Data Foundations and the DataNet Federation Consortium

Reagan Moore
School of Information and Library Science
University of North Carolina, Chapel Hill

DataNet Federation Consortium

Data Driven Science

- Implement national cyberinfrastructure
  - Federate existing discipline-specific data management systems to enable national research collaborations
- Enable collaborative research on shared data collections
  - Manage collection life cycle as the user community broadens
- Enable reproducible research
  - Manage data collections, workflows, and data flows

Cyber-infrastructure Partners:
Univ. of North Carolina, Chapel Hill
Univ. of California, San Diego
Drexel University
University of Arizona
University of Virginia
Arizona State University

Science and Engineering Initiatives:
Dataverse
Science Observatory Network - SciON
Temporal Dynamics of Learning Center
HIVE
Cyverse
Hydroshare

Federated Systems
DFC
SEAD
TerraPop
DataONE

National Science Foundation Cooperative Agreement: OCI-0940841

i·R·O·D·S

Policy-based data management

Disciplines using the iRODS data grid

1. Astrophysics	Auger supernova search	Shared collection
2. Atmospheric science	NASA Langley Atmospheric Sciences Center	Shared collection
3. Biology	Phylogenetics at CC IN2P3	Shared collection
4. Climate	NOAA National Climatic Data Center	Ingestion cache for archive
5. Cognitive Science	Temporal Dynamics of Learning Center	Shared collection
6. Computer Science	GENI experimental network	Archive
7. Cosmic Ray	AMS experiment on the International Space Station	Shared collection
8. Dark Matter Physics	Edelweiss II	Shared collection
9. Earth Science	NASA Center for Climate Simulations	Digital Library
10. Ecology	CEED Caveat Emptor Ecological Data	Digital Library
11. Engineering	CIBER-U	Digital Library
12. High Energy Physics	BaBar / Stanford Linear Accelerator	Shared collection / Archive
13. Hydrology	Institute for the Environment, UNC-CH; Hydroshare	Digital Library / portal
14. Genomics	Wellcome Trust Sanger Institute, UNC-CH	Digital Library
15. Medicine	Lineberger Cancer Institute	Patient data
16. Neuroscience	International Neuroinformatics Coordinating Facility	Shared collection
17. Neutrino Physics	T2K and dChooz neutrino experiments	Project collections
18. Oceanography	Science Observatory Network	Archive
19. Optical Astronomy	National Optical Astronomy Observatory	Archive
20. Particle Physics	Indra multi-detector collaboration at IN2P3	Project collection
21. Plant genetics	Cyverse	Collaboration environment
22. Quantum Chromodynamics	IN2P3	Project collection
23. Radio Astronomy	Cyber Square Kilometer Array, TREND, BAOradio	Digital Library
24. Seismology	Southern California Earthquake Center	Digital Library
25. Social Science	Odum Research Institute, Dataverse, TerraPop	Digital Library

Data Foundations

- Are there basic principles that govern all data management applications?
  - Can a single data fabric support all applications?
- What is the difference between a file system, a research collection, a digital library, an archive, a processing pipeline?
  - Choice of policies enforced by the system
  - Operations that are performed

Computer Actionable Definitions for Data, Information, Knowledge

Definition

• Data	objects	bits
• Information	names	metadata
• Knowledge	relationships between names	procedures
• Wisdom	relationships between relationships	policy points

Infrastructure

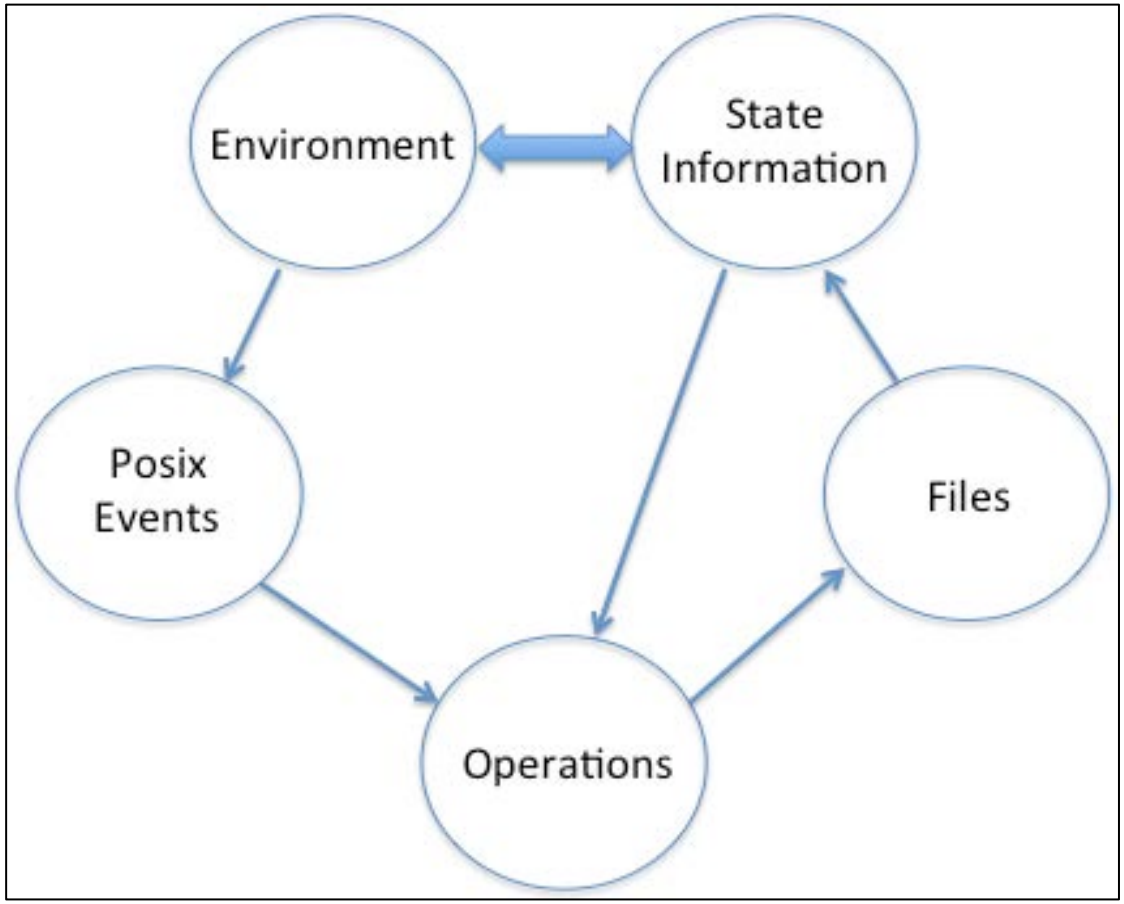
• Data	bits	File systems
• Information	metadata	Database
• Knowledge	procedures	Workflows
• Wisdom	policy enforcement points	Rule base

