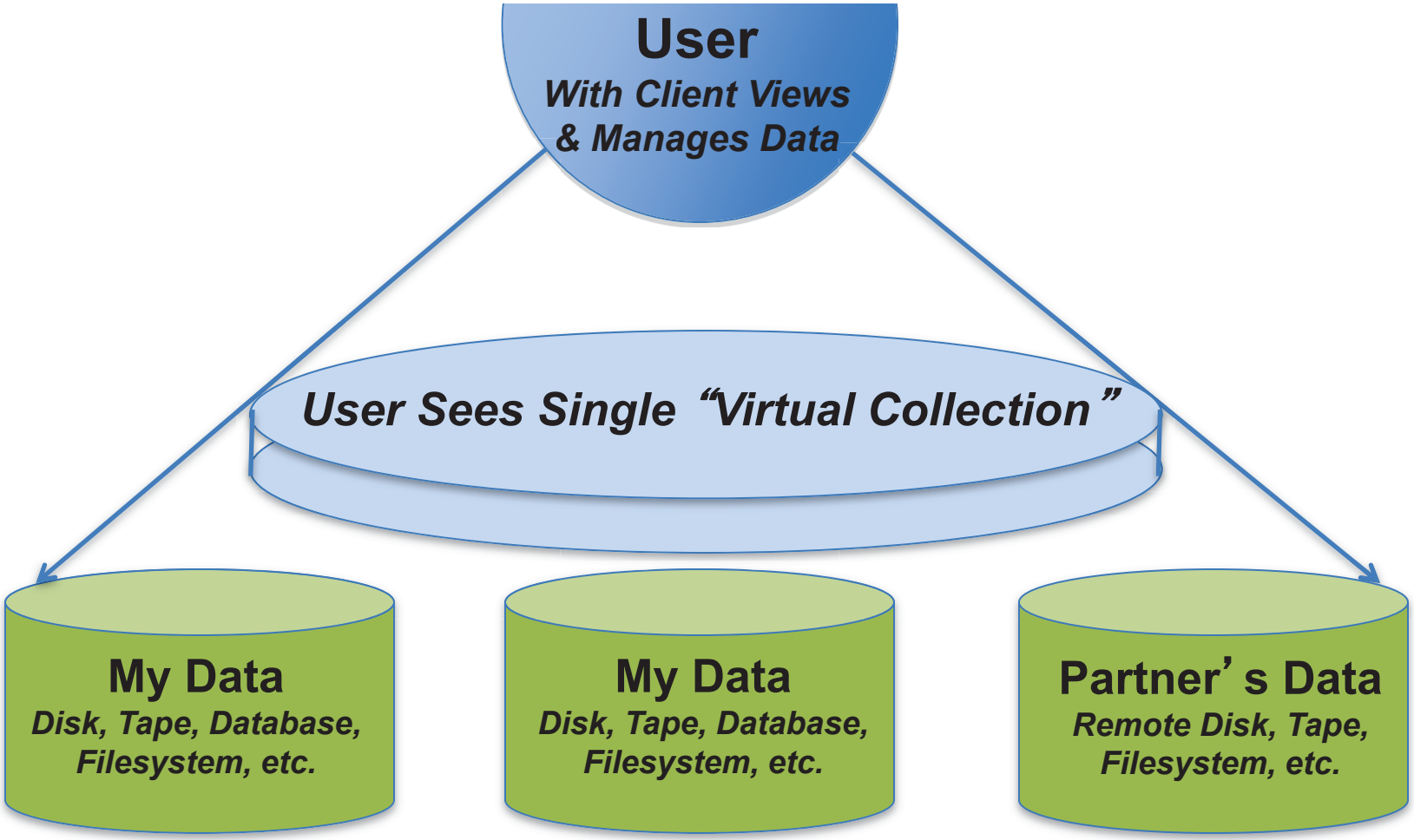
- Sitting between the applications and data providers
- Providing transparent and uniform access
- To diverse types of digital assets
  - Files, databases, streams, web, programs,…
  - Documents, images, data, sensor packets, tables,…
- From heterogeneous resources
  - File Systems, tape archives, sensor streams,…
- Distributed over a wide area network
  - Multiple administrative and security domains
- With users unaware of physical attributes of the data access
  - System addresses, paths, protocols,…

Using a Data Grid – *in Abstract*



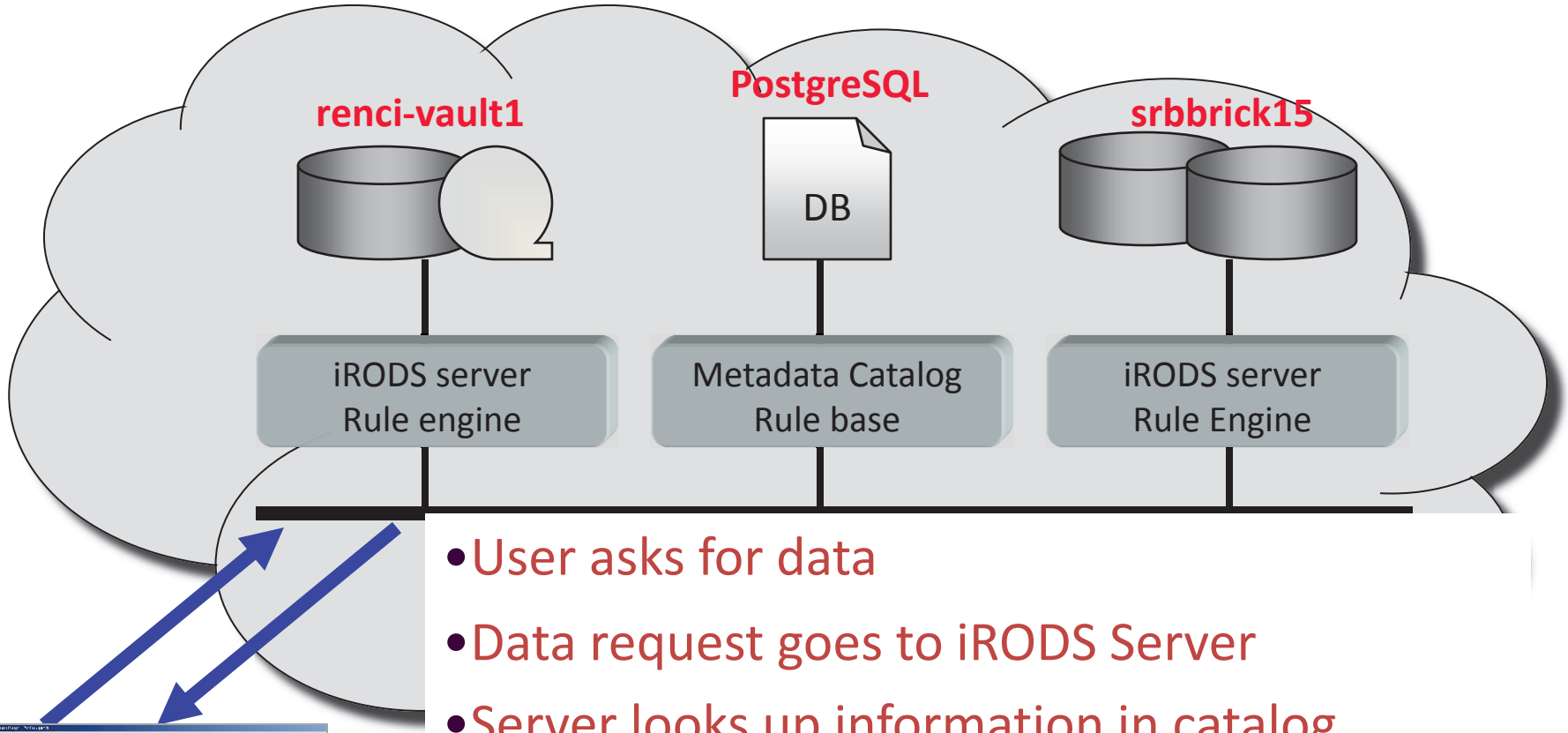
- User asks for data from the data grid
  - The data is found and returned
    - Where & how details are hidden

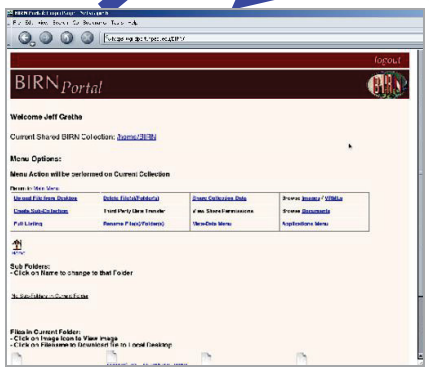
iRODS Shows Unified "Virtual Collection"



The iRODS Data System installs in a "layer" over existing or new data, letting you view, manage, and share part or all of diverse data in a unified Collection.

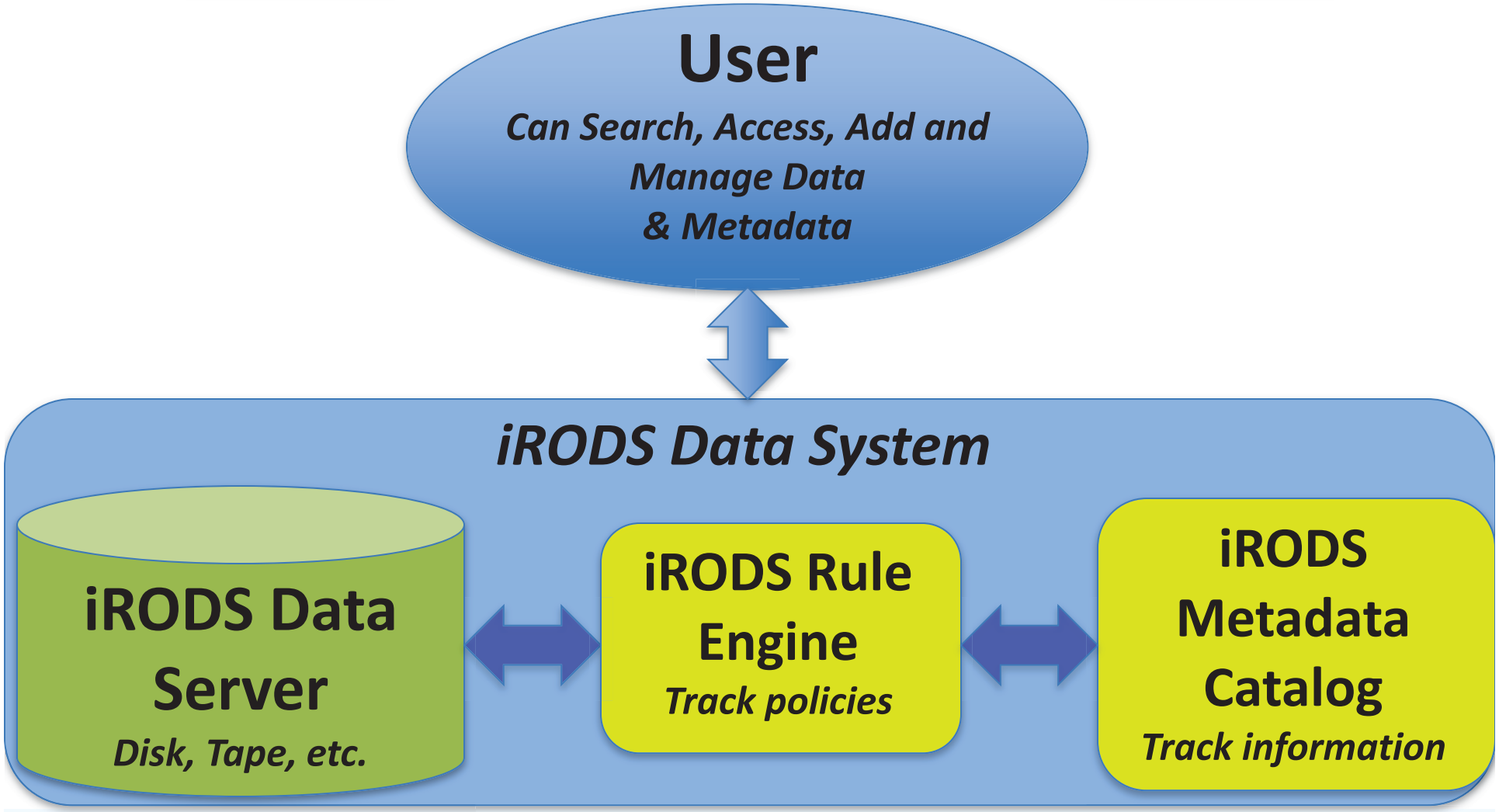
Using a Data Grid - *Details*





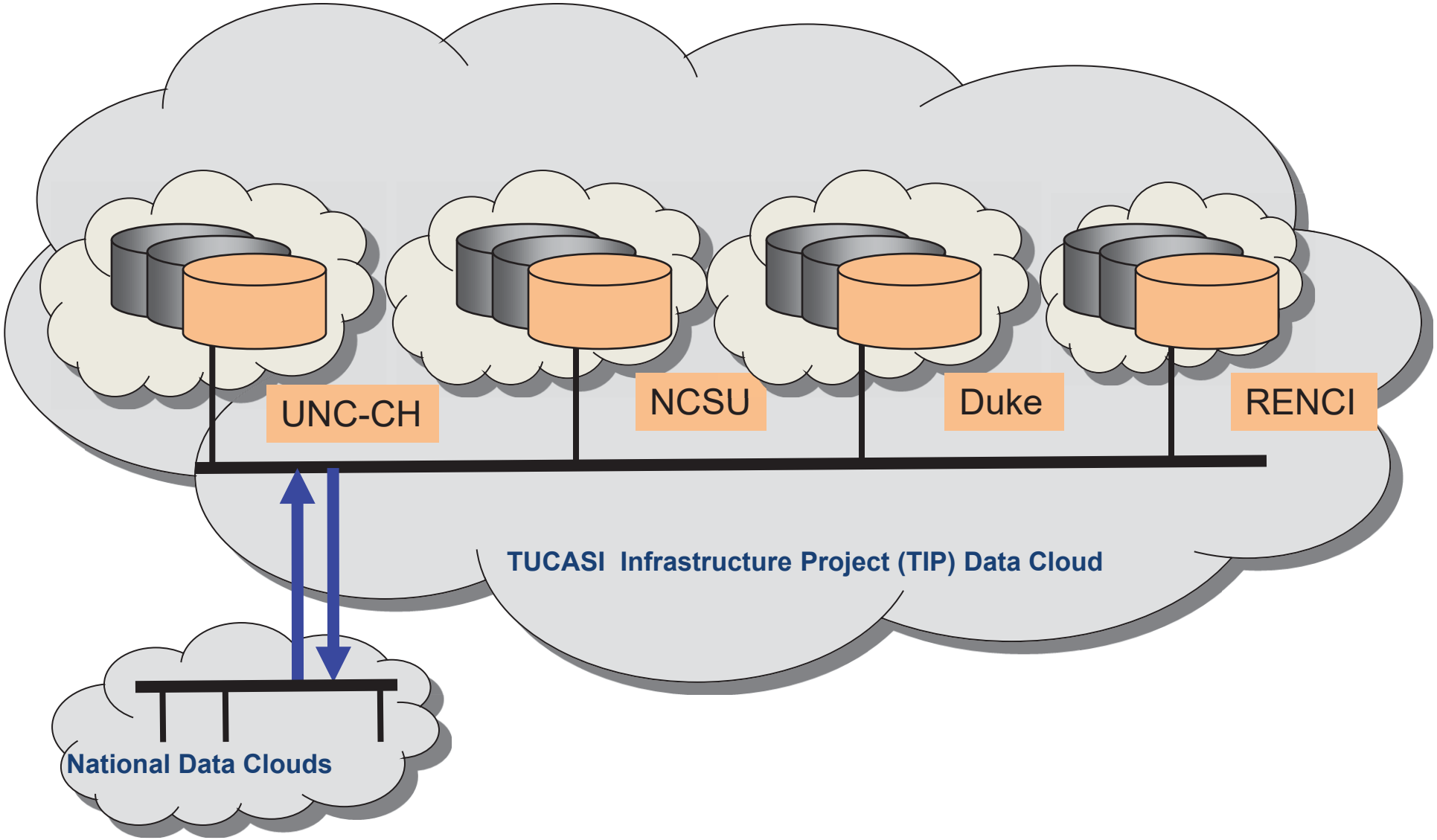
- User asks for data
- Data request goes to iRODS Server
- Server looks up information in catalog
- Catalog tells which iRODS server has data
- 1st server asks 2nd for data
- Server applies rules to control operations

Overview of iRODS Architecture



*Access data with Web-based Browser or iRODS GUI or Command Line clients.

Regional Data Infrastructure



Data Grids Are Trust Relationships

- Data-level Trust
  - Virtualization for integrity, authenticity, access provision, availability, data and metadata organization and management, community ownership and curation
- User-level Trust
  - Virtualization of authentication, authorization, auditing and accounting
- Resource-level Trust
  - Virtualization of administration and maintenance, appropriation (quota), availability and accesssibility
- These are Data Grid 1.0 level trusts

Data Grids Are Trust Relationships

- Policy-level Trust
  - Virtualization of Management, Organizational and Community Rules
- Service-level Trust
  - Virtualization of Operations and Services
- Execution-level Trust
  - Virtualization of distributed, parallel, asynchronous, delayed and/or remote execution
- These are Data Grid 2.0 level trusts

iRODS Rule

- Each rule defines
  - Event
  - Condition
  - Action chains (micro-services and rules)
  - Recovery chains
- Rule types
  - Atomic -- applied immediately
  - Deferred -- support deferred consistent constraints
  - Periodic -- typically used to validate assertions

Fundamental Data Management Concepts

- Data virtualization
  - Management of name spaces
    - Logical name space for users
    - Logical name space for storage resources
    - Logical name space for digital entities (files, URLs, SQL, tables, …)
    - Logical name space for metadata (user defined attributes)
  - Decoupling of access mechanisms from storage protocols