File Systems

File systems virtualize interactions with disk

Map from file name to a location on disk

Manage state information for each file
- Name
- Owner
- Access controls
- Size
- Creation date
- Modification date
- Directory name



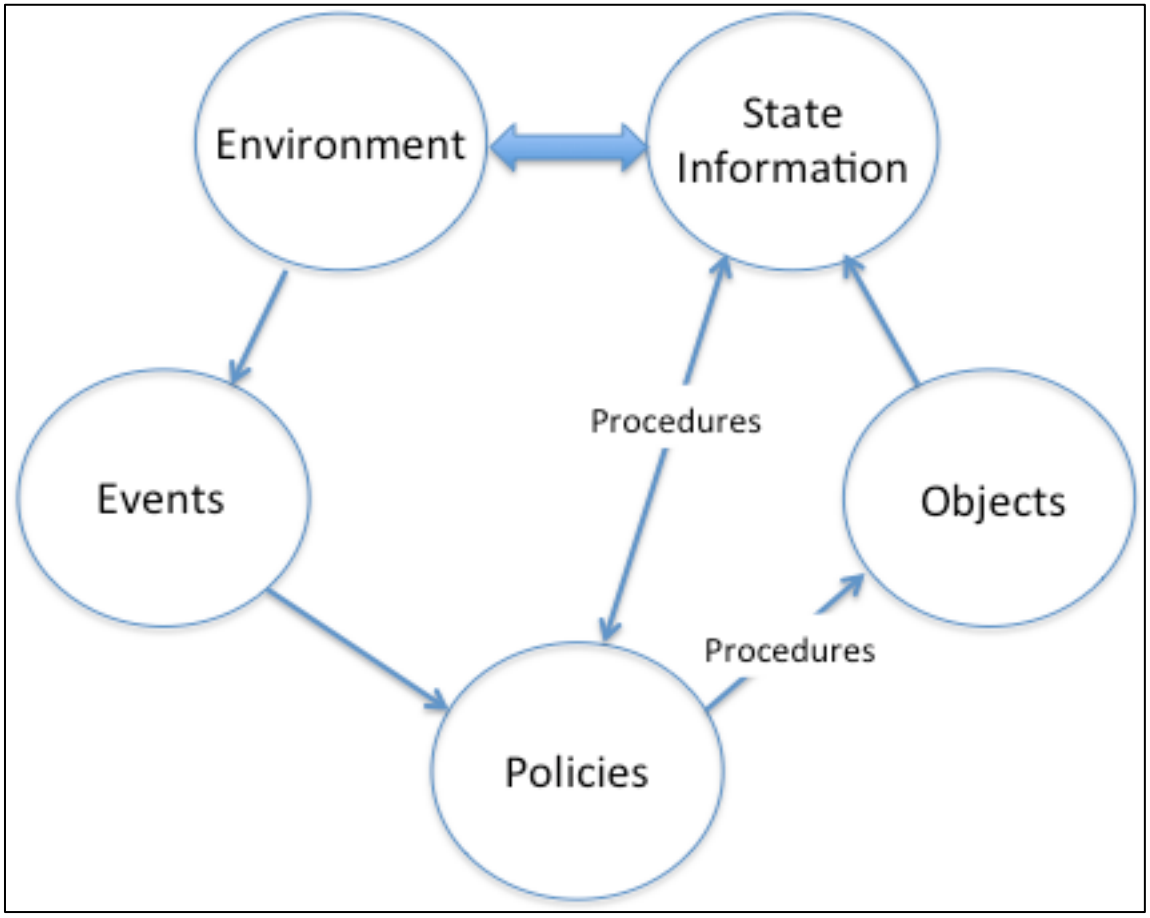
Operations: Open, close, read, write, seek, stat, mkdir