    - Standard operations for interacting with storage systems (80)
      - Posix I/O, bulk operations, latency management, registration, procedures, …
    - Standard client level operations for porting preferred interface (22)
      - C library calls, Unix commands, Java class library
      - Perl/Python/Windows load libraries, Perl/Python/Java/Windows web browsers, WSDL, Kepler workflow actors, DSpace and Fedora digital libraries, OAI-PMH, GridSphere portal, I/O redirection, GridFTP, OpenDAP, HDF5 library,Semplar MPI I/O, Cheshire
  - **Management of state information resulting from standard operations**

Fundamental Data Management Concepts

- Trust virtualization
  - Collection ownership of all deposited data
  - Users authenticate to collection, collection authenticates to remote storage system
  - Collection management of access controls
    - Roles for administration, read, write, execute, curate, audit, annotate
    - ACLs for each object
    - ACLs on metadata
    - ACLs on storage systems
    - Access controls remain invariant as data is moved within shared collection
  - Audit trails
  - End-to-end encryption

Levels of Virtualization

- Require metadata (state information, descriptive metadata) for six name spaces
  - Logical name space for users
  - Logical name space for digital entities (files, tables, URLs, SQL,…)
  - Logical name space for resources (storage systems, ORB, archives)
  - Logical name space for metadata (user defined metadata, extensible schema)
  - Logical name space for rules (assertions and constraints)
  - Logical name space for micro-services (data grid actions)
- Associate state information and descriptive information with each name space
- Virtualization of management policies

The Dialectic of Big Collaborations and the Need to Shape User Experiences

Big Data is a Big Deal

White House announcement:

http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

Big Data Across the Federal Government:

http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf

More then $200M in new commitments (NSF, HHS/NIH, DOE, DOD, DARPA, USGS)

Goal: "improve the ability to extract knowledge and insights from large and complex collections of digital data".