Policy-Based Data Management

Generalize
- Interactions – Policies
- Operations – Procedures
- Files – Objects

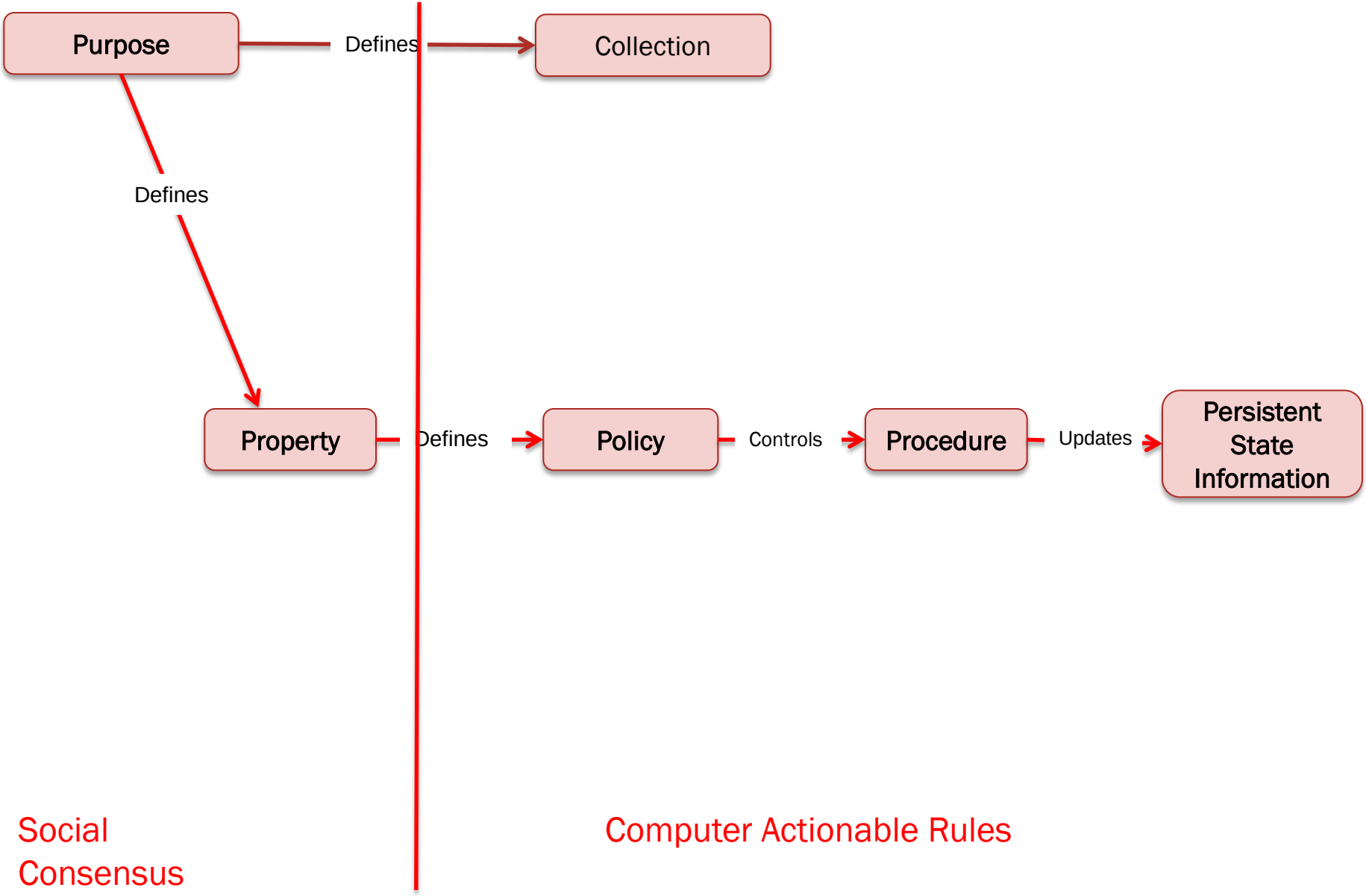
Trap events at policy enforcement points

Manage extensive set of state information about
- Files
- Users
- Storage systems
- Collections
- Policies
- Procedures
- Events

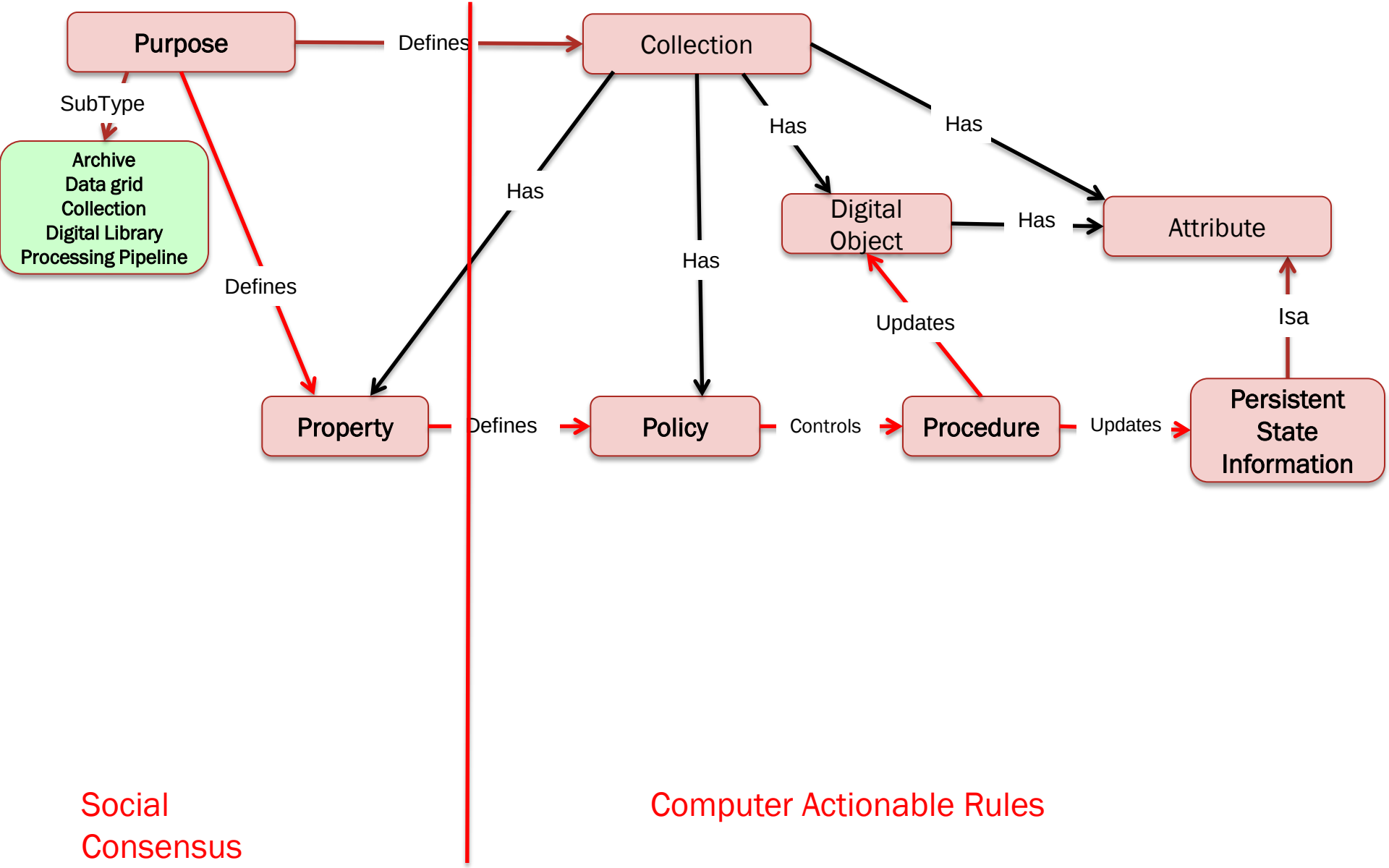


Policies: replication, retention, caching, distribution, …

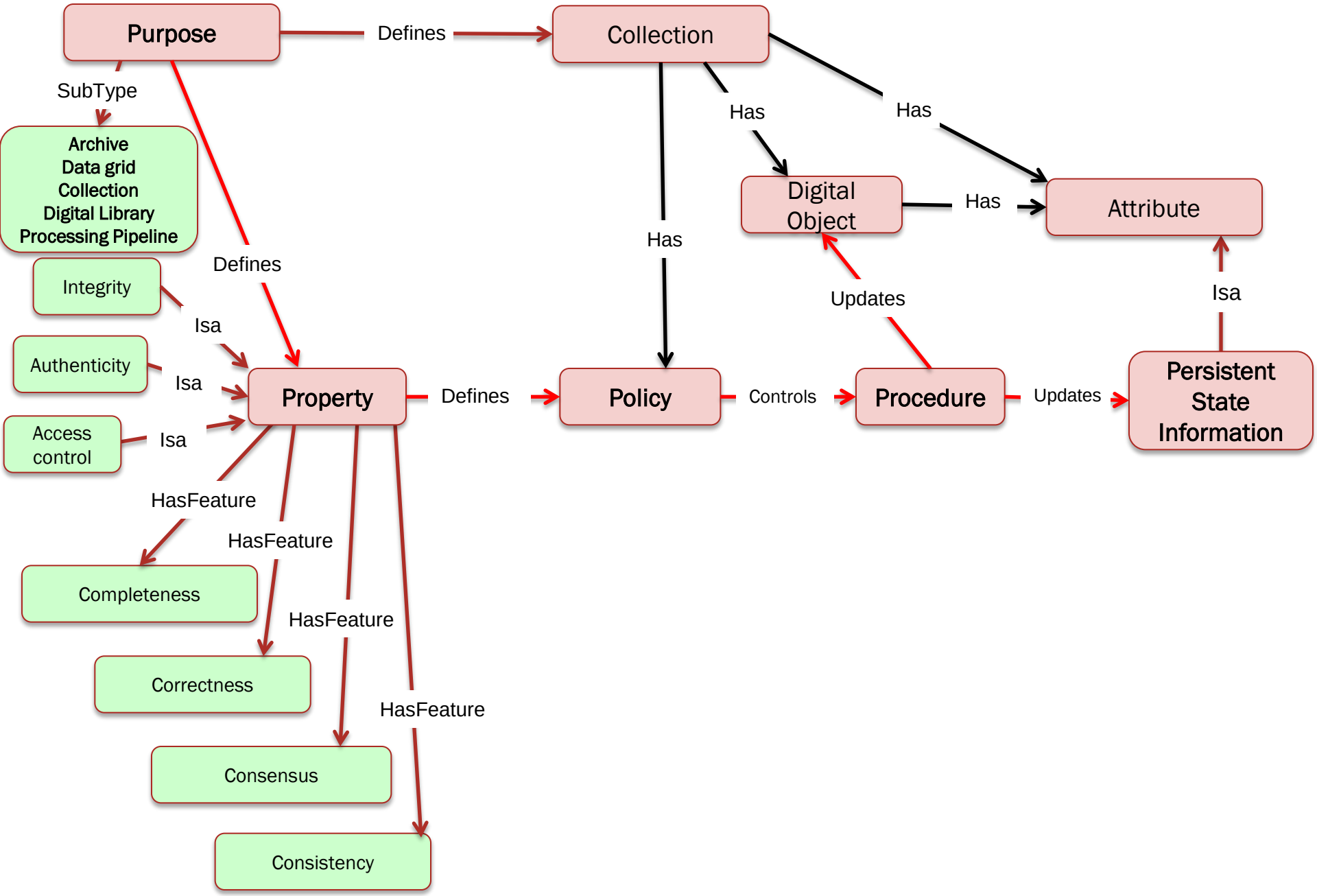
Concept Graph



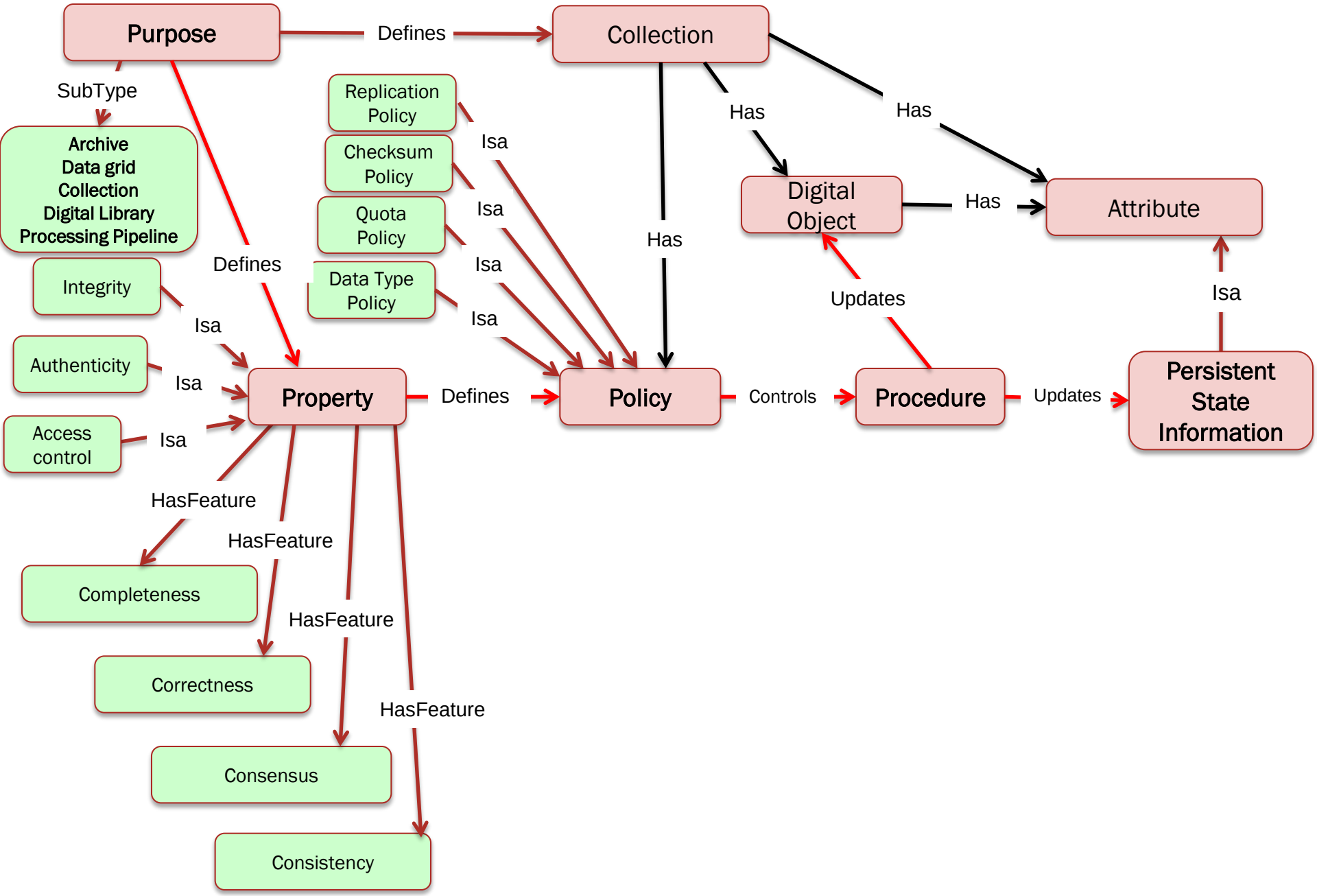
Collection


Purpose
Defines
Collection
SubType
Archive
Data grid
Collection
Digital Library
Processing Pipeline
Has
Has
Has
Digital Object
Has
Attribute
Defines
Has
Updates
Isa
Property
Defines
Policy
Controls
Procedure
Updates
Persistent State Information
Social Consensus