- **DataNet**
  - *Long-term preservation and access of data*
- **Software Infrastructure for Sustained Innovation** (SI^2)
- **Digging Into Data Challenge (NSF/NEH/IMLS & JISC)**
  - *Computational Humanities*
- **Cyber-Enabled Discovery and Innovation** (CDI)
  - *Data enabled science and engineering*
- **Core Techniques and Technologies for Advancing Big Data Science & Engineering** (BIGDATA)
- **Data Infrastructure Building Blocks** (DIBBs)
- **DataWay**
  - *National Infrastructure for Heterogeneous Data*

May 2007

Socializing CI: Networking the Humanities, Arts, and Social Sciences

CTWatch QUARTERLY

CTWatch Quarterly Current Issue | CTWatch Quarterly Archives | About CTWatch Quarterly

MAY 2007 VOLUME 3 NUMBER 2

SOCIALIZING CYBERINFRASTRUCTURE: NETWORKING THE HUMANITIES, ARTS, AND SOCIAL SCIENCES

Download this issue in PDF format (7.9 MB)

Search CTWatch Quarterly Archives

Google Search

30 funded
57 total

Community-Driven Development of Preservation Services

Funded Project Staff listed in Red and Blue

INTEGRATION & BUS DEV

UNC
- SALT
  - Richard Marciano
  - Chien-Yi Hou
- CDR
  - Dave Pcolar ++

POLICY / RULE DEV

West Virginia University
- Donald Adjeroh
- Frances Van Scoy

RENCI
- Leesa Brieger ++

DICE
- Michael Conway ++
- Reagan Moore
- Antoine de Torcy ++

UNC Libraries
- Steve Barr ++
- Greg Jansen ++

UNC Res. Comp. Svcs
- Bill Schulz ++

SILS Grad. Student team
- Heather Bowden ++
- Alex Chassanoff ++
- Christine Cheng ++
- William Miao ++
- Terrell Russell ++
- Jewel Ward ++

UNC CS Grad. Student team
- Tao Yu ++
- Hao Xu ++

STATE ARCHIVES & LIB

Michigan
- Caryn Wojcik
- Mark Harvey

North Carolina
- Kelly Eubank
- Jennifer Ricker ++
- Amy Rudersdorf ++
- Lisa Gregory ++
- Ed Southern --
- Megan Durden --
- IT
- Dean Farrell ++
- Druscie Simpson
- David Minor
- Chris Black --

Kentucky
- Glen McAninch
- Mark Myers ++

Kansas
- Scott Leonard

New York
- Bonnie Weddle
- Michael Martin ++
- Ann Marie Przybyla

California
- Chris Garmire
- Nancy Lenoil-Zimmelman
- Linda Johnson --
- Laren Metzer
- Renee Vincent-Finch --

UNIVERSITY ARCHIVES

Tufts University
- Eliot Wilczek
- Veronica Martzahl ++
- Anne Sauer

UNC Chapel Hill
- Will Owen ++
- Rich Szary ++

CULTURAL INSTITUTIONS

Getty Research Institute
- Joseph Shubitowski
- David Farneth
- Leah Prescott
- Sally Hubbard --
- Mahnaz Ghaznavi --
- Karim Boughida --

Smithsonian Institution Archives
- Riccardo Ferrante ++

SCHOOLS OF LIB & IS

UNC Chapel Hill
- Cal Lee ++

University of Wisconsin-Madison
- Kristin Eschenfelder ++

Legend	Collaborator Roles
Red	Funded
Blue	Cost-sharing
Brown	"Observer"
Black	None of the above
++	Added after project funded
--	At new institution

Topics

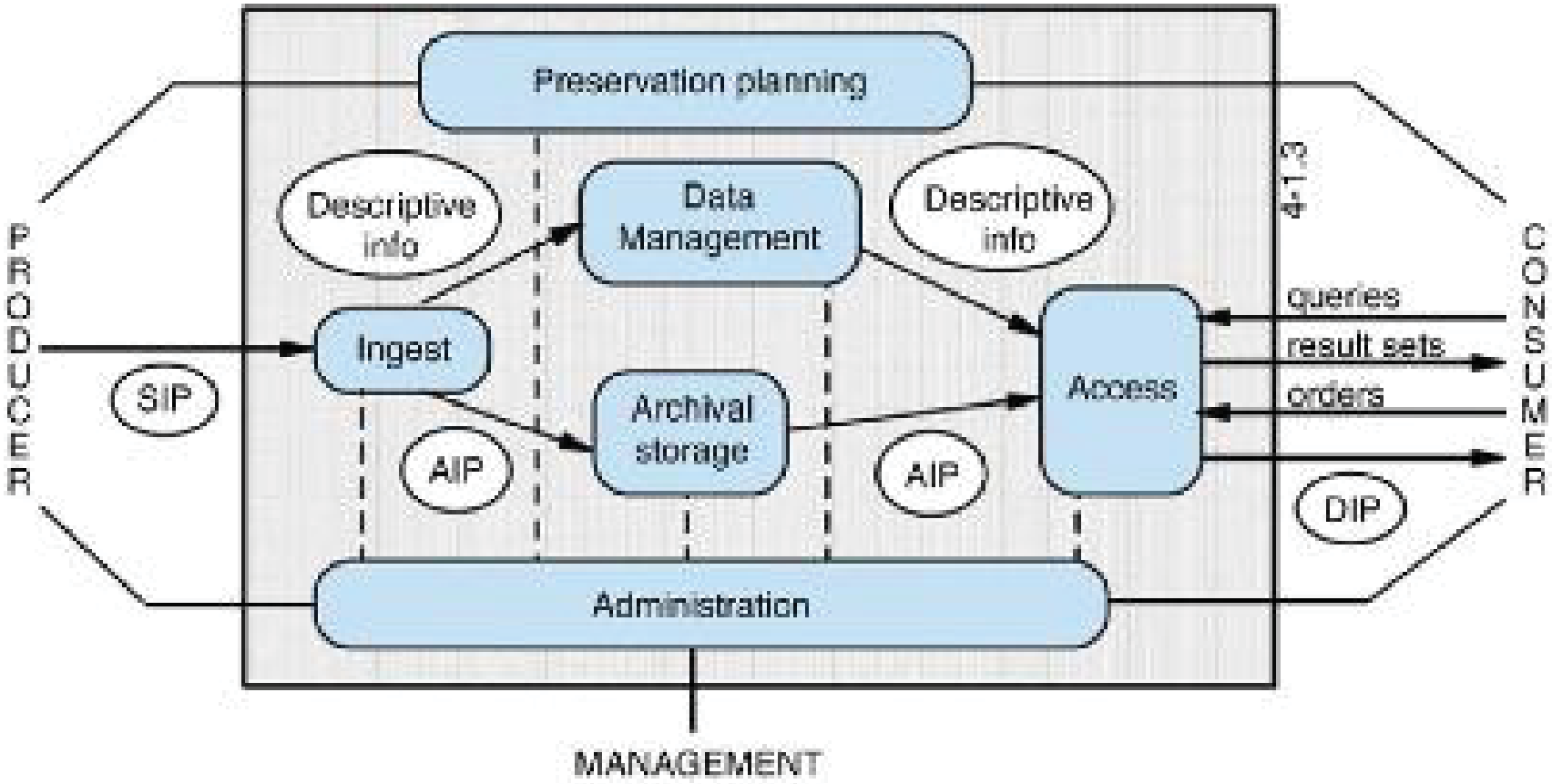
- 2003: **ISO 14271** - Open archival information system -- Reference model
- 2007: **TRAC** - Trustworthy Repositories Audit & Certification: Criteria and Checklist
- 2011: **ISO 16363** - Audit and certification of trustworthy digital repositories
- 2011: **ISO 16919** - Requirements for bodies providing audit and certification of candidate trustworthy digital repositories.