Computer Actionable Rules

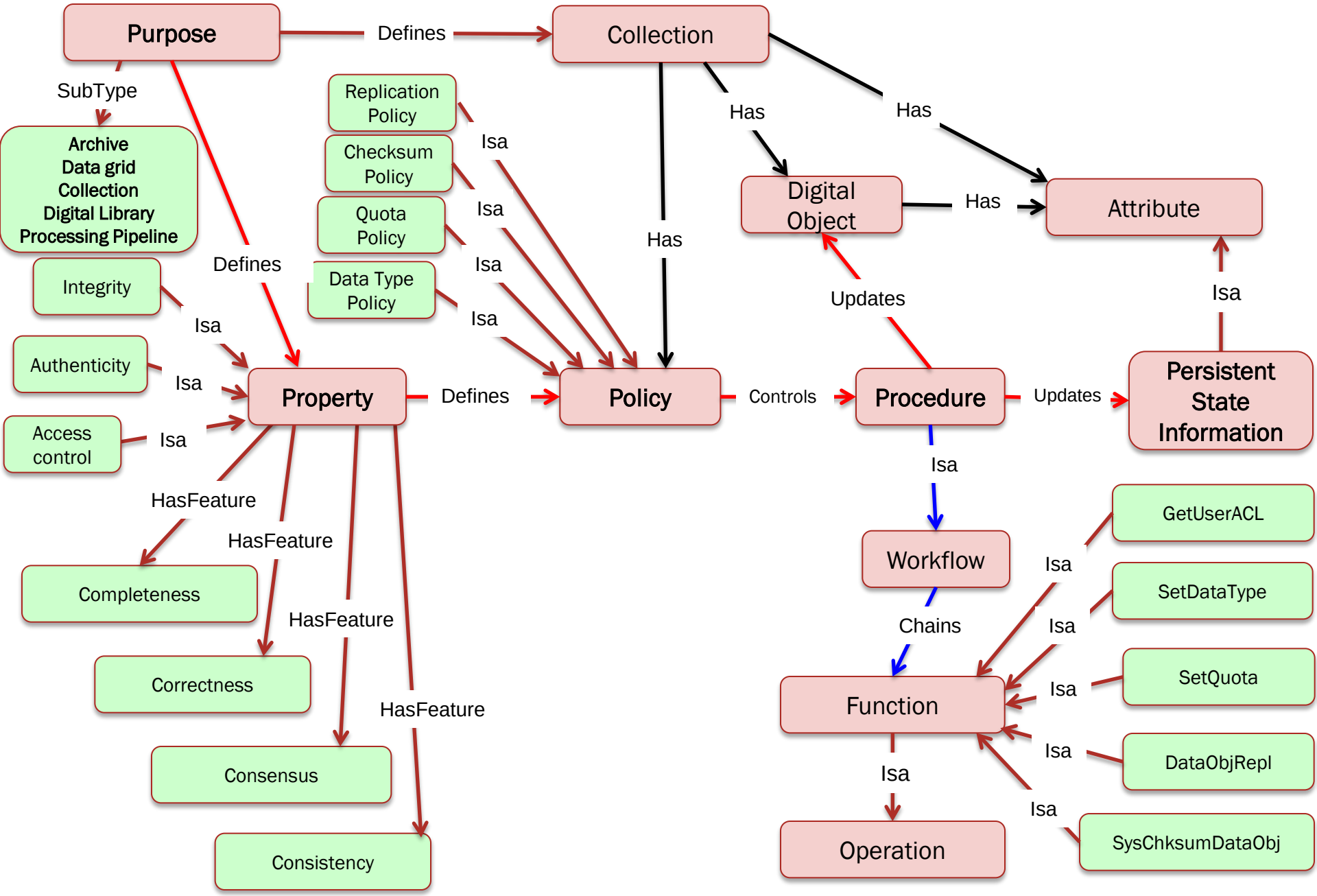
Properties


Purpose
Defines
Collection
SubType
Archive
Data grid
Collection
Digital Library
Processing Pipeline
Has
Has
Digital Object
Has
Attribute
Has
Integrity
Defines
Isa
Authenticity
Isa
Property
Defines
Policy
Controls
Procedure
Updates
Persistent State Information
Updates
Isa
Access control
Isa
HasFeature
HasFeature
HasFeature
HasFeature
Completeness
Correctness
Consensus
Consistency

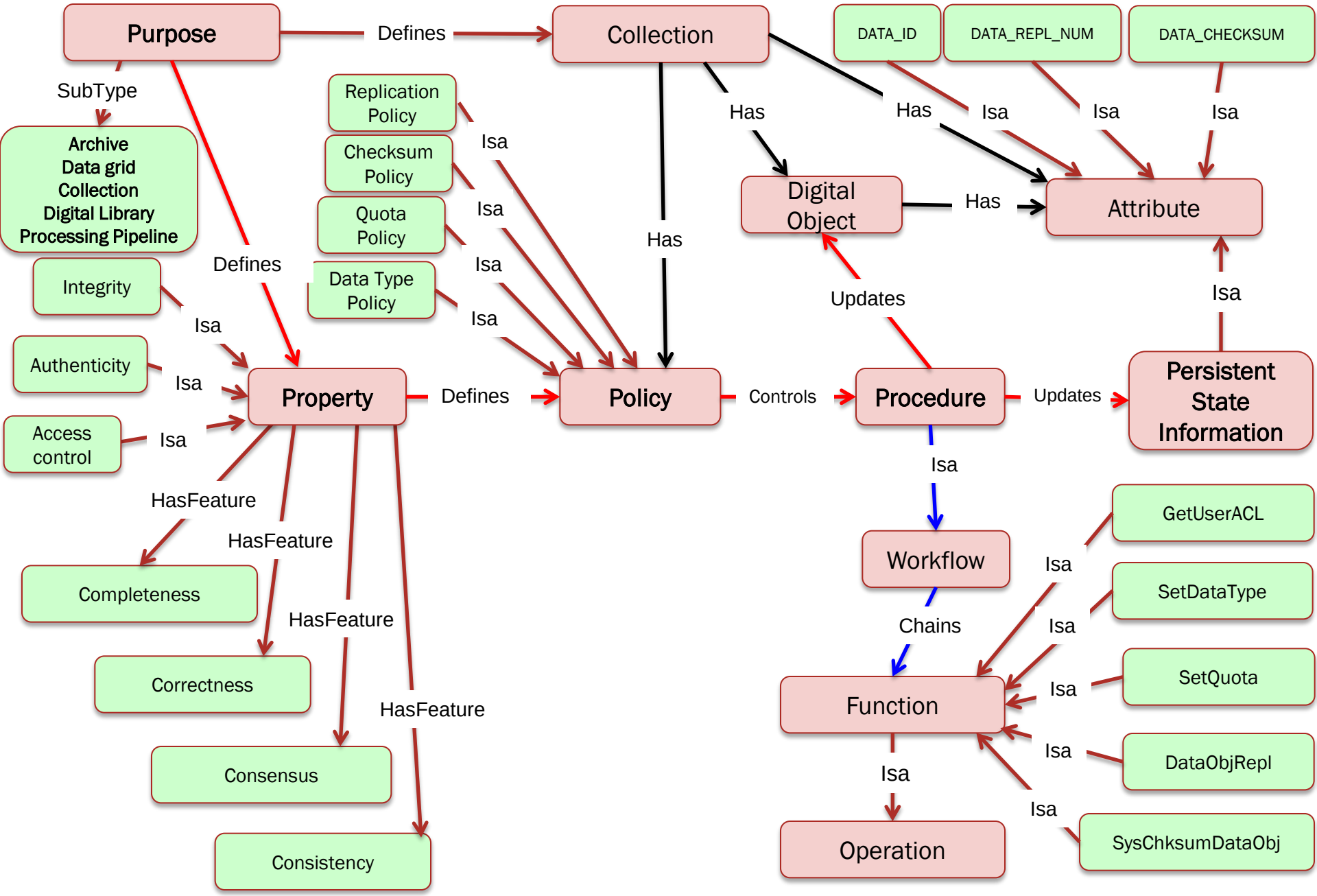
Policies


Purpose
Defines
Collection
SubType
Archive
Data grid
Collection
Digital Library
Processing Pipeline
Replication Policy
Checksum Policy
Quota Policy
Data Type Policy
Isa
Has
Has
Digital Object
Has
Attribute
Integrity
Defines
Has
Updates
Isa
Isa
Authenticity
Isa
Property
Defines
Policy
Controls
Procedure
Updates
Persistent State Information
Access control
Isa
HasFeature
HasFeature
HasFeature
HasFeature
Completeness
Correctness
Consensus
Consistency