- e.g. **DCAPE**

Motivation

- OAIS adopted by digital preservation communities
  - Institutions begin declaring themselves “OAIS-Compliant”
- Section 1.5 (*ROAD MAP FOR DEVELOPMENT OF RELATED STANDARDS*) included an item **standard(s) for accreditation of archives**

OAIS Functional Entities



Ingest: This entity provides the services and functions to accept Submission Information Packages (SIPs) from Producers (or from internal elements under Administration control) and prepare the contents for storage and management within the archive. Ingest functions include receiving SIPs, performing quality assurance on SIPs, generating an Archival Information Package (AIP) which complies with the archive's data formatting and documentation standards, extracting Descriptive Information from the AIPs for inclusion in the archive database, and coordinating updates to Archival Storage and Data Management.

OAIS Functions of the Ingest

4.1.1.2 Ingest

The functions of the Ingest entity are illustrated in figure 4-2.

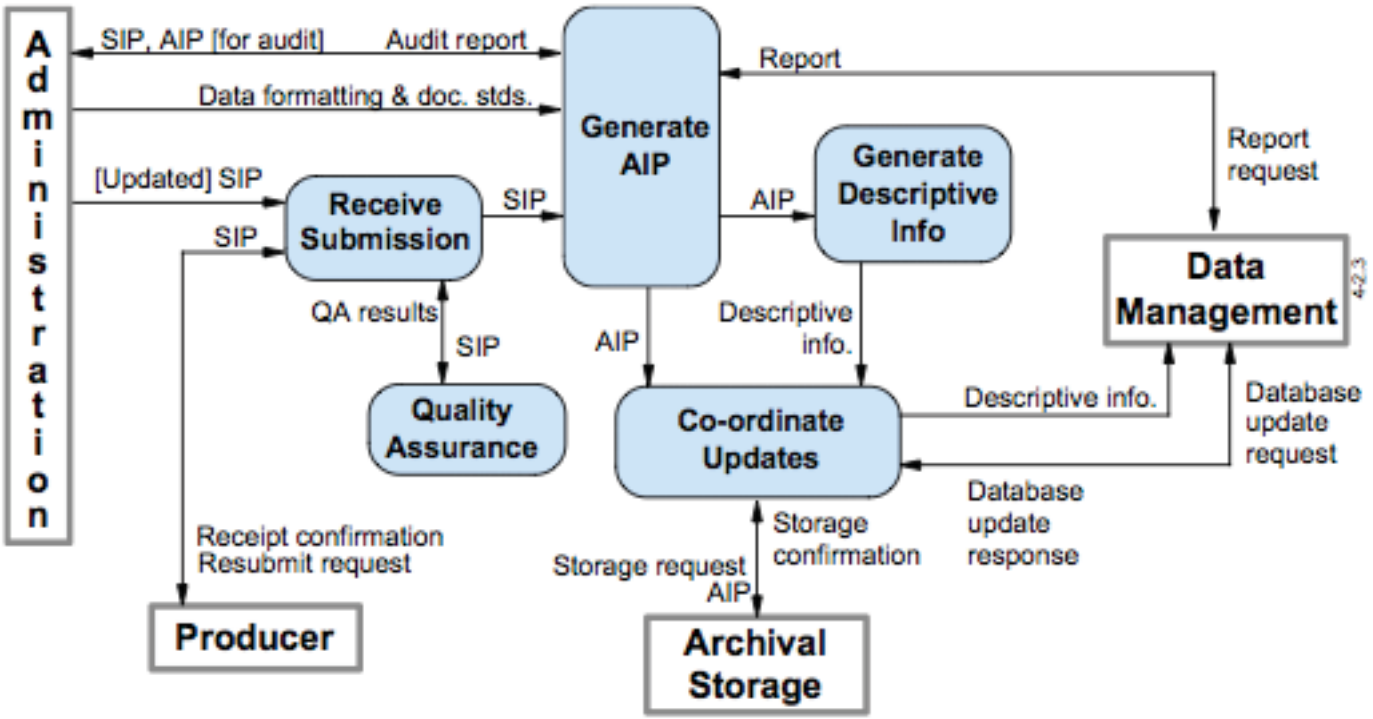


Figure 4-2: Functions of Ingest

Receive Submission function

- The **Receive Submission** function provides the appropriate storage capability or devices to receive a *SIP* from the Producer (or from Administration). Digital SIPs may be delivered via electronic transfer (e.g., FTP), loaded from media submitted to the archive, or simply mounted (e.g., CD-ROM) on the archive file system for access. Non-digital SIPs would likely be delivered by conventional shipping procedures. The Receive Submission function may represent a legal transfer of custody for the Content Information in the SIP, and may require that special access controls be placed on the contents. This function provides a *confirmation of receipt* of a SIP to the Producer, which may include a *request to resubmit* a SIP in the case of errors resulting from the SIP submission.

Quality Assurance function

- The **Quality Assurance** function validates (*QA results*) the successful transfer of the *SIP* to the staging area. For digital submissions, these mechanisms might include Cyclic Redundancy Checks (CRCs) or checksums associated with each data file, or the use of system log files to record and identify any file transfer or media read/write errors.

Generate AIP function

- The **Generate AIP** function transforms one or more *SIPs* into one or more *AIPs* that conform to the archive's *data formatting and documentation standards*. This may involve file format conversions, data representation conversions or reorganization of the content information in the SIPs. The Generate AIP function may issue *report requests* to Data Management to obtain *reports* of information needed by the Generate AIP function to produce the Descriptive Information that completes the AIP. This function sends *SIPs or AIPs for audit* to the Audit Submission function in Administration, and receives back an *audit report*.

Generate Descriptive Information function

- The **Generate Descriptive Information** function extracts Descriptive Information from the *AIPs* and collects *Descriptive Information* from other sources to provide to Coordinate Updates, and ultimately Data Management. This includes metadata to support searching and retrieving AIPs (e.g., who, what, when, where, why), and could also include special browse products (thumbnails, images) to be used by Finding Aids.