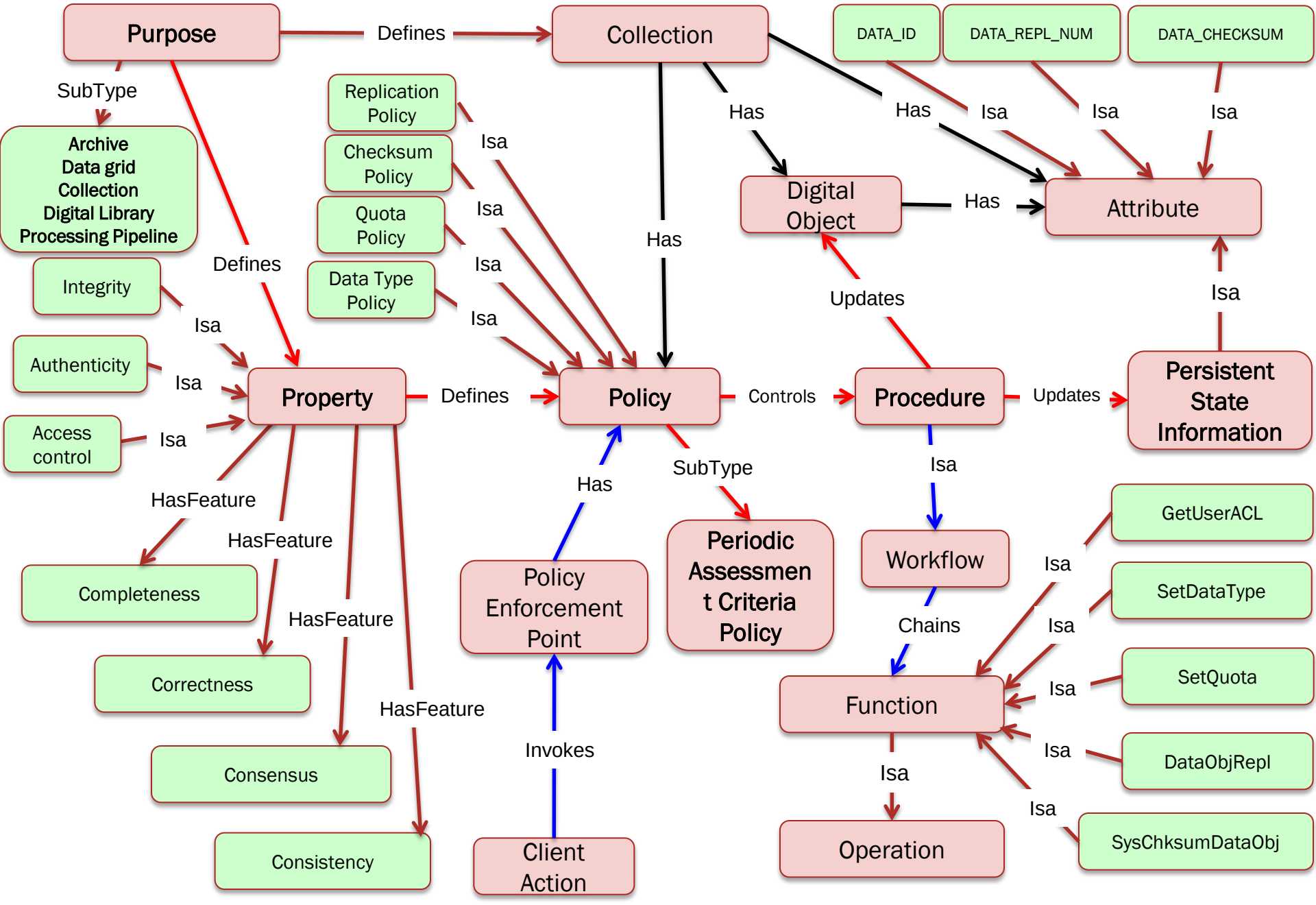
Procedures


Purpose
Defines
Collection
SubType
Archive
Data grid
Collection
Digital Library
Processing Pipeline
Replication Policy
Checksum Policy
Quota Policy
Data Type Policy
Isa
Has
Has
Digital Object
Has
Attribute
Integrity
Defines
Has
Updates
Isa
Authenticity
Isa
Property
Defines
Policy
Controls
Procedure
Updates
Persistent State Information
Access control
Isa
Isa
HasFeature
HasFeature
HasFeature
HasFeature
Completeness
Correctness
Consensus
Consistency
Workflow
Chains
Function
Isa
Operation
GetUserACL
SetDataType
SetQuota
DataObjRepl
SysChksumDataObj
Isa
Isa
Isa
Isa
Isa

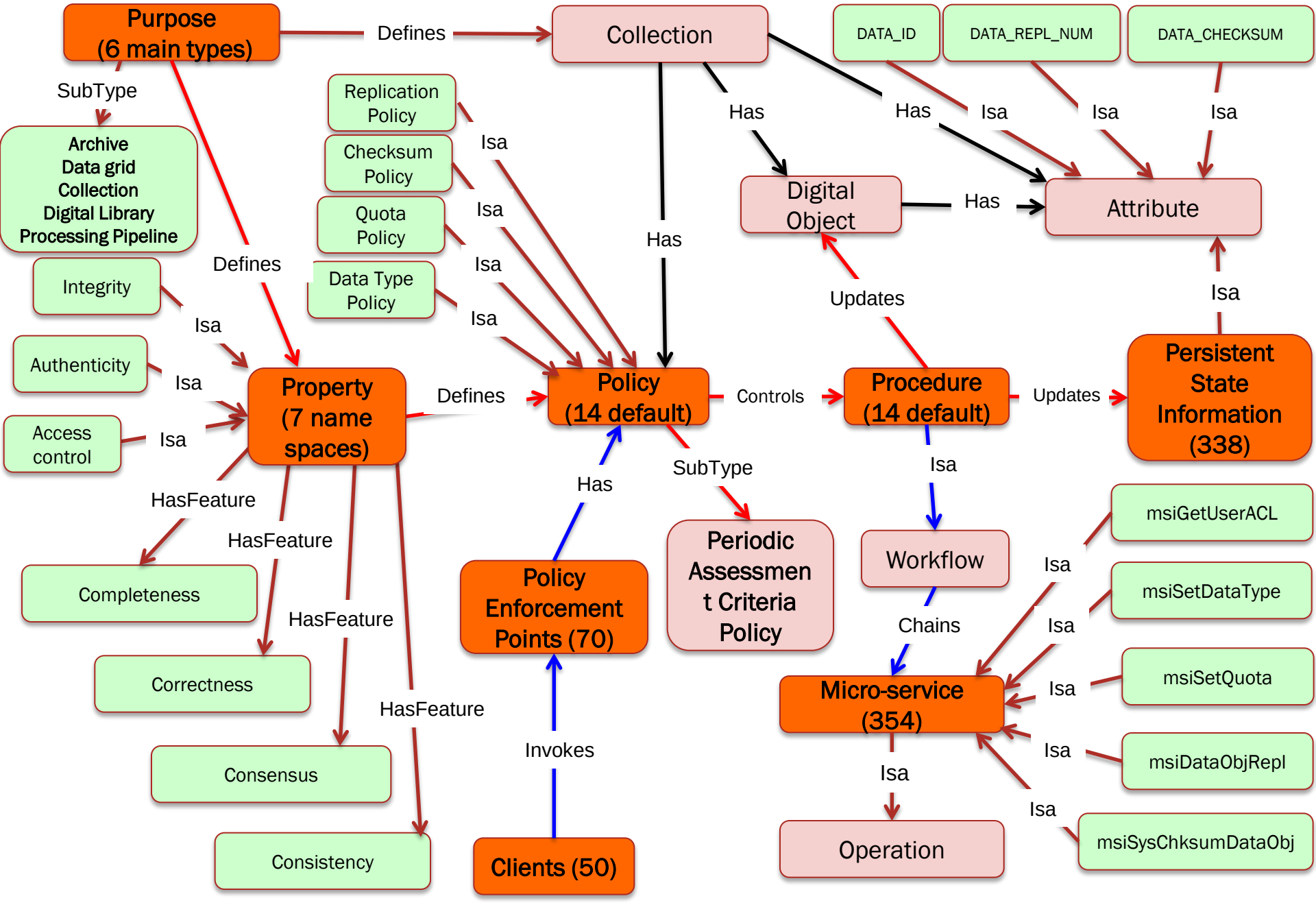
Persistent State



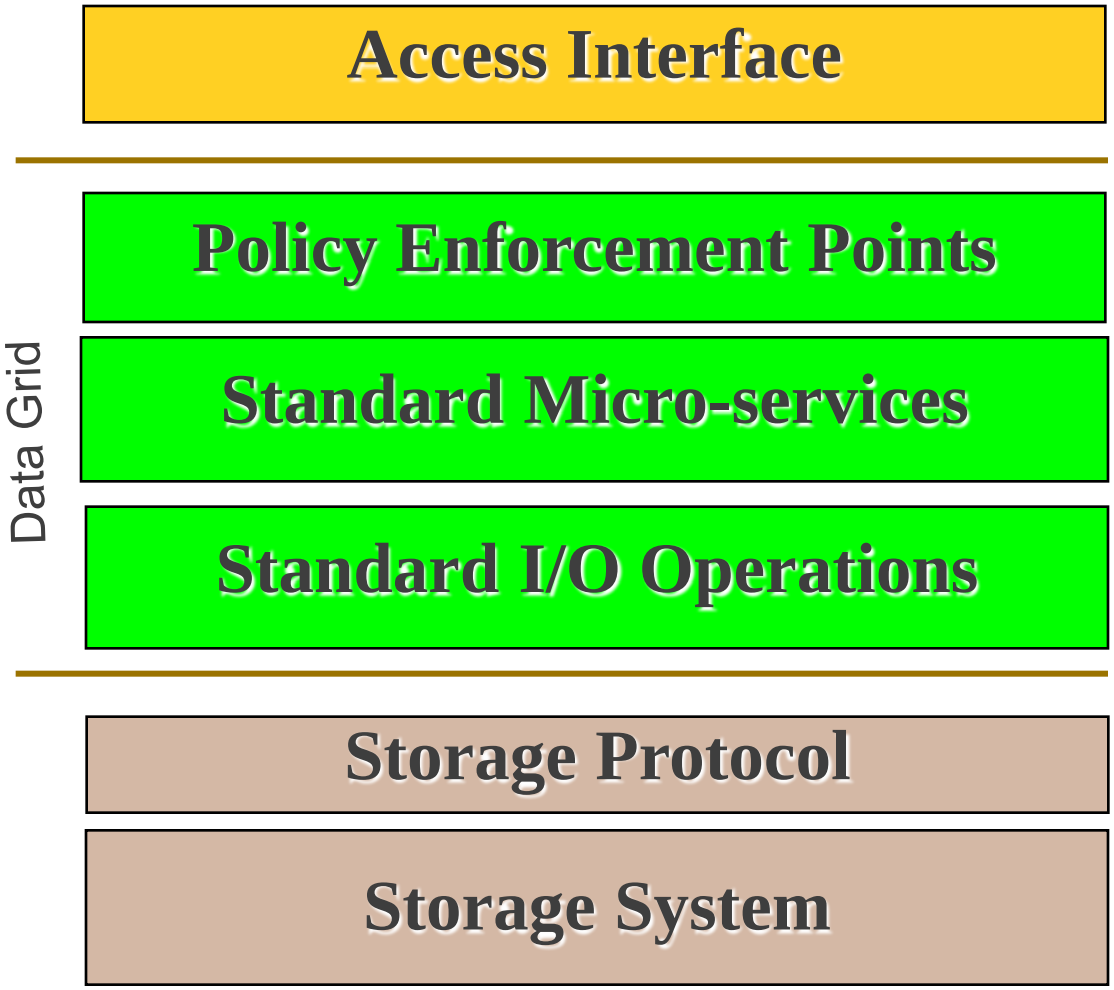
Policy Enforcement



Concept Graph



Data Workflow Virtualization



- Trap actions requested by the client at multiple policy enforcement points.
- Map from policy to standard micro-services.
- Map from micro-services to standard Posix I/O & database operations.
- Map standard I/O operations to the protocol supported by the storage system & database.

Policy Sets

- NSF Data Management Plans
  - 38 tasks to be automated
    - Computer actionable rules controlling computer executable procedures
- Protected data management
  - 51 tasks to be automated
    - Identify PII, PCI, PHI
    - Encryption, access approval flags, access controls
- ISO 16363 trustworthiness assessment
  - 133 tasks to be automated
    - Assessment reports, enforcement

National Cyberinfrastructure

- Federation mechanisms
  - Shared name spaces
    - Tightly coupled systems – user names, file names
  - Shared services
    - Loosely coupled systems – independent name space
    - Access, discover, apply service
  - Shared nothing
    - Asynchronous interactions – post to message bus
- Policy-based interaction management
  - Control user interactions
  - Control collection properties
  - Control technology interactions

Research Collaboration Infrastructure

- Discovery Environment (Cyverse)
  - Shared collections – iRODS
  - Workflow execution – Condor
  - Application virtualization – Docker
- Exploring migration of services to storage location
  - HIVE
  - Bitcurator
  - Virus scan
  - Indexing/auditing
  - GABBS
  - SciON sensor data
  - Syndicate
  - Big Data Hub

Sustainability

- iRODS policy-based data management
  - iRODS Consortium
  - Membership based support for open source software
- Southern Region Big Data Hub
  - Proposing DFC federation hub as infrastructure prototype for southern region
- Publications
  - http://datafed.org
  - https://dfcweb.datafed.org/idrop-web2/home/link?irodsURI=irods%3A%2F%2Firen2.renci.org%3A1237%2Fdfcmain%2Fhome%2FDFC-public%2FPolicy-course

More Information

Reagan Moore
rwmoore@renci.org

iRODS Consortium
http://irods.org

NSF DataNet Federation Consortium
http://datafed.org