Coordinate Updates function

- The **Coordinate Updates** function is responsible for transferring the *AIPs* to Archival Storage and the *Descriptive Information* to Data Management. Transfer of the AI*P* includes a *storage request* and may represent an electronic, physical, or a virtual (i.e., data stays in place) transfer. After the transfer is completed and verified, Archival Storage returns a *storage confirmation* indicating (or verifying) the storage identification information for the AIP. The Coordinate Updates function also incorporates the storage identification information into the Descriptive Information for the AIP and transfers it to the Data Management entity along with a *database update request*. In return, Data Management provides a *database update response* indicating the status of the update. Data Management updates may take place without a corresponding Archival Storage transfer when the SIP contains Descriptive Information for an AIP already in Archival Storage.

RLG/NARA Assessment

- Developed 105 rules that implement the TRAC assessment criteria

90	*Verify descriptive metadata and source against SIP template and set SIP compliance flag*
91	*Verify descriptive metadata against semantic term list*
92	*Verify status of metadata catalog backup (create a snapshot of metadata catalog)*
93	*Verify consistency of preservation metadata after hardware change or error*

ISO 16363

- Based largely on TRAC
- Organizational Infrastructure
  - e.g. The repository shall have a documented history of the changes to its operations, procedures, software, and hardware.
- Digital Object Management
  - e.g. The repository shall have access to necessary tools and resources to provide authoritative Representation Information for all of the digital objects it contains.
- Infrastructure and Security Risk Management
  - eg. The repository shall have procedures in place to evaluate when changes are needed to current software.

ISO 16919

- Hierarchy of ISO standards concerned with good auditing
- ISO 16919 positioned within this hierarchy
  - Ensure good practices can be applied to the evaluation of the trustworthiness of digital repositories using ISO 16363

DCAPE Goals: http://dcape.org

- The trusted digital repository infrastructure will be assembled from state-of-the-art rule-based data management systems, commodity storage systems, and sustainable preservation services.
- The software infrastructure will automate many of the administrative tasks associated with the management of archival repositories.
  - Tasks will include: authentication, replication, migration, obsolete file management, preservation metadata management, deduplication, virus checks, etc.

DCAPE Policy to Rule to Microservice mappings…

- http://dcape.org
- **Audit and Certification of Trustworthy Digital Repositories** (ISO/NP 16363). This document is currently going through the ISO review process in the same way as the OAIS Reference Model (ISO 14721), namely via ISO TC20/SC13. The wiki for the working group can be found at:
  - http://wiki.digitalrepositoryauditandcertification.org/bin/view/Main/WebHome

Initial ISO MOIMS-rac Capabilities and Mapping to *DCAPE* rules
RAC No.'s are from the "Combined Annotated document" Wiki page
http://wiki.digitalrepositoryauditandcertification.org/bin/view/Main/CombinedMetricsDocumentsFollowingFaceToFace
Accessed on Sep. 2009

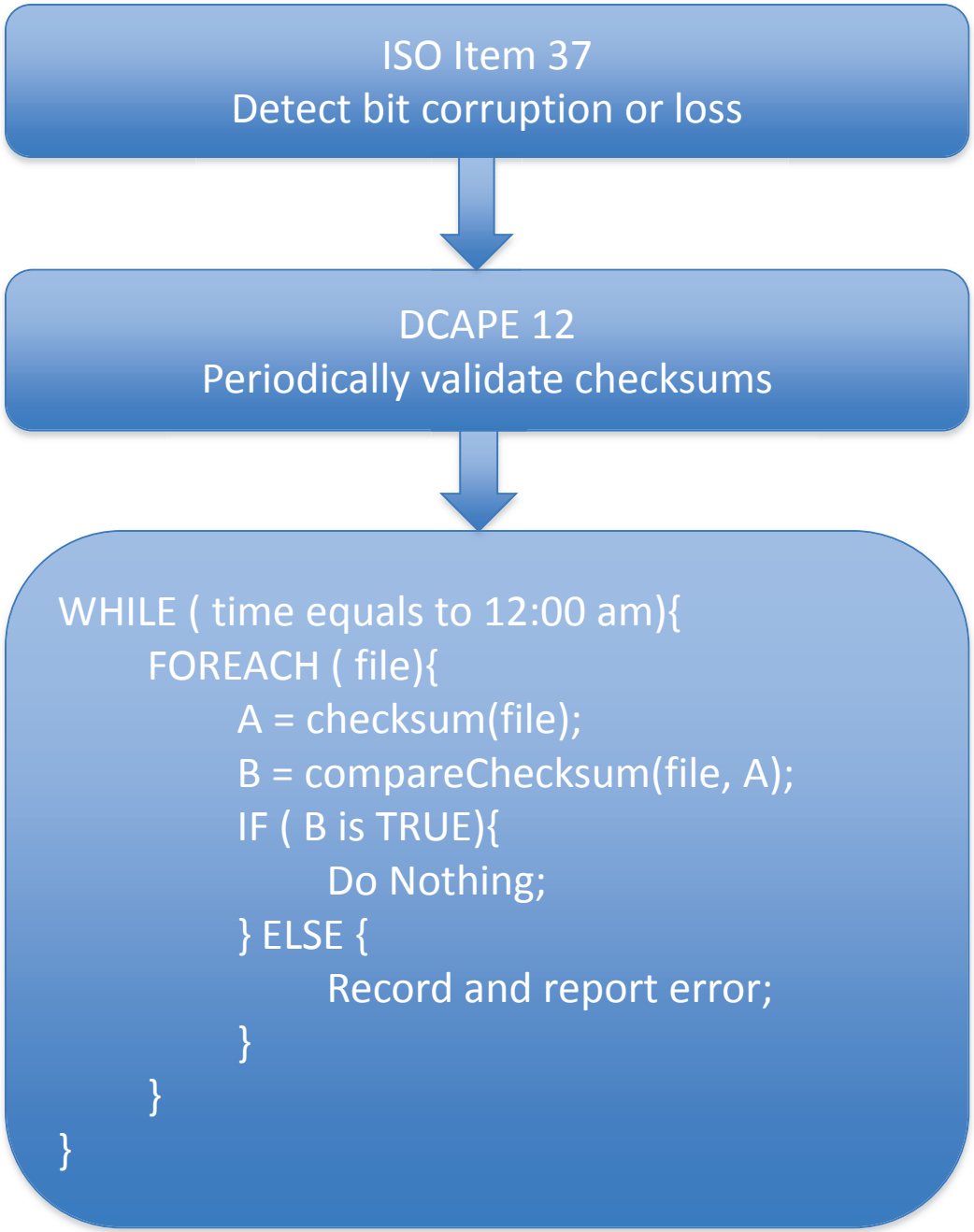
ISO Item	RAC No.	DCAPE Item	ISO Criteria	DCAPE Machine-Actionable Rule
1	A3.2.2 A5.1.3 A.5.1.4 A5.2		Address liability and challenges to ownership/rights.	Map from submission template to access and distribution controls
2	B1.1	DCAPE 4	Identify the content information and the information properties that the repository will preserve.	Define templates that specify required metadata and parameters for rules that are required to enforce properties
3	B1.1.2		Maintain a record of the Content Information and the Information Properties that it will preserve.	Link submission and policy templates to the preserved collection
4	B1.3	DCAPE 3	Specify Submission Information Package format (SIP)	Define templates that specify structure of a SIP and required content of a SIP.
5	B1.4	DCAPE 1	Verify the depositor of all materials.	Ingest data through a staging area that has a separate account for each depositor.
6	B1.5	DCAPE 6	Verify each SIP for completeness and correctness	Compare content of each SIP against template.
7	B1.6	DCAPE 8	Maintain the chain of custody during preservation.	Manage audit trails that document the identity of the archivist initiating the task
8	B1.7	DCAPE 22	Document the ingestion process and report to the producer	Send e-mail message to producer when process flags are set.
9	B1.8	DCAPE 10	Document administration processes that are relevant to content acquisition.	Maintain list of rules that govern management of the archives
10	B2.1 B2.1.1	DCAPE 13	Specify Archival Information Package format (AIP)	Define templates that specify structure of an AIP and required content of an AIP.
11	B2.1.2		Label the types of AIPs.	Store AIP type with each collection.
12	B2.2	DCAPE 13	Specify how AIPs are constructed from SIPs.	Define transformation rule based on parsing of SIP template and AIP template
13	B2.3 B2.3.1	DCAPE 14	Document the final disposition of all SIPs	Maintain an audit trail for all SIPs
14	B2.4 B2.4.1 B2.4.1.1 B2.4.1.2 B2.4.1.3		Generate persistent, unique identifiers for all AIPs.	Define unique persistent logical name for each AIP

Refined mapping process

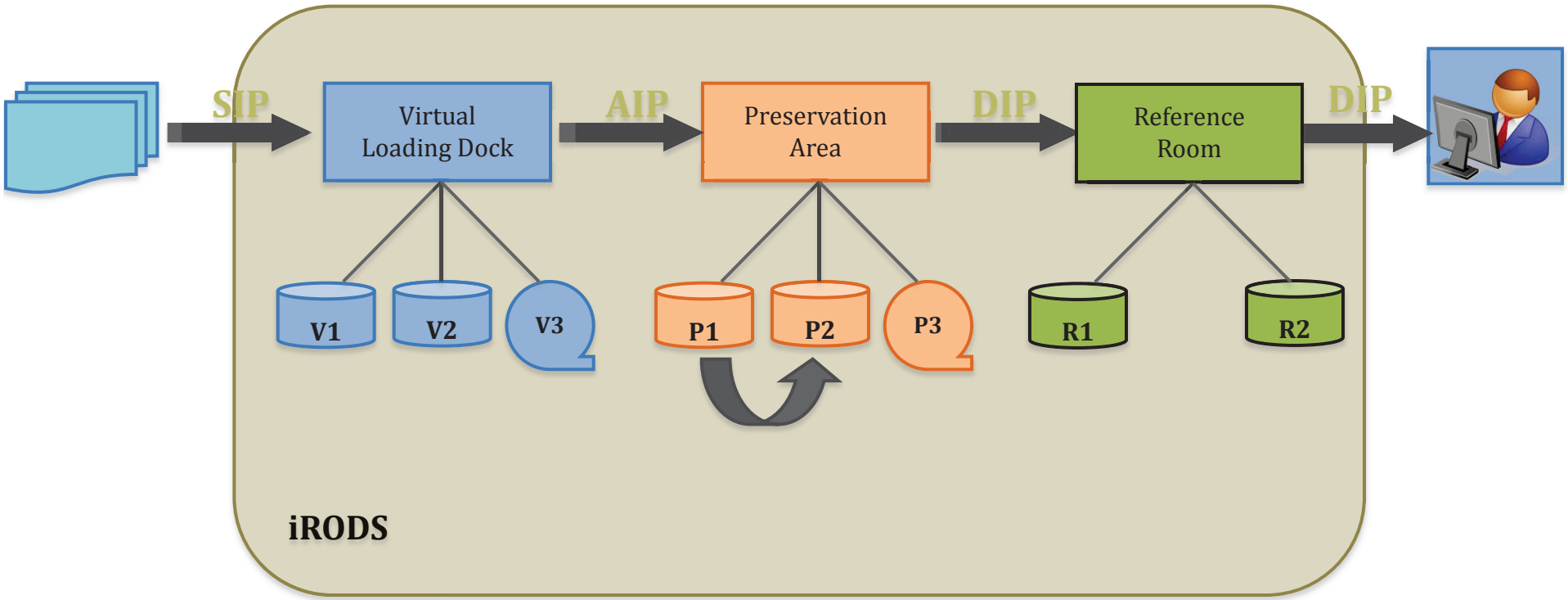
- OAIS analysis:
  - http://dcape.org/docs/Final_products/DCAPE%20iRODS%20Rules.pdf
  - See III.C.1 (Archival Storage – Manage Storage and Hierarchy – Run error checks)
- ISO analysis
  - http://wiki.digitalrepositoryauditandcertification.org/bin/view/Main/CombinedMetricsDocumentsFollowingFaceToFace#C1_1_3_Repository_has_effective
- DCAPE policy creation
  - See next slide

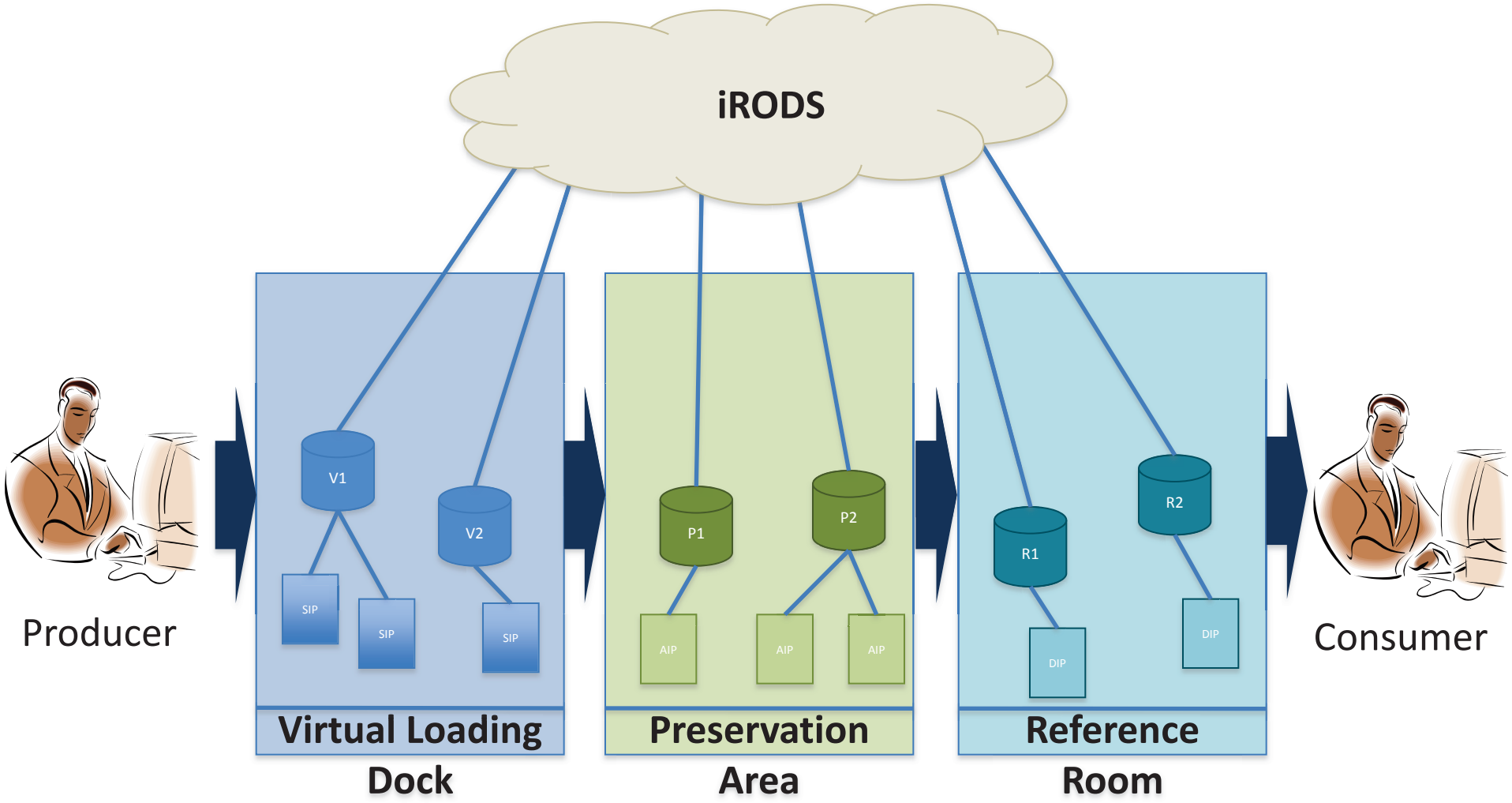
ISO Item	RAC No.	DCAPE Item	ISO Criteria	DCAPE Machine-Actionable Rule
37	C1.1.3	DCAPE 12	Repository has effective mechanisms to detect bit corruption or loss	Periodically validate checksums

Policies:

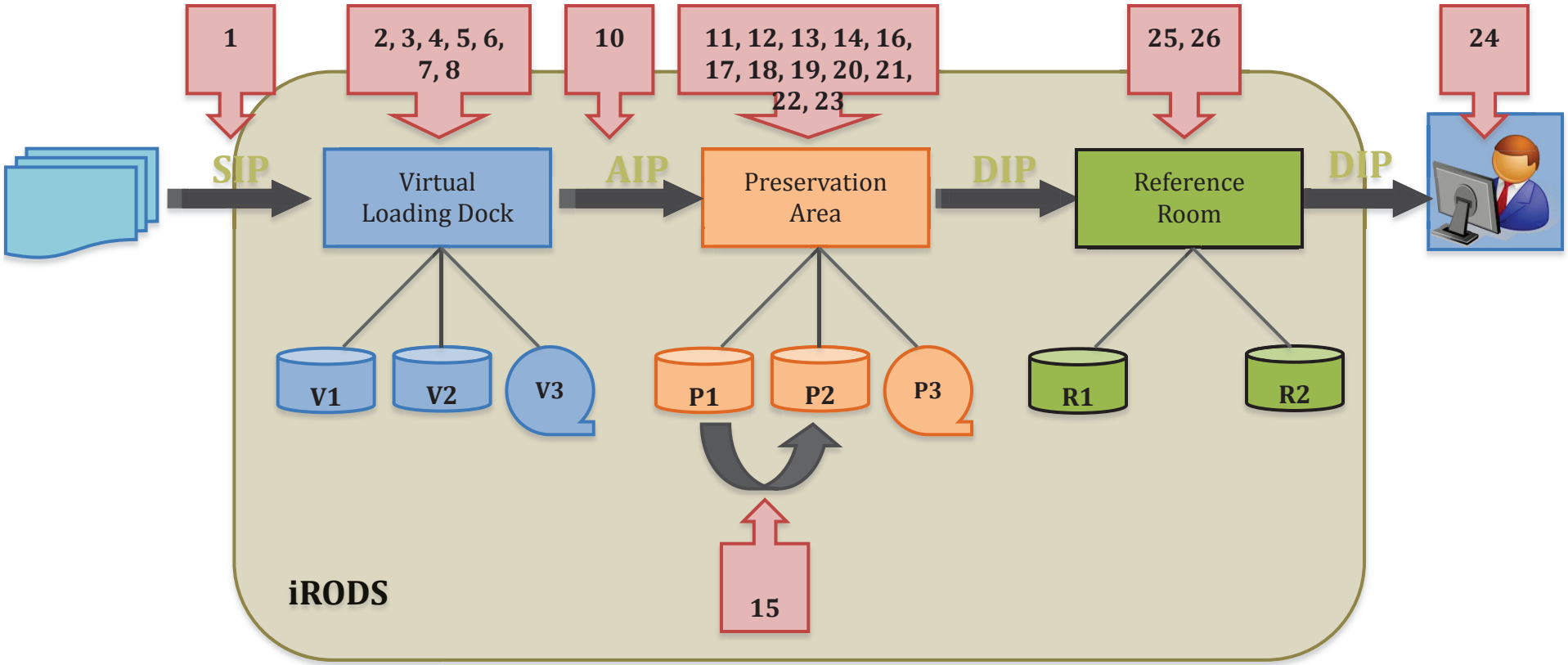

ISO Item 37
Detect bit corruption or loss
DCAPE 12
Periodically validate checksums
WHILE (time equals to 12:00 am){
FOREACH (file){
A = checksum(file);
B = compareChecksum(file, A);
IF (B is TRUE){
Do Nothing;
} ELSE {
Record and report error;
}
}
}

DCAPE Framework




iRODS
Producer
V1
V2
SIP
SIP
SIP
Virtual Loading Dock
P1
P2
AIP
AIP
AIP
Preservation Area
R1
R2
DIP
DIP
Reference Room
Consumer

DCAPE Policies



DCAPE Interface